Towards Unified and Generalizable Multimodal Foundation Models

Zhiyang Xu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Application

Lifu Huang, Chair
Xuan Wang, Co-chair
Mohit Bansal
Naren Ramakrishnan
Chandan K. Reddy

Feb 10, 2026
Blacksburg, Virginia

Keywords: Multimodal Learning, Multimodal Understanding, Multimodal Generation

Towards Unified and Generalizable Multimodal Foundation Models

Zhiyang Xu

(ABSTRACT)

Recent advances in multimodal models have reshaped multimodal learning by leveraging large language models (LLMs) as backbones and integrating them with vision encoders or diffusion models. These approaches have achieved strong performance in either multimodal understanding or multimodal generation. However, no existing system offers a unified framework capable of performing both understanding and generation across flexible input-output modality combinations while also generalizing to unseen, real-world tasks. Unification is essential not only for enabling a single model to perform traditional understanding and generation tasks, but also for enabling cross-modal tasks, such as visual storytelling, report generation, and script creation, that neither understanding nor generation models alone can accomplish. By processing and generating inputs and outputs that span multiple modalities, unified models more closely mirror the way humans naturally acquire and construct knowledge. Generalization, in turn, enables models to adapt to novel tasks in open-world environments under human-specified instructions or principles. Together, unification and generalizability constitute two fundamental pillars for advancing toward general-purpose multimodal intelligence. This dissertation advances unified and generalizable multimodal modeling through novel architectures, post-training paradigms, and reinforcement learning algorithms. It addresses two enduring challenges in multimodal foundation models: (1) the lack of modality unification, and (2) limited generalizability in open-world settings. The contributions are fourfold. First, we introduce Modality-Specialized Synergizers (MOSS), a framework that enables interleaved multimodal generation in pretrained models. Second, we propose an efficient unified architecture that bridges vision-language models and diffusion models, providing a novel pathway for joint understanding and generation. Third, we establish multimodal instruction tuning as a new post-training paradigm to improve zero-shot generalization and robustness. Finally, we extend image understanding to the spatiotemporal domain by developing a novel reinforcement learning algorithm that promotes temporal awareness, enabling vision–language models to reason effectively about videos. Extensive experiments across diverse multimodal benchmarks demonstrate that these approaches significantly enhance unification, generalizability, and overall capability. Collectively, this research strengthens the foundations of multimodal AI and outlines a pathway toward universal models that can understand, reason, and generate across modalities in complex open-world environments.

Towards Unified and Generalizable Multimodal Foundation Models

Zhiyang Xu

(GENERAL AUDIENCE ABSTRACT)

Artificial intelligence systems are increasingly able to work with different types of information, such as text, images, and videos. For example, modern AI tools can answer questions about pictures, write stories based on images, or generate new images from written descriptions. However, most existing systems are designed to perform only one type of task well—either understanding information (such as recognizing what is in an image) or generating new content (such as creating images or text). They also often struggle to adapt to new tasks that were not specifically included during training. This dissertation studies how to build more flexible AI systems that can both understand and create information across different forms of media. Instead of designing separate models for text, images, and videos, the goal is to develop unified systems that can process multiple types of data together and produce outputs in different formats. Such systems more closely resemble how humans think and communicate, since people naturally combine language, visual perception, and temporal experiences when understanding the world. To achieve this goal, this research proposes several new techniques for training and designing multimodal AI systems—AI models that can work with multiple kinds of data at the same time. First, it introduces a method that allows models to combine specialized components for different types of information while still working together as a single system. Second, it presents a new architecture that connects two previously separate types of AI models, enabling both understanding and content generation within one framework. Third, it develops a training strategy that teaches models to follow human instructions across different tasks, allowing them to better adapt to unfamiliar problems. Finally, it proposes a learning algorithm that helps AI systems better understand videos by learning how events unfold over time. Experiments show that these methods significantly improve the flexibility and adaptability of multimodal AI systems. By enabling a single model to understand and generate information across text, images, and videos, this work moves closer to building general-purpose AI tools that can assist people in a wide range of real-world tasks, such as storytelling, education, and information analysis.

Dedication

To my family for their love and support.

Acknowledgments

I would like to express my deepest gratitude to the many individuals who supported me throughout my Ph.D. journey.

First and foremost, I am sincerely grateful to my advisor, Dr. Lifu Huang, for his outstanding guidance, patience, and mentorship. His support has extended far beyond academic supervision; his advice and encouragement have profoundly shaped both my research and my personal growth. I am also thankful to my committee members for their valuable feedback, insightful suggestions, and careful guidance throughout this work. Their expertise and perspectives have played an essential role in strengthening this dissertation.

My heartfelt thanks go to the members of my lab for their collaboration, thoughtful discussions, and generous help in reviewing my papers. Their support and constructive feedback have been instrumental to my progress, and I am truly grateful for the inspiring environment they helped create.

I would also like to extend my deepest appreciation to my parents and grandparents, Xiaohong Yang, Jun Xu, Fenglian Ren, Yuzhen Zhang, Yongsheng Xu, and Zhongheng Yang, for their unwavering love, understanding, and support during this demanding period. Pursuing a Ph.D. is both exciting and challenging, and their encouragement has been a constant source of strength.

Finally, I wish to express my profound gratitude to my wife, Zexuan Li. Her patience, care, and steadfast support sustained me through many of the most difficult and important moments of this journey. I could not have completed this dissertation without her.

Contents

List of Figures x

List of Tables xii

I Introduction and Background 1

1 Introduction 2

1.1 Background . . . 2
1.2 Motivation and Solution . . . 3
1.3 Research Questions and Thesis Statement . . . 3
1.4 Contributions . . . 4
1.5 Thesis Organization . . . 5

II Unified Multimodal Modeling 7

2 Interleaved Multimodal Generation 8

2.1 Motivations . . . 8
2.2 Related Work . . . 9
2.3 Modality-Specialized Synergizers (MoSS) . . . 10
2.3.1 Background: Autoregressive Vision-Language Generalists . . . 10
2.3.2 Linear Low-Rank Adaptation (LoRA) . . . 11
2.3.3 Convolutional Low-Rank Adaptation (Convolutional LoRA) . . . 11
2.3.4 Integrating MoSS into pretrained multimodal models . . . 13
2.4 LeafInstruct . . . 13
2.4.1 Dataset Construction and Statistics . . . 14

2.4.2 Dataset construction . . . 14
2.4.3 Dataset Statistics . . . 16
2.4.4 Post-Training for Interleaved Generation . . . 16
2.5 Experiment Setup . . . 17
2.6 Results and Discussions . . . 18
2.6.1 Quantitative Results . . . 18
2.6.2 Per-task Performance . . . 19
2.6.3 Qualitative Results . . . 20
2.6.4 Comparison between **MoSS** and other PEFT Methods . . . 22
2.7 Summary . . . 23

3 Efficient Unified Architectures **25**
3.1 Motivation . . . 25
3.2 Related Work . . . 26
3.3 Preliminaries . . . 27
3.4 LaTtE-Flow . . . 28
3.4.1 LaTtE-Flow Layer Design . . . 28
3.4.2 Layerwise Timestep Experts . . . 29
3.4.3 Timestep-Conditioned Residual Attention . . . 30
3.5 Results and Discussion . . . 31
3.5.1 Image Generation and Understanding Results . . . 31
3.5.2 Ablation Studies . . . 33
3.6 Summary . . . 36

III Generalizable Multimodal Modeling **38**

4 Zero-Shot Multimodal Learning **39**
4.1 Motivation . . . 39
4.2 Related Work . . . 40

4.3 MultiInstruct . . . 41
4.3.1 Multimodal Task and Data Collection . . . 41
4.3.2 Task Instruction Creation . . . 42
4.3.3 Multimodal Instruction Formatting . . . 43
4.4 Problem Setup and Models . . . 44
4.4.1 Problem Setup . . . 44
4.4.2 Transfer Learning from Natural Instructions . . . 44
4.5 Experimental Setup . . . 45
4.6 Results and Discussion . . . 46
4.6.1 Effectiveness of Instruction Tuning on MultiInstruct . . . 46
4.6.2 Impact of Transfer Learning from Natural Instructions . . . 47
4.6.3 Impact of Increasing Multimodal Instruction Task Clusters . . . 48
4.6.4 Effect of Diverse Instructions on Instruction Tuning . . . 49
4.6.5 Effect of Fine-tuning Strategies on Model *Sensitivity* . . . 49
4.6.6 Zero-Shot Performance on NLP Tasks . . . 50
4.7 Summary . . . 50

5 Scaling Task Diversity for Robust Generalization 52
5.1 Motivation . . . 52
5.2 Related Work . . . 54
5.3 Vision-Flan . . . 54
5.3.1 Collection Pipeline . . . 54
5.3.2 Comparison with Existing Datasets . . . 56
5.4 Multi-Stage Instruction-Tuning . . . 57
5.5 Experiment Setups . . . 58
5.6 Results and Discussions . . . 59
5.6.1 Main Results . . . 59
5.6.2 Effect of Human-Labeled and GPT-4 Synthesized Datasets . . . 60
5.6.3 Single-stage Tuning on Mixed Data Vs. Two-stage Tuning . . . 61

5.6.4 Effect of Newly Created Tasks . . . 61
5.6.5 Contributions of Tasks from Different Task Groups . . . 62
5.6.6 What is Essentially Improved in VLMs during Instruction Tuning . . . 62
5.7 Summary . . . 63

6 Spatialtemporal Reasoning **65**
6.1 Motivation . . . 65
6.2 Related Works . . . 66
6.2.1 Policy Optimization Algorithms . . . 66
6.2.2 Ego-Centric Video Understanding . . . 67
6.3 Method . . . 68
6.3.1 Background on GRPO . . . 68
6.3.2 Greedy Baseline without Temporal Information . . . 69
6.3.3 Temporal Global Policy Optimization (TGPO) . . . 69
6.3.4 Prompt Engineering. . . . 70
6.3.5 Reward Modeling . . . 71
6.4 Experiments . . . 72
6.4.1 Implementation Details . . . 72
6.4.2 Evaluation . . . 73
6.4.3 Baselines . . . 73
6.5 Results . . . 74
6.5.1 Comparison with RL-based Methods . . . 74
6.5.2 System-Level Comparison . . . 75
6.6 Summary . . . 77

7 Conclusion and Future Work **78**
7.1 Conclusion . . . 78
7.2 Future Work . . . 79

Bibliography **81**

List of Figures

2.1 Failure cases of existing pretrained multimodal models (Emu2 at the top and GILL at the bottom). The output text with inferior quality is highlighted with underline. The regions that impede output images' quality are highlighted with red bounding boxes. 9

2.2 An autoregressive VLG with our proposed **MoSS** added to its linear layers. The linear LoRA on the left side is specialized to generate text tokens and the Convolutional LoRA on the right side is specialized to generate image patches. On the right handside, we show the details of convolutional operation applied to autoregressively generate image tokens. Best viewed in color. 12

2.3 Comparison between existing benchmarks and our LeafInstruct. In existing datasets such as InstructPix2Pix [12] and Mantis-Instruct [90], the outputs are in single modality, either text or image. On the contrary, the inputs and outputs of our LeafInstruct cover multiple modalities. 14

2.4 Domain distribution in LeafInstruct. 15

2.5 **Per-task performance** averaged on 5 aspects on InterleavedBench. 19

2.6 **Qualitative results** of **MoSS** based on Emu2 and open-source baselines. The <IMG> tokens denote the images' positions in the interleaved sequences. 21

2.7 Performance averaged on 5 aspects with different rank numbers. 22

3.1 **Comparison of the flow-matching process between standard diffusion / flow-matching models and our proposed LaTtE-Flow.** Unlike diffusion / flow-matching based models, which invoke the entire model at each sampling timestep, LaTtE-Flow activates only a subset of layers at each step, improving efficiency. 26

3.2 **LaTtE-Flow overall architecture**. 29

3.3 **Timestep-conditioned residual attention** 31

3.4 **Training dynamics of LaTtE-Flow vs. baselines.** FID on ImageNet 50K. 33

3.5 **Effect of group size in LaTtE-Flow Couple**. 33

3.6 **Impact of # sampling steps and CFG strength on Inception Score and FID.** 35

3.7 **Timestep-conditioned residual attention analysis.** (a) Visualization of attention behavior in Vanilla Couple and (b) learned residual gating patterns in LaTtE-Flow Couple. 35

3.8 **Timestep-conditioned residual attention gates across transformer layer in LaTtE-Flow Couple.** White regions indicate positions without gating values since residual attention is applied only within predefined layer groups. Notably, different heads exhibit distinct gating dynamics, with some emphasizing earlier timesteps, while others modulate more strongly in later layers, suggesting head-specific specialization in residual attention. 36

4.1 **Task Groups Included in MultiInstruct.** The yellow boxes represent tasks used for evaluation, while the white boxes indicate tasks used for training. . 41

4.2 **Example Instances from MultiInstruct for Four Tasks.** 42

4.3 **Model Performance as the Number of Multimodal Instruction Task Clusters Increases.** The number in the parenthesis of each cluster denotes the number of tasks. 48

4.4 **Model *Sensitivity* on Unseen Evaluation Tasks.** Lower is better. . . . 50

5.1 Sample tasks in Vision-Flan. **Instruction** denotes a task instruction crafted by annotators. **Input** means text input in the given task, and **Target** is the target response based on the instruction. 53

5.2 Comparison of task diversity between Vision-Flan and previous visual instruction tuning datasets. LLaVA and SVIT report very coarse-grained categories of tasks. Each circle represents a task category and the radius is proportional to the number of tasks in that category. The radius of circles for different datasets are comparable. 55

5.3 The left of the figure shows the LLaVA-Architecture and the right of the figure shows the two-stage visual instruction tuning pipeline. 57

5.4 Performance on four comprehensive benchmarks versus the number of training tasks. 59

5.5 Effect of the number of GPT-4 synthesized training instances on MME. The dashed gray line indicates the performance of LLaVA 1.5. 60

6.1 An overview of our proposed **TGPO**. 68

6.2 Test reward over the 3000 training steps. The reward curves are reported on four benchmarks across GRPO, GSPO, and our two **TGPO** variants. 75

List of Tables

2.1 Comparison between our LeafInstruct and existing instruction tuning datasets. . . . 16

2.2 **Main results of interleaved generation on InterleavedBench**. We show the performance of pipelines based on proprietary models (Top), open-source pretrained multimodal models (Middle), and the pretrained multimodal models trained with our **MoSS** and LeafInstruct (Bottom), respectively. Note that the scale is from 0 to 5 (5 is the best). We also report the percentage of improvement in our method over the original VLG backbone in the parentheses. The best results are highlighted in **bold**. . . . 18

2.3 Results on widely adopted multimodal understanding and text-to-image generation benchmarks. Note that the FID metric on MSCOCO is the lower the better. . . . 20

2.4 **Comparison between MoSS and existing PEFT methods**, i.e., traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA. Mixture-of-Expert LoRA uses two different sets of linear LoRA for images and text, respectively. The rank number is set to 256 for all methods in this table. . . . 22

2.5 **Human evalution** of randomly sampled instances from LeafInstruct. Note that the scale is from 0 to 3 (**Score 3 is the best**), which is different from the scale used in Table 2.2 and Table 2.4. . . . 23

3.1 **Comparison of generative models** across FID, IS, Precision, Recall, parameters, steps, and inference time on ImageNet-50K. For LaTtE-Flow, we report the number of parameters activated per timestep, given that it has a timestep-expert architecture where only a subset of layers is used at each step. We also report inference time relative to LaTtE-Flow Couple. †: taken from MaskGIT [17] . . . 32

3.2 **Results on comprehensive image understanding benchmarks.** Best scores are highlighted in **bold**. Since our LaTtE-Flow Couple is an expert architecture, we report the number of activated parameters used for image understanding. . . . 32

3.3 **Effect of time-conditioned residual attention.** . . . 34

4.1 **Zero-shot Performance on Multimodal Commonsense Reasoning.** The best performance is in **bold**. 47

4.2 **Zero-shot Performance on Question Answering and Miscellaneous.** The best performance is in **bold**. 47

4.3 **Effect of Different Number of Instructions.** Performance of $\text{OFA}_{\text{MultiInstruct}}$ finetuned on different numbers of instructions. 49

4.4 **Zero-shot Performance on NLP tasks.** The performance is reported in Rouge-L and the best performance is in **bold**. 50

5.1 Comparison between VISION-FLAN and existing visual instruction tuning datasets. 56

5.2 Comprehensive evaluation of VLMs on widely adopted benchmark datasets. CF denotes the averaged performance of VLMs on four catastrophic forgetting benchmarks. 59

5.3 Comparison of VISION-FLAN BASE trained with a fixed total amount of data instances. 60

5.4 Comparison between single-stage finetuning on mixed data and two-stage finetuning. 61

5.5 Comparison between finetuning VISION-FLAN BASE on all tasks and finetuning VISION-FLAN BASE only on existing tasks. 61

5.6 Contributions of different tasks group to the performance of VISION-FLAN CHAT. 62

5.7 Effect of tuning different modules in VISION-FLAN BASE. ✓ denotes the module is tuned and ✗ denotes the module is frozen during visual instruction tuning. 62

5.8 Results of replacing visual instruction tuned MLPs with pretrained MLPs. Gray rows show the performance of the original models and yellow rows show the performance after replacing instruction-tuned MLPs with pretrained MLPs. 63

6.1 Performance comparison of our method **TGPO** with two popular RL-based optimization methods and chain-of-thought (CoT) reasoning across egocentric benchmarks. 72

6.2 Area Under the Curve (AUC) of reward over the first 3000 training steps for different optimization methods across datasets. A higher AUC reflects faster reward improvement and improved training stability. 74

6.3 Performance comparison on EgoSchema. 76

6.4 Performance comparison on EgoPlan 2. 76

6.7 EgoTempo benchmark. 76
6.5 Performance comparison on VLM4D. 77
6.6 Performance comparison on EgoPlan. 77

Part I

Introduction and Background

Chapter 1

Introduction

1.1 Background

Recent advances in multimodal models have driven remarkable progress toward building versatile multimedia assistants, demonstrating sophisticated capabilities in understanding and generating cross-modal content such as text, images, and videos. Pretrained on large-scale internet data, these foundation models achieve state-of-the-art results on numerous benchmarks and have been applied across a wide range of real-world tasks.

Despite these successes, current multimodal foundation models face two fundamental limitations that prevent them from functioning as truly general-purpose multimodal intelligence. First, they lack modality unification. Most models are designed either for understanding (e.g., visual question answering, captioning) or for generation (e.g., text-to-image synthesis), but struggle when tasks require both modalities as inputs and outputs. For example, interleaved text–image generation tasks such as visual storytelling or multimodal script generation require a model to track narrative progression across text and images and continue generating coherent multimodal sequences, capabilities that current systems cannot fully support. Second, multimodal models exhibit limited generalizability in open-world scenarios. While they achieve strong performance on in-distribution benchmarks, these models often fail to adapt to novel tasks, robustly follow human instructions, or align outputs with human preferences. This limits their practical utility in dynamic, real-world environments.

These challenges arise from several architectural and post-training constraints. Architecturally, most multimodal models enforce single-modality outputs (either image or text), reducing flexibility. When tasks require cross-modal outputs, existing approaches often rely on multiple specialized models and decompose the problem into subtasks, leading to incoherent results. From the post-training perspective, current strategies frequently narrow models toward specific benchmarks or domains, thereby weakening their ability to generalize to unseen tasks, adapt to diverse user instructions, and sustain robust performance in broader contexts. Moreover, while these models can capture spatial relationships in static images, they lack the ability to model temporal dependencies, making it difficult to understand and reason about motion and dynamic interactions in videos.

1.2 Motivation and Solution

This dissertation address the foundamental challengs of building general-purpose multimodal intellegence by proposing architecture and training paradim innovations. Unlike existing works that develop specialize multimodal models target single-modality output or narrow tasks. Our main goal is to develope unified and generalizable multimodal foundations models that not only unify understanding and generation across multiple modalities, but generalize robustly to diverse tasks and dynamic open-world scenarios.

Our approach encompasses four critical aspects: First, the model should be able to understand and generate text and images in arbitrary sequence. Second, an efficient architecture combines vision-language models for their strong reasoning and text modeling capability, and diffusion models for their excel image synthesis capability. The unified architecture should be highly efficient at both training and inference. Third, it can easily generalize to unseen mutimodal tasks by fatihfully follow human instructions and generate output align with human-preference. Fourth, it can generalize from reasoning over static images to videos containing spatialtemporal information. Together, our aim to develop unified multimodal foundation model that can handel various modalities and generalize to real-world tasks and reasoning conditions.

This research direction represents a significant advance in the foundations of multimodal AI. Rather than relying on scaling model parameters alone, we investigate how multimodal systems can adapt their architectures, training paradigms, and temporal reasoning capabilities to unify text, images, and video under a single framework. This paradigm shift has broad implications for both methodology and application, enabling models that are more flexible, human-aligned, and capable of operating effectively in complex, open-world environments.

1.3 Research Questions and Thesis Statement

The central research question in this dissertation is: *How can we develop general-purpose multimodal intelligence that unifies modalities and generalizes across open-world tasks, enabling robust understanding and scalable generation in both textual and visual domains?* This overarching question decomposes into four interconnected research questions that guide our investigation:

- **Interleaved Multimodal Unification**: How can we enable multimodal models to seamlessly integrate text and image modalities for interleaved understanding and generation?

  The challenge is that existing pretrained multimodal models often apply identical architectures to process heterogeneous modalities, and lack the ability to generate text and images in arbitrary sequences. Addressing this requires (1) high-quality finetun-

ing data that contains both interleaved text and images as input and output; and (2) mixture of expert designs that handles varities of inputs and outputs modelities.

- **Efficient Unified Architectures**: How to design unified architectures that can jointly leverage autoregressive and diffusion paradigms to achieve both efficient image understanding and image generation?

  Autoregressive models excel at sequential reasoning, while diffusion models generate high-quality images. Yet unifying them remains an open challenge. This question seeks to explore how these paradigms can be combined into efficient architectures that deliver state-of-the-art multimodal performance.

- **Zero-Shot Multimodal Learning**: How to improve zero-shot generalization, robustness, and human-preference alignment of pretrained multimodal models via post-training innovation?

  While instruction tuning has been applied in language models, its role in multimodal models is less understood. This question investigates how multimodal instruction tuning, scaled human-labeled tasks, and multi-stage post-training frameworks can help models follow human instructions more faithfully and generalize to unseen tasks.

- **Spatiotemporal Reasoning**: How can we generalize image-based multimodal models to dynamic video sequences by incentivizing temporal awareness?

  Static image models fail to capture temporal dependencies that are crucial for video reasoning. This question examines reinforcement-style optimization strategies that encourage models to develop temporal awareness, enabling reasoning and generation in spatiotemporal domains.

Thesis Statement By developing interleaved generation, scalable autoregressive–diffusion architectures, multimodal instruction-tuning paradigms, and reinforcement-based algorisms for temporal reasoning, we can build unified multimodal foundation models that overcome the limitations of current architectures and training paradigms. This integrated approach advances both the unification of text and vision and the generalizability of models to diverse open-world tasks, enabling robust understanding, reasoning, and generation across modalities.

1.4 Contributions

This dissertation makes four major contributions that advance the development of unified and generalizable multimodal foundation models:

- **Modality-Specialized Synergizers for interleaved generation** (Chapter 2). This work studies interleaved generation. We construct LeafInstruct, the first large-scale post-training dataset for interleaved text–image generation, consisting of 184,982 high-quality instances on more than 10 diverse domains. In addition, we introduce Modality-Specialized Synergizers (MOSS), a novel design that augments pretrained multimodal architectures with modality-aware adaptation layers. MOSS enables more effective modeling of local image priors and sequential text while preserving strong cross-modal integration.

- **Efficient unified architectures bridging autoregression and diffusion paradigms** (Chapter 3). This work introduces LaTtE-Flow, a unified model that combines the strengths of autoregressive and diffusion frameworks. LaTtE-Flow leverages diffusion transformers for high-quality image generation and autoregressive vision-language models for multimodal understanding, coupled with layerwise timestep experts and residual attention. This design achieves state-of-the-art performance across diverse multimodal benchmarks.

- **A novel multimodal instruction-tuning paradigm for zero-shot generalization** (Chapter 4 and Chapter 5). We establish instruction tuning as a core paradigm for improving robustness and generalization in multimodal models. We introduce MultiInstruct, the first multimodal instruction-tuning benchmark and Vision-Flan, the most diverse large-scale visual instruction-tuning dataset to date. We further propose a two-stage instruction-tuning framework. This paradigm significantly improves zero-shot generalization and human instruction following.

- **A reinforcement learning algorithm incentivizing spatiotemporal reasoning** (Chapter 6). To extend unified modeling beyond static images, we introduce a novel reinforcement-learning algorithm that incentivizes temporal awareness in video reasoning tasks. This contribution provides one of the first systematic explorations of reinforcement-style optimization for spatiotemporal multimodal reasoning, advancing the frontier of general-purpose video–language intelligence.

1.5 Thesis Organization

The remainder of this dissertation is organized into four parts that systematically develop Unified and Generalizable Multimodal Foundation Models:

- **Part I: Introduction and Background**

  Following this introduction, Chapter I provides a comprehensive background on multimodal foundation models, including architectures for text–image understanding and

generation, instruction tuning, and extensions to temporal reasoning. This chapter establishes the technical foundations and positions our contributions within the broader landscape of multimodal AI research.

- **Part II: Unified Multimodal Modeling**

  This part focuses on unifying architectures for interleaved text–image generation. Chapter 2 introduces LeafInstruct, the first large-scale dataset for interleaved instruction tuning, and MOSS, our modality-specialized synergizers that augment unified architectures with modality-aware adaptation layers. Chapter 3 presents LaTtE-Flow, an efficient unified architecture that bridges autoregressive and diffusion paradigms, achieving state-of-the-art performance across both understanding and generation tasks.

- **Part III: Generalizable Multimodal Modeling**

  This part establishes instruction tuning and temporal reasoning as principled paradigms for enhancing generalization and robustness. Chapter 4 introduces MultiInstruct, the first multimodal instruction-tuning benchmark, while Chapter 5 presents Vision-Flan, the most diverse large-scale visual instruction-tuning dataset to date. Together, these works propose a two-stage instruction-tuning framework and provide empirical insights into human-preference alignment and zero-shot generalization. Finally, Chapter 6 extends generalization from static images to dynamic video sequences, introducing a reinforcement-learning algorithm that incentivizes temporal awareness for spatiotemporal reasoning.

Through this progression—from interleaved generation to scalable architectures, and from instruction tuning to spatiotemporal reasoning—this dissertation demonstrates that advancing multimodal AI requires not only larger models but also principled designs that specialize architectures, leverage high-quality instruction data, and incorporate temporal dependencies. By addressing the dual challenges of unification and generalizability, we move toward universal multimodal foundation models capable of understanding, reasoning, and generating across diverse modalities in open-world environments.

Part II

Unified Multimodal Modeling

Chapter 2

Interleaved Multimodal Generation

2.1 Motivations

As established in Chapter 1, multimodal foundation models are increasingly expected to support not only understanding but also generation across different modalities. A particularly important frontier is interleaved multimodal generation, where models are required to produce sequences that flexibly combine both text and images. Such capability unlocks a wide range of applications, including visual storytelling, script generation, and multimodal dialogue, which cannot be achieved by image understanding models [4, 93, 117] or image generation models alone [146, 149]. Recent works [75, 167, 171, 174] have emerged as promising candidates for this task, but they still face fundamental challenges that hinder their effectiveness.

One major challenge lies in architectural unification and post-training setup. Current pretrained multimodal models employ a single transformer backbone with a shared set of parameters to simultaneously model discrete text tokens and continuous image representations. While this design provides a unified interface, it disregards the inherent inductive biases of the two modalities. Text follows a sequential, left-to-right order, whereas images are two-dimensional, relying heavily on local spatial priors between adjacent patches. The transformer architecture, although powerful for sequence modeling, is less effective than convolutional operations in capturing local spatial dependencies [24, 239]. Consequently, existing pretrained multimodal models often produce incoherent text and distorted images. This fundamental mismatch between modality characteristics and architectural unification results in degraded performance, particularly in tasks requiring precise visual fidelity and consistent cross-modal alignment.

A second critical challenge is the lack of high quality post-training data. While many pretrained multimodal models are pretrained on large-scale multimodal corpora [246], their post-training alignment is typically limited to single-modality tasks such as text generation or image synthesis. As a result, these models frequently fail in interleaved scenarios. For example, when instructed to complete a partially multimodal sequence, they may generate repetitive or irrelevant text and produce images that fail to reflect the user's intent, as shown in Figure 4.2. This issue is exacerbated by the lack of large-scale, high-quality datasets explicitly designed for interleaved generation, making current approaches unscalable and less flexible in real-world applications.

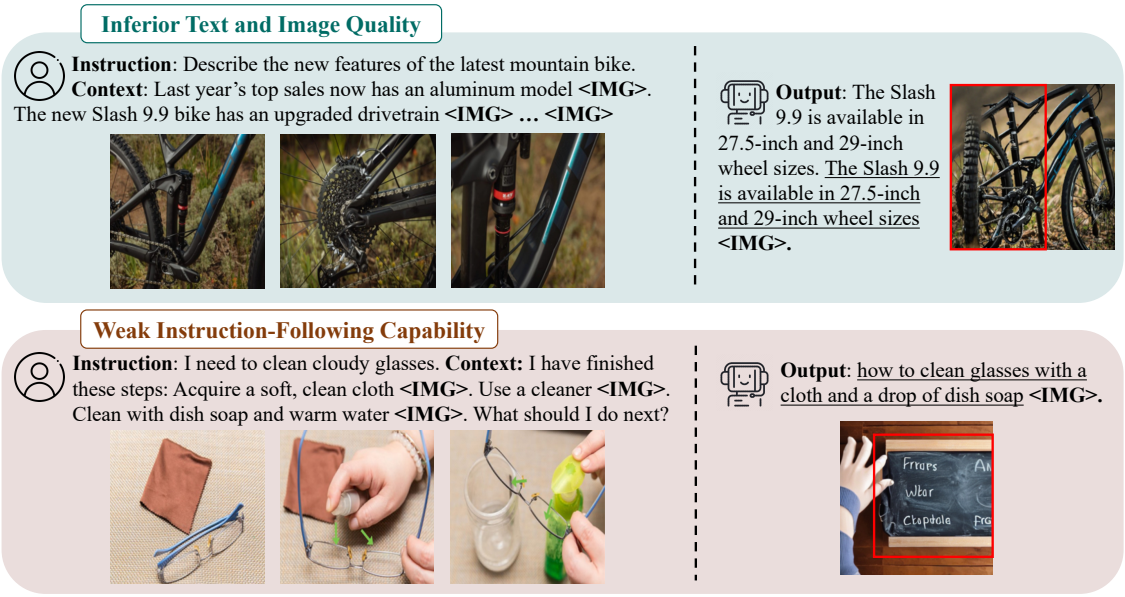


Figure 2.1: Failure cases of existing pretrained multimodal models (Emu2 at the top and GILL at the bottom). The output text with inferior quality is highlighted with underline. The regions that impede output images' quality are highlighted with red bounding boxes.

Taken together, these challenges underscore the need for a new framework in interleaved multimodal generation. One that combines architectural specialization for each modality with high-quality datasets that explicitly target interleaved tasks. Addressing these limitations is crucial for building truly generalist models that can generate coherent, high-quality text and images in arbitrary sequences, thereby extending the scope and utility of multimodal AI systems in open-world applications.

2.2 Related Work

Interleaved Vision-Language Models There are two popular formulations for VLGs: The first leverages VQGAN [35] to quantize an image into a long sequence of discrete tokens and add the vocabulary in VQGAN's codebook into the vocabulary of LLMs [1, 71, 174, 215, 221]. In this way, the LLMs are trained with a unified autoregressive objective to predict image tokens or text tokens. The predicted image tokens are fed into a VQGAN decoder to reconstruct images. The second formulation employs the CLIP image encoder to transform images into sequences of continuous embeddings [75, 101, 167, 171, 173, 177, 198, 245], which are then concatenated with text embeddings in their original order. Compared to the first approach, this formulation often requires shorter sequences to represent an image and generally yields superior performance. Our proposed method requires minimal assumptions on VLG's architectures and can be applied to many of the existing transformer-based VLGs.

Visual Instruction Tuning [205] propose MultiInstruct, the first human-label visual instruction tuning dataset to improve the generalizability of VLMs. LLaVA [117] leverages GPT-4 to convert image captions from existing annotations into three tasks, including visual dialogues, visual question answering, and detail captions. Following studies either utilize proprietary LLMs [20, 31, 86, 116, 126, 184, 217, 218, 234, 244] or human efforts [116, 203] to augment visual instruction tuning tasks. Several studies target specific aspects of VLMs' capability, such as domain and instruction bias [6, 112], object grounding [19], and OCR [64, 231]. Instruction tuning has also been widely applied to other vision-language tasks, such as image editing [13] and interleaved text-image understanding [69]. [61] finetune a model that can follow multimodal instructions to generate desired images. However, most existing instruction-tuning datasets only consider the tasks where the outputs are in a single modality, i.e., either text or image. *To facilitate the training and enhance the instruction-following capabilities for VLGs, we curated* LeafInstruct*, the first instruction-tuning dataset tailored for interleaved text-image generation across diverse domains, where the inputs and outputs can contain interleaved text and multiple images.*

Parameter-Efficient Finetuning (PEFT) PEFT methods [24, 60, 68, 70, 73, 97, 105, 114, 118, 226, 239] aim to adapt pretrained large models to various downstream tasks and have become prevalent in instruction tuning. Typically, these methods involve freezing the pretrained large models while finetuning a minimal set of newly introduced parameters. Recent studies [106, 156, 190, 225] propose to combine PEFT methods with Mixture-of-Experts to mitigate task interference and enhance performance, particularly in visual instruction tuning where models need to process inputs from two modalities. *Our proposed* MoSS *is the first PEFT method that utilizes two distinct LoRA architectures—linear and convolutional—for text and image generation within autoregressive VLGs.*

2.3 Modality-Specialized Synergizers (MoSS)

2.3.1 Background: Autoregressive Vision-Language Generalists

Existing autoregressive pretrained multimodal models can be broadly classified into two categories: those that represent each image as a sequence of *discrete tokens* [1, 174, 215], and those that represent each image as a sequence of *continuous vectors* [167, 171]. However, despite these differences in image representation, their underlying model architectures and formulations for vision-language generation remain largely similar. Thus, we do not differentiate them in the following formulation.

Model Architecture Autoregressive pretrained multimodal models typically comprise three components: an image encoder (e.g., CLIP [169] or VQ-VAE encoder [44]), a decoder-

only large language model (LLM), and an image decoder (e.g., a diffusion model [141] or VQ-VAE decoder [44]). Given a sequence of interleaved text segments and images, the image encoder processes each image into a sequence of image tokens. These image tokens are then concatenated with the text tokens in their original order and input into the LLM. The LLM autoregressively predicts the next token, which could be either text or image. Finally, the image decoder takes in the predicted image tokens and reconstructs the target image.

Training Objective The training objective of pretrained multimodal models can be loosely defined in the following unified autoregressive manner.

$$\sum_{\theta}^{\mathcal{D}} \sum_{n=1}^{N} P_\theta(s_n | s_1, s_2, ..., s_{n-1}) \tag{2.1}$$

where θ denotes the model parameters, N denotes the input sequence length, $\mathcal{D}$ denotes the training dataset, and s_i denotes a text token or an image-patch embedding. This unified objective is optimized through two types of losses: (1) If the image is represented as discrete tokens, the CrossEntropy loss is employed to minimize the divergence between the predicted probability distribution of the image or text tokens and the ground truth distribution; (2) If the image is encoded as continuous vectors, the mean-squared-error (MSE) loss is used to minimize the difference between the predicted and actual image embeddings.

