Toward Transformer-based Large Energy Models for Smart Energy Management

Yueyan Gu

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

Master of Science
in
Computer Science and Application

Xuan Wang, Chair
Farrokh Jazizadeh, Co-chair
Dawei Zhou

Nov 1, 2024
Blacksburg, Virginia

Keywords: Transformer Models, Energy Forecasting, Generalizability, Scalability, Large Model, Multivariate Time Series

Toward Transformer-based Large Energy Models for Smart Energy Management

Yueyan Gu

(ABSTRACT)

Buildings contribute significantly to global energy demand and emissions, highlighting the need for precise energy forecasting for effective management. Existing research tends to focus on specific target problems, such as individual buildings or small groups of buildings, leading to current challenges in data-driven energy forecasting, including dependence on data quality and quantity, limited generalizability, and computational inefficiency. To address these challenges, Generalized Energy Models (GEMs) for energy forecasting can potentially be developed using large-scale datasets. Transformer architectures, known for their scalability, ability to capture long-term dependencies and efficiency in parallel processing of large datasets, are considered good candidates for GEMs. In this study, we tested the hypothesis that GEMs can be efficiently developed to outperform in-situ models trained on individual buildings. To this end, we investigated and compared three candidate multivariate Transformer architectures, utilizing both zero-shot and fine-tuning strategies, with data from 1,014 buildings. The results, evaluated across three prediction horizons (24, 72, and 168 hours), confirm that GEMs significantly outperform Transformer-based in-situ (i.e., building-specific) models. Fine-tuned GEMs showed performance improvements of up to 28% and reduced training time by 55%. Besides Transformer-based in-situ models, GEMs outperformed several state-of-the-art non-Transformer deep learning baseline models in effectiveness and efficiency. We further explored the answer to a number of questions including the required data size for effective fine-tuning, as well as the impact of input sub-sequence length and pre-training dataset size on GEM performance. The findings show a significant performance boost by using larger pre-training datasets, highlighting the potential for larger GEMs using web-scale global data to move toward Large Energy Models (LEM).

Toward Transformer-based Large Energy Models for Smart Energy Management

Yueyan Gu

(GENERAL AUDIENCE ABSTRACT)

Buildings account for a large share of global energy use and emissions, which makes predicting their energy needs critical for better management. However, most research focuses on creating energy models for specific buildings or small groups, which limits their usefulness for larger-scale applications. Additionally, these models often face challenges such as relying on high-quality data, limited adaptability to different buildings, and inefficiencies when dealing with large amounts of data. This study aims to address these issues by developing Generalized Energy Models (GEMs), which use data from a large number of buildings to create more versatile and efficient energy forecasting tools. To achieve this, we used Transformer models, a type of machine learning approach known for handling large datasets efficiently and recognizing long-term patterns. We tested whether GEMs could provide better predictions than traditional models designed for individual buildings. Our analysis included data from over 1,000 buildings and used two strategies: zero-shot (using the model without further adjustments) and fine-tuning (adapting the model to specific data). The results showed that GEMs were more accurate than traditional models, improving prediction accuracy by up to 28% and reducing the time needed for training by over 50%. Additionally, GEMs outperformed other advanced methods of energy forecasting. We also examined how different factors, such as the amount of data and the length of the data sequences, influenced the model's performance. The findings suggest that using even larger datasets could lead to further improvements, opening the possibility of creating Large Energy Models (LEMs) that can make predictions on a global scale.

Dedication

To my family for their unconditional love.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Xuan Wang, and my co-advisor, Dr. Farrokh Jazizadeh, for their invaluable support, guidance, and encouragement throughout my research. I am also sincerely thankful to my committee member, Dr. Dawei Zhou, for his constructive feedback and insightful recommendations.

I am profoundly grateful to my grandmother, parents, and husband for their endless love and support, which have been a constant source of strength and motivation. Finally, I wish to extend my appreciation to my friends in Blacksburg, whose friendship and companionship have enriched this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Global energy usage is expected to grow by over 40% by 2050 [29, 46]. Buildings, as substantial contributors, account for over one-third of global energy demand and emissions [10, 16]. Effective building energy management is essential for energy efficiency, occupant comfort, and developing strategies to mitigate risks associated with climate change. Given the urgency of global climate challenges, it is imperative to intensify research efforts to develop advanced energy management strategies that address a broader spectrum of complex, large-scale problems. Effective enhancements, whether in the design or operational phases, could significantly benefit from precise energy use forecasting for informed interventions [4]. Energy use forecasting could enable (1) smart and economical energy demand management, (2) sustainable and efficient energy supply, and (3) predictive and proactive facility maintenance and operations. Smart metering infrastructure has facilitated energy forecasting applications by providing a deeper understanding of the energy use patterns and the potential drivers [35]. Machine learning (ML) models trained on smart meter data have been widely utilized for energy use forecasting, as well as for fault detection and diagnostics [7, 27, 29, 44, 59]. The ideal scenario in the field of building energy management is to establish pre-trained generalized models that are scalable and function across different buildings without the need for extensive in-situ adaptation. Recent advances in large-scale artificial intelligence model developments present a promising prospect for moving toward developing Large Energy Models. The feasibility of realizing this vision has been explored in this study.

Significant research efforts have been made to develop energy forecasting models for buildings. However, these efforts have primarily focused on individual buildings while still encountering challenges in achieving scalable forecasting models. One major limitation is the low generalizability and contextual sensitivity of these models. Existing data-driven methods have shown promising efficacy when leveraging abundant historical data in individual buildings [6] for energy forecasting, maintenance, and control (e.g., [13, 15, 19, 60]). However, limitations in the quantity of individual building data and the variability in consumption patterns constrain the generalizability of trained models across different buildings [6, 16]. Generalizability refers to an ML model's capability to adapt to new data and perform adequately [28]. This challenge stems partly from the intricate nature of energy consumption patterns, where electricity demand is heavily influenced not only by geographical and weather conditions but also by occupant behavior [16, 25]. Even identical buildings may exhibit significant variability in energy use due to their operational characteristics and occupant behavior [4, 38]. Another contributing factor could be the inherent limitations of

adopted algorithms. For instance, Miller et al. [30] leveraged data from 482 buildings and found that decision-tree-based methods showed limited generalizability. Although, in recent years, efforts have been made to increase the availability of data across different building types, the low efficiency of computing mechanisms remains a critical obstacle in tackling the increased scale of energy consumption data. Lack of efficiency may hinder advancements in large-scale modeling and model updating, both of which are critical for achieving generalized models. Thorough reviews of existing research in machine learning and deep learning for energy forecasting [3, 25] show the importance of adopting more efficient computational approaches to ensure cost-effective and practical real-world implementations.

Tackling these challenges and limitations requires leveraging large datasets that encompass buildings of varying sizes, functionalities, operational demands, and occupant dynamics. The exponential growth in smart meters and Internet of Things (IoT) infrastructures and the global efforts to make datasets available have led to the accumulation of extensive large-scale data. On the other hand, we should rely on modeling paradigms capable of handling large-scale datasets effectively and efficiently. In recent years, the proliferation of available web-scale data, computational resources, and the emergence of Transformer architectures have marked the onset of an era characterized by the dominance of effective large-scale models across a spectrum of domains, including natural language processing [5], computer vision [9]. Despite these advances, existing machine learning-based building energy studies predominantly concentrate on individual facilities or small groups of facilities, with limited attention given to "one-size-fits-all" configurations or the generalizability of models derived from such endeavors [41]. Additionally, the introduction of Transformers has transformed *efficient* sequential data processing in tasks such as natural language understanding and generation, image recognition, and time series analysis [20]. Transformers, with their deep architectures and self-attention mechanisms, excel in handling large datasets, capturing intricate relationships, and focusing on relevant features, which enable robust learning of complex patterns and generalization to new instances. Moreover, the parallel operation allows them to simultaneously capture relationships between all elements within a sequence, which is crucial for managing large-scale models [47]. The inherent parallelism and attention mechanisms make Transformers particularly suitable for processing substantial amounts of training data and capturing long-term dependencies.

This study centers around providing insight into the realization of a generalized model for building energy forecasting - i.e., a Generalized Energy Model (GEM). Given the current lack of web-scale data, we investigated the feasibility, effectiveness, and efficiency of GEMs that use Transformer-based architectures and utilized (relatively) large and diverse historical energy consumption data. Our central hypothesis is that GEMs can be developed efficiently to outperform in-situ models trained on individual target buildings. This study aims to elucidate strategies for advancing the generalizability of data-driven methods in this domain, thereby fostering more robust and scalable solutions to address the multifaceted challenges of building energy forecasting. Additionally, this work also paves the way for moving toward Large Energy Models (LEMs) utilizing web-scale data in the future. To this

end, this study contributes by pursuing the following objectives. (1) Evaluating the effectiveness and computational efficiency of Transformer architectures with different attention mechanisms for developing pre-trained generalized energy models (GEMs) using multivariate large-size historical energy consumption data. (2) Analyzing the zero-shot performance of pre-trained GEMs and comparing it with in-situ Transformer-based models and other SOTA deep learning methods. (3) Assessing the potential of fine-tuned GEMs in augmenting individual building energy forecasting performance through a comparative study with in-situ Transformer-based models and SOTA baseline models. (4) Conducting further experiments to identify the effect of several attributes including the prediction horizon, fine-tuning data size, input sub-sequence length, and pre-training data size to inform future GEM and LEM developments. In achieving these objectives, we have also answered several research questions as detailed in Section 3.4.

The remainder of the thesis [1] is organized as follows: Chapter 2 reviews the existing efforts made on building energy forecasting, with a special emphasis on Transformer-based implementations. Chapter 3 details the methodology, including data processing, model architecture configuration, training processes and schemes, hypothesis and research questions, and the case study involving approximately 1,000 buildings to explore GEMs' potential. Chapter 4 addresses the research questions by examining the zero-shot and fine-tuning performance of GEMs compared to in-situ baselines, and investigating the influence of key factors, including target building inclusion in pre-training, required data size for fine-tuning, input sub-sequence length, and pre-training dataset size. It also discusses the limitations and future directions. Finally, the last chapter summarizes the study's findings.

[1] This work has been published in [12]

Chapter 2

Review of Literature

2.1 Energy forecasting scope

Building energy forecasting research spans various scales, granularities, and prediction horizons for different tasks, addressing different end-uses, including cooling, heating, hot water, lighting, and plug loads, as well as overall electricity use [25]. Although models have been proposed for office, educational, and industrial facilities, residential and commercial buildings have garnered the most research interest. These models operate at different spatial scales, such as system, room, unit, floor, sub-building, building, and community levels with a temporal granularity ranging from 10/30-minute intervals to hourly or daily. The choice of scale and granularity depends on the specific problem, data availability, and computing methods. For instance, a study shows that the optimal monitoring for multi-family residential buildings is achieved at the floor level with hourly intervals using support vector regression [14].

An important aspect of developing forecasting models is the prediction horizon – i.e., short, medium, and long-term. The choice depends on operational and optimization objectives of energy systems in buildings and the power grid [11, 29]. Short-term forecasting, which spans up to one week ahead and primarily focuses on day-ahead forecasting, has gained significant research attention due to its pivotal role in daily power system operations. This includes activities such as energy management (optimization) within individual buildings or households, energy trading, and unit dispatching [8, 11]. Medium-term forecasting, ranging from one week to one year, is important for fuel supply planning and maintenance [11, 18]. Long-term forecasting, which anticipates load profiles extending beyond one year, is commonly used to plan energy management enhancements and to implement new appropriate production and storage systems [2].

Energy forecasting encompasses both regression and classification tasks. Regression tasks aim to predict the future trend of energy consumption, while classification tasks may involve identifying building types, locations, and uses which can be informative for energy management [31, 54]. This study focuses on short-term forecasting of overall energy consumption for non-residential buildings as a short-term prediction by using hourly data collected from smart meters.

2.2 Building energy consumption forecasting: A review of reviews

Building energy consumption forecasting has long been a prominent research area due to its significant implications. A wealth of original research and review studies has been conducted in this field. Recent competition summaries and review studies shed light on key research directions. Notable competitions, such as "ASHRAE Great Energy Predictor III" [32] and "Energy Detective" [55], emphasize the need to improve building prediction models, particularly when limited historical data is available. These competitions have also highlighted the need for generalizability of existing models, suggesting the need for further explorations [32]. Recent review studies also support these observations. For instance, Chen et al. [6] compared data-driven methods for building energy prediction problems, including linear regression, support vector machine, random forest, artificial neural networks (ANN), and recurrent neural networks (RNN), against physical modeling methods. hey suggested that data-driven methods offer the benefits of independence from domain-specific expertise and lower modeling and computing costs. However, these methods can suffer from issues such as data insufficiency and low generalizability.

Achieving effective and robust data-driven methods presents several challenges, as highlighted by various research studies. Lu et al.[25] conducted a comprehensive survey of the application of ANNs in building energy prediction based on 324 related publications. They identified key challenges, including customizing ANN architectures for specific building energy prediction scenarios, implementing efficient computing techniques for practical use, and exploring the transferability of different building contexts. Ardabili et al. [3] evaluated the performance of machine learning and deep learning techniques in terms of accuracy, reliability, and sustainability. Their findings suggest that deep learning-based approaches, particularly hybrid and ensemble methods, exhibit the highest robustness in energy consumption forecasting. Khalil et al. [16] categorized general data-driven methods—including machine learning, deep learning, and statistical analysis—for building energy forecasting based on factors such as building type and location, data components, temporal granularity, data pre-processing methods, features selection and extraction techniques, and models used. The authors highlighted the importance of addressing the reliance of data-driven methods on data quantity and quality by leveraging transfer learning. These studies collectively reveal three major challenges: (1) reliance on in-situ data quantity and quality, (2) model robustness, generalizability, transferability, and explainability, and (3) computing efficiency.

2.3 Big data and large models

In many other domains, large generalized models built upon big data have drastically changed the field and brought abundant new opportunities, offering significant improvements in ac-

curacy, contextual understanding, and adaptability across diverse tasks [34]. For instance, in natural language processing (NLP), the main concept of Large Language Models (LLMs) involves acquiring a generic, latent representation of language through a single generic task and subsequently applying it across various NLP tasks [34]. Abundant text data available for extensive self-supervised training is a necessity in this context. Well-known models, such as Llama [45] and GPT [1], have utilized large volumes of publicly available online data (i.e., web-scale data), typically incorporating trillions of tokens (i.e., units of text). Such large language models are equipped with billions or even tens of billions of trainable parameters.

