Attention as Selection: Semantic-Guided Time Series Forecasting

Xueyu Luo[1], **Qiang Lu**[1*], **Sangui Jian**[1], **Yangxue Hu**[1], **Zhengyu Ying**[1], **Ye Yu**[1], **Wenxing Lu**[2] and **Yan Qiao**[1]

[1]School of Computer Science and Information, Hefei University of Technology, Hefei, China
[2]School of Management, Hefei University of Technology, Hefei, China
xueyu@mail.hfut.edu.cn, luqiang@hfut.edu.cn, {jsg,hyx,yingzhengyu}@mail.hfut.edu.cn, {yuye,luwenxing,qiaoyan}@hfut.edu.cn

Abstract

Recent advances in time series forecasting (TSF) leverage large language models (LLMs) to provide semantic priors, enabling more robust forecasting under limited training data. However, directly fusing semantic signals into temporal features often induces modality entanglement and obscures local temporal structures when cross-modal correlations are weak, noisy, or inconsistent. To address these issues, we propose **Attention as Selection** (AAS), a dual-branch framework that decouples temporal and semantic representations. Specifically, we define cross-modal attention as a selection process, where semantic prompts are employed to induce a sparse temporal attribution distribution over temporal positions. This guides the model to focus on critical time steps without interfering with the construction of temporal representations. Furthermore, low-entropy regularization is employed alongside global cross-modal consistency constraints to regulate the selection behavior, ensuring that semantic guidance remains sparse, stable, and aligned with temporal dynamics. Extensive experiments on six real-world datasets demonstrate that AAS outperforms existing methods across various forecasting scenarios. Code is available at https://github.com/VIMLab-hfut/Attention-as-Selection.

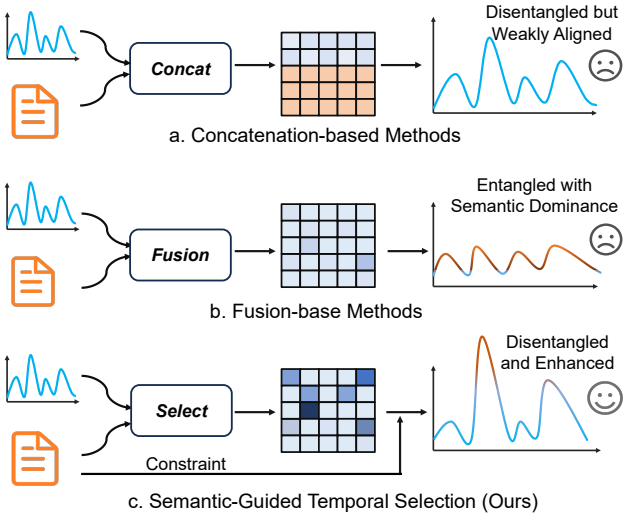

Figure 1: Comparison of cross-modal interaction methods between time series (in blue) and semantic prompts (in orange). (a) Concatenation-based methods keep modalities separate but provide weak semantic guidance, leading to poorly aligned temporal representations. (b) Fusion-based methods entangle temporal and semantic features, where semantic dominance may obscure fine-grained temporal patterns. (c) Our method employs a semantic-guided selection mechanism to modulate salient temporal features with semantic embeddings under explicit alignment constraints, yielding more discriminative and semantically aligned temporal representations.

1 Introduction

Time series forecasting (TSF) has broad applications across a wide range of domains, such as energy [Iftikhar *et al.*, 2024], environment [Vadurin *et al.*, 2025], economy [Kynigakis and Panopoulou, 2025], and other fields. Many conventional TSF models often lack explicit mechanisms to incorporate semantic information, which limits their ability to model complex real-world scenarios. In recent years, LLMs have demonstrated strong representation and reasoning abilities across diverse domains. To reduce reliance on extensive training datasets [Yang, 2025], LLMs have also been deployed in the TSF task to provide high-level semantic priors, such as structured descriptions, trend summaries and contextual knowledge [Cao *et al.*, 2024; Xue and Salim, 2023].

Existing TSF methods based on LLMs typically treat semantic representations as auxiliary features, incorporating them into temporal modeling through two primary approaches: As shown in Figure 1, concatenation-based methods directly connect semantic and temporal features at the dimensional level to maintain relative decoupling between modalities [Zhou *et al.*, 2025], which may introduce misalignment issues. Fusion-based methods achieve cross-modal interaction through feature-level fusion, such as cross-attention mechanisms [Liu *et al.*, 2025b; Liu *et al.*, 2025a], shared embedding spaces [Pan *et al.*, 2024], or joint Transformer architectures [Jin *et al.*, 2024]. Such fusion-based

*Corresponding author

methods allow semantic tokens to directly participate in constructing temporal representations, with the fused features then used for downstream prediction or generation tasks. Nevertheless, this method can lead to modal entanglement, and even local temporal structures may be obscured by semantic signals [Rehman *et al.*, 2025; Jackson, 2021]. Since accurate forecasting critically depends on preserving intricate local temporal patterns, their distortion can decrease predictive performance [Yang *et al.*, 2025].

Multimodal studies also indicate that when cross-modal correlations are weak or inconsistent, strong fusion tends to amplify noise propagation and induce modality dominance [Luo *et al.*, 2024; Zhang *et al.*, 2025; Xing *et al.*, 2024]. Driven by these limitations, multimodal studies in vision and affective computing have increasingly explored the higher-level semantic signals as guidance or constraints to preserve the integrity of core modalities, such as [Wang *et al.*, 2025b] and [Liu *et al.*, 2025c]. Nevertheless, the efficient and controlled utilization of semantic information remains underexplored in the TSF.