2.3.2 Linear Low-Rank Adaptation (LoRA)

LoRA [60] is a parameter-efficient finetuning method that freezes the pretrained model parameters and injects low-rank decomposable matrices into the layers of transformers. Formally, given the weights in a linear layer $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$, LoRA modifies the weights by adding a decomposable weight matrix $\Delta\mathbf{W}$ to $\mathbf{W}$. Thus, for a vector $\mathbf{h} \in \mathbb{R}^{d_{in}}$, the modified linear transformation $T : \mathbb{R}^{d}_{in} \rightarrow \mathbb{R}^{d}_{out}$ becomes:

$$T(\mathbf{h}) = \mathbf{h}(\mathbf{W} + \Delta\mathbf{W})^\intercal = \mathbf{h}\mathbf{W}^\intercal + \mathbf{h}\Delta\mathbf{W}^\intercal \tag{2.2}$$

$\Delta\mathbf{W}$ is decomposed into two low-rank matrices, i.e., LoRA A: $\mathbf{W}_A \in \mathbb{R}^{r \times d_{in}}$ and LoRA B: $\mathbf{W}_B \in \mathbb{R}^{d_{out} \times r}$ satisfying the low-rank constraint $r \ll min(d_{out}, d_{in})$. The final expression is

$$T(\mathbf{h}) = \mathbf{h}\mathbf{W}^\intercal + \alpha\mathbf{h}\mathbf{W}_A^\intercal\mathbf{W}_B^\intercal \tag{2.3}$$

where $\alpha \in \mathbb{R}$ is a hyper-parameter.

2.3.3 Convolutional Low-Rank Adaptation (Convolutional LoRA)

We propose Convolutional LoRA, a variant of LoRA specifically designed for modeling the local structure of image hidden states during image generation, by improving the architecture

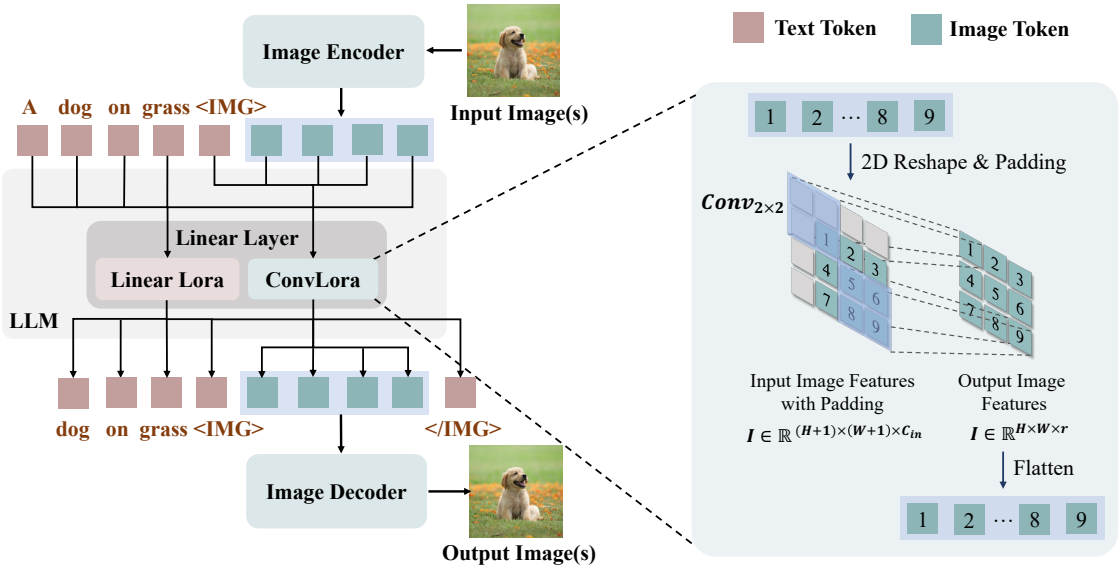


Figure 2.2: An autoregressive VLG with our proposed **MoSS** added to its linear layers. The linear LoRA on the left side is specialized to generate text tokens and the Convolutional LoRA on the right side is specialized to generate image patches. On the right handside, we show the details of convolutional operation applied to autoregressively generate image tokens. Best viewed in color.

proposed in [239]. The previous approach first reduces the dimension of input features and then performs the convolution operation within a lower-dimension space. Since dimension reduction can cause information loss, the convolution within a reduced dimension can be less effective at modeling the local priors of image patches. On the contrary, our method performs convolution in the original input feature space and the dimension is deducted during the convolution process, which alleviates the information loss issue in the previous design.

Specifically, our approach consists of a convolutional LoRA A layer, i.e., $\text{Conv}_{k\times k}$, where the kernel size is $k \times k$, the number of input channels is c_{in}, and the number of output channels is r, as well as a LoRA B: $\mathbf{W}_B \in \mathbb{R}^{C_{out}\times r}$. Given the 2D feature $\mathbf{I} \in \mathbb{R}^{H\times W\times C_{in}}$ of an image, where H denotes the height, W denotes the width, and C_{in} denotes the number of channels of $\mathbf{I}$, the convolutional LoRA A projects down its number of channels to r and simultaneously performs convolution operation. Then the LoRA B projects its number of channels up to C_{out}. The equation 2.3 becomes:

$$\tilde{T}(\mathbf{I}) = \mathbf{I}\mathbf{W}^{\intercal} + \alpha\text{Conv}_{k\times k}(\mathbf{I})\mathbf{W}_B^{\intercal} \tag{2.4}$$

where α is a hyper-parameter.

2.3.4 Integrating MoSS into pretrained multimodal models

As shown in Figure 2.2, we propose to integrate two types of adaptations into pretrained multimodal models, i.e., using **Linear LoRA** for text generation and **Convolutional LoRA** for image generation. Formally, let $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$ be the weights of any linear layer in a LLM, and let $\mathbf{H} = [\mathbf{h}_1^t, ..., \mathbf{h}_m^t, \mathbf{h}_{m+1}^i, \mathbf{h}_{m+2}^i, ..., \mathbf{h}_{m+(H\times W)}^i, ..., \mathbf{h}_N^t] \in \mathbb{R}^{N \times d_{in}}$ denotes the hidden states of a sequence of interleaved text and images, where a subscript indicates position of a hidden state and the superscript indicate if a hidden state is decoded into a text token (t) or decoded into an image-patch embedding (i). We untie $\mathbf{H}$ into text hidden states $\mathbf{H}^t = [\mathbf{h}_1^t, \mathbf{h}_2^t, ..., \mathbf{h}_m^t, \mathbf{h}_{m+(H\times W)+1}^t, ..., \mathbf{h}_N^t]$ and image hidden states $\mathbf{H}^i = [[\mathbf{h}_{m+1}^i, ..., \mathbf{h}_{m+(H\times W)}^i], [\mathbf{h}_{n+1}^i, ..., \mathbf{h}_{n+(H\times W)}^i], ...]$, where $m+1$ and $n+1$ denote the starting positions of two subsequences of image hidden states. Each subsequence of a single image has a fixed length of $H \times W$ and we reshape the hidden states of each image in $\mathbf{H}^i$ into a 2D structure. Hence, the dimension of $\mathbf{H}^i$ becomes $B \times H \times W \times C_{in}$, where B denotes the number of images in the sequence $\mathbf{H}$. We feed $\mathbf{H}^t$ into the Equation 2.3 to get $\hat{\mathbf{H}}^t = T(\mathbf{H}^t)$ and $\mathbf{H}^i$ into Equation 2.4 to get $\hat{\mathbf{H}}^i = \tilde{T}(\mathbf{H}^i)$.

It is non-trivial to integrate convolutional operation in auto-regressive model and to the best of our knowledge, we are the first to incorporate the convolutional architecture to improve interleaved generation. The right part of Figure 2.2 visualizes the convolutional operation applied to a sequence of image patches. The squares on the left denote the reshaped 2-dimensional input image patches and the larger blue squares denote the 2×2 convolution kernels. The number on each square denotes the original positions of a patch in the image sequence. For demonstration purposes, we draw image patches with $H = 3$ and $W = 3$. Note that the current hidden state of an image patch can only depend on previous hidden states since we use the autoregressive architecture. Thus, when applying the convolution operation on an image patch, the kernel only covers neighboring patches on the top and left sides of a patch. For example, the new hidden state of patch 9 is computed from patches: 5, 6, 8, and 9. To preserve the shape ($H \times W$) of the input image patches, we pad the reshaped image hidden states with zero vectors on the top and left sides, as shown by the grey squares in Figure 2.2. Finally, we assemble $\hat{\mathbf{H}}^i$ and $\hat{\mathbf{H}}^t$ back to their original sequence to form $\hat{\mathbf{H}}$.

2.4 LeafInstruct

Existing interleaved vision-language models [34, 167, 171] predominantly follow the training procedures that they are first pretrained on massive corpora of interleaved data such as MMC4 [246] and other resources and then finetuned on a mix of high-quality datasets, such as visual instruction tuning data in [117] and InstructPix2Pix [12]. However, one significant limitation of these instruction-tuning datasets is that the outputs are typically in a single modality, e.g., either text or image, which hinders the instruction-following capability of pretrained multimodal models especially in generating interleaved text and images specified

by the given instructions.

2.4.1 Dataset Construction and Statistics

To bridge the gap between limited existing resources and the practical need for improving interleaved generation models, we curated LeafInstruct, the first comprehensive instruction tuning dataset for interleaved text-and-image generation. Each instance in our dataset consists of (1) a detailed instruction, (2) an input context with interleaved text and images, and (3) a ground-truth output also with interleaved text and images. We show an example of LeafInstruct in Figure 2.3, and compare it with other representative datasets in Table 2.1.

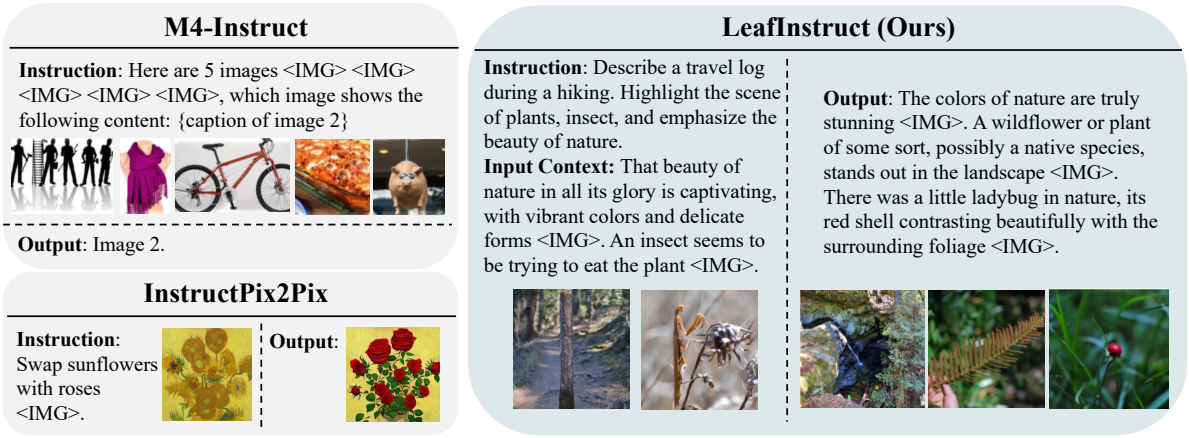


Figure 2.3: Comparison between existing benchmarks and our LeafInstruct. In existing datasets such as InstructPix2Pix [12] and Mantis-Instruct [90], the outputs are in single modality, either text or image. On the contrary, the inputs and outputs of our LeafInstruct cover multiple modalities.

2.4.2 Dataset construction

We construct a diverse instruction-tuning data collection from large-scale web resources and academic datasets, including MMDialog [39], VIST [65], WikiWeb2M [15] and YouCook2 [241]. Since the original data sources can be noisy, we meticulously devised an automatic data annotation pipeline to ensure the high quality of our curated data. We elaborate on the details of our dataset construction pipeline as follows. **Firstly**, we filter the samples based on the text length, number of images, and the coherence between text and images (measured by CLIPScore [56]). We only keep the instances with 3 to 6 images in total. We also discard the instances with more than 12 sentences to ensure a balanced ratio between the number of textual sentences and images. **Secondly**, we leverage a state-of-the-art open-sourced LLM (i.e., Llama-8B-Instruct) as a text filter to discard the instances with poor text quality. **Thirdly**,

we remove the instances with duplicate or perceptually highly similar images to ensure the diversity of the images. **Finally**, we also apply Llama3 to annotate the task instruction for each instance based on the text content and rewrite the text if it's too verbose to prevent the context length from being too long.

Details of Text Quality Filter We use Llama-8B-Instruct model to rate the text quality of an instance with the following prompt: *"Imagine you are an expert data annotator. You are given a text material and you need to evaluate its quality in terms of whether it is coherent, fluent, easy to understand, and helpful to humans. Please be critical and rate the quality as good only when the text quality is good in all four aspects. Output 1 if you think the material is good after you consider all four aspects. Output 0 if you think the material is not good enough. Here is the text material to be evaluated: {TEXT} Only output 0 or 1 and do not output anything else. Your evaluation is:"* We discard the instances if the output from Llama is 0.

Details of Image Filter We empirically found that if the images are too identical in the training instances, the trained models tend to find a shortcut to simply copy the image during generation. To this end, we design a filter to discard the instances with duplicate images to improve data quality. Specifically, we leverage the LPIPS score [230] that measures the perceptual similarity between the images. Specifically, for each instance, we enumerate each pair of images and compute their LPIPS score. If there is one pair with a score higher than 0.6, we discard the instance. We determine the threshold of 0.6 by empirical trial.

Details of Instruction Annotation We also adopt Llama-8B-Instruct to annotate the task instruction for each instance. We devise instructions to prompt the Llama3 model to rewrite the original text material in the pretraining dataset MMC4 into instruction-tuning instances. The input context length is 2048 and the output context length is 1024. We set the temperature as 1 to encourage the diversity of instructions. We use the following prompt: *"Imagine you are an expert instruction annotator. You are given a material. You need to read its content and output a brief task instruction with one sentence such that another person can recover the given the material given the instruction. The instruction you predict should be*

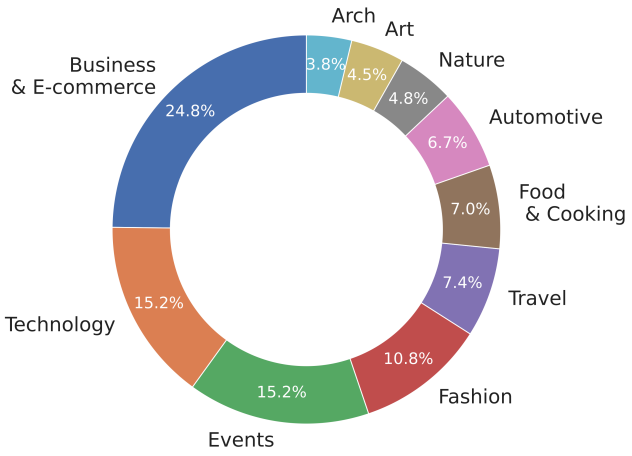


Figure 2.4: Domain distribution in LeafInstruct.

specifically tailored for creative interleaved content generation that consists of both text and images. Now you need to annotate a concise, accurate instruction for the following instance. Please only predict the instruction and do not output anything else. Please design the instruction for the multi-modal generation task interleaved with both text and images. Text: {TEXT} Instruction:".

2.4.3 Dataset Statistics

After applying our rigorous data processing pipeline, we totally obtain 184,982 high-quality instances out of more than 7 million source samples. Our dataset covers a wide range of realistic instruction-tuning tasks, including multimodal document completion, multimodal dialogue, visual storytelling, multimodal script generation, and knowledge-intensive generation. We show the domain distribution of LeafInstruct in Figure 2.4 and compare our dataset with existing datasets in Table 2.1. These analyses effectively demonstrate the diversity and the novelty of our dataset.

Dataset Name	Input Text	Input Images	Output Text	Output Images	Publicly Available
LLaVA [117]	Yes	Single	Single	No	Yes
MultiInstruct [205]	Yes	Single	Single	No	Yes
Vision-Flan [203]	Yes	Single	Single	No	Yes
InstructPix2Pix [13]	Yes	Single	No	Single	Yes
MagicBrush [229]	Yes	Single	No	Single	Yes
SuTI [21]	Yes	Multiple	No	Single	No
Instruct-Imagen [61]	Yes	Multiple	No	Single	No
Mantis-Instruct [69]	Yes	Multiple	Yes	No	Yes
LeafInstruct (Ours)	Yes	Multiple	Yes	Multiple	Yes

Table 2.1: Comparison between our LeafInstruct and existing instruction tuning datasets.

2.4.4 Post-Training for Interleaved Generation

With our curated LeafInstruct, we enable large-scale interleaved instruction tuning so that the model can learn how to follow human instructions to generate desired interleaved text and images. To preserve the VLG's capability obtained from pre-training, we only fine-tuned the modality-specialized adaptation layers, and the remaining parameters in the pretrained multimodal models are kept frozen.

Specifically, as shown on the right side of Figure 2.3, given the task instruction and the interleaved context as inputs, the model is trained to autoregressively generate interleaved text tokens and images with two alternative generation modes for text and images, respectively. We use a special token <IMG> to indicate where an image occurs in the interleaved

sequence. The training process is as follows: **(1)** The model is set to the **text generation** mode by default. During this mode, the hidden states of newly generated tokens are always routed to linear LoRA, and only the parameters in the linear LoRA are optimized. **(2)** After the <IMG> token is generated, the model switches to **image generation** mode. The VLG takes in the updated context ended with <IMG> and is trained to generate a fixed-length ($H \times W$) sequence of image patch embeddings autoregressively. All the hidden states of generated image embeddings are routed to Convolutional LoRA and only the parameters in the Convolutional LoRA are fine-tuned. **(3)** When the generation of an image is finished, the model is trained to predict an end-of-image token </IMG>, and the model will resume the text generation mode. This process will be iterated until the training on a sequence is finished.

Interleaved Inference The inference procedure of our framework is largely identical to the instruction tuning, where two generation processes iterate alternatively. The only key difference is that the fine-tuned pretrained multimodal models will automatically determine when to generate a text segment or an image at their own discretion. The iterative generation process terminates when the model produces the end-of-generation token </s> at the end of a response. Note that although our inference process is designed for interleaved generation, we can also handle the cases where the outputs only contain text or images, enabling a wide range of applications.

2.5 Experiment Setup

Evaluation Benchmarks We evaluate the interleaved generation capability of our method on InterleavedBench [119]. InterleavedBench is a comprehensive dataset specifically tailored for interleaved evaluation. InterleavedBench covers a diverse array of tasks, where the evaluation data are either curated by the authors (e.g., *document completion*), or re-annotated based on subsets of well-established academic evaluation benchmarks, including *visual storytelling* from VIST [65], *activity generation* from ActivityNet [77], *script generation* from WikiHow [213], *image editing* from MagicBrush [229], and *multi-concept image composition* from CustomDiffusion [81].

Evaluation Metrics We adopt InterleavedEval [119], a strong reference-free evaluation metric to conduct a holistic assessment of the quality of interleaved generation. InterleavedEval prompts GPT-4o to score an interleaved output from five aspects, including Text Quality, Perceptual Quality, Image Coherence, Text-Image Coherence (TIC), and Helpfulness. For each aspect, the GPT-4o outputs a discrete score from {0, 1, 2, 3, 4, 5}, where 0 is the worst and 5 is the best. We refer to the original paper [119] for a detailed definition of each score and each evaluation aspect. We also have an additional evaluation on image editing

on the *full test set* of MagicBrush using well-established metrics, including CLIPScore [56] and DINO [16].

Implementation Details To demonstrate the generalizability of our method, we adopt our **MoSS** to two representative autoregressive VLG backbones, i.e., Emu2 [167] and Chameleon [174], and fine-tune them on our **LeafInstruct** dataset. The rank number of all the LoRA is set to 256 by default. Note that for the Chameleon model, we adopt the implementation in [26] since the original model and checkpoints are not publicly available.

Baselines For fair comparisons, we primarily compare our methods with current state-of-the-art **open-source** pretrained multimodal models, including GILL [75], MiniGPT-5 [238], Pretrained Emu2, and Chameleon. We also report the performance of pipelines based on **proprietary** models, including Gemini 1.5 [148]+SDXL [141] and GPT-4o [133]+DALLE 3 [11]. For these baselines, we first prompt the VLMs (e.g., GPT-4o) to generate text along with image captions, and then feed the image captions to a separate image generation model (e.g., DALLE). We report these performances only for reference purposes.

2.6 Results and Discussions

2.6.1 Quantitative Results

Model	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness
Proprietary Models					
Gemini1.5 + SDXL	3.37	4.34	3.34	3.98	3.28
GPT-4o + DALL · E 3	3.16	4.44	3.13	4.39	3.46
Open-Source Models					
MiniGPT-5	1.31	3.44	2.06	2.66	1.76
GILL	1.44	**4.02**	2.12	2.69	1.53
Emu2	1.33	2.29	1.71	1.22	1.87
Chameleon	**3.33**	0.67	0.28	0.47	1.43
Emu2 + **MoSS** (Ours)	2.61 (+96.2%)	3.62 (+58.1%)	**3.41** (+99.4%)	**3.54** (+190.2%)	**2.71** (+44.9%)
Chameleon + **MoSS** (Ours)	2.98 (-10.5%)	2.25 (+235.8%)	1.05 (+275%)	1.7 (+261.7%)	1.82 (+27.3%)

Table 2.2: **Main results of interleaved generation on InterleavedBench**. We show the performance of pipelines based on proprietary models (Top), open-source pretrained multimodal models (Middle), and the pretrained multimodal models trained with our **MoSS** and **LeafInstruct** (Bottom), respectively. Note that the scale is from 0 to 5 (5 is the best). We also report the percentage of improvement in our method over the original VLG backbone in the parentheses. The best results are highlighted in **bold**.

Table 2.2 presents the main results of our method in comparison to the baselines. We have the

following findings. **Firstly**, our approach is highly effective and efficient when it is adapted to existing pretrained multimodal models. Applying our **MoSS** to pretrained multimodal models achieved significant improvement over their original performance on all evaluation aspects. For example, compared with the original Emu2 model, Emu2+**MoSS** achieved a performance gain of **up to 190.2%** (on Text-Image Coherence) and **97.76%** on the average of 5 aspects, almost doubling the overall performance. **Secondly**, our method beats the previous open-sourced state-of-the-art (i.e., GILL) by a large margin, i.e., **34.7%** on the average of 5 aspects. Particularly, the outputs of our method have better coherence across images (w/ 37.8% improvement in Image Coherence) and between text and images (w/ 31.6% improvement in Text-Image Coherence). Our method also exhibits better instruction-following capability and is able to generate more helpful content given the 11.5% improvement in Helpfulness. **Thirdly**, it is worth noting that the Chameleon baseline achieves good performance on Text Quality but extremely poor performance on image-related aspects. We observed that Chameleon usually generates long and comprehensive text responses with no image output, thus leading to poor performance on image-related aspects. We hypothesize the reason lies in the lack of instruction tuning on interleaved generation with both text and images. From Table 2.2, our approach improves the original Chameleon by a significant margin, especially on image-related aspects. This shows that our interleaved instruction tuning can effectively enhance a VLG that was previously poor at mixed-modal generation. **Fourthly**, there remains a notable gap between open-sourced pretrained multimodal models and the pipeline approaches based on proprietary models, indicating building a powerful and general-purpose open-sourced pretrained multimodal models is still challenging.

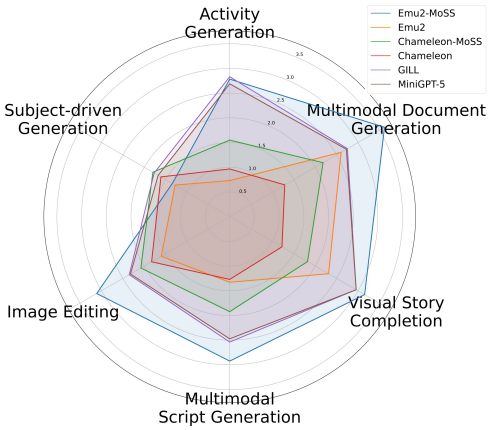


Figure 2.5: **Per-task performance** averaged on 5 aspects on InterleavedBench.

2.6.2 Per-task Performance

We also show the average performance on each task on InterleavedBench in Figure 2.5. Specifically, our method (i.e., Emu2-**MoSS**) outperforms the baselines on most tasks, often by a large margin. For subject-driven generation, the slightly lower performance of our approach compared to other baselines is due to its poorer perceptual quality.

Multimodal Understanding and Text-to-Image Generation To show that our MoSS framework can also excel on tasks requiring single modality outputs i.e., the output only contains text or an image, we evaluate its performance on widely adopted image understanding

benchmarks including MMBench, MME, MMMU, Pope, and MM-Vet, and text-to-image generation benchmarks including MSCOCO 30K [108], and GenEval [47]. For MSCOCO-30K, following the previous evaluation protocol [167], we randomly sample 30,000 captions from the validation set of MSCOCO and generate 30,000 images. We report the FID between the 30,000 generated images and real images from the validation set of MSCOCO (Note for FID, the lower the better). For other benchmarks, we adopt their official implementation of the evaluation.

Model	MMBench	MME	MMMU	Pope	MM-Vet	MSCOCO-30K FID (↓)	GenEval
Chameleon	32.7	604.5	**38.8**	59.8	9.7	26.7	**39.0**
Emu2+LoRA	54.1	1148.0	33.7	87.3	31.3	23.4	26.8
Emu2+MoE-LoRA	54.6	1170.3	34.1	88.1	31.9	22.7	28.1
Emu2+**MoSS**(Ours)	**56.0**	**1278.4**	35.8	**87.6**	**34.1**	**18.2**	28.9

Table 2.3: Results on widely adopted multimodal understanding and text-to-image generation benchmarks. Note that the FID metric on MSCOCO is the lower the better.

Since LeafInstruct mainly targets tasks with interleaved outputs, we augmented it with 500,000 instances from Vision-Flan [203], a popular visual-instruction tuning dataset targeting image understanding, and 500,000 instances from LAION-COCO[1], a standard training dataset for text-to-image generation. We finetune Emu2 with LoRA, MoE-LoRA, and MoSS on the mixed dataset. We report their performance in Table 2.3.

2.6.3 Qualitative Results

To better interpret the results, we conducted a qualitative analysis on several open-sourced baselines and our **MoSS** in Figure 2.6. Our findings are as follows. **Firstly**, our method demonstrates better helpfulness and instruction-following capabilities. For example, in the first row in Figure 2.6, our method generates a more coherent visual story with more diverse content given the input. In the third row, our approach provides a more natural and reasonable next step for the user, i.e., *have a good vegan lunch*, while other baselines either jump to *dinner* (MiniGPT-5) or stick to *breakfast* (Emu2). **Secondly**, we observed that poor text quality is a common issue for many baselines. For instance, MiniGPT-5 often fails to generate explanatory text while GILL usually generates a short caption, e.g., *the gardens* in the first row and *the tofu scramble* in the third row, instead of generating useful content to solve the task. **Thirdly**, neither GILL nor MiniGPT-5 can preserve the visual appearance of the entities and scenes in the input images. Our approach, on the contrary, faithfully retains most visual characteristics, leading to significantly better Image Coherence. **Finally**, as shown in the fourth row in Figure 2.6, for tasks such as image editing or image composition, although MiniGPT-5 and GILL can sometimes generate images with better perceptual

[1] https://huggingface.co/datasets/laion/laion-coco

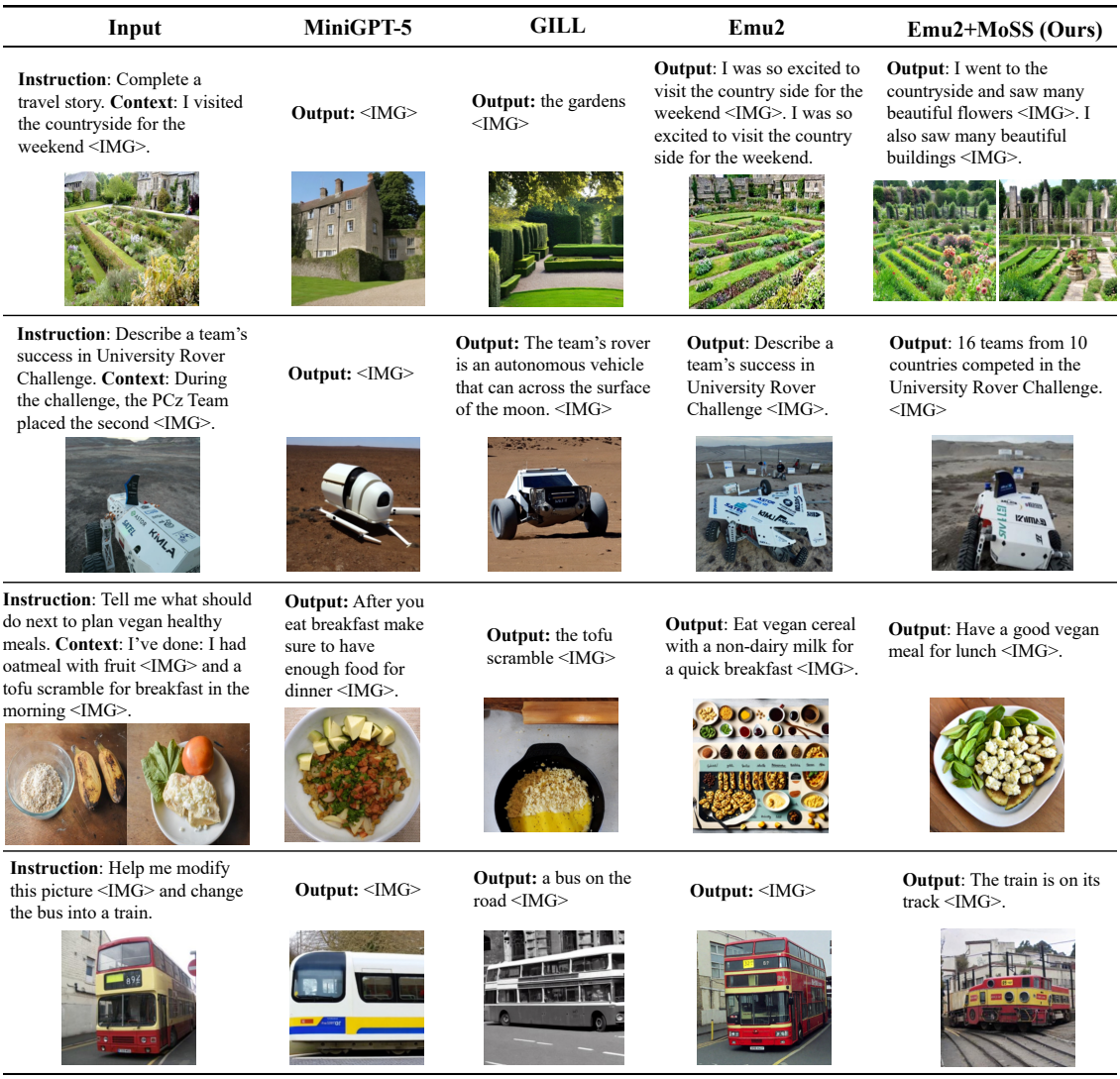


Figure 2.6: **Qualitative results** of **MoSS** based on Emu2 and open-source baselines. The <IMG> tokens denote the images' positions in the interleaved sequences.

quality, the image contents are often irrelevant to the task, ignoring input instructions and context. In contrast, our method strives to adhere to instructions and can better condition its generation on the provided image. Due to the complexity of the task, our model may produce images with lower perceptual quality and noticeable distortions. However, when taking Helpfulness into account, the images generated by our model can be considered as the better ones compared with the baselines.

Model	Text Quality	Image Quality	Image Coherence	TIC	Helpfulness
Emu2	1.33	2.29	1.71	1.22	1.87
+ LoRA	1.77	2.38	1.99	2.04	1.64
+ MoE-LoRA	1.98	3.28	2.66	2.62	2.01
+ **MoSS** (Ours)	**2.61**	**3.62**	**3.41**	**3.54**	**2.71**

Table 2.4: **Comparison between MoSS and existing PEFT methods**, i.e., traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA. Mixture-of-Expert LoRA uses two different sets of linear LoRA for images and text, respectively. The rank number is set to 256 for all methods in this table.

2.6.4 Comparison between MoSS and other PEFT Methods

To directly validate the performance improvement brought by our proposed **MoSS**, we fine-tuned Emu2 using (1) traditional linear LoRA [60] and (2) Mixture-of-Expert (MoE) LoRA [156], with the results presented in Table 2.4. In traditional linear LoRA, text and images share the same low-rank adaptation parameters, while in MoE-LoRA, two different sets of linear LoRA are used for images and text respectively. The routing strategy in MoE-LoRA is based on the output modality of each hidden state, i.e., whether the hidden state is used to generate text or image.

From Table 2.4, we effectively verify the benefits of using separate parameters for image and text. **MoSS** significantly outperforms the MoE-LoRA across all aspects, especially the image-related aspects such as Image Coherence and Text-Image Coherence (TIC). The conclusions from the results are two-fold. First, it shows that introducing modality-specialized architecture and parameters can effectively improve interleaved text-and-image generation. Second, it verifies that convolutional LoRA can improve image generation by better modeling the local priors of images.

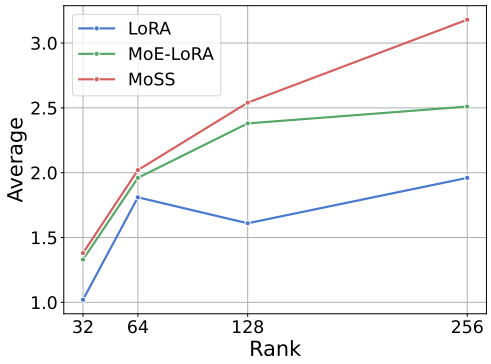


Figure 2.7: Performance averaged on 5 aspects with different rank numbers.

Effect of Rank Number To investigate how the number of rank r can affect the performance, we show the performance averaged on 5 aspects on InterleavedBench with the rank number equals $(32, 64, 128, 256)$ comparing LoRA, MoE-LoRA, and our **MoSS** in Figure 2.7. Our approach consistently outperforms LoRA and MoE-LoRA across all rank numbers, and as the rank number increases, the gap between **MoSS** and previous methods consistently grows larger. This proves the effectiveness and generalizability of **MoSS** across different rank sizes. Based on this experiment, we set the

rank number of our approach to 256 by default.

Quality Assessment of LeafInstruct To verify our LeafInstruct dataset is of high quality, we conduct a rigorous human evaluation using the multi-aspect evaluation criteria in InterleavedEval [119]. Specifically, we use a scale of 0 to 3 in the evaluation, where 0 is the lowest score while 3 is the highest. We randomly sampled 200 instances from LeafInstruct and asked two human annotators with expertise in NLP and multimodal research to rate each instance from 5 aspects. We report the averaged scores from two annotators in Table 2.5. We show that the sampled instances consistently achieved almost full scores across all 5 aspects, which effectively demonstrated that our curated dataset is of high quality.

	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness
Score	2.89	2.96	2.77	2.87	2.71

Table 2.5: **Human evalution** of randomly sampled instances from LeafInstruct. Note that the scale is from 0 to 3 (**Score 3 is the best**), which is different from the scale used in Table 2.2 and Table 2.4.