Although there are different problems with different scales, similar modeling schemes can be envisioned for energy management, specifically forecasting, which is the focus of our study. In this vision, a single GEM is trained on large datasets across different environments and types of buildings and is applied to different prediction scenarios with zero or minimal model updating. This calls for changes in conventional practices observed in the literature and the overcoming of the challenges. Publicly available online data remains limited despite recent efforts to create large benchmark datasets [4, 33]. Researchers typically work with limited or homogeneous datasets, particularly in regression-oriented energy consumption forecasting. If the modeling prospects are promising, achieving large models in this field calls for mechanisms to encourage the collection and sharing of data across a large set of facilities. Moreover, traditional approaches often use local models trained separately for each building. Among the largest studies we found, Miller et al. [30] examined 482 non-residential buildings with models trained and tested on individual buildings. To bridge the gaps toward scalable large models, we have investigated training large models across a relatively large-scale building energy dataset for improved efficacy.

2.4 State-of-the-art Transformer models for time series forecasting

Achieving the envisioned LEM modeling paradigm depends on modeling schemes that are effective and computationally efficient. The capacity of Transformers to capture long-range dependencies has been promising in time series modeling, as evident by advances in forecasting, anomaly detection, and classification [51]. Several variants have been developed based on the original Transformer model [47] with different emphases (e.g., long-term dependency, cross-dimension dependency, computing efficiency, and non-stationarity) to accommodate time series problems. These variants were developed mainly by modifications in Transformer models' positional encoding method, attention module, and model architectures [51]. Below, we have presented a few examples of recent Transformer-based architectures for time series forecasting. In doing so, we have referred to these architectures using the names of the frameworks from original publications.

Long-sequence time series forecasting (LSTF) is challenging since prediction accuracy

tends to decrease with increasing prediction horizon. To address this issue, various architectures such as Pyraformer [21], Autoformer [52], FEDformer [62], and PatchTST [36] have been developed with an emphasis on capturing long-term dependencies. This is also referred to as cross-temporal dependencies. Pyraformer architecture achieves that by leveraging a multi-resolution representation of time series through a pyramidal attention module (PAM), allowing for capturing temporal dependencies across different scales. Autoformer integrates decomposition and Auto-Correlation mechanisms directly into the original Transformer model architecture, enabling it to effectively capture long-range dependencies and complex temporal patterns in time series data. Since its introduction, the concept of inner decomposition has proven useful, leading to its adoption in several other variants, such as FEDformer, which further incorporates frequency enhancement. Moreover, PatchTST segments time series into subseries-level patches and ensures channel independence, with each channel containing a single univariate time series sharing the same embedding and Transformer weights across all series. As Nie et al. showed, it outperforms Pyraformer, Autoformer, and FEDformer in various multivariate long-term forecasting benchmark datasets [36].

In addition to cross-temporal dependency, the cross-dimension dependency is also critical for **multivariate time series forecasting**. For a specific dimension, information from associated time series in other dimensions could enhance prediction accuracy. The two different attention mechanisms for this purpose are illustrated in Figure 2.1. To incorporate the cross-dimension attention, models such as Crossformer [58] and iTransformer [23] have been developed. Crossformer integrates cross-dimension dependency and temporal dependency through a novel Dimension-Segment-Wise (DSW) embedding and a Two-Stage Attention (TSA) layer. Additionally, iTransformer employs attention and feed-forward networks on inverted dimensions to effectively capture multivariate correlations and nonlinear representations.

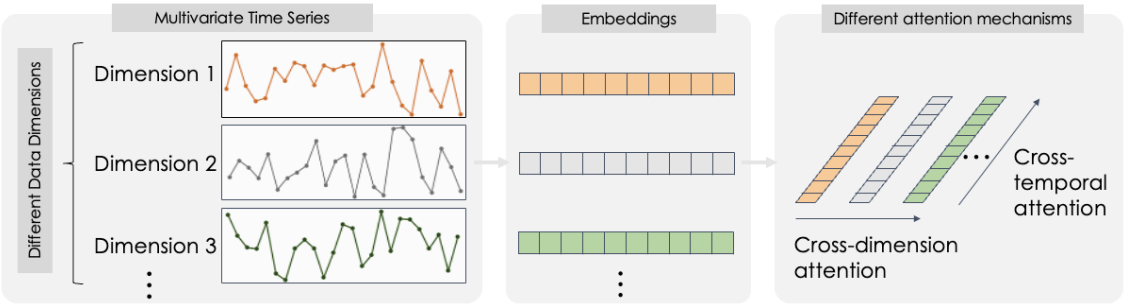


Figure 2.1: Illustration of attention mechanisms on data interpretation in the energy forecasting model

To address the performance degradation of Transformers on **non-stationary real-world data**, the Nonstationary Transformer [22] was put forward by incorporating Series Stationarization and De-stationary Attention modules, which collectively restore series predictability

while preserving intrinsic non-stationary information. Moreover, some Transformer variants lay a special focus on enhancing **efficiency**. For instance, Reformer [17] improves the efficiency through the use of locality-sensitive hashing for attention and reversible residual layers, achieving comparable performance while significantly reducing memory usage and processing time on long sequences.

2.5 Transformer-based energy consumption forecasting

Along with the popularity of Transformers in sequential data processing, they have increasingly been applied to energy forecasting problems in recent years. For instance, Song et al. [43] proposed a hierarchical multi-task learning network employing spatiotemporal attention to address multi-load (cooling, heating, and electrical load) prediction challenges and to visualize the contribution of different features and historical temporal steps. Zheng et al. [60] developed a short-term energy forecasting method by combining interpretable decomposition methods with a Temporal Fusion Transformer (TFT) model and showcased its superior accuracy and interoperability in a case study of one building. Ji et al. [15] introduced a contrastive Transformer network to predict energy consumption with limited data, tested on a large exposition center. Wang et al. [48] proposed a Multiple-Decoder Transformer model for forecasting multiple energy loads, including electricity, as well as heating and cooling loads.

Compared with applications at a larger scale [39], Transformer-based energy forecasting works have predominantly focused on the building level, as summarized in Table 2.1. These studies utilize diverse data types, including energy load type (e.g., cooling load, heating load, overall electricity load), meteorological data (e.g., temperature, dew point temperature, wind speed, cloud coverage), indoor environment information (e.g., CO_2 concentration), building features (e.g., fuel type, building use), and calendar information (e.g., day of the week, week of the month, month of the year, weekend/holiday status). Hourly historical observations are most commonly used, with a predominant emphasis on short-term forecasting over medium and long-term forecasts. The majority of these studies employ attention within an encoder-decoder structure to accomplish sequence-to-sequence prediction tasks and commonly use evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

Table 2.1: Transformer-based methods for building energy consumption forecasting

Paper	Building type and number	Data type	Date resolution	Input length and prediction horizon	Transformer structure	Baseline models	Evaluation metrics
[60]	1 building with 93 rooms	load; meteorological data; calendar information	hourly	7 days → 1 day	encoder-decoder	N-BEATS, N-HiTS, DeepAR, LSTM	MAPE
[48]	entire ASU campus	load; meteorological data; calendar information	hourly	3 days → 1 day	one-encoder and multi-decoders	SingleDeT, LSTM, Random Forest, Bag-BoostNN, SVR, ELM	MAPE
[42]	5 houses in UK	load	5/10/20/30 min	12 samples → 1 sample	encoder-decoder	ARIMA, MLP, SVM, LSTM, Deep Transformer, LSTM-SWT, CNN-LSTM	RMSE, MAE, MAPE
[19]	1 office building on campus	load; meteorological data; calendar information	hourly	24 hours → 1 hour	encoder-decoder	XGBBoost, LSTM, CLM	RMSE, MAE, R^2, uncertainty
[56]	1 residential building on campus	load; meteorological data; calendar information	hourly	4 hours → 4 hours (this study emphasizes information lag)	encoder-LSTM decoder	LSTM, GRU, Bi-LSTM, CNN-Bi-GRU-ATT, CNN-Bi-LSTM-ATT, Transformer	RMSE, MAE, R^2, time
[40]	16 commercial and residential buildings	load; building feature; meteorological data; calendar information	hourly	-	encoder-decoder	LSTM, BP, ARIMA	MAPE
[15]	1 International Expo Center	load; meteorological data	daily	7 days → 1 day	encoder-decoder	LSTM, GRU, SimCLR	RMSE, MAPE
[41]	40 buildings on campus	load; calendar information	15 min	66 hours → 6 hours	encoder	LSTM	RMSE
[37]	1 building on campus	load; meteorological data; indoor environment information	hourly	10 days → 10 days	encoder-decoder	multistep LSTM/GRU models	MAPE, MSE
[26]	20 different data streams	load; calendar information	hourly	9 experiments (12,24,36 hours → 12,24,36 hours)	encoder-decoder	S2S	MAPE, MAE, RMSE

Despite the potential of Transformers in capturing long-term dependencies and enabling parallel computing for large model establishment, their applications in energy forecasting have generally been limited in terms of data scale and forecasting horizon perspectives. The

largest-scale Transformer-based energy forecasting study to date, conducted by Rathnayaka et al. [41], focused on 40 buildings at a university campus. It was found that a global generalist model (a Transformer) trained on all buildings outperformed specialist local LSTM models trained on individual buildings in a 6-hour-ahead forecasting task. In this study, we further investigated the potential of larger Transformer-based energy forecasting models by comparing zero-shot and fine-tuned generalizable models with the in-situ model training scheme in short-term to medium-term forecasting tasks. To this end, we evaluated the performance of models by considering both the efficacy and efficiency dimensions and answered several research questions as outlined in Section 3.4.

Chapter 3

Methodology

As noted, our goal in this study is to provide insight into the feasibility of generalized energy models (GEMs) for energy forecasting that set the foundation for larger models as more data becomes available. To evaluate GEMs and their performance, we have investigated the Transformer architectures that better fit the generalized energy forecasting tasks,considering the characteristics of the times series data from smart meters and other sources. To conduct the model training and various assessments, we have utilized a publicly available large-scale dataset, which includes historical energy consumption, outdoor air temperature, and calendar information from more than 1000 buildings. In this section, we have described the adopted methodology including the data processing approach, the GEM model development scheme including pre-training and fine-tuning processes, the Transformer-based architectures for GEM, and the state-of-the-art models used for in-situ model development, which serve as baseline comparisons.

3.1 Dataset description and processing

The pubic Building Data Genome 2 (BDG2) dataset [32] was employed in this study. The BDG2 comprises data from 3,053 energy meters across 1,636 buildings of varying sizes, located at 19 sites in North America and Europe. The buildings represent diverse types, including institutional (educational), office, and public service facilities. The time series data in the BDG2 dataset covers two full years (2016 and 2017), with hourly measurements of whole-building electricity, heating and cooling water usage, steam, solar energy, as well as water and irrigation data. The dataset also includes metadata such as area, and weather conditions. The weather data from each building site includes outdoor temperature and humidity among other attributes. The diversity of the building types and climates helps us evaluate the feasibility of GEM development using different energy use patterns. For our analysis, we have processed the BDG2 dataset to prepare multivariate time series data for model training and evaluation. Considering the availability across different buildings, we focused on three dimensions of the data: (1) whole-building electricity consumption (kWh), (2)outdoor air temperature (°C), and (3) calendar information. The information from the calendar was incorporated to capture temporal context, patterns, and dependencies within the data. This was achieved by decomposing the original date-timestamps into constituent parts (month, day, weekday, and hour) and incorporating these vectors into the energy and

temperature input embeddings.

The overall missing data ratio for electricity (E) and outdoor air temperature (T) is 3.5% and 3.9%, respectively. To address missing data points, we imputed both E and T data time series (collectively denoted as "ET") using the linear interpolation approach. Time series with missing data at the beginning, which could not be imputed due to the lack of initial entries for interpolation, were excluded from the analysis. Consequently, the final processed multivariate time series data represents a total of 1014 buildings (denoted as ET_1014 set hereafter). We have further divided this dataset into two subsets to represent 914 and 100 buildings through a random sampling process. The set of 100 buildings serves as the target building dataset (denoted as ET_100 set) for the evaluation of GEMs' performance in comparison to individual in-situ models. In other words, ET_100 is used to test our hypothesis and answer our research questions. The dataset from 914 buildings serves as a relatively large dataset (denoted as ET_914 set) for pre-training of GEMs. Training the GEM models on ET_914 set is referred to as pre-training considering that during this training process, the GEM models are not fine-tuned on the data from individual buildings in the target set (ET_100). Additionally, a separate subset of 100 buildings, randomly selected from ET_914, is used to assess GEMs' performance on buildings included in the pre-training dataset. This subset is denoted as ET_100* to differentiate it from the ET_100 target set. From the ET_914 set, two subgroups of 200 and 500 buildings were also randomly sampled (denoted as ET_200 and ET_500 sets) to study the effect of the dataset size on GEM pre-training performance. The data subsets and their respective applications in this study are summarized in Table 3.1.

The metadata distributions for different subsets are depicted in Figures 3.1, 3.2, and 3.3 to demonstrate that these subsets have similar distributions. Excluding the occupancy patterns and numbers (which are not included in the BDG2 dataset), these datasets exhibit well-aligned distributions across the three groups. All data subsets span six time zones, with most buildings having an area of under 20000 square feet. To statistically assess the consistency between the datasets, we also conducted a two-sample Kolmogorov-Smirnov (KS) test on the metadata of the buildings between different pairs of subsets. The metadata comprises categorical data (space usage, sub-primary usage, time zone) and numerical data (building area, latitude, longitude). A p-value less than or equal to the significance level of 0.05 suggests significantly different distributions, while a p-value greater than 0.05 indicates no significant difference. The results, as presented in Table 3.2, indicate that the randomly selected data subsets ET_500, ET_200, and ET_100 share a statistically similar metadata distribution with ET_914.