To address these issues, we propose the **Attention as Selection (AAS)**, which interprets cross-attention as a selection operator rather than a feature fusion mechanism. The hierarchical prompt representations produced by HPE map semantic information into importance assessments for different temporal positions by correlating semantic and temporal features, thereby generating sparse temporal attribution distributions. To prevent semantic interference, this mechanism operates solely on attention weights, adjusting the response intensity of temporal features without directly participating in their construction or representation. Moreover, low-entropy regularization (LER) is introduced to shape the selection distribution, ensuring that it remains sparse and controlled. Furthermore, we incorporate a global cross-modal consistency objective to adaptively refine semantic prompts via gradient feedback from the forecasting objective. By aligning these prompts with temporally discriminative structures, AAS achieves stable and effective selection. The contributions of our work are summarized as follows:

- We propose a new semantic–temporal modeling paradigm for LLM-based TSF, where semantic information is treated as a guiding and constraining signal rather than a feature to be directly fused with temporal representations, thereby avoiding modal entanglement.
- We propose AAS, a semantically guided selection mechanism that employs attention as a selection operator. It maps semantic representations onto sparse temporal attribution distributions. AAS also modulates temporal responses solely within the weight space without interfering with temporal feature construction.
- We introduce a low-entropy regularization along with a global temporal-semantic consistency objective to promote sparse, focused, and task-aligned temporal selection, enabling controllable and interpretable semantic attribution.
- Extensive experiments on multiple real-world datasets demonstrate that AAS outperforms state-of-the-art forecasting methods.

2 Related Work

2.1 Time Series Forecasting Using LLMs

Recent work has explored the integration of LLMs into the TSF task to incorporate high-level semantic and contextual information. Early approaches treated LLMs as the primary modeling backbone, such as LLM4TS [Chang *et al.*, 2025] and AutoTimes [Liu *et al.*, 2024b], which encoded numerical time series into tokenized sequences and performed direct forecasting within the LLMs. Despite their empirical success, these methods often neglect the modality gap between time series and textual semantics [Tao *et al.*, 2025].

In contrast, recent research has increasingly adopted decoupled dual-branch architectures, where dedicated temporal models capture time-series dynamics and LLMs encode structured prompts or textual descriptions to provide complementary semantic context. Representative methods such as CALF [Liu *et al.*, 2025b] and TimeCMA [Liu *et al.*, 2025a] combine the two modalities through feature-level fusion. Despite progress, most existing LLM-based prediction frameworks still primarily treat semantic representations as auxiliary features fused with temporal embeddings, without explicitly modeling how semantic information should guide or constrain temporal reasoning. As a result, the role of semantics in controllable and interpretable temporal modeling remains insufficiently explored in the TSF.

2.2 Cross-Modal Feature Integration

Cross-attention is widely used to model cross-modal interactions in multimodal learning [Ding *et al.*, 2025; Wang *et al.*, 2024], and has been adopted in recent LLM-based TSF to integrate semantic and temporal information. Most existing approaches treat cross-attention as a feature-level fusion operator, encoding time series and textual prompts separately and aligning them to produce joint representations. Representative methods, including TimeCMA [Liu *et al.*, 2025a] and LLMTPF [Pan *et al.*, 2025], follow this dual-branch design and use the fused features for downstream forecasting [Lee *et al.*, 2025]. While these fusion-based methods enable strong integration between modalities, they also raise the risk of semantic–temporal entanglement [Sun *et al.*, 2025], and may even interfere with the preservation of intrinsic temporal structures.

To mitigate these issues, previous multimodal research has explored more regulated forms of cross-modal interaction. In computer vision, semantic-guided methods often modulate visual representations via attention reweighting, masking, or auxiliary objectives, using textual descriptions as high-level guidance rather than fused features [Liu *et al.*, 2025c]. Similar ideas appear in affective computing, where semantic graphs or knowledge structures guide temporal or emotional modeling through graph attention [Hu *et al.*, 2025], alignment, or gating mechanisms [Cen *et al.*, 2024]. However, in LLM-based TSF, Cross-attention modules typically pursue full fusion between time-series and textual embeddings [Wang *et al.*, 2025a]. Therefore, how to leverage cross-attention to enable more interpretable and regulated interaction mechanisms remains insufficiently explored in this domain.

