QIYAO-M: MULTIMODAL TIME SERIES FOUNDATION MODEL WITH ROLE-AWARE MODELING OF ENDOGENOUS AND EXOGENOUS MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing multimodal time series foundation models (TSFMs) typically model heterogeneous modalities through largely shared mechanisms, overlooking the distinct forecasting roles of endogenous and exogenous modalities. In this work, we propose **QiYao-M**, a role-aware multimodal TSFM that models the two types of modalities separately. For endogenous modalities, to capture how they evolve along with the underlying temporal dynamics, we introduce an *Endo-Multimodal Predictor* and *Endo-Multimodal Supervision* to explicitly learn their evolution from history to the future. For exogenous modalities, to generalize across domains and across various modality types and numbers under the scarcity of exo-multimodal pretraining data, we propose an *Exo-Multimodal Retrieval Enhancer* that enables rapid downstream adaptation without updating the TSFM parameters. We further introduce *Endo-Modality Proxy Training* to train this retrieval module without exogenous multimodal pretraining data. Extensive experiments across unimodal and multimodal benchmarks demonstrate strong forecasting performance in scenarios both with and without exogenous modalities.

1 INTRODUCTION

Time series forecasting is critical to real-world applications (Sezer et al., 2020; Qiu et al., 2024; Wu et al., 2025). Recent time series foundation models (TSFMs) have demonstrated strong generalization through pretraining on large-scale numerical time series corpora, learning transferable temporal patterns (Ansari et al., 2025; Fu et al., 2026; Podest et al., 2026), as illustrated in Figure 1(a). Building on this progress, recent studies have developed multimodal TSFMs that are pretrained on multimodal data and designed to jointly model time series with information such as text and images (Wu et al., 2026; Chen et al., 2025; Ahamed et al., 2026). These advances have opened up a broader paradigm of multimodal TSFMs for exploiting heterogeneous information in forecasting.

Within this emerging paradigm, prior multimodal time series studies have distinguished multimodal information by its origin into two categories (Wang et al., 2026; Razmadze & Shavit, 2026): 1) **Endogenous modalities**, which are derived from the time series itself and provide alternative views of the same underlying temporal dynamics; and 2) **Exogenous modalities**, which originate from external sources and provide additional evidence that may influence future evolution. Despite both being forms of multimodal information, their distinct information origins suggest fundamentally different roles in forecasting.

Existing multimodal TSFMs incorporate both endogenous and exogenous modalities through largely shared multimodal mechanisms (Wu et al., 2026; Ahamed et al., 2026; Wang et al., 2025a), as illustrated in Figure 1(b). However, such treatment overlooks a fundamental asymmetry in how endogenous and exogenous modalities relate to forecasting: endogenous modalities are *intrinsically coupled with the time series and evolve with its temporal dynamics*, whereas exogenous modalities face *limited multimodal pretraining data and diverse modalities in both type and number across forecasting scenarios*. Consequently, ignoring this asymmetry in role-agnostic modeling may hinder the learning of temporal evolution and the generalization across exogenous modality configurations. To bridge this gap, we aim to explicitly model endogenous and exogenous multimodality with role-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

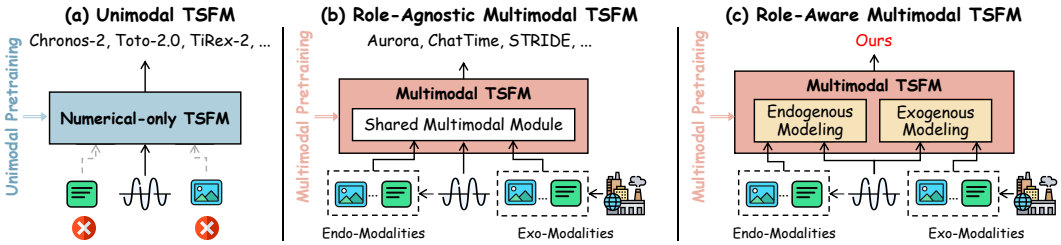


Figure 1: Illustration of different modeling paradigms for TSFMs. (a) Unimodal TSFMs on numerical time series. (b) Role-agnostic multimodal TSFMs model endogenous (endo-) and exogenous (exo-) modalities through largely shared mechanisms. (c) Our role-aware multimodal TSFM separately models endo- and exo- modalities according to their distinct characteristics.

aware strategies that better realize their distinct forecasting roles, as illustrated in Figure 1(c). This presents two key challenges:

For endogenous modalities, it remains challenging to capture how endogenous modalities evolve along with the underlying temporal dynamics. Capturing the evolution of endogenous modalities helps characterize the underlying temporal dynamics from multiple perspectives. However, numerical forecasting losses only evaluate the final forecast values, providing no modality-specific supervision for how endogenous information should evolve toward the future. Existing multimodal TSFMs (Wang et al., 2025a; Wu et al., 2026) primarily use endogenous modalities to enrich historical context, without explicitly learning how these modalities evolve over time.