Table 3.1: Description of data subsets and their application

Dataset subset	Number of buildings	Number of data points*	Application	Description
ET_1014	1014	17644614	The entire processed dataset	Buildings with no missing data in energy consumption and outdoor air temperature for 2 years after imputation
ET_914	914	15904514	GEM pretraining	Buildings randomly selected from ET_1014 subset
ET_500	500	8700500	GEM pretraining	Buildings randomly selected from ET_914 subset
ET_200	200	3480200	GEM pretraining	Buildings randomly selected from ET_914 subset
ET_100	100	1740100	Target group (for GEM evaluation and baseline comparison) – i.e., Hypothesis testing	Buildings in ET_1014 that are not included in ET_914
ET_100*	100	1740100	Target group (for GEM evaluation) – in comparison with ET_100	Buildings randomly selected from ET_914 subset

* The number of data points is calculated in the case of input length = 96 and prediction horizon = 24

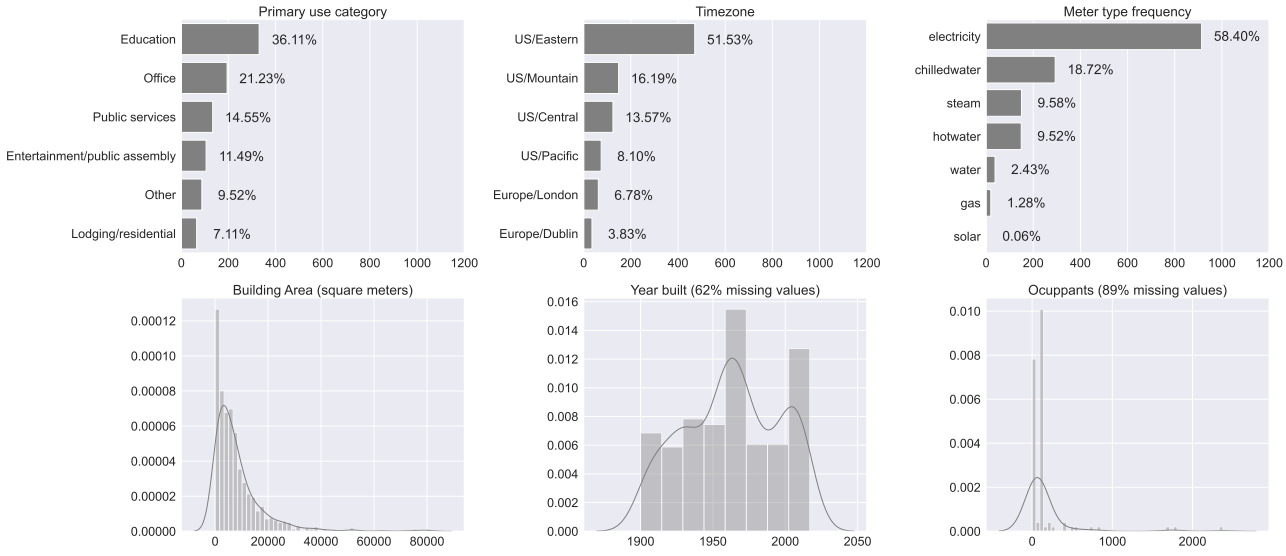


Figure 3.1: ET_914 data group metadata analysis

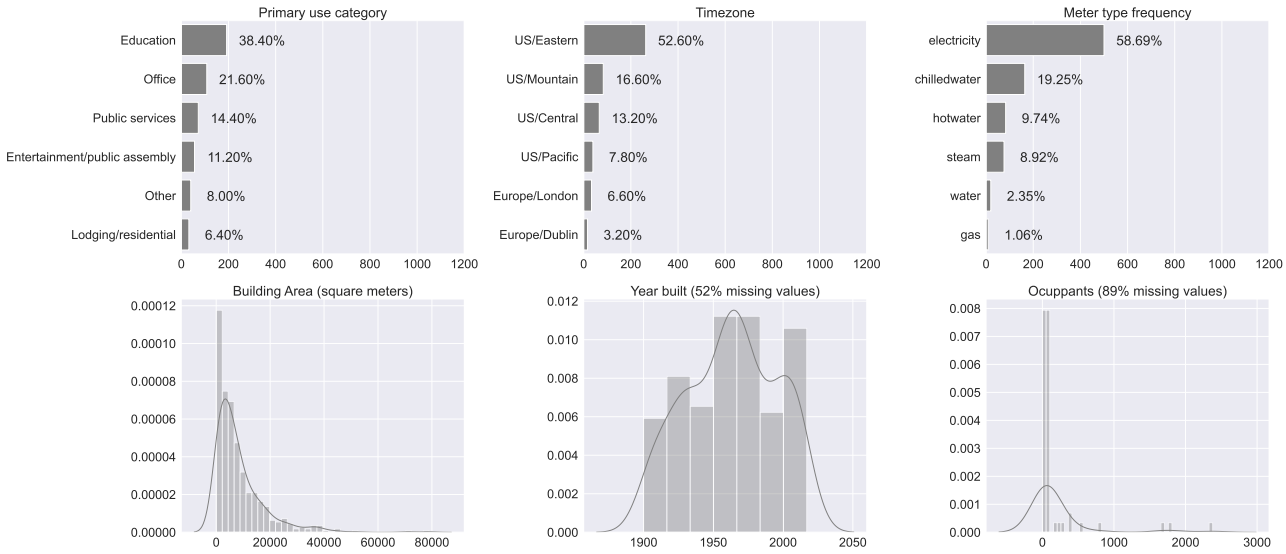


Figure 3.2: ET_500 data group metadata analysis

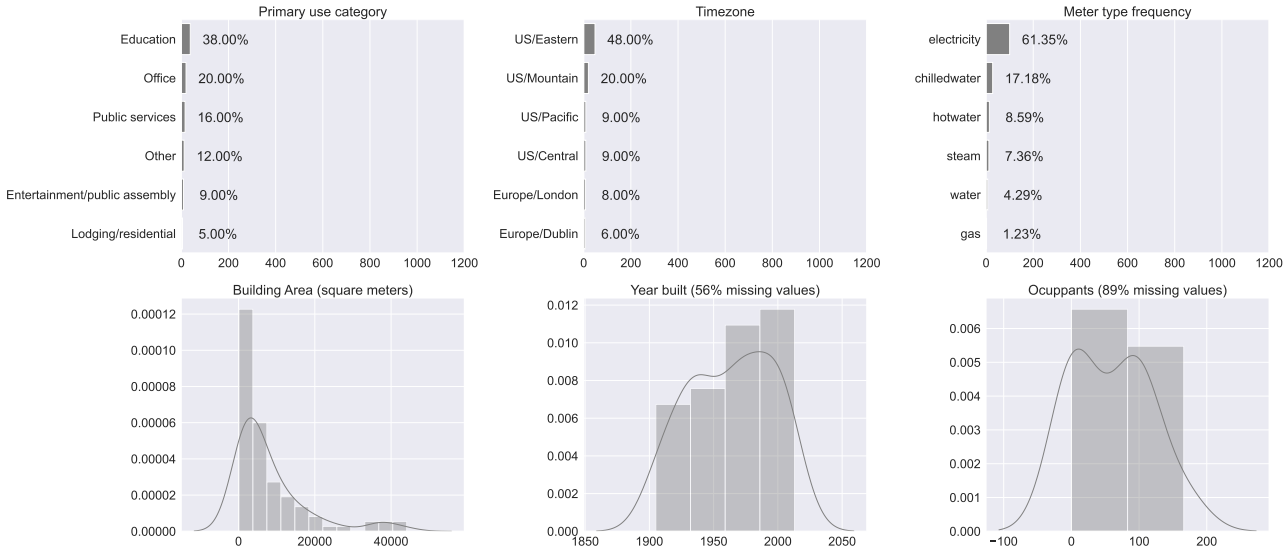


Figure 3.3: ET_100 data group metadata analysis

Table 3.2: Kolmogorov-Smirnov (KS) test results of data group pairs

Data subset pair	Statistic	p-value
ET914-ET500	0.0210	0.9690
ET914-ET200	0.0585	0.6758
ET914-ET100	0.0753	0.6818
ET500-ET100	0.0697	0.7519
ET200-ET100	0.0983	0.5794

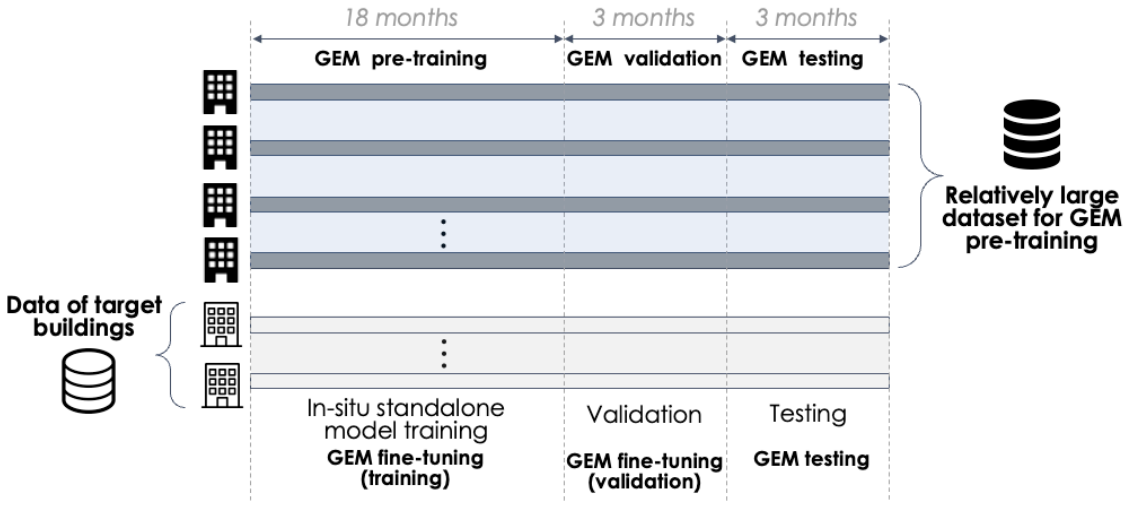


Figure 3.4: Demonstration of data split. (This figure has been designed using vectors and icons from Flaticon.com).

The time series data pre-processing also involved normalization, data splitting, subsequence extraction using a sliding window method, integration, and shuffling. First, the entire two-year energy and temperature data for each building were standardized using Z-score normalization, setting the mean value to 0 and the standard deviation to 1. As shown in Figure 3.4, for both the pre-training (ET_914) and the target group (ET_100) subsets, the two-year historical data of each building was split into a training set (the first 18 months), a validation set (from month 19 to month 21), and a testing set (from month 22 to month 24). For the in-situ training, the models were trained on the training set of an individual building in the target group and validated every few epochs on the validation set of that building. Once the validation criterion was met, the model was considered well-trained and saved for testing on the test set of the same building. Moreover, for GEM pre-training, the ET_914 subset was generated by combining all extracted subsequences from different buildings while shuffling them during the batching of the training input. The GEM model was trained and validated over the integrated training and validation sets of all buildings, and its general performance was assessed using the testing set of the ET_914 subset. Additionally, for GEM fine-tuning,

the pre-trained GEM was loaded and fine-tuned over the training and validation sets of an individual building within the target group (i.e., ET_100). The testing set of each specific building was then used to evaluate the performance of the fine-tuned GEM. In evaluating the in-situ models, zero-shot GEMs, and fine-tuned GEMs, the ground truth was the subsequent electricity time series segments following each input data point with varying lengths depending on the prediction horizon.

3.2 Transformer model architectures for GEM

The energy use patterns in buildings are affected by several factors, including occupant interactions with the environment, temperature variations, and building characteristics. The logical sequence of events in human interactions and temperature variations forms the basis for the feasibility of energy use forecasting. Temporal dependencies within a time series, such as daily and seasonal cycles, introduce regular fluctuations that must be accounted for to accurately model and predict energy use. Additionally, dependencies between different time series, such as the correlation between temperature and HVAC usage, create complex interactions that influence overall energy consumption. As discussed in Sections 2.4 and 2.5, Transformer models, with their ability to capture long-range dependencies and complex interactions, offer a promising approach to analyzing energy use patterns. These models excel at handling sequential data and can identify intricate temporal dependencies within and between time series. By leveraging self-attention mechanisms, Transformer models can evaluate the importance of different time steps and interactions, in turn enabling more accurate predictions and deeper insights into how various factors influence energy consumption. Considering our problem of interest and the data characteristics, we evaluated different attention mechanisms and adopted three architectures for comparative analysis. These attention mechanisms include series-wise (cross-temporal dependencies), dimension-wise (cross-dimension dependencies), and series-dimension-wise (cross-temporal and cross-dimension dependencies).

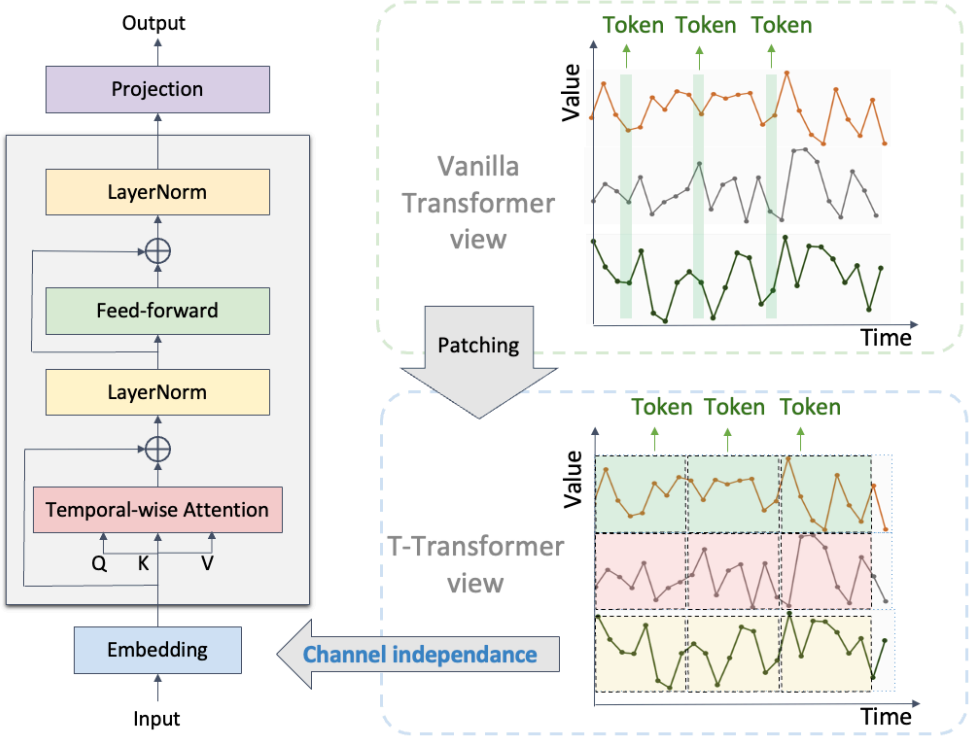


Figure 3.5: T-Transformer model architecture (modified from PatchTST [36])

Cross-temporal Transformer (T-Transformer): As illustrated in Figure 3.5, the cross-temporal Transformer models maintain the independence of different channels when using multivariate data. Despite sharing the same embedding and Transformer architecture, the data from each channel undergoes independent forward processes. This architecture aims to ensure efficient and scalable processing of multivariate time series data. By processing each channel independently, the model preserves unique channel characteristics and simplifies the overall architecture, which in turn could enhance interpretability and reduce complexity. This approach could also facilitate parallel computation, making the model more scalable and capable of handling large, high-dimensional datasets. By concentrating on temporal patterns within each channel, the model is less likely to overfit to potentially misleading inter-channel correlations, thereby improving the generalizability of the model toward new data. For this architecture, we adopted the PatchTST design [36]. This model employs a patching mechanism, as shown in Figure 3.5, similar to a sliding window that segments each univariate time series into subseries-level patches, which serve as input tokens for the Transformer. This method improves upon the original Transformer by extending the attention mechanism from point-wise to patch-wise, allowing each token to capture richer contextual information across multiple timestamps. It also reduces the number of tokens needed per input, enabling the model to effectively attend to longer historical sequences in time series data.