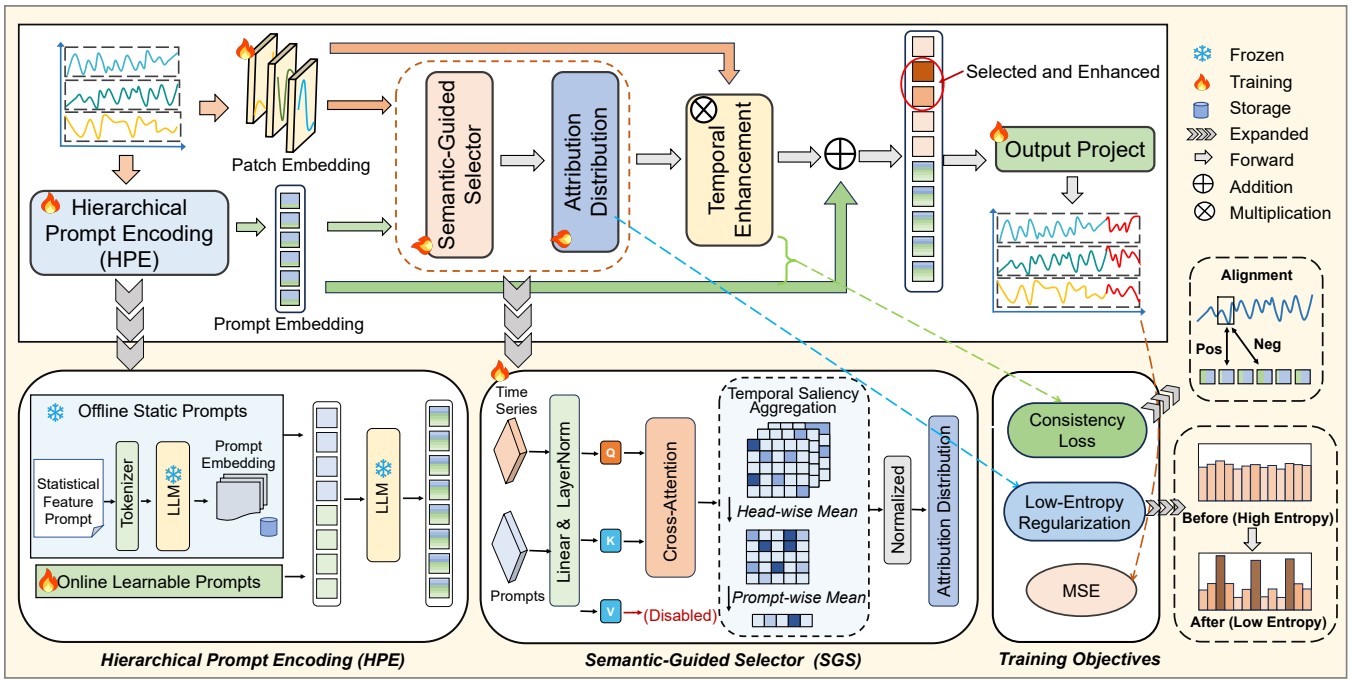

Figure 2: Overview of the AAS framework: Hierarchical Prompt Encoding module (HPE) constructs stable and adaptable semantic representations for subsequent temporal feature selection; Semantic-Guided Selection (SGS) enhances key local features through semantic guidance; and the training objectives further constrain the sparsity, concentration, and task consistency of the selection distribution, thereby achieving controllable, interpretable, and stable semantic attribution.

3 Method

Figure 2 illustrates the architecture of AAS. From semantic modeling (Section 3.2) and selection execution (Section 3.3) to constraint optimization (Section 3.4), the design forms a closed-loop structure. Semantic representations provide structured and learnable guidance for temporal selection. Additionally, explicit constraints and alignment objectives are employed to regulate the sparsity, stability, and task relevance of the outputs from the selector.

3.1 Problem Formulation

We study the task of multivariate time series forecasting (MTSF), which aims to predict future values of multiple variables based on historical observations and semantic context.

Time Series Data. Let $\mathbf{X} \in R^{T_{\text{raw}} \times N}$ denote the observed multivariate time series, where T_{raw} is the total length and N is the number of variables. At time index t, the model takes a look-back window $\mathbf{X}_{t-L+1:t}$ of length L as input and predicts the future sequence $\mathbf{Y} = \mathbf{X}_{t+1:t+H} \in R^{H \times N}$ with prediction horizon H.

Textual Data. Prompts are constructed from statistical summaries and learnable prompt tokens to provide semantic context. They are encoded by a pre-trained language model into semantic embeddings $\mathbf{E}_{\text{prompt}} \in R^{N_p \times d_{\text{llm}}}$, where N_p is the number of prompt tokens and d_{llm} is the embedding dimension.

Forecasting Objective. The goal is to learn a forecasting function f_θ that predicts future values of multivariate time series by leveraging historical temporal information together with auxiliary semantic context. Formally, the forecasting process is defined as

$$\hat{\mathbf{Y}} = f_\theta(\mathbf{X}_{t-L+1:t},\ \mathbf{E}_{\text{prompt}})\,. \tag{1}$$

3.2 Hierarchical Prompt Encoding Module (HPE)

HPE combines cached static prompts with learnable prompts to provide stable and fine-grained semantic signals for temporal selection.

Offline Static Prompts. For each variable, given a time series window $\mathbf{x} \in R^{T_{\text{raw}}}$, A set of descriptive statistical attributes is extracted and formatted into a natural language description $s(\mathbf{x})$. The description is encoded by a frozen pre-trained language model, and a static semantic embedding is obtained by mean pooling over the final-layer hidden states:

$$\mathbf{E}_{\text{static}}(\mathbf{x}) = \frac{1}{N_s}\sum_{t=1}^{N_s} \text{LLM}^{(\text{last})}\big(\text{tokenize}(s(\mathbf{x}))\big)_t \in R^{d_{\text{llm}}}, \tag{2}$$

where $\text{tokenize}(\cdot)$ denotes the tokenization function that converts text into a sequence of input tokens for the language model,$\text{LLM}^{(\text{last})}$ denotes the final-layer hidden states of the frozen language model, N_s is the number of static prompt tokens. These embeddings are computed offline and cached, providing a stable, non-trainable semantic anchor and avoiding repeated LLM forward passes during training.

Online Learnable Prompts. To enable task-specific adaptation, a set of learnable prompt embeddings $\mathbf{P}_{\text{learn}} \in$

$R^{M\times d_{\text{llm}}}$ is introduced, where M denotes the number of learnable prompt tokens. These embeddings are randomly initialized. To avoid unstable prompt drift, gradients are restricted to the learnable prompt embeddings.

$\mathbf{H}_{\text{prompt}} \in R^{N_p\times d_{\text{llm}}}$, where $N_p = M+L_p$ denotes the total number of prompt tokens. The concatenated prompt embeddings are then fed into the LLM

$$\mathbf{E}_{\text{prompt}} = \text{LLM}_{\text{last}}(\mathbf{H}_{\text{prompt}}) \in R^{N_p\times d_{\text{llm}}}. \tag{3}$$

The resulting semantic representation combines both *global stability* and *local adaptability*, mitigating the interference of semantic noise in temporal modeling.

3.3 Semantic-Guided Selection (SGS)

To decouple semantic attribution from temporal representation learning, cross-attention is restricted to operate as a selector rather than a fusion module. Specifically, SGS is constrained to: (i) generate feature attribution weights solely at temporal positions, (ii) discard all value-based feature aggregations, and (iii) enforce sparsity in the attribution distribution through explicit regularization.