2.7 Summary

This chapter has presented Modality-Specialized Synergizers, a modality-specialized adaptation framework that advances the capabilities of pretrained multimodal models in interleaved multimodal generation. We introduced MoSS, which dedicates convolutional LoRA layers for modeling local image priors and linear LoRA layers for sequential text processing, thereby addressing the fundamental discrepancy between visual and textual modalities. In addition, we constructed LeafInstruct, the first large-scale instruction-tuning dataset explicitly designed for interleaved text–image generation, comprising 184,982 high-quality instances across more than ten domains. Together, these contributions significantly improve the coherence, fidelity, and instruction adherence of interleaved outputs, establishing new state-of-the-art results among open-source pretrained multimodal models.

Our work provides two key insights. First, architectural unification alone is insufficient: modality-specific inductive biases must be explicitly modeled for high-quality interleaved generation. By assigning specialized adaptation layers to each modality, models achieve stronger representations without sacrificing cross-modal integration. Second, instruction-following ability is not automatically transferable from single-modality training to interleaved tasks. Purpose-built datasets such as LeafInstruct are critical for aligning models with human expectations in multimodal settings.

Despite these advances, several limitations remain. While modality-specialized adaptation enables both image understanding and image generation, due to the architectural limita-

tion of existing pretrained multimodel, **MoSS** still lags behind the specialized model for understanding or generation. Accordingly, Chapter 3 turns to the next stage of this dissertation: the development of efficient unified architectures, culminating in the LaTtE-Flow models. These models integrate autoregressive and diffusion approaches, coupled with layer-wise timestep experts and residual attention, advancing unified multimodal understanding and generation.

Chapter 3

Efficient Unified Architectures

3.1 Motivation

The previous chapter addressed the fundamental challenge of modality unification. While that work made notable progress on interleaved generation through the introduction of expert architectures and interleaved post-training datasets, its performance remained limited by the capacity of the underlying pretrained backbone. As a result, these models still lag behind specialized systems designed exclusively for image understanding or generation. In this chapter, we propose a novel unified architecture that combines autoregressive models with diffusion models, achieving superior performance that even surpasses specialized counterparts.

Unified multimodal foundation models promise a single architecture capable of both image understanding and image generation, thereby advancing general-purpose agents that can reason about and produce multimodal content. However, despite the rapid progress of recent quantization- and diffusion-based approaches, a persistent challenge remains: efficiency. Models that achieve strong performance in one modality often incur trade-offs in the other, and those that balance both typically require prohibitive computational costs. Diffusion- or flow-matching-based frameworks, in particular, demand dozens of full forward passes through large transformer backbones at inference, while autoregressive token-based models suffer from long sequential decoding times for high-resolution images. These inefficiencies severely constrain scalability and real-world deployment.

Motivated by these limitations, we revisit the architectural design space for unified multimodal modeling with a focus on efficiency–quality trade-offs. Our investigation centers on two critical questions: How can we reduce the inference complexity of diffusion-based unified models without degrading generative quality? Can we design lightweight mechanisms that enhance the effective capacity of a small set of transformer layers, ensuring that compact models still achieve competitive performance?

To address the first question, we introduce the Layerwise Timestep Expert architecture, which distributes the flow-matching process across groups of transformer layers, eliminating the need to repeatedly invoke the entire model at every timestep. To address the second, we propose Timestep-Conditioned Residual Attention, enabling later layers to reuse and refine attention maps from earlier layers, conditioned on the current timestep. Together, these in-

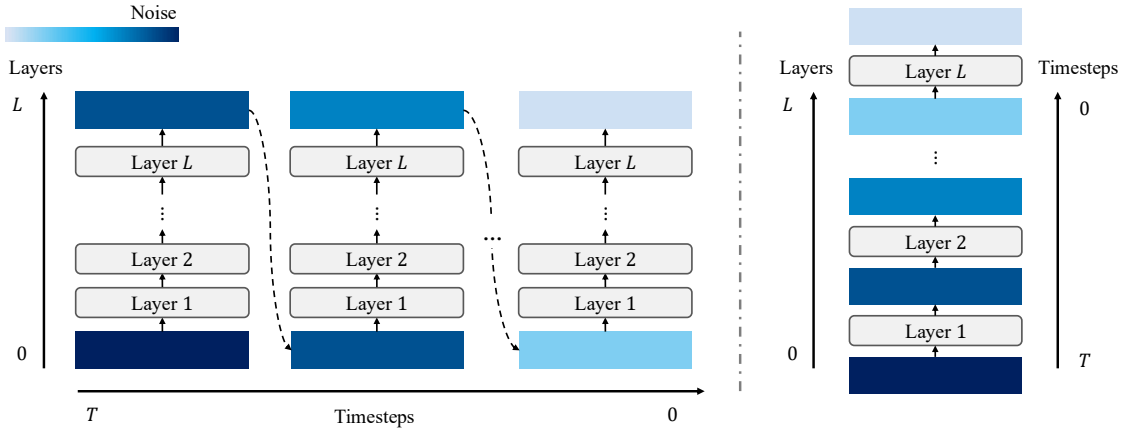


Figure 3.1: **Comparison of the flow-matching process between standard diffusion / flow-matching models and our proposed LaTtE-Flow.** Unlike diffusion / flow-matching based models, which invoke the entire model at each sampling timestep, LaTtE-Flow activates only a subset of layers at each step, improving efficiency.

novations allow LaTtE-Flow to significantly reduce computational overhead while preserving, and in some cases improving, multimodal generation and understanding performance.

3.2 Related Work

Unified Models. Unified multimodal architectures integrate multimodal understanding and generation within a single model, enabling general-purpose agents that can interpret and generate multimodal content in response to user instructions [22, 127, 159, 179, 189, 200, 240]. Existing approaches to unified modeling primarily fall into two categories: The first class of models relies on vector-quantized autoencoders [36, 180, 219] to convert images into discrete token sequences that can be processed similarly to text. These visual tokens are added to the LLM vocabulary to enable unified autoregressive training over both language and vision [22, 168, 189, 197, 199, 200]. The second class incorporates continuous generative processes, most notably diffusion models [57] or flow-matching models [109]. Some approaches connect LLMs with external diffusion modules, using the language model to guide image generation [18, 46, 135, 179, 204], while others directly train LLMs to jointly perform denoising or flow-matching steps [127, 159, 240]. Despite progress in both categories, many of these models suffer from slow image generation speeds, limiting their practical deployment in real-time or resource-constrained settings.

Multiple Experts in Diffusion Models. Recent advancements in diffusion models have increasingly adopted modular or expert-based architectures for better image genera-

tion [158, 166]. Building on this direction, several recent approaches have explored the use of expert models tailored to different diffusion timesteps [37, 85, 248]. By allocating distinct experts to specific temporal intervals, these models aim to better capture the evolving nature of the denoising process. This design is partly motivated by findings from prior work [9, 55], which show that optimization gradients from different timesteps often conflict, leading to slower convergence and degraded model performance. However, these models typically maintain a near full-parameter expert network for different timestep intervals, which leads to little or no improvement in inference efficiency under a fixed number of sampling steps. In contrast, we introduce a layerwise timestep expert architecture, which partitions the transformer layers into different groups of layers, each responsible for a specific range of timesteps. At inference time, only the corresponding group is activated, significantly reducing the number of parameters involved at each step. Moreover, our design allows all expert groups to be trained jointly, and we further integrate it within a unified model architecture, enhancing both efficiency and performance.

3.3 Preliminaries

Flow-Matching. Flow-based generative models [5, 109, 121] aim to learn a time-dependent velocity field $\boldsymbol{v}_t$ that transports samples from a simple source distribution $p_0(\boldsymbol{x})$ (e.g., standard Gaussian) to a complex target distribution $p_1(\boldsymbol{x})$ via an ordinary differential equation (ODE):

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{v}_t(\boldsymbol{x}_t), \quad \boldsymbol{x}_0 \sim p_0(\boldsymbol{x}). \tag{3.1}$$

Recently, [109] propose a simple simulation-free Conditional Flow Matching (CFM) objective by defining a conditional probability path $p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)$ and the corresponding conditional vector field $\boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)$ per sample $\boldsymbol{x}_1$. The model directly regresses the velocity $\boldsymbol{v}_t$ on a conditional vector field $\boldsymbol{u}_t(\cdot \mid \boldsymbol{x}_1)$:

$$\mathbb{E}_{t, p_1(\boldsymbol{x}_1), p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)} \left\| \boldsymbol{v}_t(\boldsymbol{x}_t, t) - \boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1) \right\|^2, \tag{3.2}$$

where $\boldsymbol{u}_t(\cdot \mid \boldsymbol{x}_1)$ uniquely determines a conditional probability path $p_t(\cdot \mid \boldsymbol{x}_1)$ towards target data sample $\boldsymbol{x}_1$. A widely adopted choice for the conditional probability path is linear interpolation between the source and target data [121]: $\boldsymbol{x}_t = t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0$. Assuming the source distribution p_0 is a standard Gaussian, this yields $\boldsymbol{x}_t \sim \mathcal{N}(t\boldsymbol{x}_1, (1-t)^2\boldsymbol{I})$. Sampling from the learned model can be obtained by first sampling $\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{x} \mid 0, 1)$ and then numerically solving the ODE in Eq. (3.1).

3.4 LaTtE-Flow

We present LaTtE-Flow (Layerwise Timestep-Expert Flow-based Transformer), a novel architecture designed for efficient and high-quality image generation and multimodal understanding, unified within a single model. Built on top of pretrained Vision-Language Models (VLMs), LaTtE-Flow leverages their powerful understanding capabilities while introducing additional flow-matching based generation components to enable scalable and effective image synthesis. To unify generation and understanding effectively, we explore two architecture designs: **LaTtE-Flow Couple** and **LaTtE-Flow Blend**, illustrated in Figure 3.2. These variants differ primarily in how the generative and understanding components are combined within the Transformer layers (Section 3.4.1).

Furthermore, we introduce two core architectural innovations applicable to both variants to enhance image generation efficiency and quality: **(1) Layerwise Timestep Experts** (Section 3.4.2), which partition the model into timestep-specialized modules to reduce sampling complexity, and **(2) Timestep-Conditioned Residual Attention** (Section 3.4.3), which injects timestep-aware residual attention into each attention layer through gating mechanisms modulated by a learned timestep embedding, improving training efficiency through effective information reuse across layers.

3.4.1 LaTtE-Flow Layer Design

LaTtE-Flow Couple preserves the pretrained VLM entirely, keeping its parameters frozen (shown in **purple** in Figure 3.2) to retain strong multimodal understanding without finetuning. To enable image generation, it introduces a trainable generative pathway alongside the frozen backbone. Specifically, each Transformer layer is augmented with a trainable replica of the original VLM layer, along with additional components for flow-matching-based generation (shown in **blue** in Figure 3.2). LaTtE-Flow Couple thus allows the model to perform image synthesis while leveraging the robust understanding capabilities of the pretrained VLM.

LaTtE-Flow Blend unifies the image generation and understanding components through a partially shared transformer layer. Here, each layer consists of task-specific submodules with separate parameters for generation and understanding, and a set of shared submodules that are used by both tasks. This design enables tighter fusion between generation and understanding signals, facilitating more effective information exchange while maintaining flexibility to specialize for each modality.

As illustrated in Figure 3.2, both LaTtE-Flow variants introduce a LaTtE-Flow Attention module to enable effective interaction between generative image latents and multimodal context. This attention module employs a hybrid positional encoding scheme, combining the original 3D Rotary Positional Embeddings (RoPE) [164], inherited from the pretrained

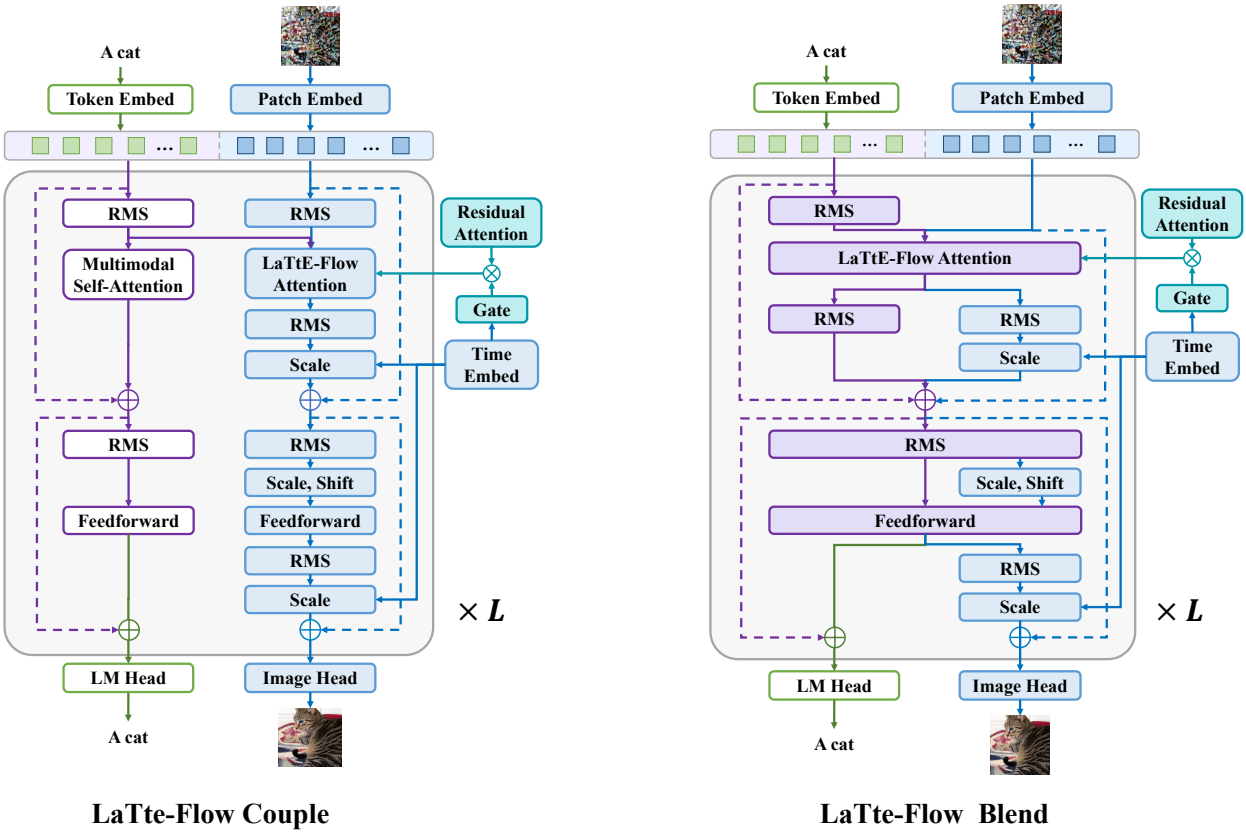


Figure 3.2: **LaTtE-Flow overall architecture**.

VLM, for encoding spatial and temporal structure in the multimodal context, with newly introduced 2D positional encodings applied to the generative image tokens.

3.4.2 Layerwise Timestep Experts

Typical sampling procedures in diffusion models [57, 163] or flow-matching models [5, 109, 121] require repeatedly invoking the full network across a large number of timesteps, leading to slow inference-time speed. For instance, consider a standard diffusion transformer (DiT) model [137] with L transformer layers. The effective computational cost for T sampling steps is $\mathcal{O}(L \times T)$, as shown in Figure 3.1 (a). To alleviate this inefficiency, we introduce a novel Layerwise Timestep Expert architecture, which reduces the effective sampling time complexity by distributing the flow-matching process across groups of transformer layers.

Specifically, instead of executing the entire model at every timestep, we partition the L transformer layers into K non-overlapping groups, where each group specializes in denoising samples within a specific timestep interval, as illustrated in Figure 3.1 (b). This design effectively enables efficient sampling, as only a subset of the network needs to be executed

at each timestep.

Let each expert group be denoted as $\mathcal{G}_k^{l,l+M} = \{l, l+1, \ldots, l+M\}$, consisting of $M = L/K$ consecutive layers (from layer l to layer $l+M$). During training, each layer group learns to predict the velocity field over its assigned timestep interval $[t_k, t_{k+1}]$ using a layerwise flow-matching loss. Specifically, each layer group $\mathcal{G}_k^{l,l+M}$ receives the noisy latent image $\boldsymbol{x}_t \in \mathbb{R}^{N_x \times d}$ along with the multimodal context $\boldsymbol{m}^l$, derived from the preceding layer $l-1$, and predicts the velocity field $\boldsymbol{s}_\theta(\boldsymbol{x}_t, \boldsymbol{m}^l, t)$. Formally, for timestep $t \in [t_k, t_{k+1}]$, the layerwise flow-matching loss is defined as:

$$\mathcal{L}_t = \mathbb{E}_{t, p_1(\boldsymbol{x}_1), p_t(\boldsymbol{x}_t|\boldsymbol{x}_1)} \left\| \mathcal{G}_k^{l,l+M}(\boldsymbol{x}_t, \boldsymbol{m}^l, t) - \boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1) \right\|^2, \quad \text{for } t \in [t_k, t_{k+1}], \tag{3.3}$$

where $\mathcal{G}_k^{l,l+M}(\cdot)$ denotes the prediction produced by the expert group and $\boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)$ is the ground-truth velocity at timestep t. By training each group exclusively on its respective timestep interval, LaTtE-Flow encourages timestep specialization, allowing the model to learn timestep-specific representations across the flow-matching process.

Inference. At inference time with T' sampling steps, we begin by precomputing the multimodal hidden states required for conditioning at each transformer layer. These multimodal representations are computed once at the start of inference and cached for reuse across all timesteps. Then, for each timestep $t \in [t_k, t_{k+1}]$, only the associated expert layer group $\mathcal{G}_k^{l,l+M}$ is activated to perform a forward pass from layer l to layer $l+M$. This process is repeated across all T' timesteps, with only $M = L/K$ layers evaluated per step. Compared to standard diffusion models or flow-matching models that execute all L layers at every step, this design significantly reduces the inference-time complexity from $\mathcal{O}(L \times T')$ to $\mathcal{O}(M \times T')$. This leads to a significant reduction in computational cost and latency during generation, without sacrificing generation quality.

3.4.3 Timestep-Conditioned Residual Attention

To facilitate information reuse across transformer layers and improve both training efficiency and generative performance, we propose Timestep-Conditioned Residual Attention, a novel mechanism that introduces adaptive residual connections between successive image attention layers based on the current timestep. The goal is to enable later layers to reuse and refine the attention patterns computed in earlier layers, while dynamically controlling the influence of past attention through the current flow-matching timestep.

Let $\boldsymbol{A}^l \in \mathbb{R}^{N_x \times N_x}$ image self-attention matrix at layer l, where N_x is the number of image tokens. In a standard self-attention layer, the attention matrix is computed as:

$$\boldsymbol{A} = \text{Softmax}\left(\frac{(\boldsymbol{h}\boldsymbol{W}^Q)(\boldsymbol{h}\boldsymbol{W}^K)^T}{\sqrt{d}}\right), \tag{3.4}$$

where $\boldsymbol{h} \in \mathbb{R}^{N_x \times d}$ denotes the hidden states of the noisy image latents, and $\boldsymbol{W}^Q, \boldsymbol{W}^K \in \mathbb{R}^{d \times d}$ are learnable query and key projection matrices.

To incorporate residual attention from the previous layer, we define the augmented self-attention matrix at layer $l+1$ as:

$$\tilde{\boldsymbol{A}}^{l+1} = \boldsymbol{A}^{l+1} + g(t) \odot \boldsymbol{A}^{l}, \quad g(t) = \tanh(\boldsymbol{h}_t \boldsymbol{W}_t), \tag{3.5}$$

where $\boldsymbol{h}_t \in \mathbb{R}^d$ is the embedding of the current flow-matching timestep t and $\boldsymbol{W}_t \in \mathbb{R}^{d \times H}$ is a trainable projection matrix, with d denoting the hidden dimension and H the number of attention heads. The head-wise gating vector $g(t) \in (-1, 1)^H$, produced by a $\tanh(\cdot)$ activation, dynamically controls the extent to which each attention head incorporates residual attention information from the previous layer. The operator $\odot$ denotes element-wise multiplication, broadcast across all attention heads. Notably, while the LaTtE-Flow Attention module jointly processes both noisy image states and multimodal hidden states, the residual attention mechanism is applied only to the self-attention map over the noisy image hidden states, as shown in Figure 3.3.

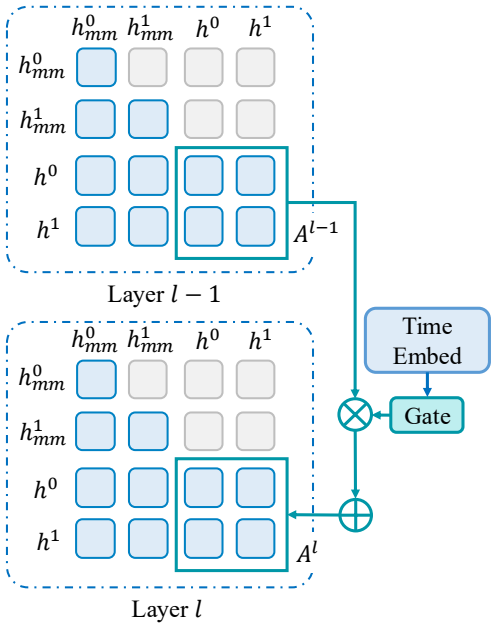


Figure 3.3: **Timestep-conditioned residual attention**

The timestep-conditioned residual attention mechanism enables the model to dynamically control how much residual attention from the previous layer is incorporated into the current layer, on a per-head basis and conditioned on the timestep. Empirically, this design accelerates convergence during training and enhances the quality of generated images.

3.5 Results and Discussion

3.5.1 Image Generation and Understanding Results

We evaluate LaTtE-Flow on both image generation (Table 3.1) and multimodal understanding (Table 3.2) tasks. Table 3.1 reports quantitative comparison between LaTtE-Flow, recent unified models, and leading image generation models. We evaluate each model in terms of generation quality, activated parameters for each inference step, and inference efficiency. All inference times are measured on a single NVIDIA L40 GPU with batch size 50. LaTtE-Flow achieves better FID scores compared to state-of-the-art unified models [22, 197, 200] that are

	Model	FID↓	IS↑	Pre↑	Rec↑	#Params	#Step	Time (s / img)	Rel. Time
Diffusion Models	ADM [33]	10.94	101.0	0.69	0.63	554M	250	9.677	168
	CDM [58]	4.88	158.7	–	–	–	8100	–	
	LDM-4-G [150]	3.60	247.7	–	–	400M	250	–	
	DiT-L/2 [137]	5.02	167.2	0.75	0.57	458M	250	1.786	31
	DiT-XL/2 [137]	2.27	278.2	0.83	0.57	675M	250	2.592	45
Masked Models	MaskGIT [17]	6.18	182.1	0.80	0.51	227M	8	0.029	0.5
	MAGE [96]	6.93	195.8	–	–	230M	–	–	
AR Models	VQVAE-2† [147]	31.11	~45	0.36	0.57	13.5B	5120	–	
	VQGAN† [36]	18.65	80.4	0.78	0.26	227M	256	1.094	19
	VQGAN [36]	15.78	74.3	–	–	1.4B	256	1.382	24
	ViT-VQGAN [219]	4.17	175.1	–	–	1.7B	1024	1.382	24
	RQTran. [84]	7.55	134.0	–	–	3.8B	68	1.210	21
Unified Models	Show-o [200]	31.26	98.7	0.55	0.69	1.3B	50	2.493	48
	Janus Pro [22]	23.68	105.2	0.58	0.49	1.5B	576	0.311	6
	Vanilla Blend (Ours)	6.12	193.7	0.78	0.69	2.0B	40	0.185	4
	LaTtE-Flow Blend (Ours)	6.03	193.9	0.77	0.68	500M	40	0.061	1
	Vanilla Couple (Ours)	6.33	192.4	0.80	0.67	2.0B	40	0.158	3
	LaTtE-Flow Couple (Ours)	5.79	213.1	0.78	0.69	500M	40	0.052	1

Table 3.1: **Comparison of generative models** across FID, IS, Precision, Recall, parameters, steps, and inference time on ImageNet-50K. For LaTtE-Flow, we report the number of parameters activated per timestep, given that it has a timestep-expert architecture where only a subset of layers is used at each step. We also report inference time relative to LaTtE-Flow Couple. †: taken from MaskGIT [17]

Model	MMBench	SEED	POPE	MM-Vet	MME-P	MMMU	RWQA	TEXTVQA
EMU2 Chat 34B [168]	-	62.8	-	48.5	-	34.1	-	66.6
Chameleon 7B [175]	19.8	27.2	19.4	8.3	202.7	22.4	39.0	0.0
Chameleon 34B [175]	32.7	-	59.8	9.7	604.5	38.8	39.2	0.0
Seed-X [46] 17B	70.1	66.5	84.2	43.0	1457.0	35.6	-	-
VILA-U 7B [199]	66.6	57.1	85.8	33.5	1401.8	32.2	46.6	48.3
EMU3 8B [189]	58.5	68.2	85.2	37.2	1243.8	31.6	57.4	64.7
MetaMorph 8B [179]	75.2	71.8	-	-	-	**41.8**	58.3	60.5
Show-o 1.3B [200]	-	-	80.0	-	1097.2	27.4	-	-
Janus 1.5B [197]	69.4	63.7	87.0	34.3	1338.0	30.5	-	-
Janus Pro 1.5B [22]	**75.5**	68.3	86.2	39.8	1444.0	36.3	-	-
LaTtE-Flow Couple 2B	74.9	**72.4**	**87.3**	**51.5**	**1501.4**	41.1	**60.7**	**79.7**

Table 3.2: **Results on comprehensive image understanding benchmarks.** Best scores are highlighted in **bold**. Since our LaTtE-Flow Couple is an expert architecture, we report the number of activated parameters used for image understanding.

pretrained on the mixture of ImageNet and other large-scale image-caption datasets, while achieving mush faster inference speed, i.e., 48× faster than Show-o [200] and 6× faster than Janus Pro [22].

Moreover, both LaTtE-Flow variants outperform their respective baselines, Vanilla Blend and Vanilla Couple, which are conceptually similar to Transfusion [240] and LMFusion [159], with much fewer activated parameters per flow-matching step and 3 to 4× faster inference speed. In addition, LaTtE-Flow exhibits competitive performance compared to diffusion models [33, 58, 137, 150], Masked Models [17, 96] and Auto-regressive (AR) models [36, 84, 147, 219] that are specialized for image generation, achieving better parameter and inference-time efficiency. These results suggest LaTtE-Flow as a promising, efficient, and effective architecture for image generation.

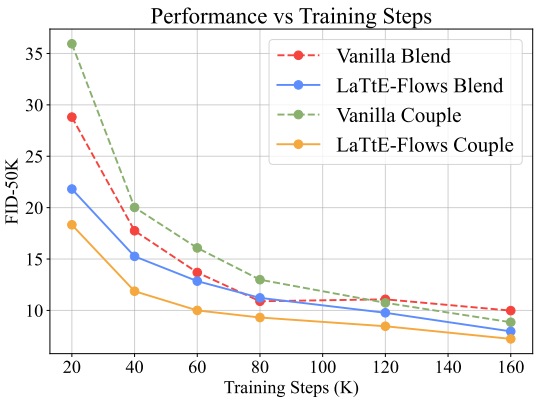


Figure 3.4: **Training dynamics of LaTtE-Flow vs. baselines.** FID on ImageNet 50K.

Table 3.2 presents results on multimodal understanding benchmarks [41, 89, 102, 122, 161, 223, 224]. LaTtE-Flow Couple achieves competitive or superior performance compared to recent unified models, demonstrating its ability to effectively leverage frozen vision-language backbones by inheriting their strong capability without additional finetuning for understanding tasks.

3.5.2 Ablation Studies

Faster Convergence Rate of LaTtE-Flow. Figure 3.4 illustrates the training dynamics of LaTtE-Flow Blend and LaTtE-Flow Couple compared to Vanilla Blend and Vanilla Couple.

We observe that both LaTtE-Flow Blend and LaTtE-Flow Couple exhibit a significantly faster convergence rate during training, reaching competitive image generation performance (lower FID) in fewer training steps. We attribute this favorable property of LaTtE-Flow to the layerwise timestep-expert architecture. As noted in prior work [9, 55], the slow convergence of diffusion models is partially due to the conflicting optimization directions of different timesteps. Optimizing for timesteps that are close can benefit each other, while optimizing timesteps that are far away can interfere with each other. Our layerwise timestep-expert architecture alleviates this challenge by distributing

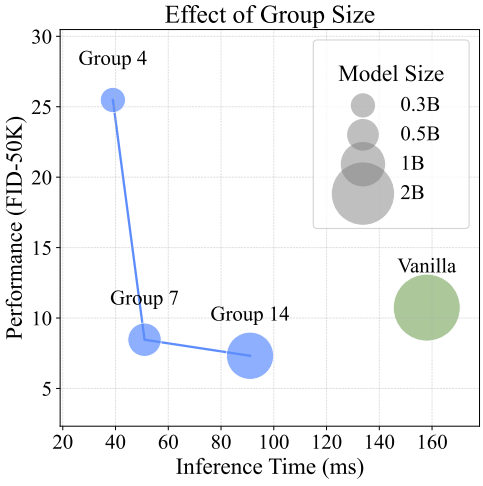


Figure 3.5: **Effect of group size in LaTtE-Flow Couple**.

timesteps across different transformer layers.

Impact of Varying Group Size. We also investigate how the timestep-expert group size M affects the trade-off between generation quality and inference efficiency. Specifically, we train LaTtE-Flow Couple with group sizes $M \in \{4, 7, 14\}$, corresponding to partitioning the transformer layers into 7, 4, and 2 expert groups, respectively. Figure 3.5 reports results at 120K training steps. We observe that larger group sizes consistently improve generation quality, as measured by FID, due to increased modeling capacity. However, this comes at the cost of reduced inference speed, since more layers are executed per timestep. Both $M=7$ and $M=14$ achieve better generation quality and efficiency compared to the baseline Vanilla Couple (Vanilla), which applies all 28 layers at every step. Thus, considering the trade-off between performance and efficiency, we select $M=7$ as the default group size in our main results in Table 3.1, which offers strong generation quality with substantial sampling speedups.

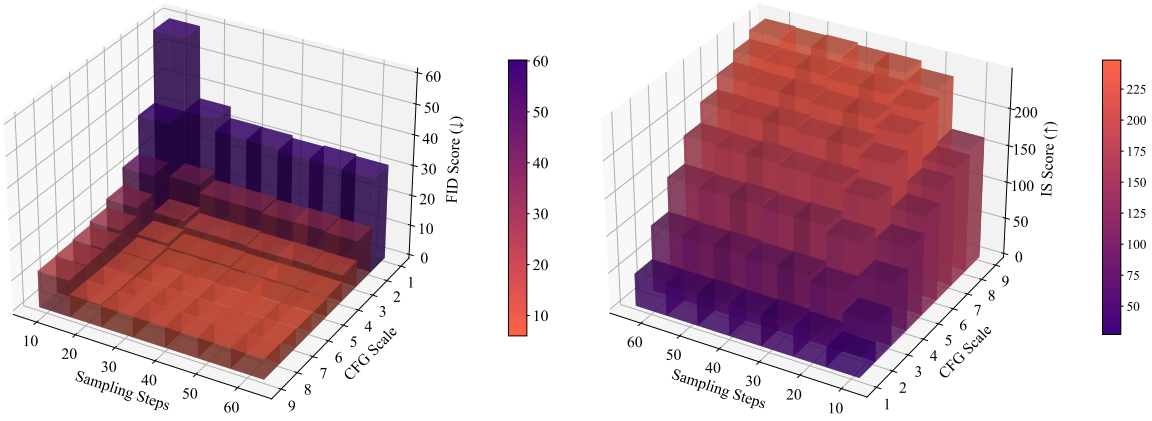
Effect of Timestep-Conditioned Residual Attention. To quantify the effect of timestep-conditioned residual attention, we compare LaTtE-Flow Couple against a variant with the timestep-conditioned residual attention removed. As shown in Table 3.3, removing residual attention leads to a notable degradation across multiple metrics, highlighting the effectiveness of time-conditioned attention across layers. Adding timestep-conditioned residual attention does not introduce additional inference time cost.

Model	FID↓	IS↑	Pre↑	Rec↑
LaTtE-Flow Couple	5.79	213.1	0.78	0.69
- w/o Residual Attention	8.26	157.0	0.75	0.61

Table 3.3: **Effect of time-conditioned residual attention.**

Effect of Sampling Steps and CFG. Figure 3.6 shows the impact of varying the number of sampling steps and classifier-free guidance scale (CFG) on image generation quality. We observe that increasing the number of steps generally improves image generation quality, leading to lower FID and higher Inception Score. However, as the number of sampling steps surpasses 40, performance improvements become marginal. In general, higher CFG leads to better Inception Score, but for FID, once the CFG goes beyond 5, performance starts to decrease slightly.

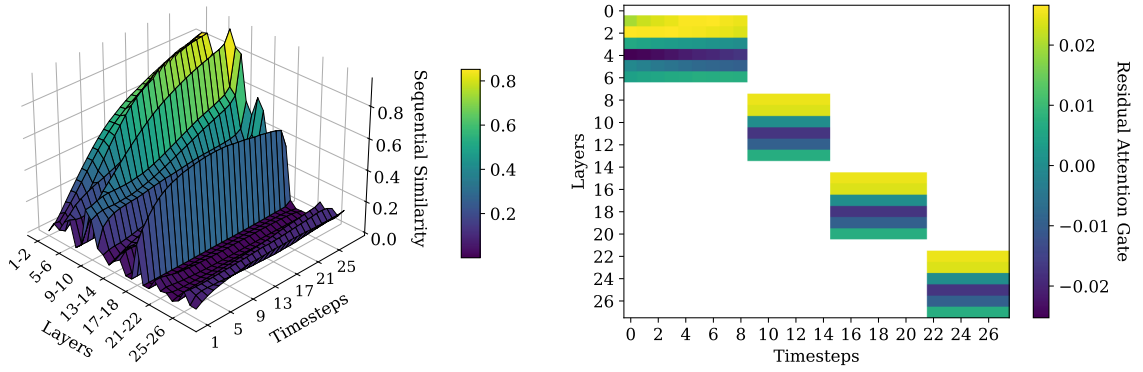
Timestep Condition in Residual Attention. To better understand the role of timestep conditioning in residual attention, we perform an in-depth analysis on both LaTtE-Flow Couple and LaTtE-Flow Blend. Specifically, we first investigate how attention patterns evolve across transformer layers and sampling timesteps in baseline models. We quantify the



(a) FID Score VS. Sampling Steps and CFG

(b) Inception Score vs Sampling Steps and CFG

Figure 3.6: **Impact of # sampling steps and CFG strength on Inception Score and FID.**



(a) Timestep-varying sequential similarity across adjacent transformer layers

(b) Timestep-conditioned residual attention gate for head 5 across transformer layers

Figure 3.7: **Timestep-conditioned residual attention analysis.** (a) Visualization of attention behavior in Vanilla Couple and (b) learned residual gating patterns in LaTtE-Flow Couple.

sequential similarity between adjacent layers at each timestep using a total variation-based metric:

$$S(\boldsymbol{A}^l, \boldsymbol{A}^{l+1}) = 1 - 0.5 \sum_i |\text{Softmax}(\boldsymbol{A}_i^l) - \text{Softmax}(\boldsymbol{A}_i^{l+1})|, \tag{3.6}$$

where $\text{Softmax}(\boldsymbol{A}_i^l)$ is the softmax-normalized i-th row of attention map $\boldsymbol{A}^l$. Higher values of S reflect greater similarity in image attention maps between successive layers.