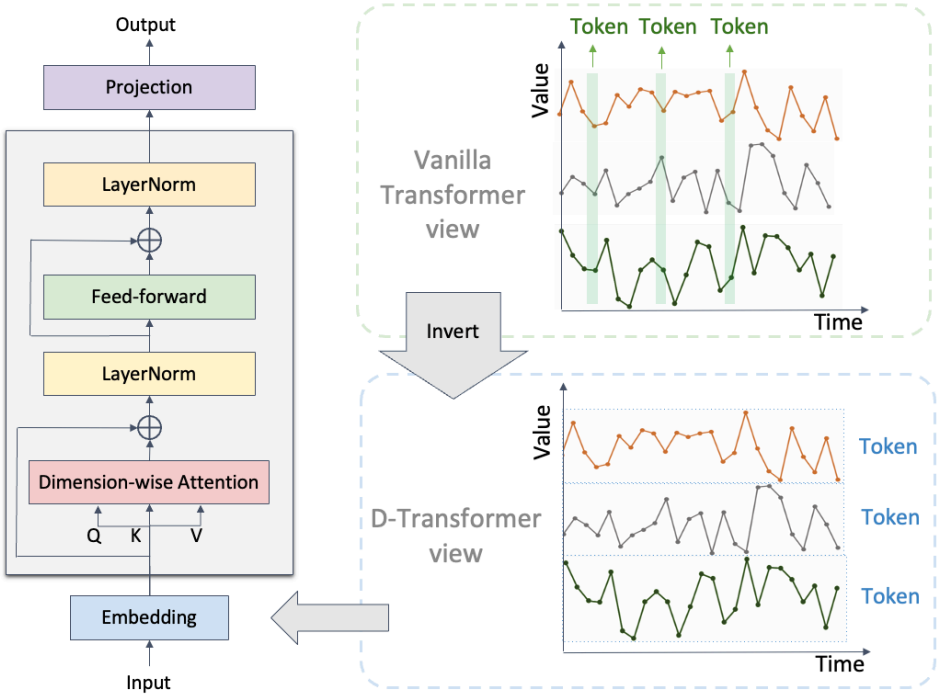


Figure 3.6: D-Transformer model architecture (modified from iTransformer [23])

Cross-dimension Transformer (D-Transformer): In the cross-dimensional Transformer model (as depicted in Figure 3.6), attention is implemented across dimensions rather than temporal tokens. In Figure 3.6, the main body on the left part adopts the encoder design of the original Transformer [47]. Cross-dimensional embeddings are processed with multi-head self-attention, normalization, and residual connections, a feed-forward network, and further normalization and residual connections to build deep contextual representations, which are finally decoded to the prediction target through a projection. As shown in the right part of Figure 3.6, raw series from various dimensions are individually embedded into tokens. Subsequently, self-attention mechanisms operate on these embedded tokens to enhance interpretability and reveal multivariate correlations. A shared feed-forward network then derives series representations for each token, followed by layer normalization to mitigate inter-variable disparities. In essence, this model design repurposes the original Transformer [47] without modifying its basic components. The only change is the input inversion in the embedding step, as shown in Figure 3.6. Unlike the original Transformer, which embeds temporal tokens representing multivariate data of each time step, this architecture introduces tokens where each time series is embedded independently as individual variables. It allows the attention mechanism to focus on correlations between different series. Meanwhile, the feed-forward network can encode information specific to each series, enhancing the model's ability to capture multivariate dependencies and nuances. For this architecture, we have adopted the iTransformer model [23].

Cross-temporal and cross-dimension Transformer (TD-Transformer): As illustrated in Figure 3.7, in this Transformer architecture, the input data is initially embedded into a 2D vector array using techniques such as Dimension-Segment-Wise (DSW) embedding to preserve both time and dimension information. This embedding approach facilitates the capture of cross-time and cross-dimension dependencies. Following embedding, cross-temporal and cross-dimension attention mechanisms, such as Two-Stage Attention (TSA) layers, are employed to efficiently model these dependencies. The hierarchical encoder-decoder structure typically includes multiple encoder layers, each leveraging cross-temporal and cross-dimension attention to capture dependencies at various scales. Upper layers typically handle longer time ranges, capturing coarser scale dependencies. The decoder component uses these representations to make final predictions by forecasting at different scales and aggregating results. This architecture could effectively model temporal and variable dependencies in multivariate time series data, potentially improving forecasting accuracy across various applications. For this architecture, we adopted the CrossFormer design and implementation [58].

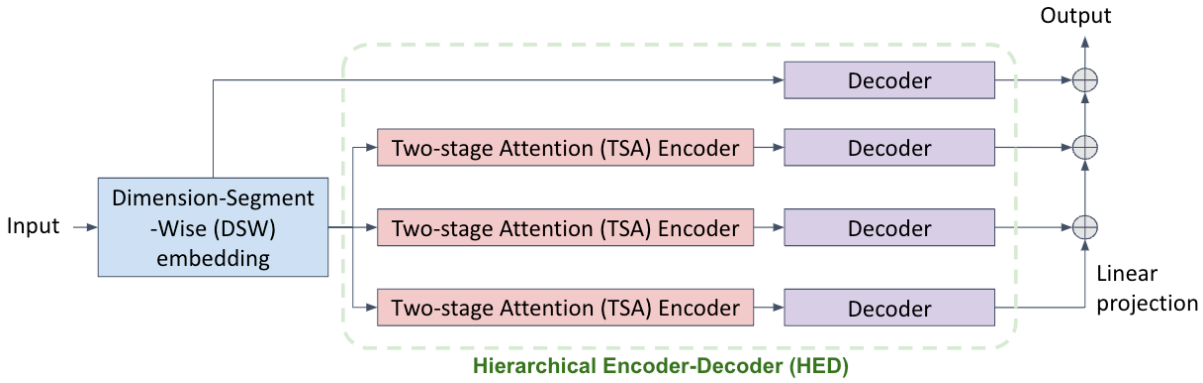


Figure 3.7: TD-Transformer architecture (modified from [58])

3.3 Model training process

The processed input data is divided into batches (smaller subsets of data points), which the models use to iteratively update the parameters during the training process. Batch training is a widely adopted technique in deep learning due to its efficiency, scalability, and effectiveness in training models on large datasets while promoting generalization and convergence. For the training loss, we used the mean squared error or MSE score (for definition see Section 3.4.2) by comparing the predicted values versus ground truth values. We also used the Adam optimizer with an initial learning rate of 1e-4. The learning rate was scaled by a factor of $0.5^{(\text{epoch}/\text{interval})}$, with the interval equal to one in our case. Decreasing the learning rate over epochs ensures efficient convergence to an optimal solution while reducing the risk of overshooting or getting trapped in local minima.

Furthermore, we utilized an early stopping criterion to ensure that the training process is halted when further training is unlikely to yield significant improvements, thus optimizing the training duration and preventing overfitting. In our experiments, the lack of significant improvements in the loss score for three epochs was used as the stopping criterion. The maximum training epochs were set to be 10 and almost all models reached the early stopping criterion. Our preliminary experiment showed that no performance improvement can be achieved by increasing the maximum number of training epochs. Moreover, to ensure the reproducibility of all results, the global random seed was set to 2024 in all experiments.

3.4 Comparative experiment schemes of GEMs and in-situ models

To test our hypothesis that a pre-trained GEM outperforms in-situ models, we considered training different models. As noted, the process of pre-training refers to the training of the GEM on a large dataset (i.e., ET_914 dataset) without exposure to the data from the ET_100 dataset. The pre-training process is the critical part of the experimental process to investigate GEM's efficacy. The integration workflow for this large-model (GEM) pre-training, as well as in-situ model training, GEM fine-tuning, and GEM zero-shot testing, is illustrated in Figure 3.8. In this context, an in-situ model training scheme involves training a specialist model (see Section 3.4.1 for baseline models we used) to learn the energy consumption patterns of an individual target building and then forecasting the future pattern for that building. As discussed in Section 2.5, most of the recent Transformer-based research on building energy forecasting adopts the in-situ model training scheme. The GEM performance was evaluated at the individual-building level using the ET_100 dataset through two schemes: (1) the zero-shot scheme, where a GEM was directly tested on individual buildings without any re-training, and (2) the fine-tuning scheme, where a GEM was fine-tuned on the training/validation dataset of the target buildings before testing on the testing set of those buildings. These schemes represent two distinct application scenarios — i.e., one without historical data available for training in a new building (zero-shot scheme) and the other with historical data available (fine-tuning scheme). We have used common metrics for evaluations as described in Section 3.4.2. Beyond testing the hypothesis, we have sought to answer the following questions:

1. Which Transformer architecture is more suitable for the development of GEM and eventually LEM models?
2. To what extent does fine-tuning a GEM model enhance its performance in a given target building that is excluded from pre-training?
3. To what extent does fine-tuning a GEM model enhance its performance in a given target building that is included in pre-training?

4. How much data is required to achieve effective fine-tuning of a pre-trained GEM model for a specific target building?

5. How does the length of the input sub-sequence affect the performance of GEMs?

6. To what extent does the size of the training dataset influence the performance of GEM models?

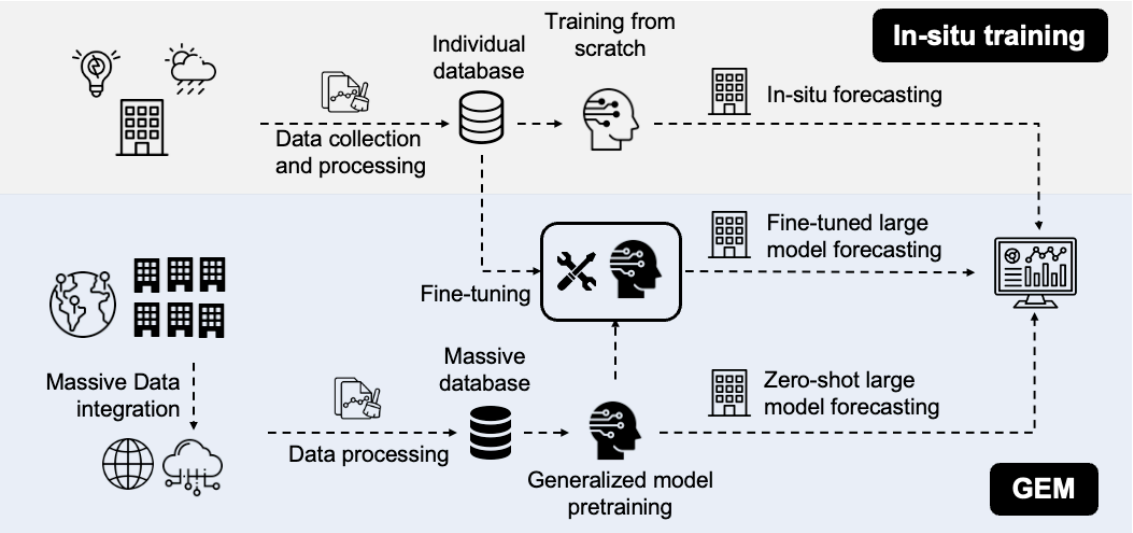


Figure 3.8: In-situ models versus zero-shot and fine-tuned GEMs for energy forecasting. (This figure has been designed using vectors and icons from Flaticon.com).

3.4.1 Baseline models for in-situ training

To demonstrate the efficacy of Transformer-based GEMs, besides the Transformer-based models, we adopted a number of the state-of-the-art deep learning models for multivariate time series forecasting including TimesNet [53], Koopa [24], FiLM [61], MICN [49], and DLinear [57]. TimesNet was proposed as a temporal 2D-variation modeling for time series analysis, addressing the intricate temporal patterns and the representational limitations of 1D time series. By converting 1D time series into 2D tensors across multiple periods, it captures both intraperiod and interperiod variations. Koopa was developed to address non-stationary time series by considering time-variant dynamics. It disentangles time-variant and time-invariant components with a Fourier Filter and employs a Koopman Predictor to model these dynamics. The model consists of stackable blocks that learn hierarchical dynamics, identifying measurement functions for Koopman embedding and using Koopman operators as linear representations of implicit transitions. FiLM , i.e., Frequency Improved Legendre Memory Model, was developed to improve time series forecasting accuracy by preserving historical data while avoiding overfitting to noise. The model utilizes Legendre Polynomial

projections to approximate historical information and employs Fourier projection to filter out noise. MICN, i.e., Multi-scale Isometric Convolution Network, integrates local features and global correlations to provide a comprehensive view of time series, capturing fluctuations and trends. The model employs a multi-scale branch structure to model various patterns separately, using down-sampled convolution for local features and isometric convolution for global correlations. DLinear is a simple linear model that was shown to outperform existing Transformer-based models on nine real-life datasets.

3.4.2 Evaluation metrics

We used several regression evaluation metrics that have been used in time series forecasting as shown in Equations 3.1-3.4, where y_i, $\hat{y}_i$, and N denote predicted values, ground truth values, and number of data points evaluated, respectively. MAE indicates the average vertical distance between the ground truth and predicted values. MSE measures the average of the squares of the deviations from the ground truth values. RMSE refers to the average magnitude of the errors between the ground truth and predicted values. MAPE measures the relative average vertical distance of the ground truth and predicted values.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \tag{3.1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.3}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{3.4}$$

Chapter 4

Results and Discussion

In this section, we have discussed the experiment results and answers the research questions listed in Section 3.4. To this end, we assessed the efficacy and efficiency by evaluating the performance of various models and detailing the training duration required for each model. Model training and testing tasks were implemented on a high-performance computing system using the Pytorch Python package. The training process was parallelized across two Nvidia DGX A100 GPUs. Each node was configured to handle two tasks simultaneously, with a maximum of 16 CPU cores allocated per task. The plots were created using the Python seaborn library [50].

4.1 GEM pre-training using different attention mechanisms

To address RQ1, which examines the effectiveness and efficiency of three Transformer architectures with distinct attention mechanisms for GEM pre-training, T-Transformer, D-Transformer, and TD-Transformer were pre-trained on ET_914, with 18 months of data as training data while the rest of data were used for validation (3 months) and testing (3 months) data. The model size, testing performance, and training time are listed in Table 4.1, where the best evaluation scores for each prediction horizon were underlined. For a fair comparison, the input sequence length was unified as 96 hours in all experiments, except in Section 4.4, where we specifically investigated the influence of input length. Three prediction horizons (24, 72, and 168 hours) were compared, which represent short to medium-term prediction. It can be seen that for each Transformer architecture except the D-Transformer for the 24-hour prediction horizon, as the prediction horizon increases, the accuracy drops and pre-training time increases. D-Transformer reaches better performance for 24 and 168-hour prediction horizons while T-Transformer presents better performance for the 72-hour prediction horizon.

Figure 4.1 compares different architectures from both effectiveness and efficiency by using the MAE metric and the required training time. Observations closer to the bottom-left origin are deemed to better fit the model development. Figure 4.1 indicates that the D-Transformer and T-Transformer qualify as good architectures for building GEMs, exhibiting both high effectiveness and high efficiency, while the TD-Transformer is less effective and efficient,

potentially attributable to the increased noise resulting from the concurrent modeling of both temporal and dimensional dependencies. The follow-up discussions on performance on an individual building basis are thus built upon the D-Transformer and T-Transformer models.