Input Representations. Given a normalized multivariate time series, patch-level temporal representations are extracted using the temporal encoder:

$$\mathbf{H}_t = \text{Embed}(\text{Patch}(\mathbf{x})) \in R^{T\times d}, \tag{4}$$

where T denotes the number of temporal patches and d is the hidden dimension of the temporal encoder. To align temporal features with the semantic embedding space of the language model, they are further projected into the LLM hidden space via a linear projection layer:

$$\tilde{\mathbf{H}}_t = \text{Proj}(\mathbf{H}_t) \in R^{T\times d_{\text{llm}}}, \tag{5}$$

The semantic representations are obtained as $\mathbf{E}_{\text{prompt}} \in R^{N_p\times d_{\text{llm}}}$ from the HPE described in Section 3.2.

Cross-Attention as Temporal Attribution. We define a semantic selector $\mathcal{S}(\cdot)$ that maps temporal and semantic representations to a normalized temporal attribution distribution:

$$\mathbf{w} = \mathcal{S}(\tilde{\mathbf{H}}_t, \mathbf{E}_{\text{prompt}}), \quad \mathbf{w} \in R^T. \tag{6}$$

The selector $\mathcal{S}(\cdot)$ is formulated as a constrained multi-head attention score operator, which computes normalized similarity scores between temporal queries and semantic keys, while explicitly abandoning all value-based feature aggregation. Specifically,

$$\mathbf{Q} = \tilde{\mathbf{H}}_t\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{E}_{\text{prompt}}\mathbf{W}_K, \tag{7}$$

where $\mathbf{W}_Q, \mathbf{W}_K \in R^{d_{\text{llm}}\times d_k}$ are learnable projection matrices and d_k denotes the per-head query/key dimension. The multi-head attention scores are computed as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \in R^{H\times T\times N_p}, \tag{8}$$

where H denotes the number of attention heads. Importantly, only the attention weights are retained and treated as semantic attribution scores for temporal positions. In this manner, semantic prompts influence temporal modeling solely through attribution, rather than directly participating in the construction of temporal features.

Temporal Saliency Aggregation. To obtain a unique attribution decision for each temporal token, the multi-head and multi-prompt attention tensor is collapsed into a single normalized attribution distribution:

$$w_t = \frac{1}{HN_p}\sum_{h=1}^{H}\sum_{p=1}^{N_p}\mathbf{A}_{h,t,p}, \quad \mathbf{w} \in R^T, \tag{9}$$

where h indexes attention heads, t indexes temporal tokens, and p indexes prompt tokens. This aggregation explicitly removes head-specific and token-specific variability, forcing the selector to produce a global and comparable temporal attribution distribution over temporal positions.

Selection-Based Temporal Enhancement. Unlike traditional attention mechanisms that generate fused representations, the attribution vector $\mathbf{w}$ is treated as a decision signal. Specifically, the selector adjusts temporal representations through constrained element-wise rescaling:

$$\widehat{\mathbf{H}}_t = \tilde{\mathbf{H}}_t \odot (1+\gamma\mathbf{w}), \tag{10}$$

where γ controls the modulation strength and $\odot$ denotes broadcasting over the feature dimension.

Importantly, this operation preserves the original temporal representation space and merely redistributes the activation energy over time, enforcing a strict separation between semantic attribution and temporal representation learning. As a result, semantic prompts influence *where* the model attends and *how important* each time step is emphasized, while the temporal encoder remains solely responsible for modeling temporal dynamics.

3.4 Training Objectives and Regularization

Low-Entropy Regularization (LER). To explicitly shape the behavior of the semantic selector and encourage sparse yet stable temporal selection, the LER term is introduced over the attribution distribution. Here, the entropy of the attribution distribution characterizes the uncertainty and dispersion of temporal selection induced by the semantic selector. Given the attribution score vector $\hat{\mathbf{w}} \in R^T$ obtained by SGS, it is normalized into a probability distribution:

$$\hat{w}_t = \frac{w_t}{\sum_{t'=1}^{T} w_{t'}}, \quad \hat{\mathbf{w}} \in R^T. \tag{11}$$

To construct a sharpened target distribution that favors peaked selections, the attribution distribution is exponentiated with an entropy control parameter $\alpha > 1$:

$$p_t = \frac{\hat{w}_t^\alpha}{\sum_{t'=1}^{T}\hat{w}_{t'}^\alpha}, \quad \mathbf{p} \in R^T. \tag{12}$$

The regularization objective minimizes the KL divergence between the normalized attribution distribution and the sharpened prior:

$$\mathcal{L}_{\text{LE}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{t=1}^{T}\hat{w}_i[t]\log\frac{\hat{w}_i[t]}{p_i[t]}, \tag{13}$$

where B denotes the batch size. This regularization encourages the model to concentrate attribution mass on a small subset of informative temporal positions. Thus, the selector can learn sparse, controllable and interpretable selection strategies, enabling semantic guidance to maintain focus without overriding the inherent temporal dynamics.

Consistency Loss. An InfoNCE-based contrastive loss is applied to encourage global consistency between temporal and semantic representations. Let $\mathbf{z}_i^t$ and $\mathbf{z}_i^s$ denote the pooled temporal and semantic embeddings of the i-th sample, respectively. The alignment objective is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp(\text{sim}(\mathbf{z}_i^t,\mathbf{z}_i^s)/\tau)}{\sum_{j=1}^{M}\exp\big(\text{sim}(\mathbf{z}_i^t,\mathbf{z}_j^s)/\tau\big)}, \quad (14)$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature hyperparameter, $(\mathbf{z}_i^t,\mathbf{z}_i^s)$ forms the positive temporal–semantic pair, and all mismatched pairs $\{(\mathbf{z}_i^t,\mathbf{z}_j^s) \mid j \neq i\}$ within the mini-batch are treated as negative pairs.

By aligning semantic prompts with the selected and enhanced temporal representations, this objective calibrates the semantic space to remain consistent with task-relevant and discriminative temporal structures. This alignment mitigates semantic drift and establishes a stable global constraint that reinforces reliable and interpretable temporal selection.