Figure 3.7 (a) shows how sequential similarity in Vanilla Couple evolves throughout the sampling process, averaged over 100 randomly selected samples. We observe that early in sam-

pling, attention maps across layers show low similarity, but as generation progresses, especially in later timesteps, similarity increases, sometimes approaching 1.0 in early layers. This motivates using residual attention for efficient reuse, with dynamic gating needed to adapt to varying similarity patterns across timesteps. Figure 3.7 (b) shows timestep-conditioned residual attention gates in LaTtE-Flow Couple, which modulate how much past-layer attention is reused. As seen across all heads (Figure 3.8), gating remains stable across timesteps within a head but varies between heads, indicating specialization. These results highlight the effectiveness of dynamic, head-specific residual attention in flow-matching generation.

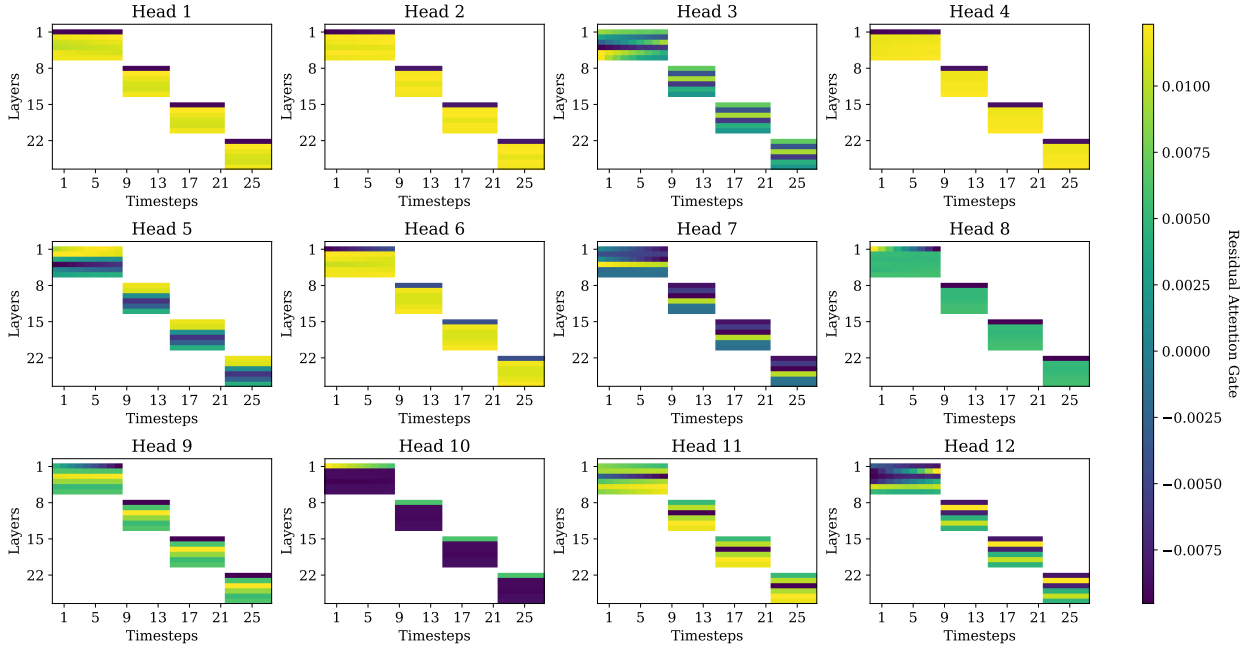


Figure 3.8: **Timestep-conditioned residual attention gates across transformer layer in LaTtE-Flow Couple.** White regions indicate positions without gating values since residual attention is applied only within predefined layer groups. Notably, different heads exhibit distinct gating dynamics, with some emphasizing earlier timesteps, while others modulate more strongly in later layers, suggesting head-specific specialization in residual attention.

3.6 Summary

In this chapter, we presented LaTtE-Flow, a unified multimodal architecture designed to balance efficiency and performance in both image understanding and generation. Unlike prior unified approaches that suffer from excessive inference costs or degraded quality, LaTtE-Flow systematically rethinks the flow-matching process and layer utilization. Specifically, we introduced two complementary innovations: the Layerwise Timestep Expert architecture, which

partitions transformer layers into timestep-specific experts to drastically reduce sampling complexity, and Timestep-Conditioned Residual Attention, a lightweight mechanism that promotes information reuse and faster convergence. Together, these designs enable LaTtE-Flow to achieve competitive accuracy while offering up to 6× faster inference compared to recent unified models.

Building on these architectural insights, LaTtE-Flow demonstrates that efficiency and high-quality multimodal generation are not mutually exclusive. Our experiments confirm that the proposed mechanisms preserve strong performance across both understanding and generation benchmarks, while making unified modeling more computationally feasible for real-world deployment. Beyond benchmark results, LaTtE-Flow also opens new directions for developing practical vision-language agents, where fast, scalable multimodal reasoning and generation are essential.

Overall, this work addresses a fundamental bottleneck in unified multimodal modeling: the high computational overhead of existing designs. However, it does not fully address the issue of generalizability, in particular, the ability to perform zero-shot transfer to unseen tasks by following diverse human instructions. In the next chapter, we propose multimodal instruction tuning, which establishes a post-training paradigm for enhancing zero-shot generalization and robustness in real-world open-world settings.

Part III

Generalizable Multimodal Modeling

Chapter 4

Zero-Shot Multimodal Learning

4.1 Motivation

The previous chapter introduced BLIP3-o, which advanced the architectural design of unified multimodal models by combining autoregressive and diffusion paradigms. While this framework significantly improved performance in both image understanding and generation, it primarily addressed the question of how to architect a unified model. However, BLIP3-o did not tackle an equally critical challenge: generalizability. In real-world applications, multimodal foundation models must perform reliably across diverse, unseen tasks under varying instructions and contexts. Without robust mechanisms to adapt to novel task specifications, even the most powerful unified architectures risk overfitting to their training distributions and failing in open-world environments.

In parallel, research in large-scale pretrained language models (PLMs) has demonstrated the effectiveness of new learning paradigms aimed at improving task generalization [14, 120, 195, 202]. Among these, instruction tuning has proven particularly successful [195]. By fine-tuning PLMs on tasks described through natural language instructions, models learn not only the task itself but also the meta-skill of interpreting instructions, enabling strong zero-shot performance on unseen tasks. This success raises an important question for multimodal learning: can instruction tuning similarly enhance the generalizability of vision–language models (VLMs) across multimodal tasks?

To answer this question, we propose MultiInstruct, the first benchmark dataset for multimodal instruction tuning. MultiInstruct contains 187 diverse tasks spanning broad categories, including Visual Question Answering [48, 165], Commonsense Reasoning [201, 227], and Visual Relationship Understanding [78]. Each task is accompanied by five expert-written instructions, ensuring diversity in phrasing and style. All tasks are reformulated into a unified sequence-to-sequence format, where input text, images, instructions, and bounding boxes are represented in a shared token space, making instruction tuning feasible across modalities.

We adopt OFA[186] as the base multimodal pre-trained model, leveraging its unified sequence-to-sequence architecture, and fine-tune it on MultiInstruct. To further explore cross-domain transfer, we incorporate Natural Instructions[131], a large-scale text-only instruction dataset, via both mixed and sequential tuning strategies. Experimental results demonstrate that instruction tuning substantially enhances zero-shot generalization on unseen multimodal tasks

and that transfer from text-only instruction data further improves performance.

Finally, motivated by the observation that PLMs are often sensitive to variations in instruction wording and length [115, 194], we introduce a new evaluation metric, Sensitivity, which quantifies a model's robustness to instruction variation. Our analysis shows that instruction tuning not only improves zero-shot performance but also reduces sensitivity, particularly when trained on diverse instructions or augmented with large-scale text-only datasets.

Together, these contributions establish multimodal instruction tuning as a principled paradigm for improving the generalizability and robustness of unified multimodal models, complementing the architectural advances from the previous chapter.

4.2 Related Work

Multimodal Pretraining Multimodal pretraining [4, 29, 91, 95, 160, 172, 186] has significantly advanced the vision-language tasks. Several recent studies [29, 124, 186, 188] also started to build a unified pre-training framework to handle a diverse set of cross-modal and unimodal tasks. Among them, VL-T5 [29] tackles vision-and-language tasks with a unified text-generation objective conditioned on multimodal inputs, while OFA [186] further extends it to image generation tasks by using a unified vocabulary for all text and visual tokens. BEIT-3 [188] utilizes a novel shared Multiway Transformer network with a shared self-attention module to align different modalities and provide deep fusion. Building on the success of multimodal pretraining, our work focuses on improving the generalization and zero-shot performance on various unseen multimodal tasks through instruction tuning.

Efficient Language Model Tuning To improve the generalizability and adaptivity of large-scale pre-trained language models, various efficient language model tuning strategies have been proposed recently. Prompt tuning [54, 98, 120, 151, 187] aims to learn a task-specific prompt by reformulating the downstream tasks to the format that the model was initially trained on and has shown competitive performance across various natural language processing applications. As a special form of prompt tuning, in-context learning [130, 202] takes one or a few examples as the prompt to demonstrate the task. Instruction tuning [195] is another simple yet effective strategy to improve the generalizability of large language models. Natural Instructions [131] is a meta-dataset containing diverse tasks with human-authored definitions, things to avoid, and demonstrations. It has shown effectiveness in improving the generalizability of language models even when the size is relatively small (e.g., BART_base) [131, 193]. InstructDial [53] applies instruction tuning to the dialogue domain and shows significant zero-shot performance on unseen dialogue tasks. While these studies have been successful in text-only domains, it has not yet been extensively explored for vision or multimodal tasks.

4.3 MultiInstruct

4.3.1 Multimodal Task and Data Collection

The MultiInstruct dataset is designed to cover a wide range of multimodal tasks that require reasoning among regions, images, and text. These tasks are meant to teach machine learning models to perform various tasks such as object recognition, visual relationship understanding, text-image grounding, and so on by following instructions so that they can perform zero-shot prediction on unseen tasks. To build MultiInstruct, we first collect 34 tasks from the existing studies in visual and multimodal learning, covering Visual Question Answering [48, 66, 78, 129, 162, 247], Commonsense Reasoning [110, 165, 201, 227], Region Understanding [78], Image Understanding [28, 72], Grounded Generation [78, 108, 220], Image-Text Matching [48, 108], Grounded Matching [78, 181, 220], Visual Relationship [78, 139], Temporal Ordering tasks that are created from WikiHow[1], and Miscellaneous [3, 32, 74, 108, 181, 214]. Each of the 34 tasks can be found with one or multiple open-source datasets, which are incorporated into MultiInstruct.

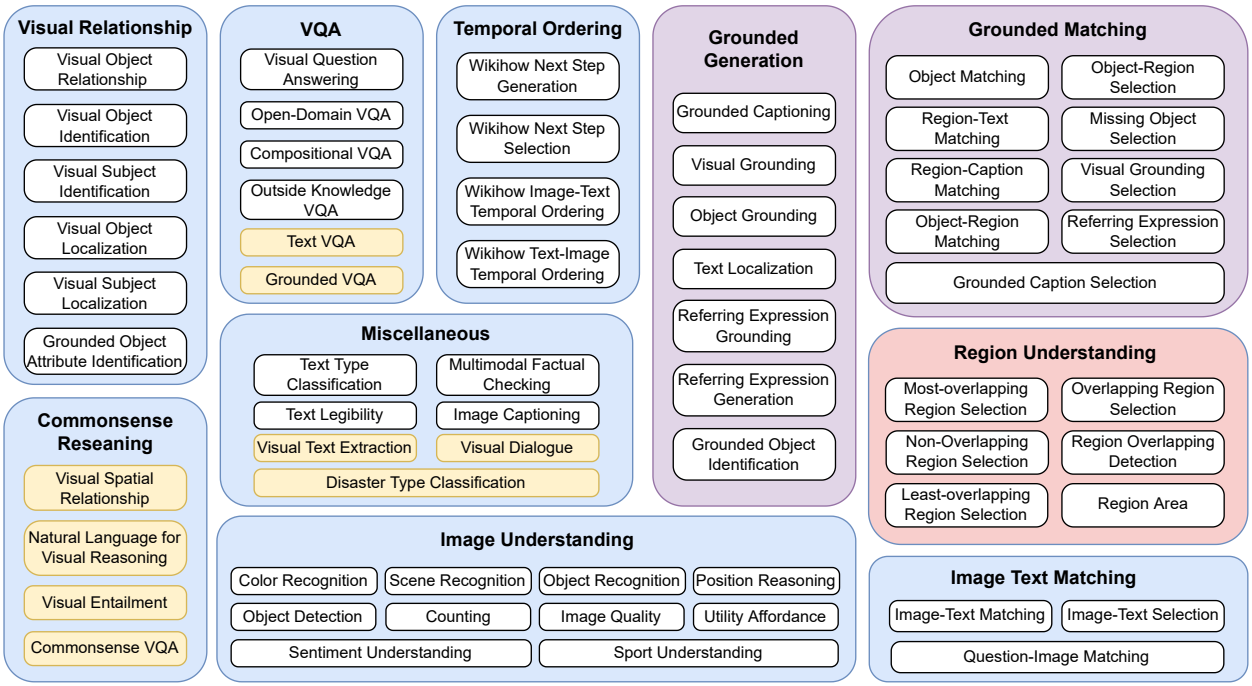


Figure 4.1: **Task Groups Included in MultiInstruct.** The yellow boxes represent tasks used for evaluation, while the white boxes indicate tasks used for training.

For each of these tasks, we further examine the possibility of deriving new tasks based on the input and output of the original task to augment the task repository. For example, *Visual*

[1] https://www.wikihow.com.

Grounding requires the model to generate a caption for a given region in the image. We derive two additional tasks from it: *Grounded Caption Selection*, which is a simpler task that requires the model to select the corresponding caption from multiple candidates for the given region, and *Visual Grounding Selection*, which requires the model to select the corresponding region from the provided candidate regions based on a given caption. Compared with *Visual Grounding*, these two new tasks require different skills based on distinct input and output information. In this way, we further derived 28 new tasks from the 34 existing tasks. We divide all 62 tasks into 10 broad categories as shown in Figure 4.1.

For the existing tasks, we use their available open-source datasets to create instances (i.e., input and output pairs) while for each new task, we create its instances by extracting the necessary information from instances of existing tasks or reformulating them. Each new task is created with 5,000 to 5M instances. We split the 62 tasks into training and evaluation based on the following criteria: (1) we take the tasks that are similar to the pre-training tasks of OFA [186] for training; and (2) we select the challenging multimodal tasks that do not overlap with the training tasks for evaluation.

4.3.2 Task Instruction Creation

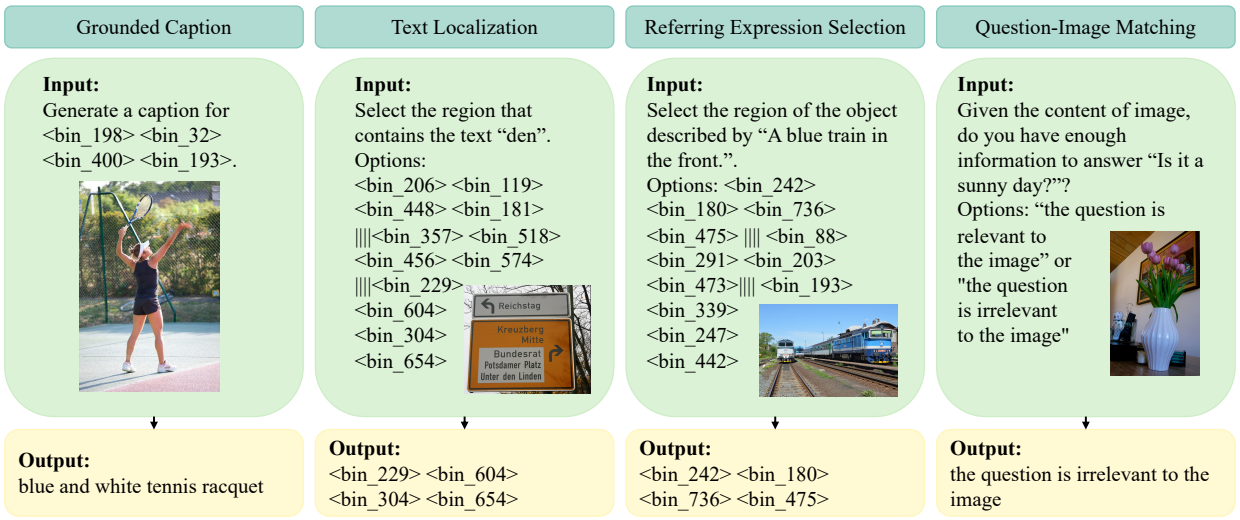


Figure 4.2: **Example Instances from MultiInstruct for Four Tasks.**

We first provide a definition for "*instruction*" used in MultiInstruct. An *instruction* is defined with a template that describes how the task should be performed and contains an arbitrary number of placeholders, including <TEXT>, <REGION> and <OPTION>, for the input information from the original task. For example, in the instruction of the Grounded Captioning task, "Generate a caption for <REGION>", <REGION> is the placeholder for region-specific information. Note that the placeholder <OPTION> is only used in classifi-

cation tasks and for some tasks, the input may also include an image that is not included in the instruction and will be fed as a separate input to the model. Figure 4.2 provides several instruction examples for the tasks included in MultiInstruct.

To produce high-quality instructions that accurately convey the intended tasks, we employ an iterative annotation process involving two expert annotators who have a thorough understanding of the task and the dataset.

Step 1: each annotator first writes 2-3 instructions for each task by giving them the specific goals of this task, the format of input data, and 10 example instances randomly sampled from the dataset. The information about the dataset is obtained from the dataset's README file or the publication that introduced the dataset. For newly derived tasks, we provide annotators with task descriptions along with 10 constructed example instances.

Step 2: to guarantee the quality of the instructions and that they effectively convey the intended tasks, we have each annotator review the instructions created by their peers, checking if they can clearly understand and identify the intended task by just reading the instruction. If any issues are identified, the reviewing annotator provides suggestions and works with the original annotator to revise the instructions.

Step 3: to ensure the consistency and avoid conflicts or repetition among instructions from different annotators, we have both annotators review the sets of instructions together, identifying any discrepancies or inconsistencies. If any are found, the annotators collaborate to resolve them and create a final set of instructions that accurately and clearly describe the task. In this way, each task will be created with 5 high-quality instructions.

Step 4: we repeat steps 1-3 to create 5 instructions for each of the training and evaluation tasks. Finally, both annotators review each task and its instructions and filter out the task that is not representative or overlaps with other tasks.

4.3.3 Multimodal Instruction Formatting

To unify the processing of various input/output data types, we follow the method from OFA [186], which involves representing images, text, and bounding box coordinates as tokens in a unified vocabulary. Specifically, we apply byte-pair encoding (BPE) [154] to encode the text input. For the target image, we apply VQ-GAN [36] to generate discrete image tokens through image quantization. To represent regions or bounding boxes of an image, we discretize the four corner coordinates into location tokens such as "<bin_242> <bin_180> <bin_736> <bin_475>" where each location token "<bin_NUM>" represents a quantized coordinate obtained by dividing the image into 1,000 bins. This approach allows us to convert different types of input into a unified vocabulary.

All tasks in MultiInstruct can then be formulated as natural language sequence-to-sequence generation problems, where the input includes: (1) an image (if there is no input image, a

black picture is used as the input); and (2) an instruction where the placeholders such as <TEXT>, <REGION> or <OPTION> are filled with specific information of each input instance. Notably, for the <OPTION> of the instructions for classification tasks, we introduce two special tokens for this field: "[Options]" to mark the beginning of the option field and "||||" to delimit the given options. We concatenate all the options with "||||" in the option field and the model will directly generate one option from them. Figure 4.2 provides several examples of the formulated input and illustrates how the original data input is combined with the instruction in the MULTIINSTRUCT.

4.4 Problem Setup and Models

4.4.1 Problem Setup

We follow the same instruction tuning setting as the previous study [195] and mainly evaluate the zero-shot learning capabilities of the fine-tuned large language models. Specifically, given a pre-trained multimodal language model M, we aim to finetune it on a collection of instruction tasks T. Each task $t \in T$ is associated with a number of training instances $\mathcal{D}^t = \{(I^t, x_j^t, y_j^t) \in \mathcal{I}^t \times \mathcal{X}^t \times \mathcal{Y}^t\}_{j=1}^N$, where x_j^t denotes the input text, image, region, and options if provided, y_j^t denotes the output of each instance, and I^t represents the set of five task instructions written by experts. The input information from x_j^t will be used to fill in the placeholders in the instruction.

We use OFA [186] as the pre-trained multimodal model due to its unified architecture and flexible input-output modalities. We finetune it on our MULTIINSTRUCT dataset to demonstrate the effectiveness of instruction tuning. Specifically, we use the transformer-based encoder of OFA to encode the instruction along with all necessary information and an optional image, and predict the output with the transformer-based decoder. Given that the training dataset contains many tasks, we mix all the training instances from these tasks and randomly shuffle them. For each instance, we also randomly sample an instruction template for each batch-based training. Note that, though some of the training tasks in MULTIINSTRUCT are similar to the pre-training tasks of OFA[2], we ensure that the evaluation tasks in MULTIINSTRUCT do not overlap with either the pre-training tasks in OFA nor the training tasks in MULTIINSTRUCT.

4.4.2 Transfer Learning from NATURAL INSTRUCTIONS

We notice that the scale of NATURAL INSTRUCTIONS [131] is significantly larger than MULTIINSTRUCT, indicating the potential of transferring the instruction learning capability from

[2]Table **??** in Appendix lists the multimodal tasks and dataset used in OFA pre-training.

the larger set of natural language tasks to multimodal tasks. We take 832 English tasks in Natural Instructions and explore several simple transfer-learning strategies:

Mixed Instruction Tuning (OFA$_{\text{MixedInstruct}}$) We combine the instances of Natural Instructions and MultiInstruct and randomly shuffle them before finetuning OFA with instructions. Note that, each task in Natural Instructions is just associated with one instruction while for each instance from MultiInstruct, we always randomly sample one instruction from the five instructions for each instance of training.

Sequential Instruction Tuning (OFA$_{\text{SeqInstruct}}$) Inspired by the Pre-Finetuning approach discussed in Armen2021prefinetuning, we propose a two-stage sequential instruction tuning strategy where we first fine-tune OFA on the Natural Instructions dataset to encourage the model to follow instructions to perform language-only tasks, and then further fine-tune it on MultiInstruct to adapt the instruction learning capability to multimodal tasks. To maximize the effectiveness of the Natural Instructions dataset, we use all instances in English-language tasks to tune the model in the first training stage.

4.5 Experimental Setup

Evaluation Metrics We report the accuracy for classification tasks and ROUGE-L [107] for all generation tasks. For the region classification task, we compute the Intersection over Union (IoU) between the generated region and all regions in the options, select the option with the highest IoU as the prediction, and compute accuracy based on this prediction. If the predicted region has no intersection with any of the regions in the options, we treat this prediction as incorrect. For classification tasks where the answer is not a single-word binary classification, we also report ROUGE-L scores following [131], which treats all tasks as text generation problems. For each task, we conduct five experiments by evaluating the model using one of the five instructions in each experiment. We report the mean and maximum performance and the standard deviation of the performance across all five experiments. We also compute the *aggregated performance* for each model based on the mean of the model's performance on all multimodal and NLP unseen tasks. We use Rouge-L as the evaluation metric for most tasks and accuracy for tasks that only have accuracy as a metric.

In addition, as instruction tuning mainly relies on the instructions to guide the model to perform prediction on various unseen multimodal tasks, we further propose to evaluate how sensitive the model is to the variety of human-written instructions in the same task, which has not been discussed in previous instruction tuning studies but is necessary to understand the effectiveness of instruction tuning. We thus further design a new metric as follows:

Sensitivity refers to the model's capability of consistently producing the same results, re-

gardless of slight variations in the wording of instructions, as long as the intended task remains the same. Specifically, for each task $t \in T$, given its associated instances with task instructions: $\mathcal{D}^t = \{(I^t, x_j^t, y_j^t) \in \mathcal{I}^t \times \mathcal{X}^t \times \mathcal{Y}^t\}_{j=1}^N$, we formally define *sensitivity* as:

$$\mathbb{E}_{t \in T}[\frac{\sigma_{i \in I^t}[\mathbb{E}_{(x,y) \in \mathcal{D}^t}[\mathcal{L}(f_\theta(i,x),y)]]}{\mu_{i \in I^t}[\mathbb{E}_{(x,y) \in \mathcal{D}^t}[\mathcal{L}(f_\theta(i,x),y)]]}]$$

where $\mathcal{L}$ denotes the evaluation metric such as accuracy or ROUGE-L, $f_\theta(\cdot)$ represents the multimodal instruction-tuned model. The standard deviation and mean of the model's performance across all instructions are denoted by $\sigma_{i \in I^t}[\cdot]$ and $\mu_{i \in I^t}[\cdot]$, respectively.

Evaluation datasets We evaluate the models on nine unseen multimodal tasks: Text VQA [162], Grounded VQA [247], Commonsense VQA [227], Visual Entailment [201], Visual Spatial Reasoning [110], Natural Language for Visual Reasoning (NLVR) [165], Visual Text Extraction [74], Visual Dialogue [32], and Disaster Type Classification [3]. These tasks belong to three task groups: Commonsense Reasoning, VQA, and Miscellaneous as shown in Figure 4.1. Tasks in the Commonsense Reasoning group have no overlap with any training task groups. Tasks in Miscellaneous do not share similarities with other tasks in the group. Although Text VQA and Grounded VQA belong to the VQA task group, they require additional skills such as extracting text from images or generating regions, making them fundamentally different from other tasks in VQA. In addition to multimodal tasks, we also evaluate the model on 20 NLP tasks collected from the test split of Natural Instructions.

Approaches for Comparison We denote the OFA finetuned on MultiInstruct as **OFA$_{\text{MultiInstruct}}$**, and compare it with the original pre-trained **OFA**[3], **OFA$_{\text{TaskName}}$** which is fine-tuned on MultiInstruct but uses the task name instead of instruction to guide the model to make predictions, and several approaches that leverage the large-scale Natural Instructions dataset, including **OFA$_{\text{NaturalInstruct}}$** which only fine-tunes OFA on Natural Instructions with instruction tuning, **OFA$_{\text{MixedInstruct}}$** and **OFA$_{\text{SeqInstruct}}$** that are specified in Section 4.4.2.

4.6 Results and Discussion

4.6.1 Effectiveness of Instruction Tuning on MultiInstruct

We evaluate the zero-shot performance of various approaches on all the unseen evaluation tasks, as shown in Table 4.1 and 4.2. Our results indicate that OFA$_{\text{MultiInstruct}}$ significantly improves the model's zero-short performance over the original pre-trained OFA model across

[3] https://ofa-beijing.oss-cn-beijing.aliyuncs.com/checkpoints/ofa_large.pt

Model	Commonsense VQA				Visual Entailment		Visual Spatial Reasoning		NLVR	
	RougeL		ACC		ACC		ACC		ACC	
	Max	Avg ± Std	Max	Avg ± Std	Max	Avg± Std	Max	Avg± Std	Max	Avg± Std
OFA	17.93	14.97 ± 4.30	0.73	0.40 ±0.29	49.99	41.86 ± 10.99	54.99	35.29 ± 22.21	56.06	52.10 ± 3.35
$\text{OFA}_{\text{TaskName}}$	48.99	-	29.01	-	55.70	-	53.76	-	55.35	-
$\text{OFA}_{\text{MultiInstruct}}$	**52.01**	**50.60** ± 1.12	**33.01**	31.17 ± 1.59	**55.96**	**55.06** ±0.76	**55.81**	**53.90** ±1.38	56.97	56.18 ± 0.95
Transfer Learning from Natural Instructions										
$\text{OFA}_{\text{NaturalInstruct}}$	27.15	14.99 ± 9.12	7.35	2.04 ± 3.01	33.28	14.86 ± 16.68	51.44	36.44 ± 20.72	56.06	35.98 ± 21.64
$\text{OFA}_{\text{MixedInstruct}}$	50.40	49.34 ± 1.04	31.31	30.27 ± 0.94	54.63	53.74 ± 0.97	55.13	52.61 ± 1.64	56.67	55.96 ± 0.48
$\text{OFA}_{\text{SeqInstruct}}$	50.93	50.07 ± 1.07	32.28	**31.23** ± 1.09	53.66	52.98 ± 0.56	54.86	53.11 ± 1.45	**57.58**	**56.63** ± 0.66

Table 4.1: **Zero-shot Performance on Multimodal Commonsense Reasoning.** The best performance is in **bold**.

Model	Text VQA		Grounded VQA		Visual Text Extraction		Visual Dialogue		Disaster Type Classification	
	RougeL		Acc		RougeL		RougeL		ACC	
	Max	Avg± Std	Max	Avg± Std	Max	Avg± Std	Max	Avg ± Std	Max	Avg ± Std
OFA	15.21	9.30 ± 5.42	0.02	0.00 ± 0.01	36.31	17.62 ± 16.82	45.46	28.71 ± 9.81	14.30	9.64 ± 4.34
$\text{OFA}_{\text{TaskName}}$	23.80	-	0.00	-	36.30	-	25.18	-	62.65	-
$\text{OFA}_{\text{MultiInstruct}}$	**27.22**	26.46 ± 0.83	**64.32**	47.22 ± 23.08	**74.35**	**62.43** ±11.56	**46.38**	32.91 ±7.59	64.88	56.00 ±12.96
Transfer Learning from Natural Instructions										
$\text{OFA}_{\text{NaturalInstruct}}$	5.59	5.40 ± 0.24	0.00	0.00 ± 0.00	5.65	1.24 ± 2.48	30.94	27.91 ± 2.16	56.64	38.21 ± 15.35
$\text{OFA}_{\text{MixedInstruct}}$	24.15	23.67 ± 0.47	63.79	**54.99** ± 18.16	62.43	46.56 ± 14.92	46.08	**38.02** ± 5.25	**68.31**	**64.31** ± 2.39
$\text{OFA}_{\text{SeqInstruct}}$	27.03	**26.67** ± 0.47	64.19	54.46 ± 15.96	71.63	60.62 ± 12.31	46.17	35.10 ± 6.92	64.46	57.89 ± 9.51

Table 4.2: **Zero-shot Performance on Question Answering and Miscellaneous.** The best performance is in **bold**.

all unseen tasks and metrics, demonstrating the effectiveness of multimodal instruction tuning on MultiInstruct. As seen in Table 4.2, OFA achieves extremely low (nearly zero) zero-shot performance on the Grounded VQA task, which requires the model to generate region-specific tokens in order to answer the question. By examining the generated results, we find that OFA, without instruction tuning, failed to follow the instruction and produce results that contain region tokens. However, by fine-tuning OFA on MultiInstruct, the model is able to better interpret and follow the instructions to properly generate the expected output. Additionally, $\text{OFA}_{\text{MultiInstruct}}$ outperforms $\text{OFA}_{\text{TaskName}}$ on all unseen tasks, particularly on the Grounded VQA task, where $\text{OFA}_{\text{TaskName}}$ achieves nearly zero performance. This suggests that the performance gain of $\text{OFA}_{\text{MultiInstruct}}$ mainly comes from instructions rather than multi-task training.

4.6.2 Impact of Transfer Learning from Natural Instructions

One key question in multimodal instruction tuning is how to effectively leverage the large-scale text-only Natural Instructions dataset to enhance the zero-shot performance on multimodal tasks. We observe that only fine-tuning OFA on Natural Instructions actually degrades the model's zero-shot performance on almost all multimodal tasks, as shown by comparing $\text{OFA}_{\text{NaturalInstruct}}$ and OFA in Table 4.1 and 4.2. One potential reason for this decline in performance is that during fine-tuning on the text-only dataset, the model learns

to focus more on text tokens and attend less to image tokens. To verify this assumption, we compare the attention of text tokens on image tokens between $\text{OFA}_{\text{NaturalInstruct}}$ and other methods and observe that text tokens attend much less to image tokens after fine-tuning on the Natural Instructions dataset.

Another observation is that although our transfer learning methods do not lead to significant performance gains over $\text{OFA}_{\text{MixedInstruct}}$, both $\text{OFA}_{\text{SeqInstruct}}$ and $\text{OFA}_{\text{MixedInstruct}}$ achieve lower standard deviation on 6 out of 9 unseen multimodal tasks compared with $\text{OFA}_{\text{MultiInstruct}}$, demonstrating the potential benefits of the much larger text-only instruction datasets to multimodal instruction tuning.

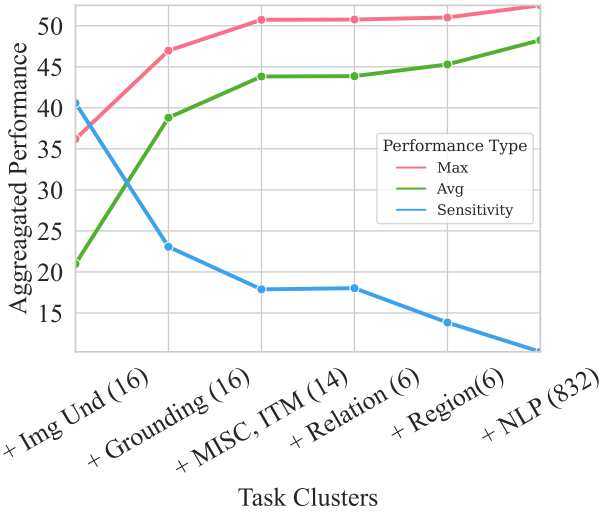


Figure 4.3: **Model Performance as the Number of Multimodal Instruction Task Clusters Increases.** The number in the parenthesis of each cluster denotes the number of tasks.

4.6.3 Impact of Increasing Multimodal Instruction Task Clusters

To evaluate the impact of the number of tasks clusters for instruction tuning, we start with the task groups shown in Figure 4.1 and group them into five larger clusters: (1) Img Und (VQA + Image Understanding), (2) Grounding (Grounded Matching + Grounded Generation), (3) MISC, ITM (Temporal Ordering + Miscellaneous + Image Text Matching), (4) Relation (Visual Relationship), (5) Region (Region Understanding), together with (6) NLP, a collection of NLP tasks from Natural Instructions. We measure the change in both the aggregated performance and *sensitivity* of $\text{OFA}_{\text{MixedInstruct}}$ as we gradually add the task clusters for training.

As we increase the number of task clusters, we observe an improvement in both the mean and maximum aggregated performance and a decrease in *sensitivity*, as shown in Figure 4.3.

Note that low *sensitivity* indicates that the model can produce consistent results despite variations in the wording of instructions. These results suggest that increasing the number of task clusters improves the model's performance on unseen tasks and leads to more consistent outputs. The results also support the effectiveness of our proposed MULTIINSTRUCT dataset.

# of Instructions	Aggregated Performance ↑	*Sensitivity* ↓
1 Instruction	42.81	24.62
5 Instructions	**47.82**	**10.45**

Table 4.3: **Effect of Different Number of Instructions.** Performance of $\text{OFA}_{\text{MultiInstruct}}$ finetuned on different numbers of instructions.

4.6.4 Effect of Diverse Instructions on Instruction Tuning

We hypothesize that using a diverse set of instructions for each task during multimodal instruction tuning can improve the model's zero-shot performance on unseen tasks and reduce its *sensitivity* to variation in the instructions. To test this hypothesis, we train an OFA model on MULTIINSTRUCT with a single fixed instruction template per task and compare its performance with OFA finetuned on 5 different instructions. As shown in Table 4.3, OFA finetuned on 5 instructions achieves much higher aggregated performance on all evaluation tasks and shows lower *sensitivity*. These results demonstrate the effectiveness of increasing the diversity of instructions and suggest that future work could explore crowd-sourcing or automatic generation strategies to create even more diverse instructions for instruction tuning.