Table 4.1: GEM pre-training performance of Transformer-based backbones

Hirizon (h)	Architecture	# Parameter	MAE	MSE	RMSE	MAPE	Training time (h)
24	D-Transformer	2.4M	.2428	.1843	.4293	2.2682	8.6
	T-Transformer	6.5M	0.3493	0.3348	0.5786	2.9351	10.8
	TD-Transformer	42.1M	0.2701	0.2085	0.4566	.0453	20.0
72	D-Transformer	2.5M	0.3469	0.3288	0.5734	3.0618	4.8
	T-Transformer	6.8M	.2583	.2010	.4483	.1372	11.7
	TD-Transformer	42.1M	0.3326	0.2936	0.5418	2.5072	21.6
168	D-Transformer	2.5M	.3102	.2722	.5217	3.0742	5.0
	T-Transformer	7.3M	0.3198	0.2860	0.5348	.8248	12.0
	TD-Transformer	42.2M	0.3660	0.3615	0.6012	2.8995	27.4

Note: The best evaluation score for each prediction horizon is underlined.

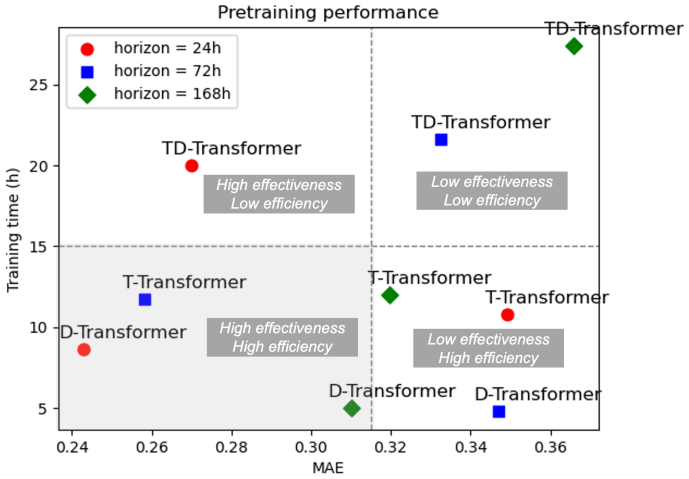
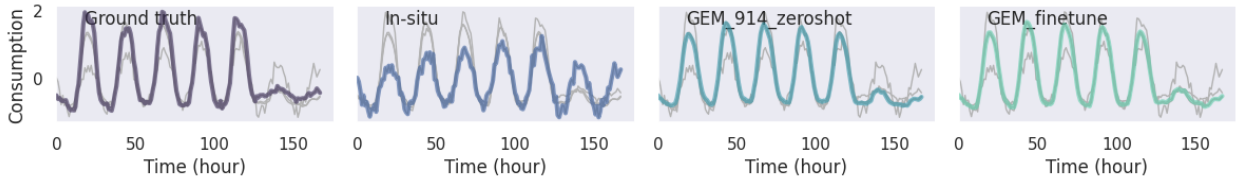


Figure 4.1: Effectiveness (MAE) and efficiency (training time) analysis of GEM pre-training performance

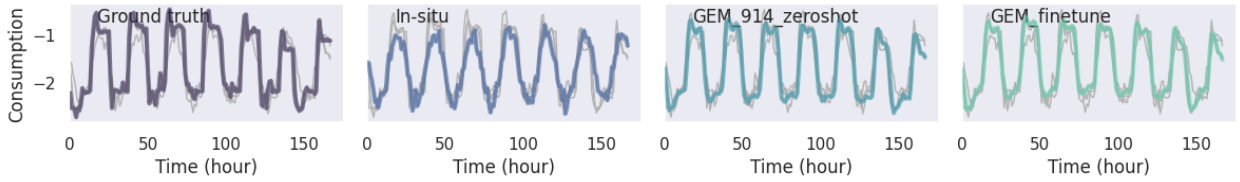
4.2 GEM performance in target buildings

To answer RQ2, which examines the impact of fine-tuning GEM models in target buildings excluded from training, and test our hypothesis, we evaluated D-Transformer- and T-Transformer-based GEMs pre-trained on ET_914 using the ET_100 dataset. We applied GEM using zero-shot and fine-tuning modes as discussed in Section 3.4. The in-situ training scheme, which employs the same Transformer-based model trained exclusively on individual buildings' data, served as the baseline for assessing GEMs' performance.

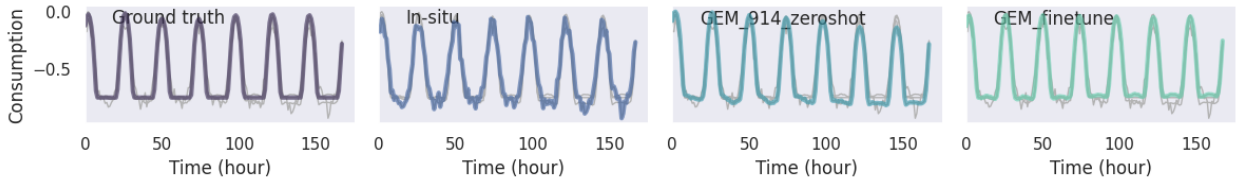
Figure 4.2 illustrates the prediction outcomes for three training/testing schemes—i.e., in-situ, GEM zero-shot, and GEM fine-tuning— for four example buildings from different types (industrial, educational, lodging, office) with a prediction horizon of 168 hours. While in-situ training from scratch can reasonably predict general trends, it often diverges from the ground truth in finer details, such as peak values and fluctuations during off-peak periods. In contrast, GEMs substantially enhance prediction accuracy compared to in-situ models. Among GEMs, fine-tuned versions offer more precise detailed predictions than zero-shot models. To better demonstrate the overall differences, the MAE scores' distributions for 100 buildings under different learning schemes of D-Transformer and T-Transformer for different prediction horizons are depicted in Figure 4.3. Moreover, Table 4.2 presents the p-values from one-tailed paired t-tests comparing GEM zero-shot/fine-tune against in-situ models and comparing GEM fine-tune against GEM zero-shot. A p-value less than 0.05 indicates that the performance enhancement is statistically significant. Notably, significant improvement is achieved by GEM zero-shot, with even more substantial enhancement observed with fine-tuning compared to the in-situ training paradigm. The potential of GEMs in improving forecasting performance through zero-shot and fine-tuning is further discussed in the following two subsections.



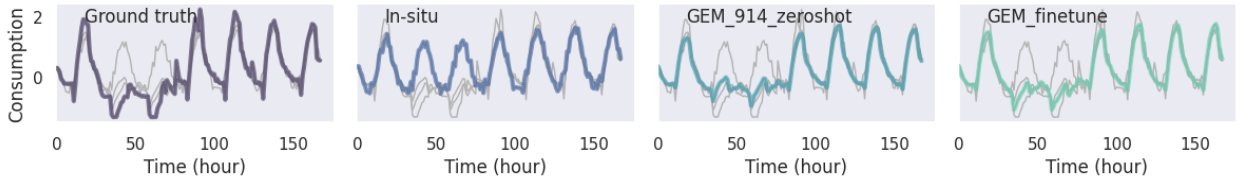
(a) An industrial building



(b) An education building



(c) A lodging building



(d) An office building

Figure 4.2: Examples of D-Transformer-based in-situ model and GEM model prediction results (Prediction horizon = 168 hours)

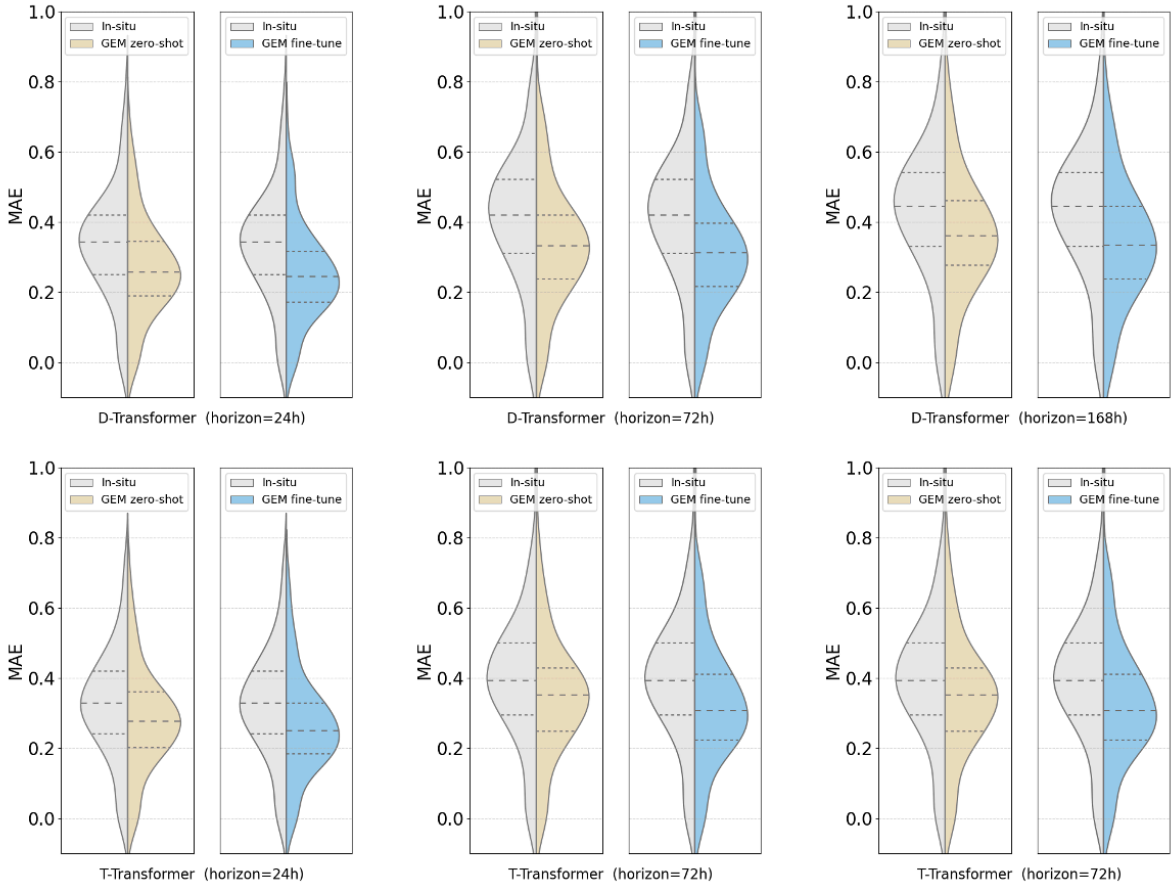


Figure 4.3: MAE score distribution of in-situ training and GEMs for individual buildings in ET_100

Table 4.2: T-test results comparing MAE metric between GEM and in-situ models using the same Transformer architectures over ET_100 dataset

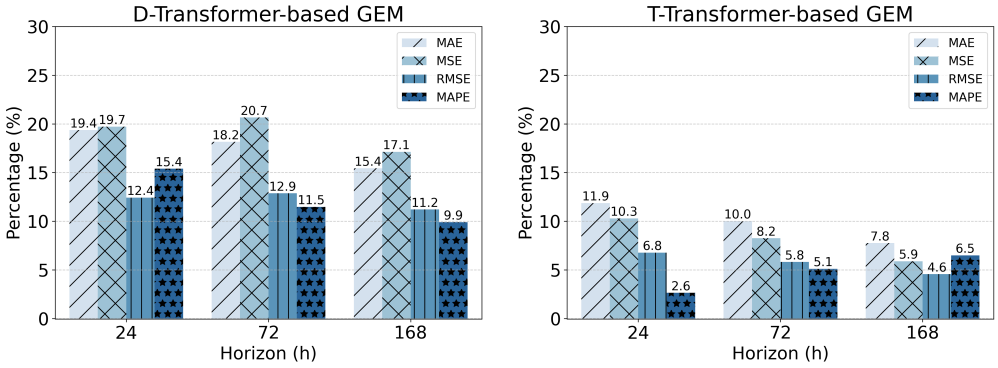
Transformer Architecture	D-Transformer			T-Transformer		
Prediction Horizon (h)	24	72	168	24	72	168
GEM zero-shot versus in-situ model*	4.7886E-19	1.6563E-14	2.6428E-16	2.0208E-11	1.4491E-08	9.9563E-11
GEM fine-tune versus in-situ model	2.8054E-24	1.584E-19	7.8221E-18	2.6616E-16	1.3629E-14	5.6856E-14
GEM fine-tune versus GEM zero-shot	4.3865E-13	3.3678E-13	1.774E-07	6.3162E-11	2.4543E-10	4.4401E-05

* The latter model is the baseline of the one-tail paired t-test. For example, in the first row, the p-value less than 0.05 indicates the performance enhancement of GEM zero-shot is significant compared with in-situ models.

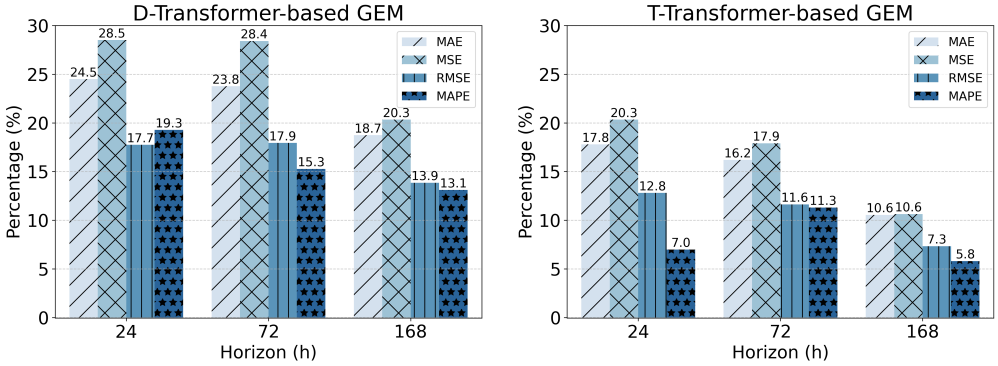
4.2.1 GEM Zero-shot Application

Zero-shot GEM refers to applying pre-trained GEMs directly on target buildings that are not included in the pre-training dataset. Figure 4.4a illustrates the average performance improvement percentage of 100 buildings facilitated by D-Transformer-based and T-Transformer-based GEMs when compared with in-situ training on individual buildings using the same D-Transformer and T-Transformer models. It is evident that GEMs effectively enhance energy forecasting performance in terms of MAE, MSE, RMSE, and MAPE. The D-Transformer demonstrated a more significant enhancement in GEM performance than the T-Transformer. Considering the MAE metric, the largest zero-shot performance enhancement was obtained when using the D-Transformer-based GEM for a prediction horizon of 24 hours, yielding 19.4%, 19.7%, 12.4%, and 15.4% decreases in MAE, MSE, RMSE, and MAPE, respectively. Additionally, a declining trend in performance enhancement is observed in both GEMs as the prediction horizon extends. This trend suggests that GEMs pre-trained on substantial datasets may yield greater improvements in short-term forecasting performance compared to longer-term tasks.