Overall Objective. The model is trained by jointly optimizing the forecasting objective together with auxiliary semantic regularization terms. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{pre}} + \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{LE}}\mathcal{L}_{\text{LE}}, \quad (15)$$

where $\mathcal{L}_{\text{pre}}$ denotes the forecasting loss, $\mathcal{L}_{\text{InfoNCE}}$ encourages cross-modal temporal–semantic consistency, $\mathcal{L}_{\text{LE}}$ enforces sparse and controlled temporal selection, and λ_{LE} is a learnable weight that is jointly optimized with the model parameters to adaptively balance the regularization strength.

4 Experimental

4.1 Experimental Setup

Datasets. We conduct experiments on multiple widely-used benchmark datasets from different domains. For long-term time series forecasting, we evaluate our method on the ETT benchmark [Zeng *et al.*, 2023], Weather and Exchange datasets [Zhou *et al.*, 2021].

Baselines. We compare our method with a diverse set of representative baselines, including: (1) LLM-based methods: Time-LLM [Jin *et al.*, 2024], TimeCMA [Liu *et al.*, 2025a], and CALF [Liu *et al.*, 2025b]; (2) Transformer-based methods: iTransformer [Liu *et al.*, 2024a] and PatchTST [Nie *et al.*, 2022]; (3) CNN-based methods: TimesNet [Wu *et al.*, 2023]; (4) Linear model-based methods: DLinear [Zeng *et al.*, 2023]. For all baselines, the official implementations or scripts released by the original authors are used.

Implementation Details. For all LLM-based methods, we adopt GPT-2 with six Transformer layers as the language model backbone, with all backbone parameters frozen during training. For fair comparison, all methods are trained under the same experimental settings. Unless otherwise specified, the input sequence length is fixed to $L = 512$ across all methods and datasets. We evaluate all models under both full-data and few-shot settings. In the few-shot setting, we randomly sample 10% of the training data, with validation and test sets unchanged. All datasets are evaluated under multiple forecasting horizons $H \in \{96, 192, 336, 720\}$ to assess robustness across different prediction scales. Each experiment of AAS is repeated three times with different random seeds, and the averaged results are reported. For additional experimental details, please refer to the supplementary material.

4.2 Long-term Forecasting Results

Table 1 summarizes the long-term forecasting performance across all datasets and horizons. The best or second-best average accuracy is achieved by AAS across most datasets, with particularly strong performance observed on the ETT benchmarks. Compared to DLinear, a strong non-LLM baseline, AAS attains an average reduction of 12.6% in MSE and 8.2% in MAE across the four ETT datasets, computed over all forecasting horizons. Against the leading LLM-based forecaster CALF, AAS consistently improves accuracy, reducing average MSE and MAE by 5.1% and 4.3%, respectively, with gains becoming more pronounced at longer horizons. On highly non-stationary datasets such as Exchange, AAS achieves stable performance by leveraging semantic-guided selection to preserve discriminative temporal structures. In addition, the global consistency objective calibrates the semantic space to remain aligned with task-relevant temporal patterns. This alignment mitigates semantic drift and suppresses the accumulation of long-horizon errors, thereby enabling robust and reliable forecasting under complex and evolving temporal dynamics.

4.3 Few-shot Forecasting Results

In the few-shot forecasting experiments, the AAS model demonstrates clear and robust superiority. As shown in Table 2, the experiments are conducted under the strict setting of using only 10% of the training data. AAS maintains consistently lower MSE and MAE than all baseline models across all ETT datasets and all forecasting horizons, indicating strong robustness and generalization under data-scarce conditions. Notably, while MSE and MAE increase for all models as the prediction horizon extends from 96 to 720, AAS exhibits the slowest error growth, indicating superior stability in long-term forecasting. In scenarios of sparse data, AAS leverages frozen LLM priors to reduce reliance on extensive training datasets, employing SGS for efficient semantically guided temporal selection. Combined with low-entropy regularization that promotes sparse and focused attribution, alongside an InfoNCE-based cross-modal alignment consistency objective, this design mitigates overfitting while preserving robust representational capabilities.

4.4 Ablation Study

To validate the effectiveness of the key components in the AAS, an ablation study is conducted across all datasets. All ablation variants were evaluated under identical experimental settings as the full model. Taking ETTh2 and Exchange as examples, we analyze the model from the following perspectives:

(1) w/o Fusion-based Cross-Attention. The SGS module is replaced by a standard cross-attention mechanism for cross-modal fusion. Temporal features serve as queries, while