4.6.5 Effect of Fine-tuning Strategies on Model *Sensitivity*

In Section 4.6.3 and 4.6.4, we have shown that the more tasks and instructions used for instruction tuning, the lower *sensitivity* the model will achieve toward the variations in instructions for each task. We further investigate the impact of fine-tuning and transfer learning strategies on model sensitivity. Figure 4.4 shows the averaged *sensitivity* of each model across all multimodal unseen tasks.

The original OFA exhibits significantly higher sensitivity to variations in instructions compared to models fine-tuned on instruction datasets, indicating that multimodal instruction tuning significantly improves the model's capability on interpreting instructions, even with varying wordings. In addition, by transferring the large-scale NATURAL INSTRUCTIONS dataset to MULTIINSTRUCT, *sensitivity* is also reduced by a large margin, highlighting the benefit of fine-tuning the model on a larger instruction dataset, regardless of different formats and modalities.

4.6.6 Zero-Shot Performance on NLP Tasks

So far, our focus has been on evaluating the zero-shot performance of multimodal tasks. In this section, we investigate the effect of multimodal instruction tuning on the performance of text-only tasks. To do this, we evaluate all our approaches on 20 natural language processing (NLP) tasks from the default test split in Natural Instructions[4].

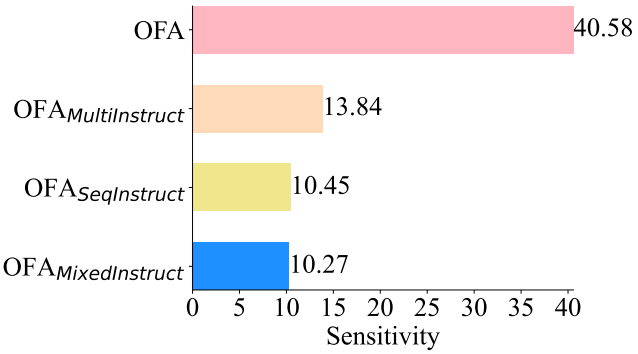


Figure 4.4: **Model *Sensitivity* on Unseen Evaluation Tasks.** Lower is better.

As shown in Table 4.4, OFA$_{\text{MultiInstruct}}$ outperforms OFA, despite the instruction tuning dataset and the unseen dataset are in different modalities. This suggests that multimodal instruction tuning can help improve the zero-shot performance on NLP tasks. In addition, we observe that OFA$_{\text{NaturalInstruct}}$ achieves the best performance on NLP tasks and OFA$_{\text{MixedInstruct}}$ is more effective in preserving the zero-shot capability gained from Natural Instructions on NLP tasks compared to OFA$_{\text{SeqInstruct}}$. Based on the results in Tables 4.1, 4.2 and 4.4, we conclude that OFA$_{\text{MixedInstruct}}$ is able to achieve overall best aggregated performance on all multimodal and NLP tasks and shows much lower *sensitivity* towards variations in the wording of instructions, making it the most promising approach.

Model	RougeL
OFA	2.25
OFA$_{\text{MultiInstruct}}$	12.18
Transfer Learning from Natural Instructions	
OFA$_{\text{NaturalInstruct}}$	**43.61**
OFA$_{\text{MixedInstruct}}$	43.32
OFA$_{\text{SeqInstruct}}$	30.79

Table 4.4: **Zero-shot Performance on NLP tasks.** The performance is reported in Rouge-L and the best performance is in **bold**.

4.7 Summary

In this chapter, we introduced MultiInstruct, the first large-scale multimodal instruction tuning benchmark designed to enhance the generalizability of vision–language models. The

[4] https://github.com/allenai/natural-instructions

dataset covers a broad range of multimodal tasks, each paired with multiple expert-written instructions to promote robustness across varying task formulations. By fine-tuning OFA [186] on MultiInstruct, we demonstrated substantial improvements in zero-shot performance on unseen multimodal tasks, establishing instruction tuning as an effective paradigm for multimodal generalization. Furthermore, we explored transfer learning techniques leveraging the large-scale text-only Natural Instructions dataset and showed that these strategies further strengthen model performance. To complement these contributions, we proposed a novel evaluation metric, Sensitivity, which quantifies robustness to instruction variations. Empirical results revealed that instruction tuning not only improves task generalization but also significantly reduces instruction sensitivity, particularly when training on a diverse set of tasks and instructions.

Despite these advances, MultiInstruct has notable limitations. First, the diversity of tasks, though broader than prior benchmarks, remains limited relative to the open-ended complexity of real-world multimodal applications. It remains unclear whether models can achieve even stronger generalization if trained on a wider spectrum of task types and domains. Second, the benchmark primarily focuses on task coverage and instruction variety but does not explicitly consider human preference alignment, an increasingly important aspect for real-world deployment. These limitations motivate the next chapter, where we introduce Vision-Flan, a scaled multimodal instruction tuning dataset that expands task diversity and incorporates strategies to align model behavior with human-preferred outputs.

Chapter 5

Scaling Task Diversity for Robust Generalization

5.1 Motivation

The previous chapter introduced MultiInstruct, which established multimodal instruction tuning as a principled paradigm to improve the generalizability of unified multimodal models. By curating a benchmark of diverse multimodal tasks and expert-written instructions, MultiInstruct significantly enhanced zero-shot performance and reduced sensitivity to instruction variations. However, its coverage of tasks remained limited, and the dataset did not explicitly consider alignment with human preferences. These limitations raise two key questions: (1) can the generalization ability of vision–language models (VLMs) be further improved by scaling task diversity beyond what MultiInstruct provides? and (2) how can we simultaneously preserve broad task competence while aligning model behavior with human-preferred responses?

Recent vision–language models (VLMs)[31, 93, 117], built upon large language models (LLMs) [27, 45] and pretrained image encoders [170], have emerged as powerful general visual assistants. These frameworks typically consist of three components: (1) a bridging module (e.g., MLP layers in LLaVA [93, 117]) that connects image encoders to LLMs, (2) large-scale image–text pairs [152] for pretraining, and (3) GPT-4 synthesized instruction datasets [87, 117] for aligning model outputs with human preferences. Despite notable progress, two significant challenges remain unaddressed.

First, limited task diversity during pretraining and instruction tuning restricts generalization. Because the majority of pretraining data is dominated by captioning tasks, VLMs often fail on tasks such as OCR or specialized reasoning [231], where relevant training instances are absent. Although recent efforts extend coverage by repurposing datasets for new tasks [64, 116, 231], their scope remains narrow compared to the full range of real-world visual tasks.

Second, most existing visual instruction tuning datasets [87, 117, 218] rely heavily on synthetic GPT-4 annotations generated from captions or dense descriptions. While this strategy produces fluent, human-like responses, it introduces spurious correlations, limited task grounding, and long-form outputs that increase hallucination risks [103, 111, 113, 243]. Moreover, models fine-tuned exclusively on synthetic data often exhibit catastrophic forgetting,

losing competence in basic perception tasks such as classification on MNIST [83] or CIFAR-10 [79].

To address these challenges, we introduce VISION-FLAN, the most diverse publicly available visual instruction tuning dataset to date. Vision-Flan consists of 187 tasks spanning a wide range of domains: perception tasks such as object detection and OCR, domain-specific tasks such as style and quality classification, and complex reasoning tasks such as graph interpretation and geometric problem solving. Each task is paired with expert-written instructions, ensuring high-quality task grounding beyond synthetic augmentation. We show some examples of VISION-FLAN in Figure 5.1. Building on this dataset, we propose a two-stage instruction tuning framework: in Stage 1, models are fine-tuned on Vision-Flan to acquire broad and diverse capabilities; in Stage 2, a small amount of GPT-4 synthesized data is used to refine the model's alignment with human preferences. This design balances diversity-driven generalization with efficient human preference alignment, while reducing reliance on large-scale synthetic data.

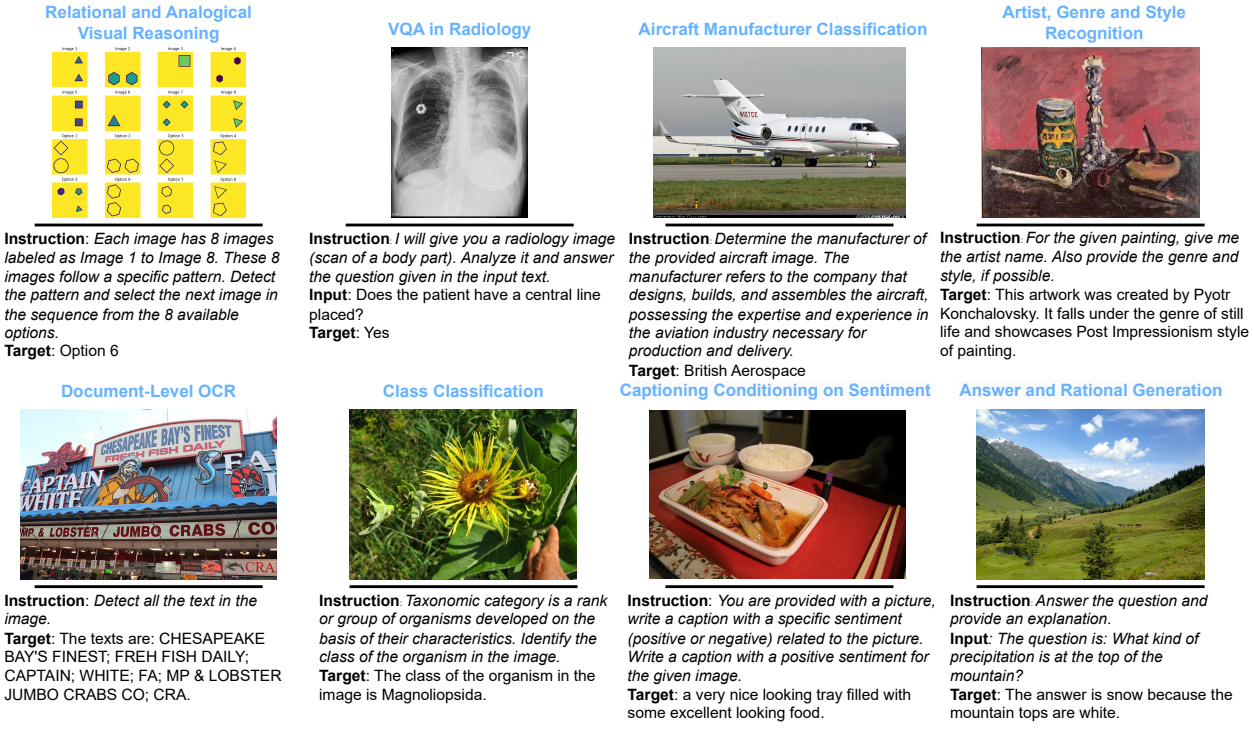


Figure 5.1: Sample tasks in VISION-FLAN. **Instruction** denotes a task instruction crafted by annotators. **Input** means text input in the given task, and **Target** is the target response based on the instruction.

5.2 Related Work

Instruction tuning [196] is first introduced in NLP and has been adapted to the visual-language domain. MultiInstruct [206] propose the first human-label multi-modal instruction tuning dataset for improving the zero-shot performance of pre-trained VLMs. LLaVA [117] leverage GPT-4 to repurpose text annotations such as captions or dense captions from existing computer-vision datasets to generate visual dialogues, Complex VQA and detail captions for visual instruction tuning. Following LLaVA, mPLUG-Owl [217], LAMM [218], MIMIC-IT [86] and Macaw-LLM [126] leverage proprietary LLMs such as GPT-4 and ChatGPT to further extend the instruction tuning tasks into 3D-domain, multiple-images and videos, and increase the amount of training instances. MiniGPT-4 [244] utilizes ChatGPT to refine output from the pre-trained VLM itself. InstructBLIP [31] and LLaVA-1.5 [116] mix the human-annotated and GPT4 synthesized datasets to enhance visual instruction tuning.

Several recent work explores different strategies to improve visual instruction tuning. StableLLaVA [100] and VPG-C [92] generate both images and texts using Stable Diffusion [150] or Blended Diffusion [6] to alleviate domain bias and encourage VLMs attend to visual details. [112] demonstrate the bias introduced by positive instructions and introduce negative instruction examples for improving robustness. Shikra [19] incorporate visual grounding tasks in visual instruction tuning to improve the VLM's referential capability. LLaVAR [231] and BLIVA [64] leverage OCR tools and GPT-4 to generate tasks helping VLMs to understand text in images. [125] and SVIT [235] empirically study the effect of scaling the size of VLMs and the size of GPT-4 synthesized dataset. Two concurrent works [20, 184] directly prompt GPT-4V with images as input to generate visual instruction tuning data and achieve superior performance. Additional related work can be found in Appendix **??**.

Unlike all prior work, our work mainly focuses on scaling human-labeled tasks in visual instruction tuning to improve VLMs' capabilities. Additionally, we perform extensive analysis to understand the characteristics of human-labeled and GPT-4 synthesized data and draw meaningful conclusions.

5.3 Vision-Flan

5.3.1 Collection Pipeline

We carefully design an annotator selection process to identify qualified annotators, which involves 2 iterations of training and testing. In the end, we hire 7 out of 21 candidates as our annotators and all of them are graduate students in computer science. To ensure the diversity and quality of the tasks in Vision-Flan, we design a rigorous annotation pipeline with four major steps:

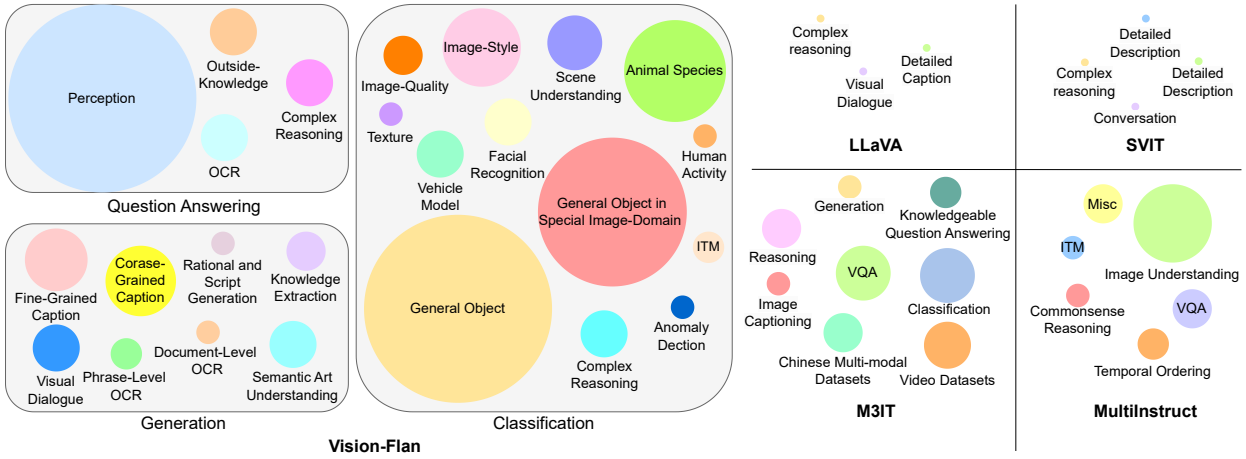


Figure 5.2: Comparison of task diversity between VISION-FLAN and previous visual instruction tuning datasets. LLaVA and SVIT report very coarse-grained categories of tasks. Each circle represents a task category and the radius is proportional to the number of tasks in that category. The radius of circles for different datasets are comparable.

Existing dataset collection and pre-processing: Two expert researchers (i.e., senior Ph.D. students in the fields of natural language processing and computer vision) search online and identify high-quality vision-language datasets. The datasets are then equally distributed to 7 annotators to download and preprocess the datasets. Each processed instance consists of an image, an instruction (the task definition from the original dataset with minor modifications), a text input if applicable, and a target output.

Creating new tasks: The two expert researchers and annotators also discuss potential new tasks that could be derived from the existing annotations. We derive new tasks by combining the annotations of two or more existing tasks on a dataset. For example, in the Concadia dataset [76], each instance consists of an image caption and a knowledge snippet related to the image. We propose a new task to predict both the caption and the background knowledge given an image, which is a free-form generation task. The new target output is formed by concatenating the caption with the knowledge snippet. We also develop new tasks by creating more basic versions of the original tasks. For example, given the object detection annotations in MSCOCO [108], we propose an object selection task in which we provide a list of objects and ask the model to select the object that appears in the image (the negative options are created by sampling objects that appear in other images but not in the given image). The expert researchers and annotators manually solve 20 instances for each newly developed task. If the human predictions match the target outputs, this new task is considered valid.

Iteratively refining the task instructions and output templates: For existing tasks, we ask annotators to write instructions based on the original task definitions with minor

modifications. For newly developed tasks, the annotators write instructions by discussing with the expert researchers. Once an annotator finishes writing a new instruction, one of the two expert researchers is randomly assigned to examine the instances and provide feedback for revising the instruction. This step iterates repeatedly until the instruction meets our requirements. We require the instruction to be *clear*, *easy to understand*, and can *be correctly executed by a human*. Each task together with its associated dataset and instruction is then added to the pool of candidate tasks for VISION-FLAN.

Verifying the quality of each task: From the candidate task pool, two expert researchers, including a native English speaker, work together to select the high-quality tasks where the instruction is fluent and effectively conveys the intended task and the task does not overlap with other tasks.

Based on these four steps, we finally collect 187 high-quality tasks, and for each task, we randomly sample 10,000 instances from its corresponding dataset. If a dataset contains less than 10,000 instances, we include all of them. We name the dataset as VISION-FLAN, consisting of 1,664,261 instances for 187 tasks in total.

5.3.2 Comparison with Existing Datasets

Dataset	Instances #	Tasks #	Source
LLaVA [117]	150K	3	Synthetic
LAMM [218]	196K	8	Synthetic
VL-Qwen [7]	350K	Unknown	Private
M^3IT [94]	2.4M	40	Synthetic
mPlug-Owl [217]	150K	3	Synthetic
Shikra [19]	156K	4	Synthetic
SVIT [235]	4.2M	4	Synthetic
MultiInstruct [206]	510K	62	Public
VISION-FLAN (Ours)	1.6M	187	Public

Table 5.1: Comparison between VISION-FLAN and existing visual instruction tuning datasets.

Table 5.1 presents a comparison between existing visual instruction tuning datasets and VISION-FLAN. For existing visual instruction tuning datasets, we directly adopt the numbers of tasks and instances reported in their original papers. The majority of these datasets are generated using proprietary language models, such as ChatGPT[1] and GPT-4[2], and exhibit a narrow range of task diversity. VL-Qwen [7] is a recently introduced large-scale dataset annotated by humans but remains inaccessible to the public. Although MultiInstruct [206] is based on publicly available datasets, it mainly focuses on visual grounding tasks and

[1] https://openai.com/blog/chatgpt
[2] https://openai.com/research/gpt-4

only contains 29 tasks that do not involve region-specific information. In contrast, VISION-FLAN encompasses a significantly more diverse array of tasks, offering a three-times increase compared to the number of tasks in MultiInstruct.

In Figure 6.1, we compare the task categories covered by VISION-FLAN and other datasets. Tasks within VISION-FLAN are first categorized into three primary groups: *Question Answering*, *Classification*, and *Generation*, and each of these primary groups is further divided into specific, fine-grained categories. For instance, within the *Classification* group, the *General Object* category involves classifying objects in images into various concepts, such as "fish", "car", and "dog". Contrastingly, the *Vehicle Model* category demands the models to accurately identify specific car brands or models, like "Toyota" and "Camry". The visualization in Figure 6.1 clearly demonstrates the superior diversity and volume of tasks in VISION-FLAN compared to existing datasets. We list tasks in each category in Appendix **??**.

5.4 Multi-Stage Instruction-Tuning

Model Architecture We adopt the same VLM architecture as LLaVA [116] and denote it as LLaVA-Architecture. As shown in Figure 5.3, it consists of a pre-trained vision encoder, a pre-trained large language model, and two layers of MLPs to connect them. In the vision-language pre-training phase of the LLaVA-Architecture, both the pre-trained vision encoder and large language model remain frozen, and only the MLP layers are trained on a large-scale image captioning dataset [152]. We leverage this pre-trained LLaVA model , without any visual instruction tuning, as our initial model and finetune it on VISION-FLAN. During visual instruction tuning, we finetune both the MLP layers and the language model while keeping the vision encoder frozen.

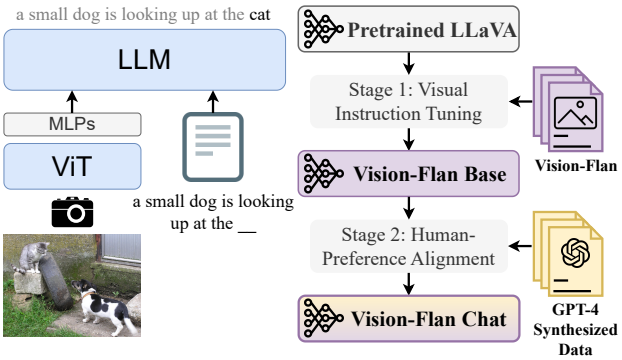


Figure 5.3: The left of the figure shows the LLaVA-Architecture and the right of the figure shows the two-stage visual instruction tuning pipeline.

Two-stage Visual Instruction Tuning Contrary to prior approaches [31, 116] that mix human-labeled data with GPT-4 synthesized data for visual instruction tuning, our study introduces a two-stage instruction tuning pipeline. As shown in Figure 5.3, in the first stage, we finetune the VLM on all 187 tasks of Vision-Flan to acquire diverse capabilities and name the resulting model as Vision-Flan Base. However, due to the brevity of target outputs presented in academic datasets, the responses from Vision-Flan Base are not in human-preferred formats. Hence, we further finetune Vision-Flan Base on GPT-4 synthesized data to align the model's outputs with human preference. We denote the yielded model as Vision-Flan Chat. This training framework requires minimal GPT-4 synthesized data while providing deep insights into the distinct contributions of human-labeled and GPT-4 synthesized data in visual instruction tuning.

Implementation Details We leverage LLaVA-Architecture with Vicuna-13B v1.5 [27], CLIP-ViT-L-336px [144] and two layers of MLP as our VLM. For the first-stage instruction tuning, we finetune the MLP layers and the language model on Vision-Flan for 1 epoch with a learning rate 2e-5 and per device batch size 16 on 8 A100 GPUs. For the second-stage instruction tuning, we further finetune the MLP layers and the language model on 1,000 instances randomly sampled from the LLaVA dataset [117] with learning rate 1e-5 and per device batch size 8 on 8 GPUs for 128 steps. In the following sections, we use LLaVA dataset and GPT-4 synthesized data interchangeably.

5.5 Experiment Setups

Evaluation Datasets We evaluate the models on several widely adopted multimodal evaluation benchmark datasets including *multiple-choice* benchmarks: **MMbench** [123], **MME** [42], and **MMMU**; *free-form generation* benchmarks: **MM-Vet** [222] and **LLaVA-Bench**; the *hallucination* benchmark: **POPE** [103], and *catastrophic forgetting* benchmarks: **CIFAR-10 and CIFAR-100** [79], **MNIST** [83], and **miniImageNet** [183].

Evaluation Protocols For MMbench, MME, MM-Vet, LLaVA-Bench, POPE and MMMU, we strictly follow their official implementations of evaluation code to evaluate the performance of each model. For datasets that do not have official evaluation codes including CIFAR-10, CIFAR-100, MNIST, and miniImageNet, we leverage the state-of-the-art open-source LLM, Vicuna 1.5 13B, to perform the evaluation and report the averaged performance on these four datasets in the CF column in Table 5.2.

Baselines We compare our models with several recent state-of-the-art vision-language models, including **BLIP-2** [93], **InstructBLIP** [31], **Shikra** [19], **LLaVA** [117], **Qwen-**

VL, **Qwen-VL-Chat** [8], and **LLaVA-1.5** [116]. The LLMs and image encoders used in all baselines are shown in Table 5.2.

5.6 Results and Discussions

5.6.1 Main Results

Model	LLM	Image Encoder	MM-Bench	MME	MMMU	LLaVA-Bench	MM-Vet	Pope	CF
BLIP-2	FlanT5-XXL	ViT-g/14	-	1293.8	34.0	-	22.4	85.3	-
InstructBlip	Vicuna-13B	ViT-g/14	36.0	1212.8	33.8	58.2	25.6	78.9	-
Mini-GPT4	Vicuna-13B	ViT-g/14	24.3	581.67	27.6	-	-	-	-
Shikra	Vicuna-13B	ViT-L/14	58.8	-	-	-	-	-	-
LLaVA	Vicuna-13B v1.5	CLIP-ViT-L-336px	38.7	1151.6	-	70.8	33.4	75.3	-
Qwen-VL	Qwen-7B	ViT-bigG	38.2	-	-	-	-	-	-
Qwen-VL-Chat	Qwen-7B	ViT-bigG	60.6	1487.5	32.9	73.6	-	-	72.1
LLaVA 1.5	Vicuna-13B v1.5	CLIP-ViT-L-336px	66.7	1531.3	33.6	70.7	35.4	83.6	73.3
VISION-FLAN BASE	Vicuna-13B v1.5	CLIP-ViT-L-336px	**69.8**	**1537.8**	**34.4**	38.5	33.4	85.9	**87.2**
Second-Stage Tuning with 1,000 GPT-4 Synthesized Instances									
VISION-FLAN CHAT	Vicuna-13B v1.5	CLIP-ViT-L-336px	67.6	1490.6	34.3	**78.3**	**38.0**	**86.1**	84.0

Table 5.2: Comprehensive evaluation of VLMs on widely adopted benchmark datasets. CF denotes the averaged performance of VLMs on four catastrophic forgetting benchmarks.

As demonstrated in Table 5.2, VISION-FLAN BASE achieves state-of-the-art performance on comprehensive evaluation benchmarks including MME, MM-Bench and MMMU, while reducing hallucination and catastrophic forgetting. However, we observe that VISION-FLAN BASE scores significantly lower on the LLaVA-Bench dataset in comparison to VLMs trained using GPT-4 synthesized data. We attribute this discrepancy to the conciseness and brevity of target outputs within academic datasets. As shown in Figure 4.2, VQA tasks frequently yield outputs comprising a single or a few words. Even outputs of many generation tasks are typically confined to one or two succinct sentences. Training on these tasks leads VISION-FLAN BASE to generate brief responses, which are not aligned with human preferences.

Conversely, through the second-stage tuning on a mere 1,000 GPT-4 synthesized data instances, VISION-FLAN CHAT achieves significant performance improvement on LLaVA-Bench, a benchmark measuring human-preference alignment, while maintaining a relatively lower rate of hallucination and catastrophic forgetting. Another finding in Table 5.2 is that compared to VISION-FLAN BASE, VISION-FLAN CHAT achieves slightly inferior performance on comprehensive evalua-

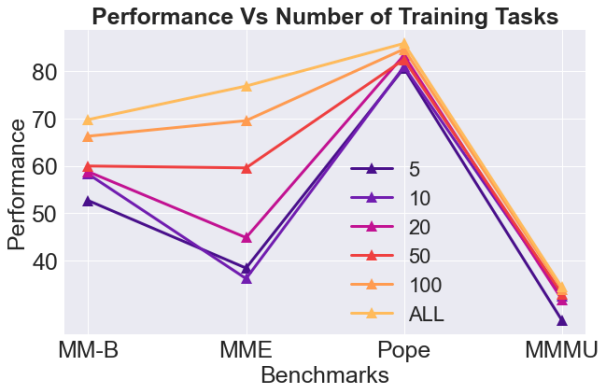


Figure 5.4: Performance on four comprehensive benchmarks versus the number of training tasks.

tion benchmarks demonstrating the bias and hallucination inevitably introduced by the GPT-4 synthesized data, which is discussed in detail in Section 5.6.2.

5.6.2 Effect of Human-Labeled and GPT-4 Synthesized Datasets

Effect of Task Diversity in Vision-Flan Figure 5.4 illustrates the relationship between the number of tasks from Vision-Flan employed during visual instruction tuning and the performance of Vision-Flan Base across four comprehensive evaluation benchmarks. It's apparent that as the number of tasks increases, the performance of Vision-Flan Base on all datasets is improved. To evaluate the impact of varying numbers of instances from different tasks, we fix the total amount of instances used for visual instruction tuning and experiment with different numbers of tasks. As demonstrated in Table 5.3, when the number of training instances is constant, augmenting the number of tasks significantly enhances model performance. These findings substantiate our hypothesis that *the diverse array of human-labeled tasks within* Vision-Flan *is essential for improving the capabilities of VLMs.*

# of Tasks	# of Instances per Task	MMB	MME	Pope	MMMU
Training with 100,000 Instances					
10	10,000	58.3	723.9	81.0	32.6
187	500	58.8	1314.3	83.3	33.3
Training with 200,000 Instances					
20	10,000	58.8	897.3	83.4	31.8
187	1,000	63.5	1373.5	83.6	33.7

Table 5.3: Comparison of Vision-Flan Base trained with a fixed total amount of data instances.

Effect of GPT-4 Synthesized Data on Comprehensive Evaluation Benchmarks Furthermore, we analyze if GPT-4 synthesized data can improve the model's performance on comprehensive evaluation benchmarks and show the results in Figure 5.5. Further tuning Vision-Flan Base on GPT-4 synthesized data instances does not lead to performance improvement. Tuning pretrained LLaVA model on a small amount of GPT-4 synthesized data (100) can improve its performance on MME but further

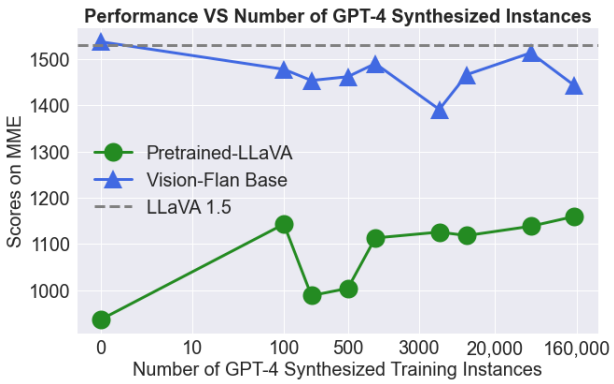


Figure 5.5: Effect of the number of GPT-4 synthesized training instances on MME. The dashed gray line indicates the performance of LLaVA 1.5.

increasing the number of training instances does not lead to any improvement. These observations are in line with recent findings in LLMs: *GPT-4 synthesized data does not improve model's capability but rather modulates the responses towards human-preferred formats* [50, 67].

5.6.3 Single-stage Tuning on Mixed Data Vs. Two-stage Tuning

In this section, we compare the performance of two training strategies based on the same pre-trained LLaVA model: (1) finetuning it on the mix of Vision-Flan and the LLaVA dataset; (2) finetuning it utilizing Vision-Flan and 1,000 instances from the LLaVA dataset with our two-stage tuning method. As illustrated in Table 5.4, the performance of VLMs finetuned on the mix of Vision-Flan and GPT-4 synthesized data is notably inferior compared to Vision-Flan Chat trained through our two-stage tuning framework.

Method	# of LLaVA	MME	LLaVA-Bench	MM-Vet
Mixed Data	1,000	1364.0	52.7	36.6
Mixed Data	158,000	1317.9	63.9	36.8
Two-stage	1,000	1490.6	78.3	38.0

Table 5.4: Comparison between single-stage finetuning on mixed data and two-stage finetuning.

5.6.4 Effect of Newly Created Tasks

In the data collection phase of Vision-Flan, we collaborate with annotators to derive 65 novel tasks from pre-existing annotations. To evaluate the impact of these newly introduced tasks, we conducted an experiment where a Vision-Flan Base was trained exclusively on the existing tasks and its performance was compared against the Vision-Flan Base trained on all tasks, including the new additions. The comparative results are presented in Table 5.5. The outcomes distinctly demonstrate the advantages of expanding the task set through the utilization of existing annotations.

Model Name	MME	MM-Bench	Pope
Vision-Flan Base	1537.8	69.8	85.9
Vision-Flan Base w/o New Tasks	1379.8	67.3	84.6

Table 5.5: Comparison between finetuning Vision-Flan Base on all tasks and finetuning Vision-Flan Base only on existing tasks.

5.6.5 Contributions of Tasks from Different Task Groups

Understanding the contribution of each task to the models' performance is crucial for visual instruction tuning. However, training a model on each task in VISION-FLAN can impose significant computational cost. Instead, we propose a group-level analysis of task contributions. As illustrated in Figure 6.1, tasks are categorized into three primary groups: Question Answering (QA), Classification, and Generation. We train three VISION-FLAN CHAT models on VISION-FLAN, each time excluding all tasks from one primary group, and show their performance in Table 5.6.

Model Name	MME	Pope	LLaVA-Bench	MM-Vet
VISION-FLAN CHAT	1537.8	86.1	78.3	38.0
w/o QA	1211.6	83.3	73.5	37.9
w/o Classification	1376.6	84.9	75.4	37.3
w/o Generation	1390.9	83.8	68.1	36.2

Table 5.6: Contributions of different tasks group to the performance of VISION-FLAN CHAT.

Our findings indicate that: (1) the inclusion of generation tasks markedly enhances the model's capability in free-form generation tasks, as evidenced by performance on LLaVA-Bench; (2) QA tasks and classification tasks contribute significantly to model's performance MME, and (3) synergistic interactions between different task groups lead to further performance enhancements.

5.6.6 What is Essentially Improved in VLMs during Instruction Tuning

LLM	MLPs	MM-Bench	MME	LLaVA-Bench	Pope
✗	✗	45.0	936.3	32.4	51.9
✗	✓	52.4	1107.3	39.1	83.3
✓	✗	69.2	1495.5	39.3	85.6
✓	✓	69.8	1537.8	38.5	85.9

Table 5.7: Effect of tuning different modules in VISION-FLAN BASE. ✓ denotes the module is tuned and ✗ denotes the module is frozen during visual instruction tuning.

In LLaVA-Architecture, the MLP layers map the visual features from a vision encoder into the embedding space of LLMs. The LLMs then interpret the visual features and follow text instructions to generate responses. In Table 5.7, we show the results of training different modules during visual instruction tuning and observe that solely tuning MLPs causes a significant performance drop compared to tuning both MLPs and LLMs during visual instruction

tuning. However, tuning LLMs with frozen MLPs results in similar performance as tuning both modules, demonstrating that visual instruction tuning mainly enables LLMs to better understand visual features while MLPs have been sufficiently learned during pretraning. To further support this claim, we replace the instruction-tuned MLPs in Vision-Flan Base and Vision-Flan Chat with the pretrained MLPs from the pre-trained LLaVA model, and show that with the pretrained MLPs, both models can retain more than 90% of performance on most tasks as shown in Table 5.8. We also compute the Pearson Correlation Coefficient between the parameters of pretrained MLPs and instruction-tuned MLPs, and find that their correlation coefficient is higher than 0.99.

Model	MMB	MME	LLaVA-Bench	Pope
Vision-Flan Base	69.8	1537.8	38.5	85.9
+ Pretrained MLP	68.0	1403.1	36.4	84.0
Vision-Flan Chat	67.6	1490.6	78.3	86.1
+ Pretrained MLP	65.7	1332.2	73.8	85.4

Table 5.8: Results of replacing visual instruction tuned MLPs with pretrained MLPs. Gray rows show the performance of the original models and yellow rows show the performance after replacing instruction-tuned MLPs with pretrained MLPs.

5.7 Summary

In this chapter, we introduced Vision-Flan, the most diverse publicly available visual instruction tuning dataset to date. Vision-Flan expands task diversity across perception, domain-specific, and complex reasoning categories, each paired with expert-written instructions. Building on this dataset, we proposed a two-stage instruction tuning framework: in the first stage, models are fine-tuned on Vision-Flan to acquire broad capabilities across diverse visual tasks; in the second stage, a small amount of GPT-4 synthesized data is used to refine alignment with human preferences. This approach not only improves generalization and reduces hallucination but also mitigates catastrophic forgetting bu scaling human-labeled tasks.