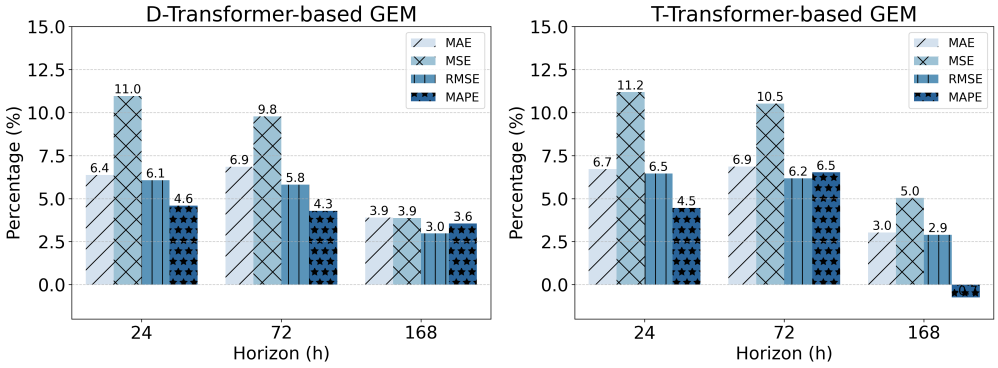
These results demonstrate the promising potential of zero-shot GEMs on unseen target buildings. Given that zero-shot GEMs require no additional data to fine-tune the model and incur no additional computing time, these findings are particularly relevant for broader applications in energy forecasting, especially in scenarios where historical data is unavailable, such as during the design and planning stages.



(a) GEM Zero-shot versus in-situ training



(b) Fine-tuned GEM versus in-situ training



(c) Fine-tuned GEM versus zero-shot GEM

Figure 4.4: Performance enhancement brought by GEMs: Comparison of zero-shot, fine-tuned, and in-situ training modalities

4.2.2 GEM Fine-tuning

In the fine-tuning scheme, GEMs were initialized with the pre-trained parameters. Training and validation were then performed using data from the first 18 months and the 19th

to 21st months of each target building, respectively. Once the early stopping criteria were met, the fine-tuned GEMs were tested on data from the 22nd to 24th months of each target building. Figure 4.4b presents the percentages of average improvement in prediction accuracy scores achieved through fine-tuned GEMs, compared to training forecasting models from scratch using individual building data. It can be observed that both D-Transformer-based and T-Transformer-based GEMs lead to significant performance enhancement for all prediction horizons. Compared with the zero-shot performance in Figure 4.4a, fine-tuning on target buildings further improved the performance of GEMs as demonstrated in Figure 4.4c. Similar to the observations in the zero-shot scheme, the D-Transformer-based GEM also achieves more substantial improvement than the T-Transformer-based GEM in the fine-tuning scheme. Moreover, fine-tuned GEMs tend to yield greater improvements in short-term forecasting performance compared to longer-term tasks. The most significant enhancement of fine-tuned GEMs was achieved by the D-Transformer for the 24-hour prediction horizon with improvements of 24.5% in MAE, 28.5% in MSE, 17.7% in RMSE, and 19.3% in MAPE.

In addition to effectiveness improvement, fine-tuning GEM also provides benefits from the efficiency perspective. Figure 4.5 shows the average percentage reduction in time required for fine-tuning GEMs compared with training a model from scratch on individual buildings. Notably, fine-tuning GEMs save more than 45% of the training time compared with in-situ training in all cases.

To address RQ4, which examines the amount of data needed for effective fine-tuning in each building, we used smaller subsets of data from individual buildings in the target group as training and validation sets for comparison. The results, summarized in Table 4.3, indicate that MAE, MSE, and RMSE increase steadily as the fine-tuning data size decreases. To achieve fine-tuned performance comparable to or better than that of an in-situ model trained from scratch on 18 months of data, at least 6 months of data is generally required. Using a fine-tuning dataset that is too small (e.g., 3 months) may lead to worse performance than zero-shot learning. This may be associated with overfitting to specific seasonal or occupancy patterns. Meanwhile, MAPE did not exhibit a consistent trend with changes in the size of the fine-tuning dataset, likely due to small actual values (i.e., small denominators in Equation 3.4) causing high percentage errors.

Table 4.3: Influence of fine-tuning data size on GEM fine-tuning performance

pred_len	months(ft_train)	months (ft_val)	**MAE**	**MSE**	**RMSE**	**MAPE**
24	18	3	**0.252**	**0.203**	**0.396**	**2.244**
24	12	3	**0.259**	**0.216**	**0.405**	3.724
24	6	2	**0.294**	0.297	**0.460**	**1.732**
24	3	1	0.348	0.429	0.552	**2.119**
72	18	3	**0.314**	**0.298**	**0.477**	**2.574**
72	12	3	**0.323**	**0.319**	**0.488**	4.339
72	6	2	**0.368**	0.446	0.556	**2.320**
72	3	1	0.438	0.648	0.670	**2.582**
168	18	3	**0.357**	**0.382**	**0.532**	**2.806**
168	12	3	**0.365**	**0.410**	**0.544**	4.959
168	6	2	**0.415**	0.565	0.616	**2.213**
168	3	1	0.496	0.839	0.745	3.122

(1) The bold numbers indicate the performance is better than the in-situ model trained from scratch on individual buildings (with 18 months of training data, 3 months of validation data, and 3 months of testing data)
(2) The underlined numbers indicate the performance is better than the GEM zero-shot performance.

In summary, utilizing pre-trained GEMs could provide a valuable shortcut at model initialization, resulting in cost savings in computational resources and facilitating convergence to a better outcome. By integrating extensive historical domain knowledge through pre-training, fine-tuning GEMs resembles leveraging innate expertise, thereby enhancing the effectiveness and efficiency of decision-making in building energy management. It is important to note, however, that effective GEM fine-tuning requires a sufficient amount of data (e.g., at least six months in our experiments) to avoid overfitting and to consistently outperform in-situ models.

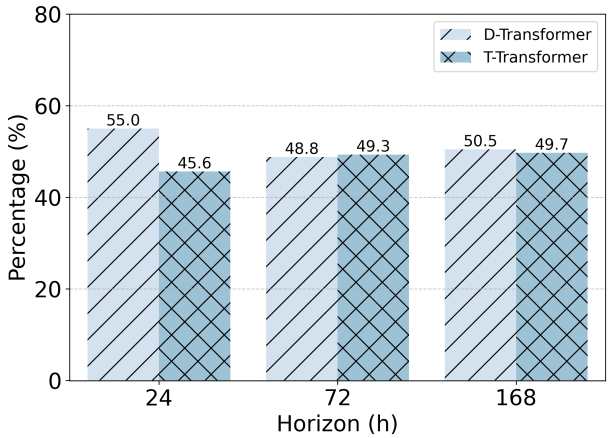


Figure 4.5: Percentage of training time consumption reduced by GEM fine-tuning compared with in-situ training

4.2.3 GEM Performance: Included vs. Excluded Buildings in Pre-training

The results presented thus far were based on observations from target buildings that were not included in the GEM pre-training dataset. To address RQ3, which examines the effect of building data inclusion in GEM training, we used another set of target buildings. To this end, we randomly selected another set of 100 buildings from the pre-training ET_914 dataset, ET_100* (shown in Table 3.1). Using the same experimental settings that were used for the ET_100 dataset, we evaluated the performance of pre-trained T-Transformer and D-Transformer GEMs over ET_100* for both zero-shot and fine-tuning schemes. A two-sample independent t-test was performed on the evaluation results of buildings in the ET_100 and ET_100* datasets. The p-values are presented in Table 4.4, with values greater than 0.05 indicating that the differences between the two groups are statistically insignificant. The results demonstrate that the performance distributions of in-situ models, zero-shot GEMs, and fine-tuned GEMs are similar across these two groups of target buildings, regardless of their inclusion in the pre-training dataset.

The average performance enhancements for both included and excluded target buildings are nearly identical (refer to Figure 4.4). Thus, for a target building that is part of the GEM pre-training data, the zero-shot GEM also provides significantly better performance than in-situ models, and fine-tuned GEM further improves prediction accuracy. Specifically, compared to in-situ models, the zero-shot GEM reduces MAE by 18.6%, 17.5%, and 14.7%, while the fine-tuned GEM reduces MAE by 24.1%, 22.9%, and 18.3% for prediction horizons of 24, 72, and 168 hours, respectively. In light of the findings from RQ2 and RQ3, it can be concluded that GEM can significantly enhance energy prediction accuracy for a target building compared to an in-situ model. The inclusion of the target building in the GEM pre-training data does not necessarily result in a statistically significant difference in performance enhancement.

Table 4.4: Two-sample independent t-test p-value comparing MAE for in-situ models and GEMs over ET_100 and ET_100* dataset

Architecture	D-Transformer			T-Transformer		
Prediction horizon (h)	24	72	168	24	72	168
In-situ	0.6406	0.6513	0.8240	0.5691	0.5687	0.7409
GEM (zero-shot)	0.6700	0.6157	0.7560	0.6177	0.7214	0.7371
GEM (fine-tune)	0.6420	0.6056	0.8025	0.5914	0.6017	0.8134

4.3 Comparison with SOTA baselines

To avoid bias and further validate the efficacy of GEMs, we extended our analysis by comparing fine-tuned GEMs against state-of-the-art (SOTA) non-Transformer multivariate time

series models. This comprehensive evaluation was conducted using the ET_100 dataset on an individual building basis. The average evaluation scores for 100 buildings and their corresponding training times are presented in Table 4.5. The parameter settings of the baseline models were manually optimized to achieve the best performance. To ensure a fair comparison, the input sequence length was set to 96 for all models. As depicted in Figure 4.6, the best performance across all three prediction horizons was achieved by the D-Transformer-based GEM. Moreover, fine-tuning pre-trained GEMs exhibited a marked efficiency advantage over training SOTA baseline models (such as TimesNet, FiLM, and Koopa) from scratch. Table 4.6 details the performance gains of the D-Transformer-based GEM compared to the best-performing baseline models. Specifically, the D-Transformer-based GEM demonstrates improvements in MAE scores of 23.9%, 9.0%, and 9.4% for the prediction horizons of 12, 24, and 72 hours, respectively, compared to the best-performing SOTA baseline models. Notably, it also reduced training time by 88%, 70%, and 47% across these prediction horizons. Overall, this comparative study further highlights the effectiveness and efficiency of GEM fine-tuning for individual building energy forecasting.

Table 4.5: Average performance of fine-tuned GEMs and SOTA in-situ baseline models (ET_100)

Prediction horizon (h)	Model	MAE	MSE	RMSE	MAPE	Training time (s)
24	TimesNet	0.402	0.380	0.557	3.007	448
	Koopa	0.344	0.334	0.509	2.844	50
	DLinear	0.352	0.301	0.503	2.604	46
	MICN	0.345	0.315	0.484	2.685	102
	FiLM	0.331	0.306	0.499	2.734	394
	D-Transformer-based GEM	0.252	0.203	0.396	2.244	46
	T-Transformer-based GEM	0.263	0.216	0.410	2.372	50
72	TimesNet	0.455	0.483	0.628	3.117	475
	Koopa	0.345	0.339	0.513	3.070	151
	DLinear	0.416	0.402	0.581	2.694	45
	MICN	0.348	0.313	0.502	2.698	78
	FiLM	0.398	0.416	0.580	2.924	557
	D-Transformer-based GEM	0.314	0.298	0.477	2.574	45
	T-Transformer-based GEM	0.325	0.311	0.491	2.528	46
168	TimesNet	0.481	0.550	0.661	3.426	551
	Koopa	-	-	-	-	-
	DLinear	0.427	0.431	0.595	2.808	47
	MICN	0.394	0.391	0.556	3.005	77
	FiLM	0.399	0.439	0.585	2.980	549
	D-Transformer-based GEM	0.357	0.382	0.532	2.806	41
	T-Transformer-based GEM	0.363	0.385	0.539	2.835	41

(1) The greener cell indicates higher accuracy or efficiency.
(2) Koopa is unable to make predictions for a prediction horizon longer than the input sequence length, so the results for a prediction horizon of 168 are omitted.

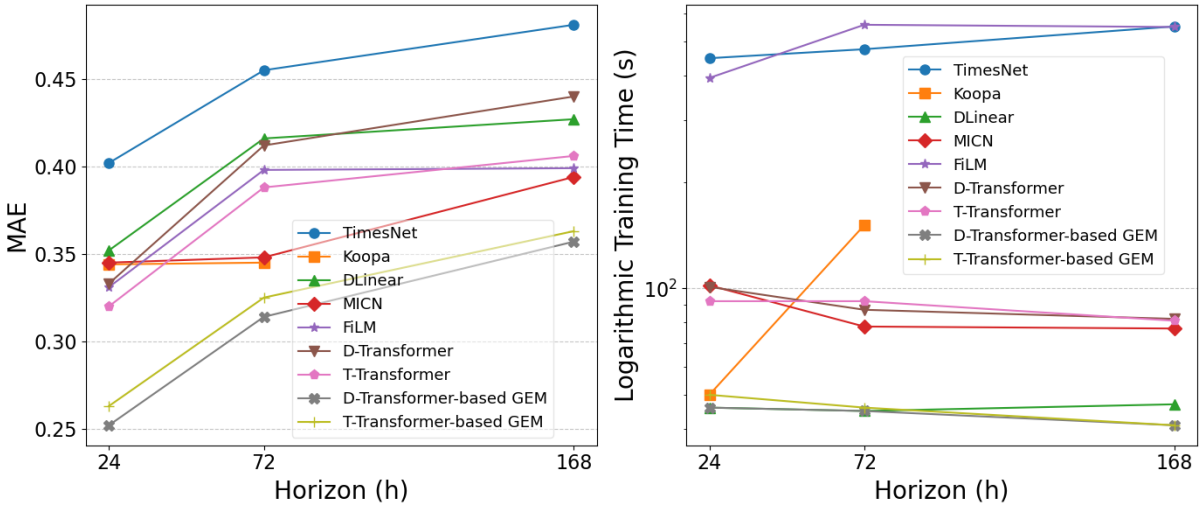


Figure 4.6: MAE score and fine-tuning/training time consumption of GEMs and SOTA in-situ baseline models

Table 4.6: Performance gains of fine-tuned GEM vs. SOTA in-situ baseline models

Prediction horizon (h)	MAE	MSE	RMSE	MAPE	Training time (s) reduction
24	23.9%	33.7%	20.6%	17.9%	88%
72	9.0%	12.1%	7.0%	16.2%	70%
168	9.4%	2.3%	4.3%	6.6%	47%

4.4 Influence of input sub-sequence length on GEM performance

To address RQ5, for D-Transformer-based GEM pre-training we examined the impact of four sub-sequence lengths to account for different activity patterns in buildings—i.e., 96 hours (4 days), 168 hours (7 days), 720 hours (30 days), and 1440 hours (60 days). Table 4.7 summarizes the evaluation scores for GEM pre-training, as well as the average scores for GEM zero-shot, fine-tuning, and in-situ models on ET_100. As sub-sequence length increases, the evaluation scores generally follow a valley trend, as illustrated in Figure 4.7, suggesting the existence of an optimal length. This optimal length depends on the specific scheme, prediction horizon, and evaluation metrics. Based on the results shown in Table 4.7, it can be concluded that 168 hours and 720 hours may be the better sub-sequence lengths because they provide better performance for GEM zero-shot and fine-tuning schemes. For practical larger-scale GEM applications, it is advisable to tune this parameter on a reserved validation subset to identify the optimal input sequence length for improved performance.