Categories		LLM-based								Transformer-based				CNN-based		Linear-based	
Models		**AAS**		TimeCMA		CALF		Time-LLM		iTransformer		PatchTST		TimesNet		Dlinear	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.364**	**0.392**	0.399	0.415	0.387	0.414	0.389	0.410	0.396	0.421	0.392	0.416	0.448	0.460	0.367	0.393
	192	**0.397**	**0.412**	0.418	0.437	0.411	0.429	0.419	0.439	0.426	0.442	0.454	0.461	0.495	0.491	0.402	0.419
	336	**0.426**	**0.435**	0.467	0.452	0.430	0.442	0.453	0.459	0.461	0.466	0.485	0.479	0.536	0.521	0.431	0.443
	720	**0.460**	0.480	0.463	**0.466**	0.476	0.484	0.445	0.465	0.681	0.598	0.672	0.576	0.578	0.539	0.475	0.497
ETTh2	96	**0.273**	**0.341**	0.375	0.408	0.297	0.353	0.303	0.356	0.302	0.359	0.315	0.371	0.361	0.408	0.300	0.365
	192	**0.328**	**0.376**	0.370	0.412	0.361	0.392	0.357	0.388	0.373	0.404	0.391	0.417	0.409	0.441	0.396	0.427
	336	**0.349**	**0.393**	0.437	0.463	0.390	0.417	0.375	0.412	0.418	0.435	0.443	0.449	0.395	0.438	0.498	0.491
	720	**0.381**	**0.426**	0.497	0.496	0.411	0.440	0.416	0.448	0.447	0.466	0.506	0.497	0.485	0.490	0.806	0.636
ETTm1	96	0.303	**0.348**	0.324	0.365	**0.300**	0.351	0.336	0.371	0.310	0.364	0.305	0.354	0.349	0.381	0.305	0.349
	192	**0.336**	**0.368**	0.373	0.393	0.337	0.376	0.361	0.386	0.348	0.387	0.359	0.394	0.480	0.449	0.337	0.368
	336	**0.366**	0.386	0.407	0.415	0.370	0.397	0.388	0.401	0.378	0.404	0.379	0.405	0.410	0.426	0.366	**0.385**
	720	**0.415**	0.423	0.468	0.448	0.424	0.426	0.433	0.428	0.440	0.441	0.451	0.449	0.468	0.462	0.420	**0.417**
ETTm2	96	**0.162**	**0.253**	0.190	0.275	0.171	0.262	0.187	0.275	0.180	0.272	0.182	0.272	0.194	0.281	0.166	0.261
	192	**0.219**	**0.295**	0.261	0.325	0.227	0.298	0.236	0.309	0.242	0.313	0.236	0.309	0.260	0.324	0.224	0.303
	336	**0.275**	**0.331**	0.317	0.360	0.288	0.335	0.283	0.341	0.294	0.347	0.293	0.346	0.319	0.369	0.296	0.359
	720	**0.367**	**0.387**	0.397	0.411	0.368	0.387	0.368	0.390	0.373	0.395	0.396	0.408	0.421	0.423	0.410	0.432
Weather	96	0.168	0.223	0.168	0.225	0.158	0.209	0.157	**0.209**	0.169	0.220	**0.155**	0.211	0.164	0.220	0.171	0.232
	192	0.202	**0.234**	0.203	0.256	0.204	0.249	**0.198**	0.247	0.212	0.257	0.209	0.259	0.221	0.270	0.215	0.274
	336	**0.245**	**0.282**	0.247	0.286	0.250	0.285	0.250	0.287	0.268	0.295	0.248	0.286	0.281	0.313	0.259	0.292
	720	0.321	0.335	**0.311**	**0.334**	0.332	0.342	0.317	0.334	0.334	0.341	0.318	0.336	0.344	0.356	0.320	0.359
Exchange	96	**0.086**	**0.205**	0.263	0.381	0.099	0.227	0.107	0.235	0.107	0.233	0.216	0.248	0.247	0.369	0.117	0.259
	192	**0.168**	**0.298**	0.629	0.603	0.191	0.315	0.252	0.362	0.208	0.327	0.209	0.327	0.375	0.456	0.234	0.368
	336	**0.351**	**0.432**	0.802	0.679	0.423	0.481	0.462	0.504	0.419	0.480	0.378	0.453	0.622	0.601	0.627	0.620
	720	**0.912**	**0.716**	2.328	1.211	1.129	0.808	1.194	0.823	1.020	0.765	1.274	0.834	1.515	0.928	1.170	0.827

Table 1: Long-term forecasting results of AAS and baseline methods. A look-back window of 512 is used with prediction horizons $H \in \{96, 192, 336, 720\}$. Performance is measured by MSE and MAE, with best values in bold and second-best underlined.

Categories		LLM-based								Transformer-based				CNN-based		Linear-based	
Models		**AAS**		TimeCMA		CALF		Time-LLM		iTransformer		PatchTST		TimesNet		Dlinear	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	Avg	**0.531**	**0.502**	0.742	0.582	0.702	0.580	0.806	0.611	0.708	0.588	0.674	0.574	0.875	0.648	0.532	0.503
ETTh2	Avg	**0.359**	**0.403**	0.439	0.460	0.459	0.464	0.473	0.473	0.426	0.447	0.450	0.460	0.478	0.485	0.415	0.445
ETTm1	Avg	**0.408**	**0.414**	0.701	0.565	0.498	0.471	0.525	0.472	0.552	0.503	0.646	0.520	0.801	0.599	0.414	0.428
ETTm2	Avg	**0.271**	**0.322**	0.344	0.384	0.303	0.338	0.314	0.356	0.296	0.349	0.328	0.367	0.372	0.399	0.298	0.352

Table 2: Few-shot forecasting results. All results are averaged over four prediction horizons $H \in \{96, 192, 336, 720\}$ with a look-back window of 512. Performance is evaluated by MSE and MAE, where the best results are in bold and the second-best are underlined.

semantic prompt embeddings function as keys and values. Unlike SGS, this variant does not introduce an explicit temporal selection mechanism, but directly produces fused temporal representations via cross-attention.

(2) w/o LER. The SGS module is retained, but the LER on the attribution weights is removed. In this setting, semantic guidance is still applied through cross-attention-based temporal selection, while the attribution distribution is unconstrained.

(3) w/o SGS. This variant serves as the backbone baseline, obtained by removing the SGS and the LER from the full model.

As shown in Table 3, the full AAS model consistently achieves the best performance on both the ETTh2 and Exchange datasets. Introducing SGS into the temporal forecasting backbone leads to the most significant performance improvement, demonstrating the effectiveness of explicitly modeling semantic-guided temporal attribution. In contrast, replacing SGS with a fusion-based alignment strategy fails to improve performance and even leads to degradation, indicating that directly integrating semantic representations into temporal features may introduce noise and interfere with temporal dynamics. This observation is further supported by Figure 3. When employing AAS, the feature response exhibits pronounced locality and heterogeneity in the temporal dimension. In contrast, standard fusion-based cross-attention produces markedly smoother feature maps, with feature responses tending toward consistency across time steps. This indicates that direct feature fusion weakens discriminative information at the temporal step level and suppresses fine-grained temporal variations.

Further adding LER on top of SGS yields additional improvements across datasets. As a regularization mechanism, LER stabilizes training dynamics and encourages more targeted temporal selection. Overall, these results indicate that semantic information is more effective than the direct feature

Method Variant	ETTh2		Exchange	
Metrics	MSE	MAE	MSE	MAE
Backbone (w/o AAS)	0.345	0.392	0.428	0.456
+ Fusion-based attention	0.356	0.398	0.454	0.462
+ SGS	0.339	0.390	0.396	0.421
+ SGS + LER	0.333	0.384	0.379	0.366

Table 3: Ablation study of different key modules on the ETTh2 and Exchange datasets. MSE and MAE are averaged over four prediction horizons (96, 192, 336, 720).