For exogenous modalities, limited multimodal pretraining data makes it challenging to generalize across domains and across various modality types and numbers. First, the same exo-modality may exhibit different forecasting effects across domains. For example, the same convective weather condition captured by satellite imagery may indicate increasing precipitation in weather forecasting (Park et al., 2025), while the same condition can reduce road traffic flow in traffic forecasting (Jia et al., 2017). Second, forecasting scenarios may differ substantially in the types and numbers of available exo-modalities. Such cross-scenario variations pose substantial challenges to model generalization. This challenge is further exacerbated by the scarcity of exo-multimodal pretraining data, making it difficult to learn transferable exogenous knowledge across diverse scenarios.

To address these challenges, we present **QiYao-M**, which adopts role-aware strategies to model endo- and exo-multimodality differently according to their distinct forecasting characteristics. **For endogenous modalities**, we introduce *Endo-Multimodal Predictor* and *Endo-Multimodal Supervision* to explicitly learn how endogenous modalities evolve along with the underlying time series during pretraining. Specifically, QiYao-M performs patch-wise fusion of endogenous modalities within the historical window. In the forecast window, instead of predicting only numerical values, the *Endo-Multimodal Predictor* explicitly predicts each endo-modality at the patch level. These predictions are further guided by *Endo-Multimodal Supervision*, which aligns the predicted endogenous representations with those derived from the ground-truth future sequence, thereby explicitly supervising their evolution from history to the future. **For exogenous modalities**, rather than relying on pretraining to acquire such generalization capabilities, we design an exo-multimodal retrieval mechanism that enables rapid downstream adaptation to unseen domains, unseen modality types, and various numbers of modalities without updating the TSFM parameters. To this end, we propose an *Exo-Multimodal Retrieval Enhancer* module and further design a dedicated training strategy, termed *Endo-Modality Proxy Training*, which trains the retrieval module using diverse endo-modalities as proxies during pretraining. Specifically, the *Exo-Multimodal Retrieval Enhancer* independently retrieves historical cases for each input exo-modality and uses their corresponding future responses as forecasting evidence, enabling adaptation to downstream tasks with various types and numbers of exo-modalities. The *Endo-Modality Proxy Training* strategy dynamically samples combinations of readily available endo-modalities to substitute for scarce exo-modalities during pretraining.

Our main contributions are summarized as follows:

- We introduce a role-aware multimodal TSFM, named QiYao-M, which treats endogenous and exogenous modalities with distinct strategies according to their different forecasting characteristics, thereby enabling more effective multimodal forecasting.

- For endogenous modalities, we introduce the *Endo-Multimodal Predictor* and *Endo-Multimodal Supervision* to explicitly model how endogenous modalities evolve along with the underlying time series.
- For exogenous modalities, we propose an *Exo-Multimodal Retrieval Enhancer* module supported by an *Endo-Modality Proxy Training* strategy, enabling the model to generalize across unseen domains and modality configurations without updating the TSFM parameters.
- Extensive experiments demonstrate that QiYao-M achieves strong performance across forecasting tasks with and without exogenous modalities.

2 Related work

2.1 Multimodal Time Series Forecasting

Recent studies have increasingly explored multimodal information for time series forecasting. One line of work bridges temporal and language representations by adapting pretrained language models to time series, including GPT4TS (Zhou et al., 2023), TEST (Sun et al., 2024), Time-LLM (Jin et al., 2024), CALF (Liu et al., 2025a), LLM-Mixer (Kowsher et al., 2025), and CC-Time (Chen et al., 2025), through reprogramming, representation alignment, or cross-modal interaction. Another line explicitly incorporates auxiliary textual information into forecasting. CMIN (Luo et al., 2023) and Modality-aware Transformer (Emami Gohari et al., 2024) jointly model financial sequences with news or textual reports, while GPT4MTS (Jia et al., 2024), Time-MMD (Liu et al., 2024b), TaTS (Li et al., 2026), and VoT (Wang et al., 2026) further explore the collection, alignment, fusion, or reasoning of contextual text with time series. Beyond text, Time-VLM (Zhong et al., 2025) introduces visual representations to complement temporal and textual information. Despite this progress, most existing approaches learn multimodal forecasting through task- or dataset-specific adaptation, rather than large-scale multimodal pretraining across diverse time series domains, limiting the transfer of multimodal knowledge to unseen domains.