Table 4.7: Influence of sub-sequence length on D-Transformer-based GEM and in-situ model performance

Scheme	Prediction horizon (h)	Sub-sequence length (h)	MAE	MSE	RMSE	MAPE
GEM pre-training	24	96	0.243	0.184	0.429	2.268
		168	.229	0.164	0.405	2.157
		720	0.232	.163	.404	.145
		1440	0.237	0.165	0.406	2.268
	72	96	0.347	0.329	0.573	3.062
		168	0.290	0.243	0.493	.787
		720	.285	.231	.481	2.804
		1440	0.292	0.238	0.488	2.852
	168	96	.310	.272	.522	3.074
		168	0.327	0.300	0.548	2.939
		720	0.327	0.295	0.543	.843
		1440	0.335	0.305	0.553	3.008
GEM (zero-shot)	24	96	0.269	0.228	0.422	2.352
		168	.253	0.204	0.396	2.218
		720	0.255	.199	.391	2.151
		1440	0.260	0.201	0.393	.128
	72	96	0.337	0.330	0.507	2.689
		168	0.316	0.297	0.476	.505
		720	.311	.281	.463	2.539
		1440	0.316	0.285	0.467	2.524
	168	96	0.372	0.398	0.549	2.910
		168	0.357	0.377	0.529	2.915
		720	.353	.360	0.518	.772
		1440	0.357	0.364	.517	2.903
GEM (fine-tune)	24	96	0.252	0.203	0.396	2.244
		168	.245	0.193	0.385	2.142
		720	0.251	.191	.384	2.121
		1440	0.259	0.195	0.388	.114
	72	96	0.314	0.298	0.477	2.574
		168	.305	0.284	0.464	2.408
		720	0.310	.274	.458	2.454
		1440	0.320	0.284	0.466	.405
	168	96	0.357	0.382	0.532	2.806
		168	.351	0.363	0.519	2.780
		720	0.353	0.351	0.513	.725
		1440	0.360	.349	.512	2.790
In-situ model	24	96	0.333	0.284	0.481	2.779
		168	.316	.257	.455	2.651
		720	0.339	0.283	0.477	.607
		1440	0.365	0.323	0.507	2.834
	72	96	0.412	0.416	0.582	3.038
		168	.383	.363	.540	2.928
		720	0.393	0.370	0.544	.877
		1440	0.410	0.398	0.561	2.978
	168	96	0.440	0.480	0.618	3.229
		168	.431	.452	.599	.157
		720	0.441	0.456	0.602	3.279
		1440	0.454	0.477	0.612	3.307

Note: The best evaluation score for each prediction horizon is underlined.

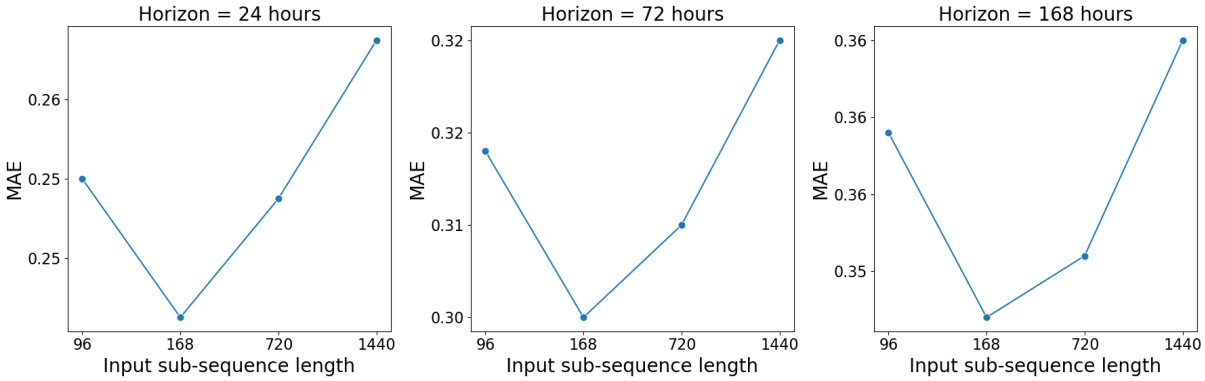
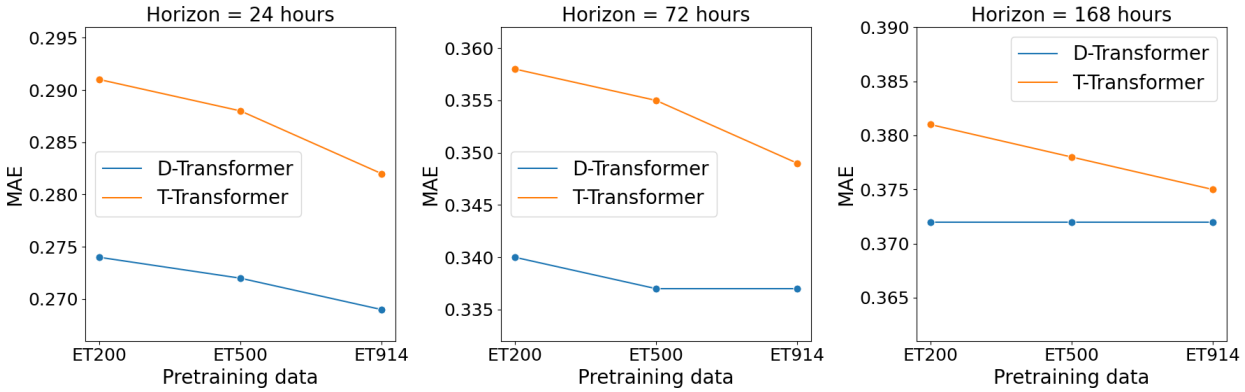


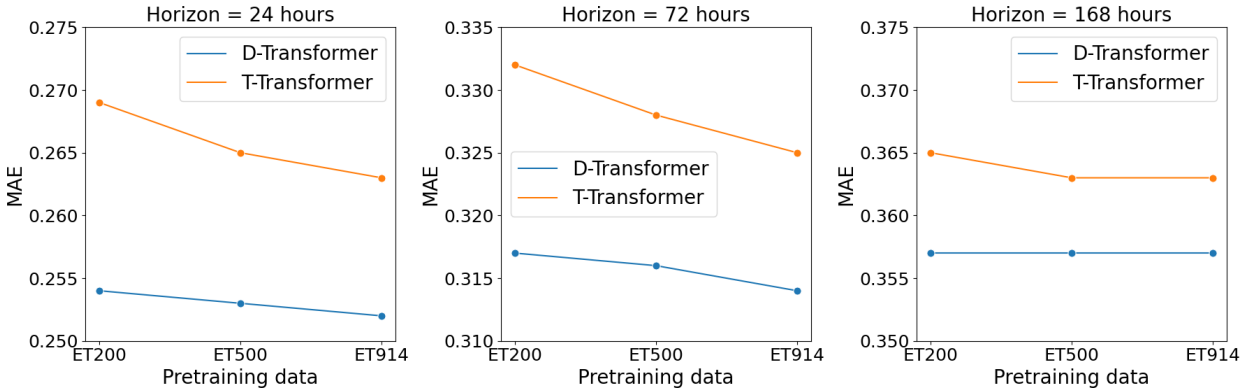
Figure 4.7: MAE scores of fine-tuned D-Transformer-based GEM with different input sub-sequence lengths

4.5 Influence of pre-training data size on GEM performance

To answer RQ6, which examines the effect of the pre-training dataset size on the performance of GEMs, we pre-trained D-Transformer-based and T-Transformer-based GEMs on ET_200 and ET_500, which are subsets of ET_914. These models are referred to as GEM200 and GEM500, respectively, compared with GEM914, which was pre-trained on the entire ET_914 dataset. The zero-shot and fine-tuning performance of GEM200 and GEM500 were evaluated in the same way used for GEM914. Figure 4.8 compares the performance of the trained models. It can be seen that, there exists a general trend of performance enhancement along the pre-training data size increase (except D-Transformer-based GEM for the 168-hour prediction horizon). Both GEM architectures witnessed steady improvement in zero-shot and fine-tuning performance at a short-term prediction horizon of 24 hours. Table 4.8 shows the result of a one-tail paired t-test comparing the GEM200 and GEM914. It can be seen that except in the case of the D-Transformer with a 168-hour prediction horizon, other scenarios all witnessed significant performance enhancements reached by GEM914 compared to GEM200.



(a) Zero-shot



(b) Fine-tune

Figure 4.8: Influence of pre-training data size on GEM performance (MAE)

Additionally, the performance enhancement percentages of GEM914 versus GEM200 are shown in Table 4.9, where a positive percentage represents that GEM914 has a better performance (i.e., lower MAE/MSE/RMSE/MAPE) compared to GEM200. In general, the larger dataset demonstrated an advantage in augmenting GEM performance. However, the extent of this improvement varied across different GEM architectures, depending on the prediction horizons. On average, the D-Transformer-based GEM saw a 0.6% improvement, while the T-Transformer-based GEM achieved a 2.3% increase in zero-shot and fine-tuning performance. Notably, the T-Transformer exhibited greater gains compared to the D-Transformer when using a larger pre-training dataset, suggesting its potential in scenarios with more extensive data availability. Additionally, shorter-term forecasts (i.e., 24 hours and 72 hours) for both models tend to benefit more from larger pre-training datasets. A more complex GEM with more parameters might better leverage a larger training dataset, potentially leading to significant future performance improvements. In summary, these findings highlight the critical role of large datasets in constructing GEMs for practical applications in building energy forecasting. Future research should incorporate even larger datasets to better understand the relationship between pre-training data size and GEM performance.

Table 4.9: Performance enhancement of GEM 914 compared with GEM 200

GEM model	Prediction horizon (h)	Zero-shot				Fine-tune				**Ave.**
		MAE	MSE	RMSE	MAPE	MAE	MSE	RMSE	MAPE	
D-Transformer	24	2%	2%	1%	0%	1%	0%	0%	1%	**0.8%**
	72	1%	1%	0%	0%	1%	1%	1%	1%	**0.8%**
	168	0%	0%	0%	1%	0%	-1%	0%	1%	**0.1%**
	Ave.	**0.8%**	**1.3%**	**0.5%**	**0.2%**	**0.4%**	**0.1%**	**0.1%**	**1.0%**	**0.6%**
T-Transformer	24	3%	4%	2%	2%	2%	2%	1%	3%	**2.5%**
	72	2%	3%	2%	5%	2%	3%	2%	7%	**3.3%**
	168	2%	2%	1%	3%	1%	1%	0%	0%	**1.2%**
	Ave.	**2.4%**	**3.0%**	**1.7%**	**3.4%**	**1.6%**	**1.8%**	**1.2%**	**3.3%**	**2.3%**

Table 4.8: Paired t-test results of GEM200 and GEM914 over 100 buildings (ET100)

Architecture	D-Transformer			T-Transformer		
Prediction horizon (h)	24	72	168	24	72	168
Zero-shot	2.09E-09	1.29E-02	3.23E-01	5.11E-11	1.53E-06	3.21E-08
Fine-tune	1.38E-03	5.11E-04	1.92E-01	1.88E-11	1.29E-08	3.66E-02

4.6 Discussion

The experimental results above confirm our hypothesis that both zero-shot and fine-tuned GEMs outperform in-situ models with the same Transformer architectures. Fine-tuned GEMs provide significant performance improvements and reduced training times but require adequate data to achieve these benefits. Moreover, input sub-sequence length has a non-negligible influence and needs to be tuned in real-world applications to reach optimal performance depending on each specific task. Additionally, the potential for greater performance gains with larger pre-training datasets was demonstrated accentuating the potential for moving toward developing Large Energy Models with very large datasets. While this study demonstrates the significant potential of GEMs, there are limitations to be addressed in future directions of this research. Although this experiment utilizes one of the largest publicly available datasets to train and evaluate Transformer models for energy forecasting, it remains limited compared to the vast datasets used in generalized models in other domains. Moreover, due to the high proportion of missing data in other dimensions of the dataset, only three input dimensions—energy demand, temperature, and calendar information—were utilized. Incorporating additional dimensions could potentially improve GEMs' performance. From a model perspective, we have adopted state-of-the-art architectures that have been designed for time series processing while future efforts can delve deeper into the impact of fine-tuning those architectures for energy forecasting problems. Finally, from the lens of functional diversity, GEMs assessed in this study are limited to single-horizon forecast-

ing, whereas broader functionalities can be taken into account to meet the diverse demands for real-world applications, such as multi-horizon forecasting, classification problems, and generative tasks.

Future research directions can focus on addressing the limitations including downstream tasks based on pre-trained GEMs for fault detection and data imputation. This integration could help build versatile models that better accommodate real-world applications for monitoring and control of building or grid systems. Currently, the presented models were designed for fixed prediction horizons, and adapting to different horizons requires separate model pre-training, which can be inefficient. Future work could explore developing multi-scale GEMs or adaptable-horizon GEMs that can efficiently handle varying prediction horizons without the need for additional pre-training. Another essential direction is the use of larger and more complex pre-training datasets. Future studies could investigate the impact of higher-dimensional data and larger datasets on the performance of GEMs, aiming to leverage the full potential of web-scale global data for more accurate and robust energy forecasting. However, this calls for formalizing the efforts in compiling global data that facilitate the development of GEMs and eventually LEMs. By addressing these research areas, the community can further enhance the capabilities and applicability of GEMs, contributing to more reliable and efficient energy management systems. The knowledge formalized by GEMs can also be transferred to domains related to building energy consumption prediction, such as renewable energy generation prediction, occupancy analysis, and electric vehicle charging prediction.