Our experiments demonstrate that Vision-Flan significantly enhances the capabilities of VLMs, enabling them to perform more robustly across a wide range of tasks. Moreover, the two-stage framework achieves stronger human preference alignment with substantially less synthetic data compared to prior methods, striking an effective balance between data diversity and efficiency. Through extensive analyses, we further clarified the complementary roles of human-labeled versus GPT-4 synthesized data and highlighted the impacts of different training strategies.

While Vision-Flan establishes a strong foundation for scaling instruction tuning in static image understanding, it remains limited to static image scenarios. In real-world applications,

however, vision–language models must reason over temporal sequences in video. To address this gap, the next chapter introduces our method for incentivizing temporal reasoning via a novel reinforcement learning algorithm, extending generalization beyond static images into the spatiotemporal domain.

Chapter 6

Spatialtemporal Reasoning

6.1 Motivation

The previous chapter demonstrated that large-scale instruction tuning with Vision-Flan can substantially improve the generalizability and human preference alignment of vision–language models. However, these post-training efforts were restricted to static image understanding, where tasks are defined over a single image or an image–text pair. In contrast, many real-world applications require reasoning over video, where information unfolds dynamically over time and models must integrate both spatial and temporal cues to generate coherent and accurate responses. Extending unified multimodal modeling into the spatiotemporal domain is therefore a critical step toward building general-purpose multimodal foundation models.

Recent advances in Multimodal Large Language Models (MLLMs) [30, 132, 185, 192, 208, 233] have significantly improved visual reasoning capabilities. Much of this progress [52, 88, 185, 208], however, remains centered on static images or short image sequences. While such models demonstrate strong spatial understanding, they often struggle in video-based scenarios [43, 63, 209, 236], where effective reasoning requires temporal awareness, i.e., the ability to localize, integrate, and reason about events across time [209, 212]. Existing video–language modeling approaches typically adapt image-based architectures through frame-based or clip-level representations. Although effective for short-range perception, these strategies often fail to capture long-horizon dependencies and the causal structure of events, limiting their applicability to complex, time-sensitive tasks.

These challenges are particularly pronounced in egocentric video understanding [49]. Unlike third-person videos with relatively stable viewpoints, egocentric recordings are characterized by rapid viewpoint changes, severe partial observability, and strong dependencies between past and future frames. Tasks such as action recognition, temporal grounding, and intention prediction [23, 128, 140] inherently require models to maintain temporal coherence and reason over extended temporal contexts. Despite recent progress [136, 138, 228], existing MLLMs frequently exhibit hallucinated events, temporally inconsistent reasoning, and degraded performance on egocentric temporal benchmarks.

We identify three fundamental factors that limit temporal reasoning in current video–language models. First, post-training paradigms such as supervised fine-tuning and reinforcement

learning methods, including Group Relative Policy Optimization (GRPO) [155], primarily optimize response quality without explicitly incentivizing temporal consistency or causal understanding [40]. Second, many training datasets contain questions that can be answered from a single frame, allowing models to rely on spatial shortcuts rather than learning genuine temporal dynamics [38]. Third, existing benchmarks rarely provide high-quality, long-horizon temporal reasoning chains; instead, they typically offer only video–answer pairs, as collecting temporally grounded rationales is costly and difficult [182]. Together, these limitations encourage degenerate solutions that treat videos as unordered collections of images, leading to temporally incoherent predictions and explanations.

To address these challenges, this chapter introduces Temporal Global Policy Optimization (**TGPO**), a reinforcement learning algorithm explicitly designed to incentivize temporal awareness in MLLMs. Rather than relying solely on static supervision, **TGPO** calibrates reinforcement learning rewards based on the model's sensitivity to temporal order. During training, the model is evaluated on both ordered video inputs and temporally shuffled counterparts. The performance gap between these settings is used as a temporally calibrated reward, which penalizes solutions that ignore temporal structure. By integrating this calibration mechanism into policy optimization objectives, **TGPO** encourages models to rely on temporally grounded reasoning rather than spatial shortcuts.

We integrate **TGPO** with two widely used policy optimization frameworks, GRPO and Group Sequence Policy Optimization (GSPO) [237], and apply them to Qwen2.5-VL models [208]. Following a cold-start training regime inspired by DeepSeek-R1-Zero [51], models are trained directly with reinforcement learning without supervised fine-tuning. Extensive evaluations on five egocentric video benchmarks demonstrate that **TGPO** consistently improves temporal reasoning performance over prior reinforcement learning approaches, highlighting the effectiveness of explicitly incentivizing temporal awareness for robust egocentric video understanding.

6.2 Related Works

6.2.1 Policy Optimization Algorithms

A wide range of policy-optimization algorithms has been explored for post-training large language models (LLMs). Proximal Policy Optimization (PPO) [153] introduces a clipped surrogate objective that stabilizes policy updates and has become a foundational baseline in reinforcement learning from human feedback (RLHF) [134]. Direct Preference Optimization (DPO) [145] instead learns directly from preference pairs without training a reward model, simplifying alignment by optimizing likelihood ratios between preferred and disfavored responses. More recent critic-free approaches include ReMax [104], which adapts REINFORCE with a greedy baseline for simplicity, and RLOO [2], which reduces variance by subtracting

a leave-one-out baseline computed from other sampled responses. REINFORCE++ [62] further advances this line of work by using a batch-normalized reward baseline for advantage estimation, avoiding an explicit critic while improving robustness and generalization across reward models and long chain-of-thought settings. Group-based algorithms such as GRPO [155] normalize rewards within each sample group to estimate relative advantages without a value network, while GSPO [237] extends this idea by performing optimization at the sequence level to better match sequence-level rewards. In contrast, Single-stream Policy Optimization (SPO) [207] revisits policy-gradient learning from a non-grouped perspective, replacing per-group baselines with a persistent KL-adaptive value tracker and globally normalized advantages, enabling stable, low-variance learning signals and significantly improved scalability.

6.2.2 Ego-Centric Video Understanding

Ego4D [49] introduced a large-scale egocentric video dataset with 3,670 hours of multimodal first-person recordings and a comprehensive benchmark suite covering episodic memory, interaction understanding, and future activity forecasting, establishing a foundational resource for egocentric perception research. EgoVLPv2 [142] advanced egocentric video–language pretraining by integrating cross-modal fusion directly into the backbone networks, enabling more unified representations and reducing downstream fine-tuning costs. GroundNLQ [59] proposed a multi-scale multimodal grounding framework for long egocentric videos and achieved state-of-the-art performance in the Ego4D Natural Language Queries Challenge through specialized egocentric feature extraction. EMQA [10] introduced episodic memory–based video question answering, constraining models to maintain constant-sized memory representations and releasing the large-scale QAEGO4D dataset to study long-horizon egocentric reasoning. EgoVideo [138] presented an egocentric foundation model tailored for Ego4D and EPIC-Kitchens challenges, demonstrating strong generalization across diverse egocentric tasks, including moment retrieval and action anticipation. MM-EGO [216] explored building egocentric multimodal LLMs by generating 7M QA pairs from Ego4D, proposing a memory pointer prompting mechanism to improve long-video comprehension and de-biased evaluation. EgoLife [210] introduced a life-oriented egocentric dataset and QA benchmark spanning daily activities over extended time horizons, along with an integrated assistant system combining multimodal modeling and retrieval for long-context reasoning. EgoVLM [182] applied Group Relative Policy Optimization to directly align vision–language models with egocentric reasoning behaviors, demonstrating substantial gains over general-purpose VLMs and introducing a keyframe-based reward for temporal grounding. Ego-R1 [178] proposed a Chain-of-Tool-Thought framework with an RL-trained agent to reason over ultra-long egocentric videos, enabling modular tool invocation for temporal retrieval and multimodal understanding across week-long time spans. EgoVITA [80] introduced a reinforcement learning framework that alternates between egocentric planning and exocentric verification, improving causal and visually grounded reasoning for first-person video understanding. Exo2Ego [228]

leveraged large-scale synchronized ego–exo video–text data to transfer exocentric knowledge into egocentric domains through progressive mapping, significantly improving egocentric performance while highlighting limitations of existing MLLMs.

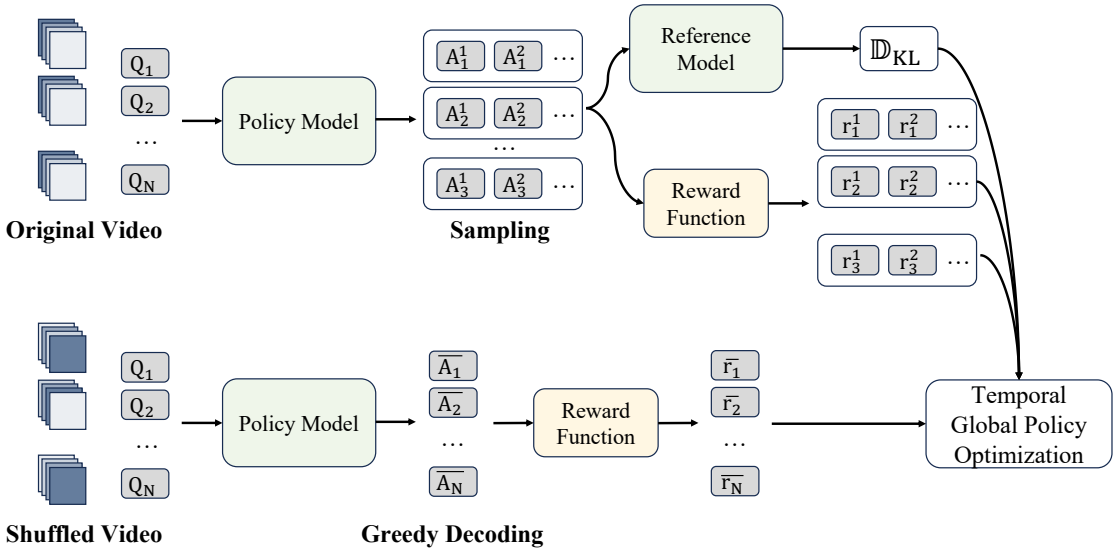


Figure 6.1: An overview of our proposed **TGPO**.

6.3 Method

6.3.1 Background on GRPO

In the GRPO framework [155], given an input prompt s and a video clip c as context, the MLLM samples a group of $|G|$ candidate responses $\{y_1, \ldots, y_{|G|}\}$. A reward function $r(\cdot)$ assigns a scalar score to each response, producing $\{r(y_1), \ldots, r(y_{|G|})\}$. GRPO optimizes the policy by maximizing a group-normalized advantage estimator:

$$\hat{A}(s, c; \theta) = \frac{1}{|G|} \sum_{i=1}^{|G|} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \frac{\pi_\theta(y_{i,t} \mid s, c)}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid s, c)} \cdot \frac{r(y_i) - \mu_G}{\sigma_G},$$

where $\pi_\theta(y_{i,t} \mid s, c)$ denotes the log-probability of generating token $y_{i,t}$ under the current parameters θ, and $\pi_{\theta_{\text{old}}}$ corresponds to a recently updated policy used to form the importance ratio. Here $\mu_G = \text{mean}(\{r(y_i)\}_{i=1}^{|G|})$ and $\sigma_G = \text{std}(\{r(y_i)\}_{i=1}^{|G|})$ are computed within the sampled group for the given (s, c).

To stabilize training and prevent excessive drift from a reference policy π_{ref}, GRPO includes a KL regularizer. The resulting objective is

$$\max_{\theta} \; \mathbb{E}_{(s,c)\sim\mathcal{D}}[\hat{A}(s,c;\theta) \; - \; \beta \, D_{\text{KL}}(\pi_\theta \,||\, \pi_{\text{ref}})],$$

where β controls the strength of the regularization.

6.3.2 Greedy Baseline without Temporal Information

We introduce a temporally calibrated reward for video understanding that explicitly encourages MLLMs to exploit temporal cues, rather than relying on static per-frame shortcuts.

Given a prompt s and a video c, we first sample a response from the current policy, $y \sim \pi_\theta(\cdot \mid s, c)$. We then construct a temporal-calibrated baseline by shuffling the video frames and generate a response with greedy decoding:

$$\hat{y} \sim \pi_\theta(\cdot \mid s, \text{shuffle}(c)),$$

We define the temporally calibrated reward as

$$r_{\text{T}}(y) \; = \; r(y) \; - \; r(\hat{y}).$$

Compared to $r(y)$, the temporally calibrated reward takes into consideration whether temporal reasoning is used in the model's generation. Specifically, when $r_{\text{T}}(y) > 0$, the model performs better with the temporally coherent video than with shuffled frames, indicating that it leverages temporal dependencies; we therefore assign a positive training signal. Conversely, $r_{\text{T}}(y) \leq 0$ suggests that the prediction does not benefit from temporal information (e.g., the model treats the input as a set of independent frames), and the resulting non-positive signal discourages such shortcut behavior.

6.3.3 Temporal Global Policy Optimization (TGPO)

We next show that **TGPO** can be incorporated into existing policy-optimization objectives. We present two variants based on GRPO and GSPO.

Integration with GRPO. Standard GRPO normalizes rewards within each group of size $|G|$ for each prompt s_j in a mini-batch B. In contrast, **TGPO** performs normalization across all group samples in the mini-batch, which prevents low-variance groups (often corresponding to temporally insensitive instances) from being artificially amplified by within-group normalization.

Concretely, let $|B|$ be the batch size and $|G|$ the group size. For instance j, if the rewards of the group samples $\{r(y_{j,1}), \ldots, r(y_{j,|G|})\}$ are close to the baseline reward $r(\hat{y}_j)$, then the

calibrated rewards $r_{\mathrm{T}}(y_{j,i})$ are near zero, indicating that temporal information is largely unnecessary for answering s_j. With only in-group normalization, such low-variance groups can still yield large normalized advantages and thus contribute disproportionately to training, encouraging temporally insensitive behavior. Global normalization mitigates this issue: if other instances in the same mini-batch achieve large $r_{\mathrm{T}}(\cdot)$, the batch-level mean increases and the normalized advantage for temporally insensitive instances becomes small or negative, reducing their influence. The resulting **TGPO** (GRPO) advantage estimator is

$$\hat{A}(B;\theta) = \frac{1}{|B|}\sum_{j=1}^{|B|}\frac{1}{|G|}\sum_{i=1}^{|G|}\frac{1}{|y_{j,i}|}\sum_{t=1}^{|y_{j,i}|}\frac{\pi_\theta(y_{j,i,t} \mid s_j, c_j)}{\pi_{\theta_{\text{old}}}(y_{j,i,t} \mid s_j, c_j)} \cdot \frac{r_{\mathrm{T}}(y_{j,i}) - \mu_B}{\sigma_B},$$

where $\mu_B = \text{mean}(\{r_{\mathrm{T}}(y_{j,i})\}_{j=1,i=1}^{|B|,|G|})$ and $\sigma_B = \text{std}(\{r_{\mathrm{T}}(y_{j,i})\}_{j=1,i=1}^{|B|,|G|})$ are computed over all $|B|\times|G|$ samples in the mini-batch.

Integration with GSPO. GSPO [237] replaces token-level importance ratios with a sequence-level likelihood ratio, which empirically improves stability while achieving performance comparable to GRPO. Under GSPO, we define the sequence-level importance ratio as

$$\rho_{j,i}(\theta) = \exp\Big(\frac{1}{|y_{j,i}|}\sum_{t=1}^{|y_{j,i}|}\log\frac{\pi_\theta(y_{j,i,t} \mid s_j, c_j)}{\pi_{\theta_{\text{old}}}(y_{j,i,t} \mid s_j, c_j)}\Big),$$

then the **TGPO** (GSPO) advantage estimator becomes

$$\hat{A}(B;\theta) = \frac{1}{|B|}\sum_{j=1}^{|B|}\frac{1}{|G|}\sum_{i=1}^{|G|}\rho_{j,i}(\theta) \cdot \frac{r_{\mathrm{T}}(y_{j,i}) - \mu_B}{\sigma_B},$$

6.3.4 Prompt Engineering.

During rollout, we prompt the model to generate not only the final answer but also a reasoning trace, formatted in a required structure. Following prior work [51], we design the input prompt to encourage the model to produce an answer accompanied by step-by-step reasoning. The exact prompts are provided below.

I provide you with a question about the given video.
Provide a detailed thinking process within <think> </think> tags and then output the answer within the <answer> </answer> tags. Within <think> tags, provide a detailed, step-by-step reasoning process. First, describe in detail what happens in the video that is relevant to the question. Then, explain how you arrive at your answer by referencing specific evidence or events from the video. Make sure your reasoning is clear, logical, and closely tied to what is shown in the video. Within <answer> tags, clearly state your chosen answer, ensuring you select it from the provided options.
The question is: [EVENT]

6.3.5 Reward Modeling

We employ a composite reward that evaluates both *answer correctness* and *output compliance*. Since our datasets use multiple-choice video question answering (VQA), each model response is expected to contain (i) a reasoning segment enclosed by <think>...</think> and (ii) a final choice enclosed by <answer>...</answer>. The overall reward is computed per sampled response and used for policy optimization.

Accuracy Reward. Let $\bar{\alpha}$ denote the ground-truth option for a given question, and let a be the option extracted from the model output (i.e., the content inside <answer>...</answer> after normalization). We define a binary accuracy reward:

$$r_{\text{Accu}}(a) = \begin{cases} 0, & \text{if } a \neq \bar{\alpha}, \\ 1, & \text{if } a = \bar{\alpha}. \end{cases}$$

Format Reward. In addition to correctness, we compute reward adherence to the required response structure. Specifically, we assign:

$$r_{\text{Form}}(a) = \begin{cases} 0, & \text{not follow the required format}, \\ 1, & \text{follow the required format}. \end{cases}$$

A response is considered well-formatted if it contains exactly one <think>...</think> block and one <answer>...</answer> block; and the <answer> block is non-empty and contains a valid option from the provided candidate set (after normalization).

Combined Reward. We combine correctness and formatting as

$$r(a) = r_{\text{Accu}}(a) + \lambda r_{\text{Form}}(a),$$

where λ controls the strength of the format signal.

Model	EgoSchema↑	EgoPlan↑	EgoPlan 2↑	VLM4D↑	EgoTempo↑
Qwen2.5-VL 3B					
+ CoT	20.4	22.2	23.5	37.1	31.6
+ GRPO	46.5	36.5	37.1	46.8	40.0
+ GSPO	45.4	36.3	32.7	47.4	42.0
+ TGPO (GRPO)	49.6	36.8	42.3	49.6	45.2
+ TGPO (GSPO)	49.7	36.7	41.1	48.6	42.6

Table 6.1: Performance comparison of our method **TGPO** with two popular RL-based optimization methods and chain-of-thought (CoT) reasoning across egocentric benchmarks.

6.4 Experiments

6.4.1 Implementation Details

Base Model and RL Framework. We adopt Qwen2.5-VL-3B [208] as our base model due to its strong performance on video understanding tasks and its demonstrated potential for RL–based optimization [40, 99, 191]. All experiments are conducted using this backbone.

We train the model using **verl** [157], a flexible and efficient reinforcement learning framework designed for large-scale model training. At the time of our experiments, **verl** did not natively support video-based RL training. We therefore extend its implementation to enable video input processing and video understanding with **vLLM** [82]. In addition, we implement the GSPO baseline following the original formulation in prior work [237].

Unless otherwise specified, we use the same training hyper-parameters across all experiments: format-reward weight $\lambda = 0.1$, learning rate 1×10^{-6} with a constant scheduler, KL regularization coefficient 1×10^{-4}, weight decay 0.01, number of rollouts 8, sampling temperature 1.0, micro-batch size of 4 per GPU, and mini-batch size 64. Training is performed on 8 nodes, each equipped with 8 NVIDIA A100 GPUs with 40GB memory.

Training Data. We use EgoIT99K [211] as the training dataset. Since our method and all baselines (Section 6.4.3) rely on verifiable rewards, we restrict training to the subsets of EgoIT99K that contain multiple-choice and yes/no questions.

For each video, we uniformly sample 32 frames. Increasing the number of frames did not yield noticeable improvements in our preliminary experiments. Due to the lack of large-scale, high-quality supervised finetuning data for egocentric video understanding, we follow the cold-start training paradigm of DeepSeek-R1-Zero [51], training the model directly with reinforcement learning without a supervised pretraining stage.

6.4.2 Evaluation

Evaluation Datasets and Metrics. We evaluate all models on five egocentric video question-answering benchmarks designed to assess temporal understanding. All evaluation datasets consist of multiple-choice questions. Prior work [38] has shown that the full EgoSchema [128] benchmark contains instances that can often be solved using language priors or static spatial cues from a single frame. To more faithfully evaluate temporal reasoning, we adopt the temporal and *others* splits filtered by [38]. EgoPlan Bench [23] and EgoPlan2 Bench [143] evaluate planning and anticipation capabilities, requiring models to observe ongoing actions in egocentric videos and predict the next step. EgoPlan2 further increases diversity and realism by grounding tasks in more complex real-world scenarios. VLM4D [242] consists of carefully curated egocentric video–question pairs that emphasize translational and rotational motion, perspective awareness, and motion continuity. Finally, EgoTempo [140] focuses on holistic temporal understanding, where correct answers cannot be inferred from a single frame or common-sense reasoning alone. Following [228], we modify the answer of instances in EgoTempo [140] from direct answer generation to multiple-choice QA for a more stable and reliable evaluation. To construct high-quality distractors, we leverage a state-of-the-art vision-language model (VLM). Specifically, we provide the Gemini 2.5 model with the video instances and prompt it to generate plausible incorrect options based on the visual context and temporal proximity to the ground-truth answer.

6.4.3 Baselines

Our main contribution is the proposed RL algorithm **TGPO** for better temporal understanding and reasoning. Hence, our main comparison is focusing on comparing our approach against two RLVR baselines, including GRPO [155] and GSPO [237]. GRPO has been widely adopt for training strong video reasoning models including Video-R1 [40], Time-R1 [191], and Video-Chat-R1, [99]. In our implementation, we keep everything for GRPO and GSPO, including training data and hyperparameters, the same as our methods for a fair comparison.

In addition, we benchmark against strong MLLMs including proprietary models: Gemini-1.5-Pro [176], Gemini-2.5-Pro [30], and Claude-Sonnet-4 [1]; open-source models: Qwen2.5-VL [208], LLaVA-Video [233], LLaVA-NeXT-Video [232], Video-LLaMA-2 [25], Llava-OneVision [88], and InternVideo2 [192]; and the egocentric-specialized model EgoVLM [182], which has been trained with GRPO.

[1] https://www.anthropic.com/claude/sonnet

Method	VLM4D	EgoPlan	EgoPlan2	EgoSchema
Baseline				
GRPO	1265.13	898.25	784.04	1159.03
GSPO	1305.99	693.21	552.01	1118.76
Ours				
TGPO (GRPO)	**1339.49**	**923.93**	**786.97**	**1306.32**
TGPO (GSPO)	1333.75	905.82	544.95	1144.84

Table 6.2: Area Under the Curve (AUC) of reward over the first 3000 training steps for different optimization methods across datasets. A higher AUC reflects faster reward improvement and improved training stability.

6.5 Results

6.5.1 Comparison with RL-based Methods

Overall Performance. As shown in Table 6.1, TGPO consistently outperforms existing RL-based approaches across all benchmarks. While standard CoT prompting performs poorly on temporally demanding tasks, GRPO and GSPO yield substantial improvements, particularly on EgoSchema and EgoPlan 2. However, these methods do not explicitly enforce temporal consistency, limiting their effectiveness on benchmarks requiring long-horizon temporal reasoning. Integrating TGPO on top of GRPO or GSPO leads to further and consistent gains. In particular, Qwen2.5-VL 3B + TGPO (GRPO) achieves the strongest overall performance, reaching 49.6 on EgoSchema, 36.8 on EgoPlan, 42.3 on EgoPlan 2, 49.6 on VLM4D, and 45.2 on EgoTempo. The largest improvements are observed on EgoSchema and EgoPlan 2, indicating that temporally calibrated rewards effectively discourage single-frame shortcuts and promote reasoning over event order and temporal dependencies. Overall, these results demonstrate that TGPO substantially enhances temporal reasoning in MLLMs under a fully reinforcement-learning-based training regime, outperforming existing RL methods.

Training Dynamics. We present the training dynamics (reward versus training steps) on four benchmarks in Figure 6.2. Darker curves denote smoothed performance for improved readability, while lighter curves show raw reward trajectories. We exclude EgoTempo from this analysis, as its long video sequences substantially slow down training. As shown in the figure, both TGPO variants consistently achieve higher reward than their baseline counterparts throughout most of the training process across the evaluated benchmarks. In addition, TGPO exhibits faster early-stage reward growth and smoother training trajectories, particularly on VLM4D and EgoSchema, indicating improved learning efficiency and more stable optimization dynamics. Notably, GSPO shows a clear late-stage performance degradation on EgoPlan, whereas TGPO (GRPO) maintains a steady improvement trend. On the more challenging EgoPlan2 benchmark, we observe pronounced reward degradation for GSPO and

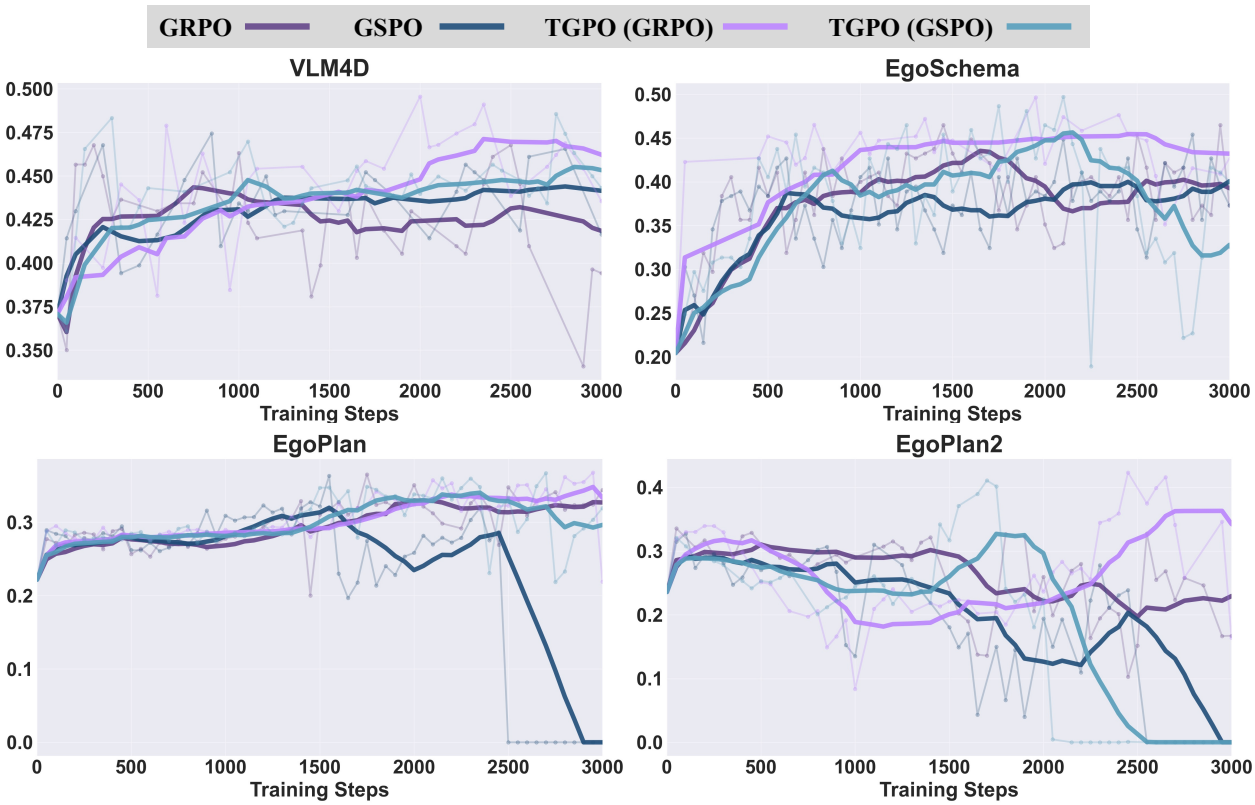


Figure 6.2: Test reward over the 3000 training steps. The reward curves are reported on four benchmarks across GRPO, GSPO, and our two **TGPO** variants.

GRPO, as well as for TGPO (GSPO), while TGPO (GRPO) largely avoids this failure mode and preserves stable reward accumulation throughout training.

To quantitatively assess learning efficiency, we report the Area Under the Curve (AUC) of reward over the first 3000 training steps in Table 6.2. In reinforcement learning, a higher AUC reflects faster reward improvement and sustained performance throughout training, capturing overall sample efficiency rather than isolated peak performance. Notably, TGPO (GRPO) improves AUC over GRPO from 1159.03 to 1306.32 on EgoSchema (+12.7%) and from 1265.13 to 1339.49 on VLM4D (+5.9%). TGPO (GSPO) also yields consistent gains, improving AUC over GSPO from 1305.99 to 1333.75 on VLM4D (+2.1%). Overall, these results demonstrate that **TGPO** enables more effective and stable reward accumulation over time, leading to faster convergence and improved training efficiency across benchmarks.

6.5.2 System-Level Comparison

Tables 6.3–6.7 report performance on five egocentric video question answering benchmarks that emphasize temporal understanding and long-horizon reasoning. Across all datasets,

Model	EgoSchema↑
Qwen2.5-VL (3B)	20.4
LLaVA-Video (7B)	34.1
LLaVA-Video (72B)	42.7
Llava-OneVision (72B)	36.2
EgoVLM GRPO (3B)*	46.5
Qwen2.5-VL 3B	
+TGPO	**49.7**

Table 6.3: Performance comparison on EgoSchema.

Model	EgoPlan 2↑
Proprietary	
GPT-4o	32.6
Open-source MLLMs	
LLaVA-Video (7B)	25.3
LLaVA-NeXT-Video (7B)	23.3
Video-LLaMA-2 (7B)	23.0
EgoVLM GRPO (3B)*	37.1
Qwen2.5-VL 3B	
+TGPO	**42.3**

Table 6.4: Performance comparison on EgoPlan 2.

TGPO consistently achieves the strongest performance among open-source models, outperforming supervised fine-tuning and standard reinforcement learning baselines under comparable model scales.

On EgoSchema (Table 6.3), TGPO improves upon the strongest open-source baseline (EgoVLM GRPO, 3B) by a clear margin (49.7 vs. 46.5), and surpasses substantially larger models such as LLaVA-Video (72B). A similar trend is observed on EgoPlan and EgoPlan 2 (Tables 6.6 and 6.4), where TGPO consistently outperforms both supervised fine-tuning and reinforcement learning baselines initialized from the same backbone. On EgoPlan 2 in particular, TGPO achieves a large gain over EgoVLM GRPO (42.3 vs. 37.1), indicating that temporally calibrated optimization is especially beneficial in more challenging planning scenarios. Notably, TGPO also outperforms proprietary systems such as GPT-4o on EgoPlan 2, despite the latter's significantly larger scale and access to private training data.

On VLM4D (Table 6.5), which evaluates temporal understanding in translational and rotational motions and perspective awareness, TGPO again yields consistent improvements over strong open-source baselines, including Qwen2.5-VL (7B) and EgoVLM GRPO (3B). While large proprietary models such as Gemini-2.5-Pro still achieve the highest performance, TGPO narrows the gap substantially, outperforming several proprietary models and highlighting the effectiveness of targeted temporal reinforcement learning.

Finally, on EgoTempo (Table 6.7), TGPO achieves the best overall performance among all models, including proprietary systems. Compared to EgoVLM GRPO, TGPO yields a notable improve-

Model	EgoTempo↑
Proprietary	
Gemini-Flash	39.1
GPT-4o	40.1
Claude-3.5-Sonnet	13.1
Open-source MLLMs	
Qwen2-VL (7B)	26.1
Qwen2-VL (72B)	28.4
LLaVA-OneVision (7B)	23.3
LLaVA-OneVision (72B)	26.5
LLaVA-NeXT-Video (34B)	15.7
EgoVLM GRPO (3B)*	40.8
Qwen2.5-VL 3B	
+TGPO	**45.2**

Table 6.7: EgoTempo benchmark.

Model	VLM4D↑
Proprietary	
Gemini2.5-Pro	64.6
GPT-4o	55.5
Claude-Sonnet-4	52.6
Grok-2-Vision	48.8
Open-source MLLMs	
Qwen2.5-VL (3B)	37.1
Qwen2.5-VL (7B)	42.3
InternVideo2 (8B)	35.6
Llava-OneVision (7B)	36.8
EgoVLM GRPO (3B)*	46.8
Qwen2.5-VL 3B	
+TGPO	**49.6**

Table 6.5: Performance comparison on VLM4D.

Model	EgoPlan↑
Proprietary	
Gemini1.5-Pro	32.8
GPT-4o	32.8
Open-source MLLMs	
Qwen2.5-VL (3B)	32.9
Qwen2.5-VL (7B)	33.0
LLaVA-Video (7B)	33.6
EgoVLM SFT (3B)*	32.1
EgoVLM Dr. GRPO (3B)	33.0
EgoVLM GRPO (3B)*	36.5
Qwen2.5-VL 3B	
+TGPO	**36.8**

Table 6.6: Performance comparison on EgoPlan.

ment (45.2 vs. 40.8), underscoring its advantage in tasks that require fine-grained temporal alignment rather than static visual recognition. This result further supports the hypothesis that explicitly optimizing temporal decision-making is critical for egocentric video understanding.

Overall, these results demonstrate that TGPO delivers consistent and robust gains across diverse egocentric benchmarks, and that temporally calibrated reinforcement learning can substantially reduce the performance gap between open-source and proprietary multimodal large language models without increasing model scale.

6.6 Summary

This work highlights the importance of explicitly modeling temporal structure in multimodal large language models for egocentric video understanding. By framing temporal awareness as a learnable and incentivized capability rather than an emergent property of post-training, we introduce TGPO as a principled reinforcement learning approach that directly targets causal and temporal reasoning. Our results demonstrate that contrastive temporal rewards and cold-start RL training can effectively guide MLLMs toward coherent, temporally grounded reasoning without reliance on supervised finetuning. We hope this perspective encourages future research to move beyond static visual reasoning and toward learning-based frameworks that better align multimodal models with the dynamic, causal nature of real-world perception.

Chapter 7

Conclusion and Future Work

This dissertation investigates the fundamental challenge of building general-purpose multimodal intelligence systems that can seamlessly unify text, images, and video, and robustly generalize across diverse tasks in open-world settings. Rather than focusing on isolated modalities or narrowly defined benchmarks, this work advances a unified perspective on multimodal foundation models by rethinking architectural design, training paradigms, and temporal reasoning mechanisms. The overarching goal is to move beyond task-specific multimodal systems toward models that exhibit flexible understanding, scalable generation, and strong alignment with human instructions across modalities.

This goal is achieved by addressing four interconnected research challenges: (1) enabling interleaved multimodal understanding and generation via modality-specialized synergizers, (2) designing efficient unified architectures that bridge autoregressive and diffusion paradigms, (3) improving zero-shot generalization and human-preference alignment through multimodal instruction tuning, and (4) extending static multimodal reasoning to dynamic spatiotemporal video understanding.

7.1 Conclusion

First, this dissertation advances interleaved multimodal unification by addressing a core limitation of existing vision–language models: their inability to flexibly generate and reason over text and images in arbitrary sequences. Through the construction of LeafInstruct, the first large-scale post-training dataset specifically designed for interleaved text–image generation, this work provides the necessary data foundation for studying this problem at scale. Building on this dataset, the proposed Modality-Specialized Synergizers (MOSS) introduce a principled architectural augmentation that enables modality-aware adaptation while preserving strong cross-modal interaction. This contribution demonstrates that unification does not require treating all modalities identically; instead, carefully designed specialization within a unified framework leads to more robust and controllable multimodal behavior.