2.2 Time Series Foundation Models

Time series foundation models (TSFMs) leverage large-scale pretraining to learn transferable temporal patterns and enable zero-shot generalization across domains. Representative models, including UniTS (Gao et al., 2024), TimesFM (Das et al., 2024), LightGTS (Wang et al., 2025b), PatchTST-FM (Wen et al., 2026), Sundial (Liu et al., 2025b), and ROSE (Wang et al., 2025c), explore diverse architectures and pretraining objectives for general-purpose forecasting. More recent models further broaden their capabilities: Toto 2.0 (Khwaja et al., 2026) investigates large-scale model scaling, Chronos-2 (Ansari et al., 2025) supports multivariate and covariate-informed forecasting, TiRex-2 (Podest et al., 2026) enables efficient recurrent and streaming forecasting, and ZEUS (Fu et al., 2026) extends pretraining toward multiple time series tasks. Recently, ChatTime (Wang et al., 2025a), STRIDE (Ahamed et al., 2026) and Aurora (Wu et al., 2026) further introduce multimodal pretraining and modeling into TSFMs. However, these models primarily support exogenous textual information and struggle to generalize to scenarios with various types and numbers of exo-modalities. More importantly, existing multimodal TSFMs typically adopt role-agnostic modeling, using shared mechanisms across heterogeneous modalities without explicitly modeling endogenous and exogenous modalities according to their distinct forecasting characteristics.

3 QiYao-M

In this work, we develop a role-aware multimodal TSFM, named QiYao-M. As shown in Figure 2, we design dedicated multimodal modeling mechanisms and training strategies for endo- and exo-modalities according to their distinct forecasting characteristics.

In terms of model architecture, after normalization and patching, the input time series is processed through three main modules: 1) *Endo-Multimodal Fusion* constructs patch-level endo-modalities and fuses them within each patch to capture fine-grained temporal dynamics; 2) After the Transformer backbone, we introduce an optional *Exo-Multimodal Retrieval Enhancer*, activated only

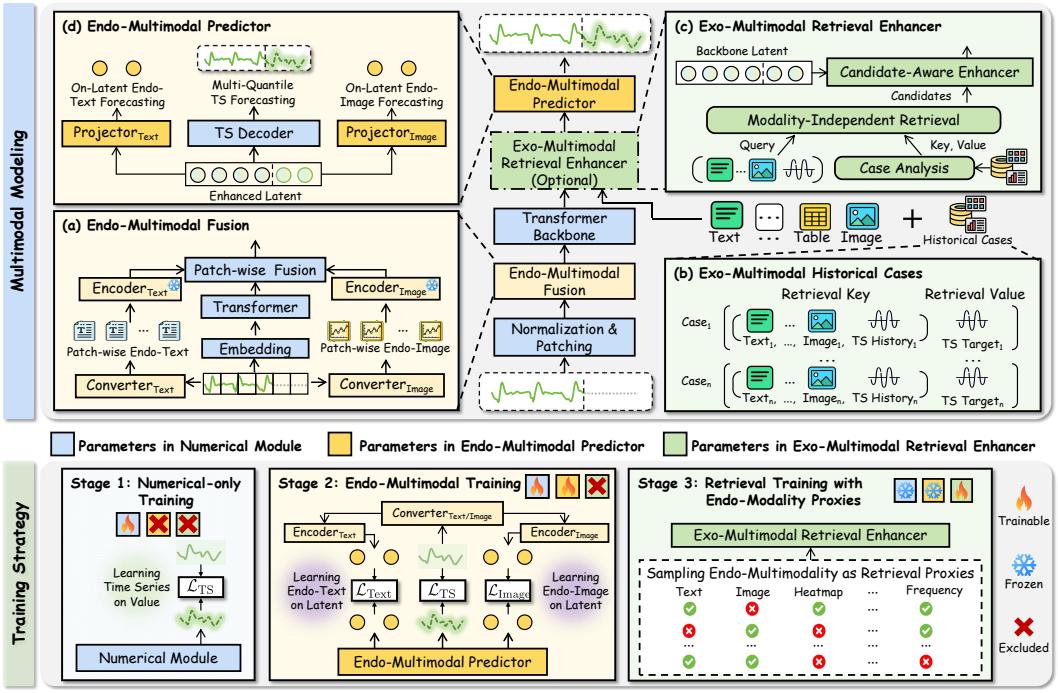


Figure 2: The Multimodal Modeling and Training Strategy of QiYao-M.