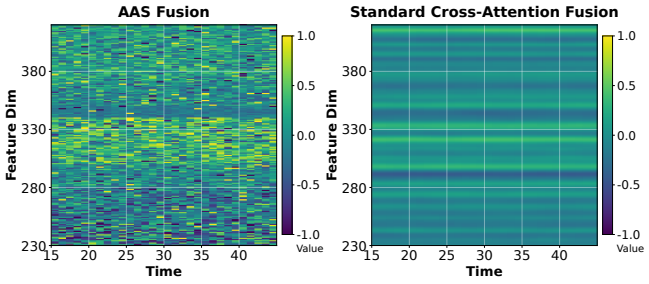

Figure 3: Impact of different fusion strategies on temporal feature representations on the ETTh2 dataset.

fusion in guiding temporal selection, and the combination of SGS and LER achieves the best overall performance.

4.5 Hyperparameter Sensitivity

Look-back Window. The impact of the look-back window size is further analyzed under different forecasting horizons, with ETTh2 used as a representative dataset. As illustrated in Figure 4, increasing the look-back window consistently improves performance across most prediction lengths, indicating that a richer historical context is beneficial for modeling long-range temporal dependencies. However, the best results are achieved when the window size is set to 512, while further extending the window leads to marginal or negative performance gains. This trend suggests that although longer input sequences provide additional temporal information, excessively large windows may introduce redundant or irrelevant history, especially under long-horizon forecasting settings. Overall, the results demonstrate that a moderately large look-back window strikes a favorable balance between contextual richness and effective temporal modeling.

Entropy Control Parameter. This section analyzes the sensitivity of the proposed LER with respect to the entropy control parameter α in the KL-based attention prior. α is varied over $1, 2, 4, 8$ while all other training configurations are fixed. Figure 5 reports forecasting performance on the ETTh2 and Weather datasets in terms of MSE and MAE under different values of α. Across all datasets, the model consistently achieves better performance when α is set to a moderate value (e.g., $\alpha = 4$), indicating that a fixed moderate setting provides a reliable inductive bias without requiring dataset-specific tuning. This setting encourages the attribution distribution to become sufficiently concentrated, while still preserving discriminative temporal patterns. When α is too small, the sharpened target distribution degenerates into the original attribution distribution, resulting in a weaker regularization signal and diminished effectiveness of semantic-guided selection. On the other hand, excessively large values of α impose overly sharp target distributions, which can overly constrain the selector and reduce its flexibility in capturing diverse temporal dependencies. Overall, the performance variation across different values of α remains relatively limited, indicating that the proposed regularization strategy is robust to the choice of α within a reasonable range.

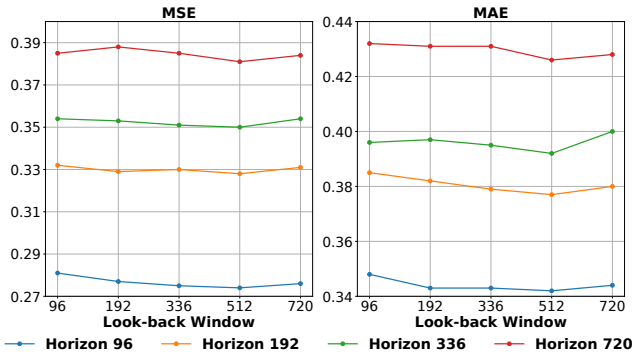

Figure 4: Forecasting performance of AAS with varying look-back windows on the ETTh2 dataset.

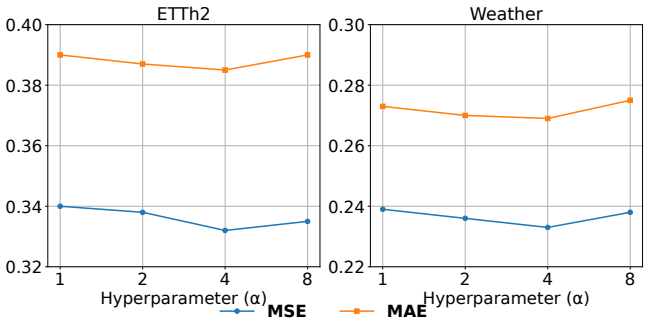

Figure 5: Forecasting performance of AAS with varying entropy control parameters on the ETTh2 and Weather datasets. MSE and MAE are averaged over four prediction horizons.

5 Conclusion

In this work, we propose AAS, an attention-as-selection framework for semantic-guided time series forecasting that explicitly redefines cross-attention as a temporal selection mechanism. Through sparse attribution-based modulation and low-entropy regularization, AAS enables controlled semantic guidance while preserving intrinsic temporal structure. This design mitigates modality interference and maintains fine-grained temporal dynamics. Extensive experiments demonstrate strong performance in both few-shot and long-term forecasting scenarios. These results indicate that treating attention as selection provides a principled and effective paradigm for integrating LLM semantics into time series prediction. Beyond predictive accuracy, AAS further improves interpretability by explicitly revealing how semantic information modulates temporal attribution and selection patterns.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62372153, No.62572168) and the Anhui Provincial Natural Science Foundation (No.2308085MF216, No.2508085MF151).

References

[Cao *et al.*, 2024] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. TEMPO: Prompt-based generative pre-trained transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

[Cen *et al.*, 2024] Jinglun Cen, Chunmei Qing, Haochun Ou, Xiangmin Xu, and Junpeng Tan. Masanet: Multi-aspect semantic auxiliary network for visual sentiment analysis. *IEEE Transactions on Affective Computing*, 15:1439–1450, 2024.

[Chang *et al.*, 2025] Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–20, 2025.