Second, this dissertation proposes efficient unified architectures that reconcile the complementary strengths of autoregressive and diffusion-based modeling. While autoregressive models excel at sequential reasoning and language-centric tasks, diffusion models remain the dominant paradigm for high-fidelity image synthesis. By introducing LaTtE-Flow, a unified

architecture that integrates diffusion transformers with autoregressive vision–language models through layerwise timestep experts and residual attention, this work shows that these paradigms can coexist within a single, scalable framework. The resulting model achieves strong performance across both multimodal understanding and generation tasks, illustrating that architectural unification—rather than paradigm replacement—is a viable path toward general-purpose multimodal systems.

Third, this dissertation establishes multimodal instruction tuning as a central mechanism for improving zero-shot generalization, robustness, and human-preference alignment. Through the introduction of MultiInstruct and Vision-Flan, this work significantly expands the scale, diversity, and coverage of multimodal instruction-following data. The proposed two-stage instruction-tuning paradigm demonstrates that instruction tuning is not merely an extension of language-model training, but a foundational component for aligning multimodal models with human intent across unseen tasks. Empirical results across a wide range of benchmarks show consistent improvements in generalization, reinforcing the role of instruction tuning as a unifying post-training strategy for multimodal intelligence.

Finally, this dissertation extends unified multimodal modeling from static images to spatiotemporal video reasoning, addressing a critical gap in existing multimodal systems. It introduces temporal global policy optimization (TGPO), a reinforcement-learning algorithm that explicitly incentivizes temporal awareness. Specifically, TGPO contrasts model outputs generated from temporally ordered versus shuffled video frames to derive calibrated, globally normalized reward signals that explicitly favor temporally coherent reasoning. The results demonstrate that temporal reasoning does not naturally emerge from static-image training alone; instead, it must be explicitly encouraged through objective design. This contribution highlights the importance of temporal structure in advancing multimodal intelligence toward real-world, dynamic environments.

Taken together, this dissertation presents a cohesive framework for unified and generalizable multimodal foundation models. By jointly addressing interleaved generation, architectural efficiency, instruction-following generalization, and temporal reasoning, the work moves beyond incremental improvements to existing models and instead redefines how multimodal systems can be designed, trained, and evaluated. It shows that progress in multimodal AI requires not only larger datasets and models, but also principled architectural specialization, scalable post-training paradigms, and explicit incentives for temporal understanding.

7.2 Future Work

Unified image encoders for joint understanding and generation. A natural next step is the development of a single, unified image encoder that simultaneously captures high-level semantic abstractions and preserves reconstructable visual details. While current multimodal systems often rely on separate representations for perception and synthesis, this

separation introduces a structural gap between understanding and generation. Future work will explore representation learning strategies that maintain semantic compactness while retaining sufficient low-level information for faithful reconstruction. Such shared representations would enable both discriminative and generative tasks to operate within a common feature space, reducing modality fragmentation and improving robustness under cross-domain transfer and adversarial distribution shifts.

Long-horizon, self-evolving, and reflective multimodal agents. Building upon unified multimodal foundations, future research will investigate agentic systems capable of long-horizon reasoning, persistent memory, and self-evolution through experience, while maintaining reliability, safety, and alignment as autonomy increases. Rather than executing fixed pipelines, such agents would dynamically construct and adapt workflows to solve open-ended tasks over extended temporal horizons, leveraging memory mechanisms, experience replay, and continual learning objectives to progressively refine strategies, representations, and decision policies. To support trustworthy behavior, this direction will also integrate self-critique and reflective feedback loops that enable agents to evaluate their own outputs, summarize acquired skills, detect failure modes, and correct erroneous or unsafe actions via internal consistency checks, retrospective reasoning, and safety-oriented objectives. Together, these advances aim to move multimodal models from reactive inference engines toward adaptive problem solvers that improve through interaction while remaining transparent, resilient, and dependable in real-world deployments.

Physical consistency and long-term dynamics modeling. Future work will pursue world modeling from raw video that learns dynamics without explicit action supervision, treating large-scale passive video as the primary signal for motion, causality, and temporal structure. A central goal is to learn predictive representations that support long-horizon forecasting while remaining physically consistent, by capturing object interactions, motion constraints, and conservation-like regularities rather than relying on short-range correlations across frames. Building on such physics-aware video models, an additional challenge is to bridge high-dimensional pixels to low-dimensional actions and plans, enabling the system to translate perceptual states into abstract decision variables and to reason about outcomes under candidate strategies. Together, these directions aim to couple video-only learning with physically grounded temporal coherence and actionable abstractions, moving unified multimodal models toward reliable long-term prediction and decision-making in dynamic environments.

Bibliography

[1] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520, 2022.

[2] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12248–12267. Association for Computational Linguistics, 2024.

[3] Firoj Alam, Tanvirul Alam, Md Hasan, Abul Hasnat, Muhammad Imran, Ferda Ofli, et al. Medic: a multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, pages 1–24, 2022.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. 2022.

[5] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023.

[6] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18187–18197. IEEE, 2022.

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou,

Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, abs/2309.16609, 2023.

[8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[9] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[10] Leonard Bärmann and Alexander H. Waibel. Where did i leave my keys? — episodic-memory-based question answering on egocentric videos. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1559–1567, 2022.

[11] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions.

[12] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[13] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18392–18402. IEEE, 2023.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[15] Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. Wikiweb2m: A page-level multimodal wikipedia dataset. *CoRR*, abs/2305.05432, 2023.

[16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[17] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022.

[18] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025.

[19] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195, 2023.

[20] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793, 2023.

[21] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[22] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[23] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *CoRR*, abs/2312.06722, 2023.

[24] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[25] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024.

[26] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. ANOLE: an open, autoregressive, native large multimodal models for interleaved image-text generation. *CoRR*, abs/2407.06135, 2024.

[27] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[28] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3646–3656, 2020.

[29] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[30] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi,

David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash Katariya, David Rim, Mojtaba Seyedhosseini, Sean Ammirati, Jonas Valfridsson, Mahan Malihi, Timothy Knight, Andeep Toor, Thomas Lampe, Abe Ittycheriah, Lewis Chiang, Chak Yeung, Alexandre Fréchette, Jinmeng Rao, Huisheng Wang, Himanshu Srivastava, Richard Zhang, Rocky Rhodes, Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gimeno, Rachana Fellinger, Pierre Marcenac, José Leal, Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sheryl Luo, Dan Garrette, Vera Axelrod, Sorin Baltateanu, David Barker, Dongkai Chen, Horia Toma, Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yujing Zhang, Hongbin Liu, Chao Wang, Martin Polacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, Megan Barnes, Ishaan Malhi, Anfal Siddiqui, Qixuan Feng, Mihai Damaschin, Daniele Pighin, Andreas Steiner, Samuel Yang, Ramya Sree Boppana, Simeon Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika, Da Huang, Christopher A. Choquette-Choo, Mohak Patel, Tianhe Yu, Toni Creswell, Jerry, Liu, Catarina Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton, Binbin Xiong, Jiaqi Pan, Thomas Strohmann, Tolly Powell, Babi Seal, Doug DeCarlo, Pranav Shyam, Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Immanuel Odisho, Josef Broder, Oscar Chang, Arun Nair, Artem Shtefan, Maura O'Brien, Manu Agarwal, Sahitya Potluri, Siddharth Goyal, Amit Jhindal, Saksham Thakur, Yury Stuken, James Lyon, Kristina Toutanova, Fangxiaoyu Feng, Austin Wu, Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman, Disha Shrivastava, Chongyang Shi, Hamish Tomlinson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer, Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga, Vincent Perot, Nuo Wang Pierse, Feng Han, Yoel Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Yeung, Dave Dopson, Yi ting Chen, Apoorv Kulshreshtha, Tongfei Guo, Philip Pham, Tal Schuster, Junquan Chen, Alex Polozov, Jinwei Xing, Huanjie Zhou, Praneeth Kacham, Doron Kukliansky, Antoine Miech, Sergey Yaroshenko, Ed Chi, Sholto Douglas, Hongliang Fei, Mathieu Blondel, Preethi Myla, Lior Madmoni, Xing Wu, Daniel Keysers, Kristian Kjems, Isabela Albuquerque, Lijun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu, Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Florian Hartmann, Subha Puttagunta, Hugo Song, Ehsan Amid, Anca Stefanoiu, Andrew Lee, Paul Pucciarelli, Emma Wang, Amit Raul, Slav Petrov, Isaac Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max Schumacher, Grace Vesom, Alex Panagopoulos, Konstantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan Zhang, Bill Rosgen, Matija Kecman, Matthew Tung, Alexandra Belias, Noah Goodman, Paul Covington, Brian Wieder, Nikita Saxena, Elnaz Davoodi, Muhuan Huang, Sharath Maddineni, Vincent Roulet, Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xintian, Wu, Guangda Lai, Paul Collins, Alex Haig, Vytenis Sakenas, Xiaowei Xu, Marissa Giustina, Laurent El Shafey, Pichi Charoenpanit, Shefali Garg, Joshua Ainslie, Boone Severson, Montse Gonzalez Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng, Michiel Bakker, Sheng Li, Nevan Wichers, Jamie Rogers, Xinyang Geng, Yeqing Li, Rolf Jager-

man, Chao Jia, Nadav Olmert, David Sharon, Matthew Mauger, Sandeep Mariserla, Hongxu Ma, Megha Mohabey, Kyuyeun Kim, Alek Andreev, Scott Pollom, Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu, Andrew Leach, Irina Blok, Ganesh Poomal Girirajan, Roee Aharoni, Benigno Uria, Andrei Sozanschi, Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro, Martin Zlocha, Vighnesh Birodkar, Sami Lachgar, Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg, Fei Zheng, Gregory Dibb, Emily Graves, Swachhand Lokhande, Gabriel Rasskin, George-Cristian Muraru, Corbin Quick, Sandeep Tata, Pierre Sermanet, Aditya Chawla, Itay Karo, Yan Wang, Susan Zhang, Orgad Keller, Anca Dragan, Guolong Su, Ian Chou, Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wilson, Ruibo Liu, Shibo Wang, Georgie Evans, David Du, Alfonso Castaño, Gautam Prasad, Mona El Mahdy, Sebastian Gerlach, Machel Reid, Jarrod Kahn, Amir Zait, Thanumalayan Sankaranarayana Pillai, Thatcher Ulrich, Guanyu Wang, Jan Wassenberg, Efrat Farkash, Kiran Yalasangi, Congchao Wang, Maria Bauza, Simon Bucher, Ting Liu, Jun Yan, Gary Leung, Vikas Sindhwani, Parker Barnes, Avi Singh, Ivan Jurin, Jichuan Chang, Niket Kumar Bhumihar, Sivan Eiger, Gui Citovsky, Ben Withbroe, Zhang Li, Siyang Xue, Niccolò Dal Santo, Georgi Stoyanov, Yves Raimond, Steven Zheng, Yilin Gao, Vít Listík, Sławek Kwasiborski, Rachel Saputro, Adnan Ozturel, Ganesh Mallya, Kushal Majmundar, Ross West, Paul Caron, Jinliang Wei, Lluis Castrejon, Sharad Vikram, Deepak Ramachandran, Nikhil Dhawan, Jiho Park, Sara Smoot, George van den Driessche, Yochai Blau, Chase Malik, Wei Liang, Roy Hirsch, Cicero Nogueira dos Santos, Eugene Weinstein, Aäron van den Oord, Sid Lall, Nicholas FitzGerald, Zixuan Jiang, Xuan Yang, Dale Webster, Ali Elqursh, Aedan Pope, Georges Rotival, David Raposo, Wanzheng Zhu, Jeff Dean, Sami Alabed, Dustin Tran, Arushi Gupta, Zach Gleicher, Jessica Austin, Edouard Rosseel, Megh Umekar, Dipanjan Das, Yinghao Sun, Kai Chen, Karolis Misiunas, Xiang Zhou, Yixian Di, Alyssa Loo, Josh Newlan, Bo Li, Vinay Ramasesh, Ying Xu, Alex Chen, Sudeep Gandhe, Radu Soricut, Nikita Gupta, Shuguang Hu, Seliem El-Sayed, Xavier Garcia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt, Lu Huang, Alex Gurney, Zhiying Zhang, Alexander Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhargav Kanagal Shamanna, Felix Fischer, Josef Dean, Karan Gill, Ross Mcilroy, Abhishek Bhowmick, Jeremy Selier, Antoine Yang, Derek Cheng, Vladimir Magay, Jie Tan, Dhriti Varma, Christian Walder, Tomas Kocisky, Ryo Nakashima, Paul Natsev, Mike Kwong, Ionel Gog, Chiyuan Zhang, Sander Dieleman, Thomas Jimma, Andrey Ryabtsev, Siddhartha Brahma, David Steiner, Dayou Du, Ante Žužul, Mislav Žanić, Mukund Raghavachari, Willi Gierke, Zeyu Zheng, Dessie Petrova, Yann Dauphin, Yuchuan Liu, Ido Kessler, Steven Hand, Chris Duvarney, Seokhwan Kim, Hyo Lee, Léonard Hussenot, Jeffrey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gaurav Singh Tomar, Keyvan Amiri, Du Phan, Fabian Fuchs, Tobias Weyand, Nenad Tomasev, Alexandra Cordell, Xin Liu, Jonathan Mallinson, Pankaj Joshi, Andy Crawford, Arun Suggala, Steve Chien, Nick Fernando, Mariella Sanchez-Vargas, Duncan Williams, Phil Crone, Xiyang Luo, Igor Karpov, Jyn Shan, Terry Thurk, Robin Strudel, Paul Voigtlaender, Piyush Patil,

Tim Dozat, Ali Khodaei, Sahil Singla, Piotr Ambroszczyk, Qiyin Wu, Yifan Chang, Brian Roark, Chaitra Hegde, Tianli Ding, Angelos Filos, Zhongru Wu, André Susano Pinto, Shuang Liu, Saarthak Khanna, Aditya Pandey, Siobhan Mcloughlin, Qiujia Li, Sam Haves, Allan Zhou, Elena Buchatskaya, Isabel Leal, Peter de Boursac, Nami Akazawa, Nina Anderson, Terry Chen, Krishna Somandepalli, Chen Liang, Sheela Goenka, Stephanie Winkler, Alexander Grushetsky, Yifan Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric Li, Ruichao Li, Tomer Golany, Dawid Wegner, Tao Jiang, Omer Barak, Yuan Shangguan, Eszter Vértes, Renee Wong, Jörg Bornschein, Alex Tudor, Michele Bevilacqua, Tom Schaul, Ankit Singh Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng, Cory McLean, Jonathan Lai, Jennifer Beattie, Nate Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang, Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra, Myriam Khan, Megha Goel, Rob Willoughby, David Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Kementsietsidis, Arjun Khare, James Keeling, Keyang Xu, Nathan Waters, Florent Altché, Ashok Popat, Bhavishya Mittal, David Saxton, Dalia El Badawy, Michael Mathieu, Zheng Zheng, Hao Zhou, Nishant Ranka, Richard Shin, Qingnan Duan, Tim Salimans, Ioana Mihailescu, Uri Shaham, Ming-Wei Chang, Yannis Assael, Nishanth Dikkala, Martin Izzard, Vincent Cohen-Addad, Cat Graves, Vlad Feinberg, Grace Chung, DJ Strouse, Danny Karmon, Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham, Jon Blanton, Alex Vasiloff, Jarred Barber, Mark Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang, Lei Zhang, Himanshu Gupta, Matt Young, Julia Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan, Beer Changpinyo, Filip Pavetić, Amy Coyle, Yasuhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou, Harish Rajamani, Blake Hechtman, Eddie Cao, Da-Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dimple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-Chih Hung, Steven Pecht, Georgi Todorov, Nikhil Khadke, Pramod Gupta, Preethi Lahoti, Arnaud Autef, Karthik Duddu, James Lee-Thorp, Alexander Bykovsky, Tautvydas Misiunas, Sebastian Flennerhag, Santhosh Thangaraj, Jed McGiffin, Zack Nado, Markus Kunesch, Andreas Noever, Amir Hertz, Marco Liang, Victor Stone, Evan Palmer, Samira Daruki, Arijit Pramanik, Siim Põder, Austin Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter, Avraham Ruderman, Wenlei Zhou, Chirag Nagpal, Kiran Vodrahalli, George Necula, Paul Barham, Ellie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao, Maciej Mikuła, Tom Eccles, Hidetoshi Shimokawa, Kanav Garg, Luke Vilnis, Hanwen Chen, Ilia Shumailov, Kuang-Huei Lee, Abdelrahman Abdelhamed, Meiyan Xie, Vered Cohen, Ester Hlavnova, Dan Malkin, Chawin Sitawarin, James Lottes, Pauline Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang, Aroma Mahendru, Zafarali Ahmed, James Martens, Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin, Arseniy Klimovskiy, Mary Phuong, Danny Vainstein, Jin Xie, Bhuvana Ramabhadran, Nathan Howard, Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw, James Thornton, Adi Gerzi Rosenthal, Shane Gu, Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio, Aarush Selvan, Joel Wee, Chenxi Liu,

Maryam Majzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty, Renke Pan, Hoang Nguyen, Qiong, Hu, Sarah Perrin, Abhinav Arora, Emily Pitler, Weiyi Wang, Kaushik Shivakumar, Flavien Prost, Ben Limonchik, Jing Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan, Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha Vasudevan, Eli Collins, Vasilisa Bashlovkina, Tony Lu, Doug Fritz, Jongbin Park, Yanqi Zhou, Chen Su, Richard Tanburn, Mikhail Sushkov, Mitchelle Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li, Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang, Andrew Dai, Phuong Dao, Mike Burrows, Henry Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoesund, Art Khurshudov, Zhenkai Zhu, Albert Webson, Elizabeth Kemp, Tat Tan, Saurabh Agrawal, Susie Sargsyan, Liqun Cheng, Jim Stephan, Tom Kwiatkowski, David Reid, Arunkumar Byravan, Assaf Hurwitz Michaely, Nicolas Heess, Luowei Zhou, Sonam Goenka, Viral Carpenter, Anselm Levskaya, Bo Wang, Reed Roberts, Rémi Leblond, Sharat Chikkerur, Stav Ginzburg, Max Chang, Robert Riachi, Chuqiao, Xu, Zalán Borsos, Michael Pliskin, Julia Pawar, Morgane Lustman, Hannah Kirkwood, Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran Milan, Sean Augenstein, Anna Goldie, Laurel Prince, Karthik Raman, Yanhua Sun, Vivian Xia, Aaron Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis, Lukas Zilka, David Vilar Torres, Lisa Patel, Sho Arora, Betty Chan, Jonas Adler, Kareem Ayoub, Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baumgartner, Haitian Sun, Yael Karov, Yaroslav Akulov, Hui Zheng, Irene Cai, Claudio Fantacci, James Rubin, Alex Rav Acha, Mengchao Wang, Nina D'Souza, Rohit Sathyanarayana, Shengyang Dai, Simon Rowe, Andrey Simanovsky, Omer Goldman, Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg, Tania Rojas-Esponda, Praneet Dutta, Amy Zeng, Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq Iqbal, Becca Roelofs, Ga-Young Joung, Parker Beak, Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei, Guy Bensky, Timo Denk, Harish Ganapathy, Jeff Stanway, Pratik Joshi, Francesco Bertolini, Jessica Lo, Olivia Ma, Zachary Charles, Geta Sampemane, Himanshu Sahni, Xu Chen, Harry Askham, David Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur Bražinskas, Li Zhong, Zhichun Wu, Mark Epstein, Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Goldshtein, Alex Ruiz, Mohammed Badawi, Matthias Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo, Theophane Weber, Haichuan Yang, Pan-Pan Jiang, Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou, Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan, Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko, Salil Deshmukh, Apoorv Jindal, James Svensson, Divya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vincent Tsang, Bowen Liang, Fernando Pereira, Suhas Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thiagarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil Jain, Ishita Dasgupta, Taylan Bilal, David Reitter, Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio Guadarrama, Zaheer Ab-

bas, Stefani Karp, Petko Georgiev, Chun-Sung Ferng, Marc Brockschmidt, Liqian Peng, Christoph Hirnschall, Vikas Verma, Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu, Phil Wallis, Randall Parker, Qifei Wang, Yang Xu, Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon Kim, Andrea Gesmundo, Mackenzie Thomas, Sergey Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst, Sam Conway-Rahman, Helen Ran, Kay McKinney, Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapathy, Jane Labanowski, Zachary Garrett, Ben Vargas, Sean Sun, Bryan Gale, Thomas Buschmann, Guillaume Desjardins, Nimesh Ghelani, Palak Jain, Mudit Verma, Chulayuth Asawaroengchai, Julian Eisenschlos, Jitendra Harlalka, Hideto Kazawa, Don Metzler, Joshua Howland, Ying Jian, Jake Ades, Viral Shah, Tynan Gangwani, Seungji Lee, Roman Ring, Steven M. Hernandez, Dean Reich, Amer Sinha, Ashutosh Sathe, Joe Kovac, Ashleah Gill, Ajay Kannan, Andrea D'olimpio, Martin Sevenich, Jay Whang, Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang, Shuba Lall, Yossi Matias, Bill Jia, Abe Friesen, Sara Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu, Anna Shekhawat, Safeen Huda, Peter Grabowski, Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi Hassen, Parker Schuh, Vedrana Milutinovic, Chris Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe Barth-Maron, Justin Frye, Natalie Axelsson, Tao Zhu, Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang Ye, Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sampathkumar, Ronny Huang, Victor Lavrenko, Ahmed Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy, Bidisha Samanta, Domagoj Ćevid, Xin Wang, Fangtao Li, Michael Voznesensky, Matt Hoffman, Andreas Terzis, Vikash Sehwag, Gil Fidel, Luheng He, Mu Cai, Yanzhang He, Alex Feng, Martin Nikoltchev, Samrat Phatale, Jason Chase, Rory Lawton, Ming Zhang, Tom Ouyang, Manuel Tragut, Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li, Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing Han, Thomas Anthony, Sneha Kudugunta, Viorica Patraucean, Aniket Ray, Xinyun Chen, Xiaochen Yang, Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija Ražnatović, Bethanie Brownfield, James An, Sheng Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua Kessinger, James Noraky, Siqi Liu, Keran Rong, Petar Veličković, Keith Rush, Alex Goldin, Fanny Wei, Shiva Mohan Reddy Garlapati, Caroline Pantofaru, Okwan Kwon, Jianmo Ni, Eric Noland, Julia Di Trapani, Françoise Beaufays, Abhijit Guha Roy, Yinlam Chow, Aybuke Turker, Geoffrey Cideron, Lantao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak, Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Milad Nasr, Chester Kwak, Suraj Satishkumar Sheth, Alex Kaskasoli, Ankesh Anand, Balaji Lakshminarayanan, Sammy Jerome, David Bieber, Chun-Te Chu, Alexandre Senges, Tianxiao Shen, Mukund Sridhar, Ndaba Ndebele, Benjamin Beyret, Shakir Mohamed, Mia Chen, Markus Freitag, Jiaxian Guo, Luyang Liu, Paul Roit, Heng Chen, Shen Yan, Tom Stone, JD Co-Reyes, Jeremy Cole, Salvatore Scellato, Shekoofeh Azizi, Hadi Hashemi, Alicia Jin, Anand Iyer, Marcella Valentine, András György, Arun Ahuja, Daniel Hernandez Diaz, Chen-Yu Lee, Nathan Clement, Weize Kong, Drew Garmon, Ishaan Watts, Kush Bhatia, Khyatti Gupta,

Matt Miecnikowski, Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi, James Atwood, Jo Chick, Mark Collier, Fotis Iliopoulos, Ryan Trostle, Beliz Gunel, Ramiro Leal-Cavazos, Arnar Mar Hrafnkelsson, Michael Guzman, Xiaoen Ju, Andy Forbes, Jesse Emond, Kushal Chauhan, Ben Caine, Li Xiao, Wenjun Zeng, Alexandre Moufarek, Daniel Murphy, Maya Meng, Nitish Gupta, Felix Riedel, Anil Das, Elijah Lawal, Shashi Narayan, Tiberiu Sosea, James Swirhun, Linda Friso, Behnam Neyshabur, Jing Lu, Sertan Girgin, Michael Wunder, Edouard Yvinec, Aroonalok Pyne, Victor Carbune, Shruti Rijhwani, Yang Guo, Tulsee Doshi, Anton Briukhov, Max Bain, Ayal Hitron, Xuanhui Wang, Ashish Gupta, Ke Chen, Cosmo Du, Weiyang Zhang, Dhruv Shah, Arjun Akula, Max Dylla, Ashyana Kachra, Weicheng Kuo, Tingting Zou, Lily Wang, Luyao Xu, Jifan Zhu, Justin Snyder, Sachit Menon, Orhan Firat, Igor Mordatch, Yuan Yuan, Natalia Ponomareva, Rory Blevins, Lawrence Moore, Weijun Wang, Phil Chen, Martin Scholz, Artur Dwornik, Jason Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan Song, Matt Miller, Uday Kalra, Adam Raveret, Oscar Akerlund, Felix Wu, Andrew Nystrom, Namrata Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao, Buhuang Liu, Avi Caciularu, Lauren Lax, Urvashi Khandelwal, Victoria Langston, Eric Bailey, Silvio Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mondal, Guru Guruganesh, Nan Hua, Ofir Roval, Paweł Wesołowski, Rishikesh Ingale, Jonathan Halcrow, Tim Sohn, Christof Angermueller, Bahram Raad, Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timothy Lillicrap, Austin Huang, Lydia Lihui Zhang, Dominik Paulus, Clement Farabet, Alex Wertheim, Bing Wang, Rishabh Joshi, Chu ling Ko, Yonghui Wu, Shubham Agrawal, Lily Lin, XiangHai Sheng, Peter Sung, Tyler Breland-King, Christina Butterfield, Swapnil Gawde, Sumeet Singh, Qiao Zhang, Raj Apte, Shilpa Shetty, Adrian Hutter, Tao Li, Elizabeth Salesky, Federico Lebron, Jonni Kanerva, Michela Paganini, Arthur Nguyen, Rohith Vallu, Jan-Thorsten Peter, Sarmishta Velury, David Kao, Jay Hoover, Anna Bortsova, Colton Bishop, Shoshana Jakobovits, Alessandro Agostini, Alekh Agarwal, Chang Liu, Charles Kwong, Sasan Tavakkol, Ioana Bica, Alex Greve, Anirudh GP, Jake Marcus, Le Hou, Tom Duerig, Rivka Moroshko, Dave Lacey, Andy Davis, Julien Amelot, Guohui Wang, Frank Kim, Theofilos Strinopoulos, Hui Wan, Charline Le Lan, Shankar Krishnan, Haotian Tang, Peter Humphreys, Junwen Bai, Idan Heimlich Shtacher, Diego Machado, Chenxi Pang, Ken Burke, Dangyi Liu, Renga Aravamudhan, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly Fu, Robin Alazard, Barak Meiri, Daniel Winter, Charlie Chen, Mingda Zhang, Jens Heitkaemper, John Lambert, Jinhyuk Lee, Alexander Frömmgen, Sergey Rogulenko, Pranav Nair, Paul Niemczyk, Anton Bulyenov, Bibo Xu, Hadar Shemtov, Morteza Zadimoghaddam, Serge Toropov, Mateo Wirth, Hanjun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex Kurakin, Chansoo Lee, Kalesha Bullard, Nicolas Serrano, Ivana Balazevic, Yang Li, Johan Schalkwyk, Mark Murphy, Mingyang Zhang, Kevin Sequeira, Romina Datta, Nishant Agrawal, Charles Sutton, Nithya Attaluri, Mencher Chiang, Wael Farhan, Gregory Thornton, Kate Lin, Travis Choma, Hung Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia Comşa, Michael Riley, Arjun Pillai, Basil Mustafa, Ben Golan, Amir Zandieh, Jean-Baptiste Lespiau, Billy Porter,

David Ross, Sujeevan Rajayogam, Mohit Agarwal, Subhashini Venugopalan, Bobak Shahriari, Qiqi Yan, Hao Xu, Taylor Tobin, Pavel Dubov, Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Sebastian Borgeaud, Lucio Dery, Shanthal Vasanth, Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh, Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron Archer, Chris Rawles, Elinor Davies, Alexey Svyatkovskiy, Tomy Tsai, Yoni Halpern, Christian Reisswig, Bartek Wydrowski, Bo Chang, Joan Puigcerver, Mor Hazan Taege, Jian Li, Eva Schnider, Xinjian Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha Subramaniam, Aviral Kumar, Pete Blois, Zhuyun Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian Xie, Anja Hauth, Alexandru Ţifrea, Yuqi Li, Sam El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding, Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai, Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen, Alec Go, Pablo Sprechmann, Nithish Kannen, Alen Carin, Paramjit Sandhu, Isabel Edkins, Leslie Nooteboom, Jai Gupta, Loren Maggiore, Javad Azizi, Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny Tarlow, Duncan Smith, Desi Ivanov, Mohammad Babaeizadeh, Ankita Goel, Satish Kambala, Grace Chu, Matej Kastelic, Michelle Liu, Hagen Soltau, Austin Stone, Shivani Agrawal, Min Kim, Kedar Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel Soh, Arvind Kannan, DY Kim, Blake JianHang Chen, Afief Halumi, Sudeshna Roy, Yulong Wang, Olcan Sercinoglu, Gena Gibson, Sijal Bhatnagar, Motoki Sano, Daniel von Dincklage, Qingchun Ren, Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl Doersch, Jilei, Wang, Bingyuan Liu, Qijun Tan, Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl Lebsack, Josh Dillon, Rajiv Mathews, Tom Cobley, Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop Nath, Tara Sainath, Alexei Bendebury, Ryan Julian, Bharath Mankalale, Daria Ćurko, Paulo Zacchello, Adam R. Brown, Kiranbir Sodhia, Heidi Howard, Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna Bulanova, Lesley Katzen, Roman Goldenberg, Anton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam Upadhyay, Cheng Li, Soroush Radpour, Marcello Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan, Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani, Trevor Yacovone, Ryan Kappedal, Kostas Andriopoulos, Li Lao, YaGuang Li, Grigory Rozhdestvenskiy, Kazuma Hashimoto, Andrew Audibert, Sophia Austin, Daniel Rodriguez, Anian Ruoss, Garrett Honke, Deep Karkhanis, Xi Xiong, Qing Wei, James Huang, Zhaoqi Leng, Vittal Premachandran, Stan Bileschi, Georgios Evangelopoulos, Thomas Mensink, Jay Pavagadhi, Denis Teplyashin, Paul Chang, Linting Xue, Garrett Tanzer, Sally Goldman, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan Mullins, Alexey Guseynov, Chung-Ching Chang, Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham, Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith, Alex Irpan, Carey Radebaugh, Summer Yue, Lijie Fan, Victor Ungureanu, Christina Sorokin, Hannah Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd Wang, Chu-Cheng Lin, Hui Peng, Megan Shum, Goran Petrovic, Demetra Brady, Richard Nguyen, Klaus

Macherey, Zhihao Li, Harman Singh, Madhavi Yenugula, Mariko Iinuma, Xinyi Chen, Kavya Kopparapu, Alexey Stern, Shachi Dave, Chandu Thekkath, Florence Perot, Anurag Kumar, Fangda Li, Yang Xiao, Matthew Bilotti, Mohammad Hossein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang, Ela Gruzewska, Anand Rao, Sindhu Raghuram, Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia, Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena Pankov, Steven Schwarcz, Andrea Burns, Christine Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, Antoine He, Amy Stuart, Arun Narayanan, Yukun Zhu, Christian Frank, Bahar Fatemi, Amit Sabne, Oran Lang, Indro Bhattacharya, Shane Settle, Maria Wang, Brendan McMahan, Andrea Tacchetti, Livio Baldini Soares, Majid Hadian, Serkan Cabi, Timothy Chung, Nikita Putikhin, Gang Li, Jeremy Chen, Austin Tarango, Henryk Michalewski, Mehran Kazemi, Hussain Masoom, Hila Sheftel, Rakesh Shivanna, Archita Vadali, Ramona Comanescu, Doug Reid, Joss Moore, Arvind Neelakantan, Michaël Sander, Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani, JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shitao Weng, Chia-Hua Ho, John Karro, Kalpesh Krishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel Eppens, Marco Andreetto, Navin Sarma, Silvano Bonacina, Burcu Karagol Ayan, Megha Nawhal, Zhihao Shan, Mike Dusenberry, Shantanu Thakoor, Sagar Gubbi, Duc Dung Nguyen, Reut Tsarfaty, Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-Juen Chen, Alessandro Epasto, Georgi Stephanov, Ye Jin, Samuel Gehman, Aida Amini, Jack Weber, Feryal Behbahani, Shawn Xu, Miltos Allamanis, Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski, Hang Qi, David Greene, Xinyi Wu, Abodunrinwa Toki, Daniel Vlasic, Jane Shapiro, Ragha Kotikalapudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cassirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming Mao, João Gabriel Oliveira, Zoltan Egyed, Bernd Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juliette Pluto, Petros Maniatis, Steve Li, Yaohui Guo, Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatterjee, Julie Jin, Katrina, Xu, Jennimaria Palomaki, Séb Arnold, Madhavi Sewak, Federico Piccinini, Mohit Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin Vaswani, Chongyan Chen, Matheus Wisniewski, Qin Cao, John Aslanides, Nguyet Minh Phu, Maximilian Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn, Marco Selvi, Anders Andreassen, Krishan Subudhi, Prem Eruvbetine, Oliver Woodman, Tomas Mery, Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr, Shixin Luo, Riccardo Patana, Matthew Watson, Dani Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel Hoogeboom, Sharon Silver, Emily Caveness, Xiaokai Zhao, Mikel Rodriguez, Maxine Deines, Libin Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue, Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria Shen, Elijah Peake, Remi Crocker, Shubha Srinivas Raghvendra, Danny Swisher, Woohyun Han, Richa Singh, Ling Wu, Vladimir Pchelin, Tsendsuren Munkhdalai, Dana Alon, Geoff Bacon, Efren Robles, Jannis Bulian, Melvin Johnson, George Powell, Felipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing, Mihajlo Velimirović, Hubert Soyer, William Kong, Tony, Nguyên, Zhen Yang, Jeremiah Liu, Joost van Amersfoort, Daniel Gillick, Baochen Sun,