when exo-modalities are available, which retrieves historical cases with similar exogenous conditions and uses their future responses as forecasting evidence; 3) *Endo-Multimodal Predictor* predicts both time series and endo-modalities for each future patch, providing explicit training signals for learning their temporal evolution from history to the future. The details are presented in Section 3.1.

We further design a three-stage training strategy to fully exploit the capabilities of each module and alleviate the scarcity of exo-multimodal data, as detailed in Section 3.2.

3.1 Multimodal Modeling

Our method explicitly distinguishes endogenous and exogenous multimodal information and models them separately through dedicated modules to achieve role-aware multimodal modeling. Following the channel-independence strategy (Nie et al., 2023), we aim to predict the future H time steps at Q quantile levels, $\hat{\boldsymbol{Y}} \in \mathbb{R}^{H\times Q}$, from the historical sequence $\boldsymbol{X} \in \mathbb{R}^{L}$ of length L, optionally augmented with available exogenous multimodal information.

3.1.1 Endo-Multimodal Fusion

To better capture the temporal dynamics of the underlying system from multiple perspectives, we construct endogenous multimodal information for each patch. To balance efficiency and expressiveness, we use statistical-feature descriptions and line plots of multi-order differences as the endogenous modalities, with their construction details provided in Appendix A.2.

Temporal Encoding. Following existing TSFMs (Khwaja et al., 2026; Ansari et al., 2025), we zero-pad the historical sequence to the forecast horizon, resulting in $\tilde{\boldsymbol{X}} \in \mathbb{R}^{L+H}$ and augment it with a normalized time index $\boldsymbol{J} \in \mathbb{R}^{L+H}$ and an observation mask $\boldsymbol{M} \in \mathbb{R}^{L+H}$. The resulting sequence is patchified and encoded as follows:

$$\boldsymbol{H}_0 = \mathrm{MLP}\Big(\mathrm{Patch}\Big([\tilde{\boldsymbol{X}};\boldsymbol{J};\boldsymbol{M}]\Big)\Big) \in \mathbb{R}^{N_{L+F}\times d}, \quad \boldsymbol{H}^{\mathrm{TS}} = \mathrm{Transformer}(\boldsymbol{H}_0) \in \mathbb{R}^{N_{L+F}\times d}, \tag{1}$$

where $\tilde{\boldsymbol{X}} \in \mathbb{R}^{L+H}$ is the zero-padded sequence, $\boldsymbol{J} = \left[-\frac{L}{C}, -\frac{L-1}{C}, \ldots, 0, \ldots, \frac{H-1}{C}\right]$ denotes the normalized time index, and $\boldsymbol{M} = [\mathbf{1}^L, \mathbf{0}^H]$ distinguishes observed and forecast positions. N_{L+F} denote the numbers of patches in the historical and forecast windows.

Endo-Multimodal Construction & Encoding. For each patch, we construct statistical feature descriptions and multi-order difference plots as the Endo-Text and Endo-Image, respectively. A frozen pretrained CLIP (Radford et al., 2021) encodes these modalities, and the resulting token-level features are average-pooled to obtain patch-level representations $\boldsymbol{H}^{\text{Text}}, \boldsymbol{H}^{\text{Image}} \in \mathbb{R}^{N \times d_{\text{Clip}}}$.