[Ding *et al.*, 2025] Liwei Ding, Kowei Shih, Hairu Wen, Xinshi Li, and Qin Yang. Cross-attention transformer-based visual-language fusion for multimodal image analysis. *International Journal of Applied Science*, 2025.

[Hu *et al.*, 2025] R. Hu, Jizheng Yi, Lijiang Chen, and Ze Jin. Graph reconstruction attention fusion network for multimodal sentiment analysis. *IEEE Transactions on Industrial Informatics*, 21:297–306, 2025.

[Iftikhar *et al.*, 2024] Hasnain Iftikhar, Salvatore Mancha Gonzales, Justyna Zywiołek, and J. L. López-Gonzales. Electricity demand forecasting using a novel time series ensemble technique. *IEEE Access*, 12:88963–88975, 2024.

[Jackson, 2021] Rebecca L Jackson. The neural correlates of semantic control revisited. *NeuroImage*, 224:117444, 2021.

[Jin *et al.*, 2024] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[Kynigakis and Panopoulou, 2025] Iason Kynigakis and E. Panopoulou. Modeling the distribution of key economic indicators in a data-rich environment: new empirical evidence. *Journal of the Operational Research Society*, 76:2071 – 2090, 2025.

[Lee *et al.*, 2025] Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18082–18090, 2025.

[Liu *et al.*, 2024a] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. 2024:11116–11140, 2024.

[Liu *et al.*, 2024b] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Autotimes: Autoregressive time series forecasters via large language models. *Advances in Neural Information Processing Systems*, 37:122154–122184, 2024.

[Liu *et al.*, 2025a] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *AAAI*, 2025.

[Liu *et al.*, 2025b] Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18915–18923, 2025.

[Liu *et al.*, 2025c] Shipeng Liu, Zhonglin Zhang, Deng feng Chen, and Liang Zhao. Describe-to-score: Text-guided efficient image complexity assessment. *ArXiv*, abs/2509.16609, 2025.

[Luo *et al.*, 2024] Yuanyi Luo, Rui Wu, Jiafeng Liu, and Xianglong Tang. Balanced sentimental information via multimodal interaction model. *Multimedia Systems*, 30, 2024.

[Nie *et al.*, 2022] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[Pan *et al.*, 2024] Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. $\acute{s}$2 ip-llm: Semantic space informed prompt learning with llm for time series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

[Pan *et al.*, 2025] Qihong Pan, Haofei Tan, Guojiang Shen, Xiangjie Kong, Mengmeng Wang, and Chenyang Xu. Llm-tpf: Multiscale temporal periodicity-semantic fusion llms for time series forecasting. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6030–6038, 2025.

[Rehman *et al.*, 2025] Mohammad Zia Ur Rehman, Devraj Raghuvanshi, Umang Jain, Shubhi Bansal, and Nagendra Kumar. A multimodal-multitask framework with cross-modal relation and hierarchical interactive attention for semantic comprehension. *Information Fusion*, page 103628, 2025.

[Sun *et al.*, 2025] Siming Sun, Kai Zhang, Xuejun Jiang, Wenchao Meng, and Qinmin Yang. Enhancing llms for time series forecasting via structure-guided cross-modal alignment. *ArXiv*, abs/2505.13175, 2025.

[Tao *et al.*, 2025] Xiaoyu Tao, Shilong Zhang, Mingyue Cheng, Daoyu Wang, Tingyue Pan, Bokai Pan, Changqing

Zhang, and Shijin Wang. From values to tokens: An llm-driven framework for context-aware time series forecasting via symbolic discretization. 2025.

[Vadurin *et al.*, 2025] K. Vadurin, Andrii Perekrest, V. Bakharev, V. Shendryk, Yuliia Parfenenko, and S. Shendryk. Towards digitalization for air pollution detection: Forecasting information system of the environmental monitoring. *Sustainability*, 2025.

[Wang *et al.*, 2024] Xixi Wang, Xiao Wang, Bo Jiang, Jin Tang, and Bin Luo. Mutualformer: Multi-modal representation learning via cross-diffusion attention. *International Journal of Computer Vision*, 132(9):3867–3888, 2024.

[Wang *et al.*, 2025a] Lei Wang, Keyao Dong, and Xiaoyong Zhao. A novel llm time series forecasting method based on integer-decimal decomposition. *Scientific Reports*, 15(1):23004, 2025.

[Wang *et al.*, 2025b] Xue Wang, Jiatong Wu, Pengfei Zhang, and Zhongjun Yu. Language-driven cross-attention for visible–infrared image fusion using clip. *Sensors (Basel, Switzerland)*, 25, 2025.

[Wu *et al.*, 2023] 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.

[Xing *et al.*, 2024] Ling Xing, Hongyu Qu, Rui Yan, Xiangbo Shu, and Jinhui Tang. Locality-aware cross-modal correspondence learning for dense audio-visual events localization. *ArXiv*, abs/2409.07967, 2024.

[Xue and Salim, 2023] Hao Xue and Flora D Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6851–6864, 2023.

[Yang *et al.*, 2025] Zhiqiang Yang, Mengxiao Yin, Junjie Liao, Fancui Xie, Peizhao Zheng, Jiachao Li, and Bei Hua. Fftnet: Fusing frequency and temporal awareness in long-term time series forecasting. *Electronics*, 2025.

[Yang, 2025] Janghoon Yang. Context information can be more important than reasoning for time series forecasting with a large language model. *2025 22nd International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 1–6, 2025.

[Zeng *et al.*, 2023] 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.

[Zhang *et al.*, 2025] Shunxiang Zhang, Jiajia Liu, Yixuan Jiao, Yulei Zhang, Lei Chen, and Kuan Ching Li. A multimodal semantic fusion network with cross-modal alignment for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21:1 – 22, 2025.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[Zhou *et al.*, 2025] Shiqiao Zhou, Holger Schöner, Huanbo Lyu, Edouard Fouché, and Shuo Wang. Balm-tsf: Balanced multimodal alignment for llm-based time series forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4498–4508, 2025.