Nathalie Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou, Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis Rouillard, Hongji Li, Amol Mandhane, Nova Fallen, Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Brennan, Jenny Lee, Jerry Chang, Martin Sundermeyer, Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fabrikant, William Bono, Jindong Gu, Ryan Foley, Yiran Mao, Manolis Delakis, Dhruva Bhaswar, Roy Frostig, Nick Li, Avital Zipori, Cath Hope, Olga Kozlova, Swaroop Mishra, Josip Djolonga, Craig Schiff, Majd Al Merey, Eleftheria Briakou, Peter Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen Kallakuri, Paige Kunkle, Hannah Brennan, Tom Lieber, Hassan Mansoor, Julian Walker, Bing Zhang, Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex Druinsky, Donghyun Cho, Rui Yao, Ferjad Naeem, Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang, James Zhao, Andrew Over, Abhishek Chakladar, Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale Cong, Prakash Shroff, Carl Saroufim, Sobhan Miryoosefi, Mohamed Hammad, Tajwar Nasir, Weijuan Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu, Martin Matysiak, Naina Raisinghani, Huiyu Wang, Lexi Baugher, Rahul Sukthankar, Minh Giang, John Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine Smith, Andy Ly, Yicheng Wang, Rishabh Bansal, Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh Keshava, Deepak Sharma, Marvin Chow, Kunal Lad, Komal Jalan, Simon Osindero, Craig Swanson, Jacob Scott, Anastasija Ilić, Xiaowei Li, Siddhartha Reddy Jonnalagadda, Afzal Shama Soudagar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jarrett, Naveen Kumar, Maulik Shah, Matt Lawlor, Austin Waters, Mark Graham, Rhys May, Sabela Ramos, Sandra Lefdal, Zeynep Cankara, Nacho Cano, Brendan O'Donoghue, Jed Borovik, Frederick Liu, Jordan Grimstad, Mahmoud Alnahlawi, Katerina Tsihlas, Tom Hudson, Nikolai Grigorev, Yiling Jia, Terry Huang, Tobenna Peter Igwe, Sergei Lebedev, Xiaodan Tang, Igor Krivokon, Frankie Garcia, Melissa Tan, Eric Jia, Peter Stys, Shikhar Vashishth, Yu Liang, Balaji Venkatraman, Chenjie Gu, Anastasios Kementsietsidis, Chen Zhu, Junehyuk Jung, Yunfei Bai, Mohammad Javad Hosseini, Faruk Ahmed, Aditya Gupta, Xin Yuan, Shereen Ashraf, Shitij Nigam, Gautam Vasudevan, Pranjal Awasthi, Adi Mayrav Gilady, Zelda Mariet, Ramy Eskander, Haiguang Li, Hexiang Hu, Guillermo Garrido, Philippe Schlattner, George Zhang, Rohun Saxena, Petar Dević, Kritika Muralidharan, Ashwin Murthy, Yiqian Zhou, Min Choi, Arissa Wongpanich, Zhengdong Wang, Premal Shah, Yuntao Xu, Yiling Huang, Stephen Spencer, Alice Chen, James Cohan, Junjie Wang, Jonathan Tompson, Junru Wu, Ruba Haroun, Haiqiong Li, Blanca Huergo, Fan Yang, Tongxin Yin, James Wendt, Michael Bendersky, Rahma Chaabouni, Javier Snaider, Johan Ferret, Abhishek Jindal, Tara Thompson, Andrew Xue, Will Bishop, Shubham Milind Phal, Archit Sharma, Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shomrat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mihai Dorin Istin, Sam Sobell, Yang Lu, Emily Nottage, Dorsa Sadigh, Jeremiah Willcock, Tingnan Zhang, Steve Xu, Sasha Brown, Katherine Lee, Gary Wang, Yun Zhu, Yi Tay, Cheolmin Kim, Audrey Gutierrez, Abhanshu Sharma, Yongqin Xian, Sungy-

ong Seo, Claire Cui, Elena Pochernina, Cip Baetu, Krzysztof Jastrzębski, Mimi Ly, Mohamed Elhawaty, Dan Suh, Eren Sezener, Pidong Wang, Nancy Yuen, George Tucker, Jiahao Cai, Zuguang Yang, Cindy Wang, Alex Muzio, Hai Qian, Jae Yoo, Derek Lockhart, Kevin R. McKee, Mandy Guo, Malika Mehrotra, Artur Mendonça, Sanket Vaibhav Mehta, Sherry Ben, Chetan Tekur, Jiaqi Mu, Muye Zhu, Victoria Krakovna, Hongrae Lee, AJ Maschinot, Sébastien Cevey, HyunJeong Choe, Aijun Bai, Hansa Srinivasan, Derek Gasaway, Nick Young, Patrick Siegler, Dan Holtmann-Rice, Vihari Piratla, Kate Baumli, Roey Yogev, Alex Hofer, Hado van Hasselt, Svetlana Grant, Yuri Chervonyi, David Silver, Andrew Hogue, Ayushi Agarwal, Kathie Wang, Preeti Singh, Four Flynn, Josh Lipschultz, Robert David, Lizzetth Bellot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate Olszewska, Kevin Hui, Akanksha Maurya, Nikos Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kelley, Anirudh Baddepudi, Johannes Mauerer, Gregory Shaw, Alex Siegman, Lin Yang, Shravya Shetty, Subhrajit Roy, Yunting Song, Wojciech Stokowiec, Ryan Burnell, Omkar Savant, Robert Busa-Fekete, Jin Miao, Samrat Ghosh, Liam MacDermed, Phillip Lippe, Mikhail Dektiarev, Zach Behrman, Fabian Mentzer, Kelvin Nguyen, Meng Wei, Siddharth Verma, Chris Knutsen, Sudeep Dasari, Zhipeng Yan, Petr Mitrichev, Xingyu Wang, Virat Shejwalkar, Jacob Austin, Srinivas Sunkara, Navneet Potti, Yan Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Bloniarz, Divyansh Shukla, Elizabeth Cole, Jong Lee, Sheng Zhang, Sushant Kafle, Siddharth Vashishtha, Parsa Mahmoudieh, Grace Chen, Raphael Hoffmann, Pranesh Srinivasan, Agustin Dal Lago, Yoav Ben Shalom, Zi Wang, Michael Elabd, Anuj Sharma, Junhyuk Oh, Suraj Kothawade, Maigo Le, Marianne Monteiro, Shentao Yang, Kaiz Alarakyia, Robert Geirhos, Diana Mincu, Håvard Garnes, Hayato Kobayashi, Soroosh Mariooryad, Kacper Krasowiak, Zhixin, Lai, Shibl Mourad, Mingqiu Wang, Fan Bu, Ophir Aharoni, Guanjie Chen, Abhimanyu Goyal, Vadim Zubov, Ankur Bapna, Elahe Dabir, Nisarg Kothari, Kay Lamerigts, Nicola De Cao, Jeremy Shar, Christopher Yew, Nitish Kulkarni, Dre Mahaarachchi, Mandar Joshi, Zhenhai Zhu, Jared Lichtarge, Yichao Zhou, Hannah Muckenhirn, Vittorio Selo, Oriol Vinyals, Peter Chen, Anthony Brohan, Vaibhav Mehta, Sarah Cogan, Ruth Wang, Ty Geri, Wei-Jen Ko, Wei Chen, Fabio Viola, Keshav Shivam, Lisa Wang, Madeleine Clare Elish, Raluca Ada Popa, Sébastien Pereira, Jianqiao Liu, Raphael Koster, Donnie Kim, Gufeng Zhang, Sayna Ebrahimi, Partha Talukdar, Yanyan Zheng, Petra Poklukar, Ales Mikhalap, Dale Johnson, Anitha Vijayakumar, Mark Omernick, Matt Dibb, Ayush Dubey, Qiong Hu, Apurv Suman, Vaibhav Aggarwal, Ilya Kornakov, Fei Xia, Wing Lowe, Alexey Kolganov, Ted Xiao, Vitaly Nikolaev, Steven Hemingray, Bonnie Li, Joana Iljazi, Mikołaj Rybiński, Ballie Sandhu, Peggy Lu, Thang Luong, Rodolphe Jenatton, Vineetha Govindaraj, Hui, Li, Gabriel Dulac-Arnold, Wonpyo Park, Henry Wang, Abhinit Modi, Jean Pouget-Abadie, Kristina Greller, Rahul Gupta, Robert Berry, Prajit Ramachandran, Jinyu Xie, Liam McCafferty, Jianling Wang, Kilol Gupta, Hyeontaek Lim, Blaž Bratanič, Andy Brock, Ilia Akolzin, Jim Sproch, Dan Karliner, Duhyeon Kim, Adrian Goedeckemeyer, Noam Shazeer, Cordelia

Schmid, Daniele Calandriello, Parul Bhatia, Krzysztof Choromanski, Ceslee Montgomery, Dheeru Dua, Ana Ramalho, Helen King, Yue Gao, Lynn Nguyen, David Lindner, Divya Pitta, Oleaser Johnson, Khalid Salama, Diego Ardila, Michael Han, Erin Farnese, Seth Odoom, Ziyue Wang, Xiangzhuo Ding, Norman Rink, Ray Smith, Harshal Tushar Lehri, Eden Cohen, Neera Vats, Tong He, Parthasarathy Gopavarapu, Adam Paszke, Miteyan Patel, Wouter Van Gansbeke, Lucia Loher, Luis Castro, Maria Voitovich, Tamara von Glehn, Nelson George, Simon Niklaus, Zach Eaton-Rosen, Nemanja Rakićević, Erik Jue, Sagi Perel, Carrie Zhang, Yuval Bahat, Angéline Pouget, Zhi Xing, Fantine Huot, Ashish Shenoy, Taylor Bos, Vincent Coriou, Bryan Richter, Natasha Noy, Yaqing Wang, Santiago Ontanon, Siyang Qin, Gleb Makarchuk, Demis Hassabis, Zhuowan Li, Mandar Sharma, Kumaran Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren, Chen, Manaal Faruqui, Jialin Wu, Slavica Andačić, Szabolcs Payrits, Daniel McDuff, Tom Hume, Yuan Cao, MH Tessler, Qingze Wang, Yinan Wang, Ivor Rendulic, Eirikur Agustsson, Matthew Johnson, Tanya Lando, Andrew Howard, Sri Gayatri Sundara Padmanabhan, Mayank Daswani, Andrea Banino, Michael Kilgore, Jonathan Heek, Ziwei Ji, Alvaro Caceres, Conglong Li, Nora Kassner, Alexey Vlaskin, Zeyu Liu, Alex Grills, Yanhan Hou, Roykrong Sukkerd, Gowoon Cheon, Nishita Shetty, Larisa Markeeva, Piotr Stanczyk, Tejas Iyer, Yuan Gong, Shawn Gao, Keerthana Gopalakrishnan, Tim Blyth, Malcolm Reynolds, Avishkar Bhoopchand, Misha Bilenko, Dero Gharibian, Vicky Zayats, Aleksandra Faust, Abhinav Singh, Min Ma, Hongyang Jiao, Sudheendra Vijayanarasimhan, Lora Aroyo, Vikas Yadav, Sarah Chakera, Ashwin Kakarla, Vilobh Meshram, Karol Gregor, Gabriela Botea, Evan Senter, Dawei Jia, Geza Kovacs, Neha Sharma, Sebastien Baur, Kai Kang, Yifan He, Lin Zhuo, Marija Kostelac, Itay Laish, Songyou Peng, Louis O'Bryan, Daniel Kasenberg, Girish Ramchandra Rao, Edouard Leurent, Biao Zhang, Sage Stevens, Ana Salazar, Ye Zhang, Ivan Lobov, Jake Walker, Allen Porter, Morgan Redshaw, Han Ke, Abhishek Rao, Alex Lee, Hoi Lam, Michael Moffitt, Jaeyoun Kim, Siyuan Qiao, Terry Koo, Robert Dadashi, Xinying Song, Mukund Sundararajan, Peng Xu, Chizu Kawamoto, Yan Zhong, Clara Barbu, Apoorv Reddy, Mauro Verzetti, Leon Li, George Papamakarios, Hanna Klimczak-Plucińska, Mary Cassin, Koray Kavukcuoglu, Rigel Swavely, Alain Vaucher, Jeffrey Zhao, Ross Hemsley, Michael Tschannen, Heming Ge, Gaurav Menghani, Yang Yu, Natalie Ha, Wei He, Xiao Wu, Maggie Song, Rachel Sterneck, Stefan Zinke, Dan A. Calian, Annie Marsden, Alejandro Cruzado Ruiz, Matteo Hessel, Almog Gueta, Benjamin Lee, Brian Farris, Manish Gupta, Yunjie Li, Mohammad Saleh, Vedant Misra, Kefan Xiao, Piermaria Mendolicchio, Gavin Buttimore, Varvara Krayvanova, Nigamaa Nayakanti, Matthew Wiethoff, Yash Pande, Azalia Mirhoseini, Ni Lao, Jasmine Liu, Yiqing Hua, Angie Chen, Yury Malkov, Dmitry Kalashnikov, Shubham Gupta, Kartik Audhkhasi, Yuexiang Zhai, Sudhindra Kopalle, Prateek Jain, Eran Ofek, Clemens Meyer, Khuslen Baatarsukh, Hana Strejček, Jun Qian, James Freedman, Ricardo Figueira, Michal Sokolik, Olivier Bachem, Raymond Lin, Dia Kharrat, Chris Hidey, Pingmei Xu, Dennis Duan, Yin Li, Muge Ersoy, Richard Everett, Kevin Cen, Rebeca Santamaria-Fernandez, Amir

Taubenfeld, Ian Mackinnon, Linda Deng, Polina Zablotskaia, Shashank Viswanadha, Shivanker Goel, Damion Yates, Yunxiao Deng, Peter Choy, Mingqing Chen, Abhishek Sinha, Alex Mossin, Yiming Wang, Arthur Szlam, Susan Hao, Paul Kishan Rubenstein, Metin Toksoz-Exley, Miranda Aperghis, Yin Zhong, Junwhan Ahn, Michael Isard, Olivier Lacombe, Florian Luisier, Chrysovalantis Anastasiou, Yogesh Kalley, Utsav Prabhu, Emma Dunleavy, Shaan Bijwadia, Justin Mao-Jones, Kelly Chen, Rama Pasumarthi, Emily Wood, Adil Dostmohamed, Nate Hurley, Jiri Simsa, Alicia Parrish, Mantas Pajarskas, Matt Harvey, Ondrej Skopek, Yony Kochinski, Javier Rey, Verena Rieser, Denny Zhou, Sun Jae Lee, Trilok Acharya, Guowang Li, Joe Jiang, Xiaofan Zhang, Bryant Gipson, Ethan Mahintorabi, Marco Gelmi, Nima Khajehnouri, Angel Yeh, Kayi Lee, Loic Matthey, Leslie Baker, Trang Pham, Han Fu, Alex Pak, Prakhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao, Patrik Zochbauer, Andreea Marzoca, Noam Velan, Junhao Zeng, Gilles Baechler, Danny Driess, Divya Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir Levine, Jon Schneider, Erika Gemzer, Samuel Petit, Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles, Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vicente Franco, Adrian Collister, Hao Zhang, Renshen Wang, Ruizhe Zhao, Leandro Kieliger, Kurt Shuster, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes, Abhishek Bapna, Jasper Snoek, Weel Yang, Puranjay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li, SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Karmarkar, Tammo Spalink, Tao Huang, Lucas Gonzalez, Duc-Hieu Tran, Averi Nowak, John Palowitch, Martin Chadwick, Ellie Talius, Harsh Mehta, Thibault Sellam, Philipp Fränken, Massimo Nicosia, Kyle He, Aditya Kini, David Amos, Sugato Basu, Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sampath, Ada Ma, Candice Schumann, Aditya Siddhant, Rohin Shah, John Youssef, Rishabh Agarwal, Natalie Dabney, Alessio Tonioni, Moran Ambar, Jing Li, Isabelle Guyon, Benny Li, David Soergel, Boya Fang, Georgi Karadzhov, Cristian Udrescu, Trieu Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal Gupta, Mara Finkelstein, Denis Petek, Lihao Liang, Greg Billock, Pei Sun, David Wood, Yiwen Song, Xiaobin Yu, Tatiana Matejovicova, Regev Cohen, Kalyan Andra, David D'Ambrosio, Zhiwei Deng, Vincent Nallatamby, Ebrahim Songhori, Rumen Dangovski, Andrew Lampinen, Pankil Botadra, Adam Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kuncoro, Toshihiro Yoshino, Ankit Bhagatwala, Marcáurelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai Ye, Obaid Sarvana, John Nham, Chenkai Kuang, Isabel Gao, Jinoo Baek, Shubham Mittal, Ayzaan Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie Muir, Pascal Lamblin, Wolfgang Macherey, Ethan Dyer, Logan Kilpatrick, Víctor Campos, Mukul Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei Severyn, Kleopatra Chatziprimou, Oleksandr Ferludin, Mason Dimarco, Aditya Kusupati, Joe Heyward, Dan Bahir, Kevin Villela, Katie Millican, Dror Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas Roth, Zichuan Wei, Siddharth Gopal, Deepanway Ghoshal, Edward Lee, Sharon Lin, Jennie Lees, Dayeong Lee, Anahita Hosseini, Connie Fan, Seth Neel, Marcus Wu, Yasemin Altun, Honglong Cai, Enrique Piqueras,

Josh Woodward, Alessandro Bissacco, Salem Haykal, Mahyar Bordbar, Prasha Sundaram, Sarah Hodkinson, Daniel Toyama, George Polovets, Austin Myers, Anu Sinha, Tomer Levinboim, Kashyap Krishnakumar, Rachita Chhaparia, Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu, Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali Eslami, Mariano Schain, Michiel Blokzijl, David Welling, Dave Orr, Levent Bolelli, Nicolas Perez-Nieves, Mikhail Sirotenko, Aman Prasad, Arjun Kar, Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz, Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka, Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana Berman, Maggie Tran, Steven Baker, Ewa Andrejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi Mirzazadeh, Thais Kagohara, Shimu Wu, Christian Schallhart, Bernett Orlando, Chen Wang, Alban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer, Nolan Ramsden, Shuo yiin Chang, Sidharth Mudgal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq Ahmad, Ambrose Slone, Amy Hua, Natan Potikha, Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran, Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom Funkhouser, Yuanzhen Li, Varun Yerram, Ramya Ganeshan, Daniel Formoso, Karol Langner, Tian Shi, Huijian Li, Yumeya Yamamori, Amayika Panda, Alaa Saade, Angelo Scorza Scarpati, Chris Breaux, CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav Tulsyan, Sanghyun Woo, Evgenii Eltyshev, Will Grathwohl, Chanel Parks, Seth Benjamin, Rina Panigrahy, Shenil Dodhia, Daniel De Freitas, Chris Sauer, Will Song, Ferran Alet, Jackson Tolins, Cosmin Paduraru, Xingyi Zhou, Brian Albert, Zizhao Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir Globerson, Owen Xiao, James Manyika, Tom Hennigan, Rong Rong, Josip Matak, Anton Bakalov, Ankur Sharma, Danila Sinopalnikov, Andrew Pierson, Stephen Roller, Geoff Brown, Mingcen Gao, Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vassigh, Iain Barr, Zhicheng Wang, Anna Korsun, Rajesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan, Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz, Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian Yu, Abhishek Nayyar, Bryce Petrini, Will Truong, Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti, Wei Wang, Jingcao Hu, Vahab Mirrokni, Ananth Balashankar, Avia Aharon, Aahil Mehta, Ahmet Iscen, Joseph Kready, Lucas Manning, Anhad Mohananey, Yuankai Chen, Anshuman Tripathi, Allen Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml, Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila, Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tatineni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela Rosca, Adrià Puigdomènech, Zohar Yahav, Sanjay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts, Aditya Srikanth Veerubhotla, Dan Zheng, Zihang Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hombaiah, Romal Thoppilan, Sashank Reddi, Lukasz Lew, Li Li, Kellie Webster, KP Sawhney, Lampros Lamprou, Siamak Shakeri, Mayank Lunayach, Jianmin Chen, Sumit Bagri, Alex Salcianu, Ying Chen, Yani Donchev, Charlotte Magister, Signe Nørly, Vitor Rodrigues, Tomas Izo, Hila Noga, Joe Zou, Thomas

Köppe, Wenxuan Zhou, Kenton Lee, Xiangzhu Long, Danielle Eisenbud, Anthony Chen, Connor Schenck, Chi Ming To, Peilin Zhong, Emanuel Taropa, Minh Truong, Omer Levy, Danilo Martins, Zhiyuan Zhang, Christopher Semturs, Kelvin Zhang, Alex Yakubovich, Pol Moreno, Lara McConnaughey, Di Lu, Sam Redmond, Lotte Weerts, Yonatan Bitton, Tiziana Refice, Nicolas Lacasse, Arthur Conmy, Corentin Tallec, Julian Odell, Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang, Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov, Rich Galt, Matthew Denton, Ben Murdoch, Caitlin Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim McConnell, Luis C. Cobo, James Qin, Thi Avrahami, Daniel Balle, Yu Watanabe, Annie Louis, Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie Rives, Charles Yoon, Andrei Rusu, James Cobon-Kerr, Chris Hahn, Jiaming Luo, Yuvein, Zhu, Niharika Ahuja, Rodrigo Benenson, Raphaël Lopez Kaufman, Honglin Yu, Lloyd Hightower, Junlin Zhang, Darren Ni, Lisa Anne Hendricks, Gabby Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaiche, Sheena Panthaplackel, Klemen Kloboves, Sarthak Jauhari, Canfer Akbulut, Todor Davchev, Evgeny Gladchenko, David Madras, Aleksandr Chuklin, Tyrone Hill, Quan Yuan, Mukundan Madhavan, Luke Leonhard, Dylan Scandinaro, Qihang Chen, Ning Niu, Arthur Douillard, Bogdan Damoc, Yasumasa Onoe, Fabian Pedregosa, Fred Bertsch, Chas Leichner, Joseph Pagadora, Jonathan Malmaud, Sameera Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen, Miaosen Wang, Roopal Garg, Jing Chen, Utku Evci, Jonathan Lee, Leon Liu, Koji Kojima, Masa Yamaguchi, Arunkumar Rajendran, AJ Piergiovanni, Vinodh Kumar Rajendran, Marco Fornoni, Gabriel Ibagon, Harry Ragan, Sadh MNM Khan, John Blitzer, Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane Park, Arunachalam Narayanaswamy, Chengda Wu, Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi Kumar, Laura Graesser, Yichi Zhang, Richard Killam, Vincent Zhuang, Mai Giménez, Wael Al Jishi, Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sieniek, Roshan Sharma, Tom Paine, Anelia Angelova, Sravanti Addepalli, Carolina Parada, Kingshuk Majumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Artiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah York, Félix de Chaumont Quitry, Jiazhong Nie, Dee Cattle, Alok Gunjan, Bilal Piot, Waleed Khawaja, Seojin Bang, Simon Wang, Siavash Khodadadeh, Raghavender R, Praynaa Rawlani, Richard Powell, Kevin Lee, Johannes Griesser, GS Oh, Cesar Magalhaes, Yujia Li, Simon Tokumine, Hadas Natalie Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan Belov, Rahul Arya, Phoenix Meadowlark, Shlomi Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki, Praseem Banzal, Gan Song, Pranavaraj Ponnuramu, Fred Zhang, George Scrivener, Salah Zaiem, Alif Raditya Rochman, Kehang Han, Badih Ghazi, Kate Lee, Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta, Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang, Daniil Mirylenka, Emma Cooney, Edward Chou, Mingyang Ling, Cindy Fan, Ben Coleman, Guilherme

Tubone, Ravin Kumar, Jason Baldridge, Felix Hernandez-Campos, Angeliki Lazaridou, James Besley, Itay Yona, Neslihan Bulut, Quentin Wellens, AJ Pierigiovanni, Jasmine George, Richard Green, Pu Han, Connie Tao, Geoff Clark, Chong You, Abbas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin Chaugule, Angad Chandorkar, Altaf Rahman, Will Thompson, Penporn Koanantakool, Mike Bernico, Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej Kula, Yizhong Liang, Dahun Kim, Yangsibo Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley Helmholz. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

[31] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023.

[32] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[33] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. 2021.

[34] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2024.

[35] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE, 2021.

[36] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[37] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Remix-dit: Mixing diffusion transformers for multi-expert denoising. 2024.

[38] Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video LLM benchmarks: Knowledge, spatial perception, or true temporal understanding? *CoRR*, abs/2505.14321, 2025.

[39] Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. MMDialog: A large-scale multi-turn dialogue dataset

towards multi-modal open-domain conversation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada, July 2023. Association for Computational Linguistics.

[40] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *CoRR*, abs/2503.21776, 2025.

[41] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[42] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.

[43] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE, 2025.

[44] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, volume 13675 of *Lecture Notes in Computer Science*, pages 89–106. Springer, 2022.

[45] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter V2: parameter-efficient visual instruction model. *CoRR*, abs/2304.15010, 2023.

[46] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[47] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[48] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[49] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18973–18990. IEEE, 2022.

[50] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *CoRR*, abs/2305.15717, 2023.

[51] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, Hao Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L.

Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, Tao Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638, 2025.

[52] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, Shixiong Zhao, Shuai Peng, Shuangye Li, Sihang Yuan, Sijin Wu, Tianheng Cheng, Weiwei Liu, Wenqian Wang, Xianhan Zeng, Xiao Liu, Xiaobo Qin, Xiaohan Ding, Xiaojun Xiao, Xiaoying Zhang, Xuanwei Zhang, Xuehan Xiong, Yanghua Peng, Yangrui Chen, Yanwei Li, Yanxu Hu, Yi Lin, Yiyuan Hu, Yiyuan Zhang, Youbin Wu, Yu Li, Yudong Liu, Yue Ling, Yujia Qin, Zanbo Wang, Zhiwu He, Aoxue Zhang, Bairen Yi, Bencheng Liao, Can Huang, Can Zhang, Chaorui Deng, Chaoyi Deng, Cheng Lin, Cheng Yuan, Chenggang Li, Chenhui Gou, Chenwei Lou, Chengzhi Wei, Chundian Liu, Chunyuan Li, Deyao Zhu, Donghong Zhong, Feng Li, Feng Zhang, Gang Wu, Guodong Li, Guohong Xiao, Haibin Lin, Haihua Yang, Haoming Wang, Heng Ji, Hongxiang Hao, Hui Shen, Huixia Li, Jiahao Li, Jialong Wu, Jianhua Zhu, Jianpeng Jiao, Jiashi Feng, Jiaze Chen, Jianhui Duan, Jihao Liu, Jin Zeng, Jingqun Tang, Jingyu Sun, Joya Chen, Jun Long, Junda Feng, Junfeng Zhan, Junjie Fang, Junting Lu, Kai Hua, Kai Liu, Kai Shen, Kaiyuan Zhang, and Ke Shen. Seed1.5-vl technical report. *CoRR*, abs/2505.07062, 2025.

[53] Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P. Bigham. Improving zero and few-shot generalization in dialogue through instruction tuning, 2022.

[54] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. *AI Open*, 2022.

[55] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng,

and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *ICCV*, 2023.

[56] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[57] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

[58] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. 2022.

[59] Zhijian Hou, Lei Ji, Difei Gao, Wanjun Zhong, Kun Yan, Chao Li, Wing-Kwong Chan, Chong-Wah Ngo, Nan Duan, and Mike Zheng Shou. Groundnlq @ ego4d natural language queries challenge 2023. *CoRR*, abs/2306.15255, 2023.

[60] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

[61] Hexiang Hu, Kelvin C. K. Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William W. Cohen, Ming-Wei Chang, and Xuhui Jia. Instruct-imagen: Image generation with multi-modal instruction. *CoRR*, abs/2401.01952, 2024.

[62] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization, 2025.

[63] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *CoRR*, abs/2501.13826, 2025.

[64] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. BLIVA: A simple multimodal LLM for better handling of text-rich visual questions. *CoRR*, abs/2308.09936, 2023.

[65] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1233–1239. The Association for Computational Linguistics, 2016.

[66] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[67] Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks, 2023.

[68] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, volume 13693 of *Lecture Notes in Computer Science*, pages 709–727. Springer, 2022.

[69] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: interleaved multi-image instruction tuning. *CoRR*, abs/2405.01483, 2024.

[70] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *CoRR*, abs/2207.07039, 2022.

[71] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. *CoRR*, abs/2309.04669, 2023.

[72] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973, 2017.

[73] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021.

[74] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020.

[75] Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. Generating images with multimodal language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[76] Elisa Kreiss, Fei Fang, Noah D. Goodman, and Christopher Potts. Concadia: Towards image-based text generation with a purpose, 2022.

[77] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

[78] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[79] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[80] Yogesh Kulkarni and Pooyan Fazli. Egovita: Learning to plan and verify for egocentric video reasoning. *CoRR*, abs/2511.18242, 2025.

[81] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023.

[82] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace, editors, *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM, 2023.

[83] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[84] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022.

[85] Yunsung Lee, JinYoung Kim, Hyojun Go, Myeongho Jeong, Shinhyeok Oh, and Seungtaek Choi. Multi-architecture multi-expert diffusion models. In *AAAI*, 2024.

[86] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425, 2023.

[87] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726, 2023.

[88] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025.

[89] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024.

[90] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895, 2024.

[91] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, and Jifeng Dai. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. *CoRR*, abs/2211.09808, 2022.

[92] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. 2023.

[93] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. 202:19730–19742, 2023.

[94] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *CoRR*, abs/2306.04387, 2023.

[95] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022.

[96] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *CVPR*, 2023.

[97] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.

[98] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors,

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 4582–4597. Association for Computational Linguistics, 2021.

[99] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *CoRR*, abs/2504.06958, 2025.

[100] Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *CoRR*, abs/2308.10253, 2023.

[101] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *CoRR*, abs/2403.18814, 2024.

[102] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. 2023.

[103] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[104] Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[105] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[106] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947, 2024.

[107] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[108] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in

context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[109] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

[110] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022.

[111] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models. *CoRR*, abs/2310.14566, 2023.

[112] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565, 2023.

[113] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2023.

[114] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *CoRR*, abs/2205.05638, 2022.

[115] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *CoRR*, abs/2205.05638, 2022.

[116] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023.

[117] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[118] Minqian Liu and Lifu Huang. Teamwork is not always good: An empirical study of classifier drift in class-incremental information extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2241–2257, Toronto, Canada, July 2023. Association for Computational Linguistics.

[119] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[120] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[121] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

[122] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *ECCV*, 2024.

[123] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[124] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks, 2022.

[125] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023.

[126] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *CoRR*, abs/2306.09093, 2023.

[127] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *CVPR*, 2025.

[128] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[129] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[130] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

[131] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[132] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman,

Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024.

[133] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-

05-26.

[134] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[135] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. *arXiv preprint arXiv:2504.06256*, 2025.

[136] Alkesh Patel, Vibhav Chitalia, and Yinfei Yang. Advancing egocentric video question answering with multimodal large language models. *CoRR*, abs/2504.04550, 2025.

[137] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[138] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, and Yu Qiao. Egovideo: Exploring egocentric foundation model and downstream adaptation. *CoRR*, abs/2406.18070, 2024.

[139] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021.

[140] Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24129–24138. Computer Vision Foundation / IEEE, 2025.

[141] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023.

[142] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 5262–5274. IEEE, 2023.

[143] Lu Qiu, Yuying Ge, Yi Chen, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *CoRR*, abs/2412.04447, 2024.

[144] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[145] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[146] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[147] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[148] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024.

[149] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[150] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[151] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[152] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[153] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[154] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.

[155] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

[156] Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. Multimodal instruction tuning with conditional mixture of lora. *CoRR*, abs/2402.15896, 2024.

[157] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. *CoRR*, abs/2409.19256, 2024.

[158] Minglei Shi, Ziyang Yuan, Haotian Yang, Xintao Wang, Mingwu Zheng, Xin Tao, Wenliang Zhao, Wenzhao Zheng, Jie Zhou, Jiwen Lu, et al. Diffmoe: Dynamic token selection for scalable diffusion transformers. *arXiv preprint arXiv:2503.14487*, 2025.

[159] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

[160] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[161] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE, 2019.

[162] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[163] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. 2019.

[164] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

[165] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017.

[166] Haotian Sun, Tao Lei, Bowen Zhang, Yanghao Li, Haoshuo Huang, Ruoming Pang, Bo Dai, and Nan Du. Ec-dit: Scaling diffusion transformers with adaptive expert-choice routing. In *ICLR*, 2025.

[167] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14398–14409, June 2024.

[168] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[169] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023.

[170] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023.

[171] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pre-training in multimodality. *CoRR*, abs/2307.05222, 2023.

[172] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[173] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context, interleaved, and interactive any-to-any generation. *CoRR*, abs/2311.18775, 2023.

[174] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

[175] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[176] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya,

Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer,

Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux,

Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal,

Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki

Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[177] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *CoRR*, abs/2401.10208, 2024.

[178] Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *CoRR*, abs/2506.13654, 2025.

[179] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

[180] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. 2017.

[181] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[182] Ashwin Vinod, Shrey Pandit, Aditya Vavre, and Linshen Liu. Egovlm: Policy optimization for egocentric video understanding. *CoRR*, abs/2506.03097, 2025.

[183] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *CoRR*, abs/1606.04080, 2016.

[184] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting GPT-4V for better visual instruction tuning. *CoRR*, abs/2311.07574, 2023.

[185] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[186] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.

[187] Sijia Wang, Mo Yu, and Lifu Huang. The art of prompting: Event detection based on type specific prompts. *arXiv preprint arXiv:2204.07241*, 2022.

[188] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *CoRR*, abs/2208.10442, 2022.

[189] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[190] Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5744–5760. Association for Computational Linguistics, 2022.

[191] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding, 2025.

[192] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416. Springer, 2024.

[193] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022.

[194] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics.

[195] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021.

[196] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. 2022.

[197] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

[198] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. *CoRR*, abs/2309.05519, 2023.

[199] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model

integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[200] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[201] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[202] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *CoRR*, abs/2111.02080, 2021.

[203] Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024.

[204] Zhiyang Xu, Minqian Liu, Ying Shen, Joy Rimchala, Jiaxin Zhang, Qifan Wang, Yu Cheng, and Lifu Huang. Modality-specialized synergizers for interleaved vision-language generalists. In *ICLR*, 2025.

[205] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11445–11465. Association for Computational Linguistics, 2023.

[206] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11445–11465. Association for Computational Linguistics, 2023.

[207] Zhongwen Xu and Zihan Ding. Single-stream policy optimization, 2025.

[208] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024.

[209] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 10632–10643. Computer Vision Foundation / IEEE, 2025.

[210] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. Egolife: Towards egocentric life assistant. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 28885–28900. Computer Vision Foundation / IEEE, 2025.

[211] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. Egolife: Towards egocentric life assistant. *CoRR*, abs/2503.03803, 2025.

[212] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. *CoRR*, abs/2511.04670, 2025.

[213] Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. Visual goal-step inference using wikiHow. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2167–2179, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[214] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *arXiv preprint arXiv:2205.12487*, 2022.

[215] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR, 2023.

[216] Hanrong Ye, Haotian Zhang, Erik A. Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. Mm-ego: Towards building egocentric multimodal llms. *CoRR*, abs/2410.07177, 2024.

[217] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023.

[218] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *CoRR*, abs/2306.06687, 2023.

[219] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *ICLR*, 2022.

[220] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[221] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *CoRR*, abs/2309.02591, 2023.

[222] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[223] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. 2024.

[224] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*. IEEE, 2024.

[225] Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermis, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *CoRR*, abs/2309.05444, 2023.

[226] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022.

[227] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[228] Haoyu Zhang, Qiaohui Chu, Meng Liu, Yunxiao Wang, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, Yaowei Wang, and Liqiang Nie. Exo2ego: Exocentric knowledge guided MLLM for egocentric video understanding. *CoRR*, abs/2503.09143, 2025.

[229] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[230] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[231] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *CoRR*, abs/2306.17107, 2023.

[232] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024.

[233] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025, 2025.

[234] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models, 2024.

[235] Bo Zhao, Boya Wu, and Tiejun Huang. SVIT: scaling up visual instruction tuning. *CoRR*, abs/2307.04087, 2023.

[236] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: measuring expert-level multi-discipline video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 8475–8489. Computer Vision Foundation / IEEE, 2025.

[237] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *CoRR*, abs/2507.18071, 2025.

[238] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *CoRR*, abs/2310.02239, 2023.

[239] Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution meets lora: Parameter efficient finetuning for segment anything model. *CoRR*, abs/2401.17868, 2024.

[240] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

[241] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598. AAAI Press, 2018.

[242] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. VLM4D: towards spatiotemporal awareness in vision language models. *CoRR*, abs/2508.02095, 2025.

[243] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models, 2023.

[244] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023.

[245] Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. VL-GPT: A generative pre-trained transformer for vision and language understanding and generation. *CoRR*, abs/2312.09251, 2023.

[246] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[247] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.

[248] Shaobin Zhuang, Yiwei Guo, Yanbo Ding, Kunchang Li, Xinyuan Chen, Yaohui Wang, Fangyikang Wang, Ying Zhang, Chen Li, and Yali Wang. Timestep master: Asymmetrical mixture of timestep lora experts for versatile and efficient diffusion models in vision. *arXiv preprint arXiv:2503.07416*, 2025.