Chapter 5

Conclusions

To address the knowledge gaps in data-driven energy forecasting—i.e., dependence on data for specific contexts and limited generalizability and efficiency issues—and motivated by the potential of large models in other domains, we investigated Generalized Energy Models (GEMs) for more effective and efficient energy forecasting to avoid the need for extensive in-situ training. We investigated Transformer architectures for GEMs given their attributes including capturing long-term dependencies and their enhanced efficiency due to parallel computing. Based on the comparison of different Transformer architectures with series-wise (T-Transformer), dimension-wise (D-Transformer), and series-dimension-wise (TD-Transformer) attention mechanisms, we identified T-Transformer and D-Transformer as the most suitable architectures for developing GEMs. Two GEM application schemes of zero-shot and fine-tuned were explored by pre-training GEM models using data from 1014 buildings (with 914 buildings for pre-training and 100 buildings as the target group). The models were tasked with predicting future energy demand for 24, 72, and 168hour horizons based on date-timestamp information, outdoor air temperature, and energy consumption records.

We validated a central hypothesis that GEMs can be developed efficiently to outperform in-situ (i.e., building-specific) models trained on individual target buildings. The experimental results demonstrated that both zero-shot and fine-tuned applications of GEMs outperform in-situ models trained on target buildings' data from scratch, regardless of whether the building was included in pre-training. For target buildings excluded from pre-training, the GEM with dimension-wise attention yielded the most significant performance improvement over in-situ models, reducing MSE by 20% and 28% for a 24-hour prediction horizon in zero-shot and fine-tuned scenarios, respectively, while also cutting training time by 55%. However, it is crucial to have sufficient data to gain benefits from fine-tuning and to consistently surpass zero-shot GEMs and in-situ models. In our analyses, a minimum of data collected over six months was necessary to observe these gains. In addition to outperforming Transformer-based in-situ models, fine-tuned GEMs also surpassed other state-of-the-art multivariate time series forecasting baseline models trained using the in-situ training paradigm. Moreover, this study suggests the existence of the optimal sub-sequence length for GEM, which should be appropriately selected based on the prediction horizon and evaluation criteria. Additionally, GEMs showed an average significant improvement of 0.6% and 2.3% in zero-shot and fine-tuning performance with dimension-wise (D-Transformer) and series-wise (T-transformer) attention mechanisms, respectively, when larger pre-training datasets were used. The potential for greater performance gains with larger pre-training datasets suggests that using

web-scale global data to develop more extensive GEMs could potentially lead to the advancement of Large Energy Models for energy forecasting.

Bibliography

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Tanveer Ahmad, Huanxin Chen, Ronggeng Huang, Guo Yabin, Jiangyu Wang, Jan Shair, Hafiz Muhammad Azeem Akram, Syed Agha Hassnain Mohsan, and Muhammad Kazim. Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment. *Energy*, 158:17–32, 2018.

[3] Sina Ardabili, Leila Abdolalizadeh, Csaba Mako, Bernat Torok, and Amir Mosavi. Systematic Review of Deep Learning and Machine Learning for Building Energy. *Frontiers in Energy Research*, 10:786027, March 2022. ISSN 2296-598X. doi: 10.3389/fenrg.2022.786027. URL https://www.frontiersin.org/articles/10.3389/fenrg.2022.786027/full.

[4] Ada Canaydin, Chun Fu, Attila Balint, Mohamad Khalil, Clayton Miller, and Hussain Kazmi. Interpretable domain-informed and domain-agnostic features for supervised and unsupervised learning on building energy demand data. *Applied Energy*, 360:122741, April 2024. ISSN 03062619. doi: 10.1016/j.apenergy.2024.122741. URL https://linkinghub.elsevier.com/retrieve/pii/S0306261924001247.

[5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[6] Yongbao Chen, Mingyue Guo, Zhisen Chen, Zhe Chen, and Ying Ji. Physical energy and data-driven models in building energy prediction: A review. *Energy Reports*, 8: 2656–2671, November 2022. ISSN 23524847. doi: 10.1016/j.egyr.2022.01.162. URL https://linkinghub.elsevier.com/retrieve/pii/S2352484722001615.

[7] Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, July 2017. ISSN 13640321. doi: 10.1016/j.rser.2017.02.085. URL https://linkinghub.elsevier.com/retrieve/pii/S1364032117303155.

[8] Yavuz Eren and İbrahim Küçükdemiral. A comprehensive review on deep learning approaches for short-term load forecasting. *Renewable and Sustainable Energy Reviews*, 189:114031, 2024.

[9] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, et al. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] M González-Torres, Luis Pérez-Lombard, Juan F Coronel, Ismael R Maestre, and Da Yan. A review on buildings energy information: Trends, end-uses, fuels and drivers. *Energy Reports*, 8:626–637, 2022.

[11] Arne Groß, Antonia Lenders, Friedhelm Schwenker, Daniel A Braun, and David Fischer. Comparison of short-term electrical load forecasting methods for different building types. *Energy Informatics*, 4(Suppl 3):13, 2021.

[12] Yueyan Gu, Farrokh Jazizadeh., and Xuan Wang. Toward large energy models for forecasting: A comparative study of transformers' efficacy. *Applied Energy*, 2025. doi: 10.1016/j.apenergy.2025.125358. URL https://doi.org/10.1016/j.apenergy.2025.125358. [Accepted].

[13] Junhui Huang and Sakdirat Kaewunruen. Forecasting Energy Consumption of a Public Building Using Transformer and Support Vector Regression. *Energies*, 16 (2):966, January 2023. ISSN 1996-1073. doi: 10.3390/en16020966. URL https://www.mdpi.com/1996-1073/16/2/966.

[14] Rishee K. Jain, Kevin M. Smith, Patricia J. Culligan, and John E. Taylor. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy*, 123:168–178, June 2014. ISSN 03062619. doi: 10.1016/j.apenergy.2014.02.057. URL https://linkinghub.elsevier.com/retrieve/pii/S0306261914002013.

[15] Wenxian Ji, Zeyu Cao, and Xiaorun Li. Small Sample Building Energy Consumption Prediction Using Contrastive Transformer Networks. *Sensors*, 23(22):9270, November 2023. ISSN 1424-8220. doi: 10.3390/s23229270. URL https://www.mdpi.com/1424-8220/23/22/9270.

[16] Mohamad Khalil, A. Stephen McGough, Zoya Pourmirza, Mehdi Pazhoohesh, and Sara Walker. Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption — A systematic review. *Engineering Applications of Artificial Intelligence*, 115:105287, October 2022. ISSN 09521976. doi: 10.1016/j.engappai.2022.105287. URL https://linkinghub.elsevier.com/retrieve/pii/S0952197622003372.

[17] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[18] Elias Kyriakides and Marios Polycarpou. Short term electric load forecasting: A tutorial. *Trends in neural computation*, pages 391–418, 2007.

[19] Long Li, Xingyu Su, Xianting Bi, Yueliang Lu, and Xuetao Sun. A novel Transformer-based network forecasting method for building cooling loads. *Energy and Buildings*, 296:113409, October 2023. ISSN 03787788. doi: 10.1016/j.enbuild.2023.113409. URL https://linkinghub.elsevier.com/retrieve/pii/S0378778823006394.

[20] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3:111–132, 2022.

[21] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*, 2021.

[22] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35:9881–9893, 2022.

[23] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[24] Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Chujie Lu, Sihui Li, and Zhengjun Lu. Building energy prediction using artificial neural networks: A literature survey. *Energy and Buildings*, 262:111718, May 2022. ISSN 03787788. doi: 10.1016/j.enbuild.2021.111718. URL https://linkinghub.elsevier.com/retrieve/pii/S0378778821010021.

[26] Alexandra L'Heureux, Katarina Grolinger, and Miriam A. M. Capretz. Transformer-Based Model for Electrical Load Forecasting. *Energies*, 15(14):4993, July 2022. ISSN 1996-1073. doi: 10.3390/en15144993. URL https://www.mdpi.com/1996-1073/15/14/4993.

[27] Neda Maleki, Oxana Lundström, Arslan Musaddiq, John Jeansson, Tobias Olsson, and Fredrik Ahlgren. Future energy insights: Time-series and deep learning models for city load forecasting. *Applied Energy*, 374:124067, 2024.

[28] Massimiliano Manfren, Patrick Ab. James, and Lamberto Tronchin. Data-driven building energy modelling – An analysis of the potential for generalisation through interpretable machine learning. *Renewable and Sustainable Energy Reviews*, 167: 112686, October 2022. ISSN 13640321. doi: 10.1016/j.rser.2022.112686. URL https://linkinghub.elsevier.com/retrieve/pii/S1364032122005779.

[29] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, and F. Santos García. A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis. *Journal of Building Engineering*, 33:101692, January 2021. ISSN 23527102. doi: 10.1016/j.jobe.2020.101692. URL https://linkinghub.elsevier.com/retrieve/pii/S2352710220310627.

[30] Clayton Miller. More buildings make more generalizable models—benchmarking prediction methods on open electrical meter data. *Machine Learning and Knowledge Extraction*, 1(3):974–993, 2019.

[31] Clayton Miller and Forrest Meggers. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. *Energy and Buildings*, 156:360–373, 2017.

[32] Clayton Miller, Pandarasamy Arjunan, Anjukan Kathirgamanathan, Chun Fu, Jonathan Roth, June Young Park, Chris Balbach, Krishnan Gowri, Zoltan Nagy, Anthony D. Fontanini, and Jeff Haberl. The ASHRAE Great Energy Predictor III competition: Overview and results. *Science and Technology for the Built Environment*, 26(10):1427–1447, November 2020. ISSN 2374-4731, 2374-474X. doi: 10.1080/23744731.2020.1795514. URL https://www.tandfonline.com/doi/full/10.1080/23744731.2020.1795514.

[33] Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition. *Scientific Data*, 7(1):368, October 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00712-x. URL https://www.nature.com/articles/s41597-020-00712-x.

[34] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[35] Ramyar Rashed Mohassel, Alan Fung, Farah Mohammadi, and Kaamran Raahemifar. A survey on advanced metering infrastructure. *International Journal of Electrical Power & Energy Systems*, 63:473–484, 2014.

[36] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023.

[37] Hugo S. Oliveira and Helder P. Oliveira. Transformers for Energy Forecast. *Sensors*, 23(15):6840, August 2023. ISSN 1424-8220. doi: 10.3390/s23156840. URL https://www.mdpi.com/1424-8220/23/15/6840.

[38] Jieyang Peng, Andreas Kimmig, Dongkun Wang, Zhibin Niu, Xiufeng Liu, Xiaoming Tao, and Jivka Ovtcharova. Energy consumption forecasting based on spatio-temporal behavioral analysis for demand-side management. *Applied Energy*, 374:124027, 2024.

[39] Peng Ran, Kun Dong, Xu Liu, and Jing Wang. Short-term load forecasting based on ceemdan and transformer. *Electric Power Systems Research*, 214:108885, 2023.

[40] Zhuyi Rao and Yunxiang Zhang. Transformer-based power system energy prediction model. In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pages 913–917, Chongqing, China, June 2020. IEEE. ISBN 978-1-72814-323-1. doi: 10.1109/ITOEC49072.2020.9141649. URL https://ieeexplore.ieee.org/document/9141649/.

[41] Prabod Rathnayaka, Harsha Moraliyage, Nishan Mills, Daswin De Silva, and Andrew Jennings. Specialist vs Generalist: A Transformer Architecture for Global Forecasting Energy Time Series. In *2022 15th International Conference on Human System Interaction (HSI)*, pages 1–5, Melbourne, Australia, July 2022. IEEE. ISBN 978-1-66546-822-0. doi: 10.1109/HSI55341.2022.9869463. URL https://ieeexplore.ieee.org/document/9869463/.

[42] Lyes Saad Saoud, Hasan Al-Marzouqi, and Ramy Hussein. Household Energy Consumption Prediction Using the Stationary Wavelet Transform and Transformers. *IEEE Access*, 10:5171–5183, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3140818. URL https://ieeexplore.ieee.org/document/9672113/.

[43] Cairong Song, Haidong Yang, Jianyang Cai, Pan Yang, Hao Bao, Kangkang Xu, and Xian-Bing Meng. Multi-energy load forecasting via hierarchical multi-task learning and spatiotemporal attention. *Applied Energy*, 373:123788, 2024.

[44] Ying Sun, Fariborz Haghighat, and Benjamin C.M. Fung. A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings*, 221:110022, August 2020. ISSN 03787788. doi: 10.1016/j.enbuild.2020.110022. URL https://linkinghub.elsevier.com/retrieve/pii/S0378778819339313.

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[46] U.S. Energy Information Administration. International energy outlook 2019. Technical report, U.S. Department of Energy, September 24 2019. URL https://www.eia.gov/outlooks/ieo/pdf/ieo2019.pdf. Accessed: 2024-06-03.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[48] Chen Wang, Ying Wang, Zhetong Ding, Tao Zheng, Jiangyi Hu, and Kaifeng Zhang. A Transformer-Based Method of Multienergy Load Forecasting in Integrated Energy System. *IEEE Transactions on Smart Grid*, 13(4):2703–2714, July 2022. ISSN 1949-3053, 1949-3061. doi: 10.1109/TSG.2022.3166600. URL https://ieeexplore.ieee.org/document/9756020/.

[49] Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

[50] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

[51] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

[52] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, 2022.

[53] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

[54] Tong Xiao, Peng Xu, Renrong Ding, and Zhe Chen. An interpretable method for identifying mislabeled commercial building based on temporal feature extraction and ensemble classifier. *Sustainable Cities and Society*, 78:103635, 2022.

[55] Tong Xiao, Peng Xu, Ruikai He, and Huajing Sha. Status quo and opportunities for building energy prediction in limited data Context—Overview from a competition. *Applied Energy*, 305:117829, January 2022. ISSN 03062619. doi: 10.1016/j.apenergy.2021.117829. URL https://linkinghub.elsevier.com/retrieve/pii/S0306261921011570.

[56] Qin Yan, Zhiying Lu, Hong Liu, Xingtang He, Xihai Zhang, and Jianlin Guo. An improved feature-time Transformer encoder-Bi-LSTM for short-term forecasting of user-level integrated energy loads. *Energy and Buildings*, 297:113396, October 2023. ISSN 03787788. doi: 10.1016/j.enbuild.2023.113396. URL https://linkinghub.elsevier.com/retrieve/pii/S0378778823006266.

[57] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[58] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2022.

[59] Yang Zhao, Chaobo Zhang, Yiwen Zhang, Zihao Wang, and Junyang Li. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment*, 1 (2):149–164, April 2020. ISSN 26661233. doi: 10.1016/j.enbenv.2019.11.003. URL https://linkinghub.elsevier.com/retrieve/pii/S2666123319300121.

[60] Peijun Zheng, Heng Zhou, Jiang Liu, and Yosuke Nakanishi. Interpretable building energy consumption forecasting using spectral clustering algorithm and temporal fusion transformers architecture. *Applied Energy*, 349:121607, November 2023. ISSN 03062619. doi: 10.1016/j.apenergy.2023.121607. URL https://linkinghub.elsevier.com/retrieve/pii/S0306261923009716.

[61] Tian Zhou, Ziqing Ma, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, Rong Jin, et al. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in Neural Information Processing Systems*, 35:12677–12690, 2022.

[62] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pages 27268–27286. PMLR, 2022.