Patch-wise Fusion. To bridge the representation-space gap between the pretrained CLIP encoder and the temporal module, we employ multi-stage MLPs to fuse the time-series $\boldsymbol{H}^{\text{TS}}$, text $\boldsymbol{H}^{\text{Text}}$, and image $\boldsymbol{H}^{\text{Image}}$ representations at the patch level, which is formulated as follows:

$$\boldsymbol{H}_1 = \text{MLP}_1\left(\text{Concat}\left(\boldsymbol{H}^{\text{TS}}, \boldsymbol{H}^{\text{Text}}\right)\right), \quad \boldsymbol{H}_2 = \text{MLP}_2\left(\text{Concat}\left(\boldsymbol{H}^{\text{TS}}, \boldsymbol{H}^{\text{Image}}\right)\right), \tag{2}$$

$$\boldsymbol{H} = \text{MLP}_3\left(\boldsymbol{H}_1 + \boldsymbol{H}_2 + \boldsymbol{H}^{\text{TS}}\right), \tag{3}$$

where $\text{MLP}_1 := \mathbb{R}^{d+d_{\text{Clip}}} \to \mathbb{R}^d$, $\text{MLP}_2 := \mathbb{R}^{d+d_{\text{Clip}}} \to \mathbb{R}^d$, and $\text{MLP}_3 := \mathbb{R}^d \to \mathbb{R}^d$ denote learnable multilayer perceptrons operating independently on each patch. $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ fuse the time-series representation with the text and image representations, respectively, and are further aggregated by MLP_3 to obtain the final patch-level representation $\boldsymbol{H}$.

3.1.2 Exo-Multimodal Retrieval Enhancer

We introduce this module to accommodate various downstream exogenous modality types and numbers through retrieval from historical cases without updating TSFM parameters. The module is pretrained with an *Endo-Modality Proxy Training* strategy, which dynamically samples combinations of available endo-modalities as retrieval proxies.

To reduce the noise introduced by irrelevant modalities, we first design an *Exo-Multimodal Cases Analysis* module (Figure 3(a)), which analyzes the provided historical cases before retrieval and identifies the forecasting contribution of each modality.

To handle various modality types and numbers, we propose *Modality-Independent Retrieval* (Figure 3(b)), where each modality independently retrieves top-K historical cases as candidates, which are then merged and re-ranked by similarity scores and modality forecasting contributions. The selected candidates are incorporated into the forecasting representation through the *Candidate-Aware Enhancer*.

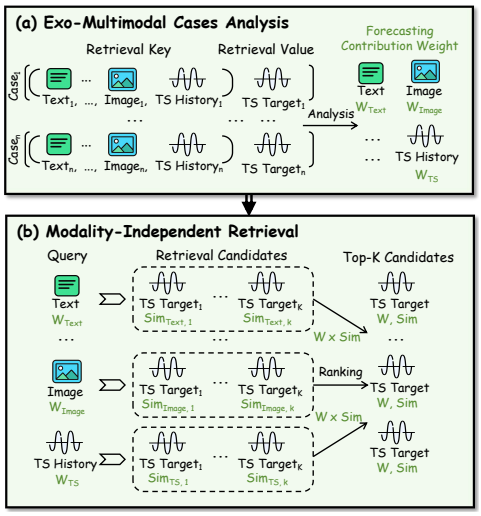


Figure 3: Key steps of the *Exo-Multimodal Retrieval Enhancer*.

Exo-Multimodal Cases Analysis. Given N_R historical cases $\boldsymbol{R} = \{\boldsymbol{r}_i\}_{i=1}^{N_R}$, each case is represented as $\boldsymbol{r}_i = (\boldsymbol{H}_i^{\text{his}}, \{\boldsymbol{E}_i^m\}_{m\in\mathcal{M}}, \boldsymbol{Y}_i^{\text{fut}})$, where $\boldsymbol{H}_i^{\text{his}}$ and $\boldsymbol{Y}_i^{\text{fut}}$ denote its historical time-series representation and future sequence, respectively. $\mathcal{M}$ denotes the set of available exo-modalities with $M = |\mathcal{M}|$, and $\boldsymbol{E}_i^m$ denotes the representation of modality $m \in \mathcal{M}$.

As illustrated in Figure 3(a), we estimate the *Forecasting Contribution Weight* $\boldsymbol{W} = (W^1, \ldots, W^M) \in \mathbb{R}^M$ by evaluating how effectively each modality retrieves historical cases whose future responses are informative for forecasting.

Specifically, for each reference case $\boldsymbol{r}_i$ and modality $m \in \mathcal{M}$, we retrieve the top-K other reference cases according to the modality-specific similarity $s_{ij}^m = s(\boldsymbol{E}_i^m, \boldsymbol{E}_j^m)$ and aggregate their future sequences into a retrieval-based future estimate:

$$\mathcal{N}_i^m = \text{TopK}_{j\neq i}(s_{ij}^m), \quad \hat{\boldsymbol{Y}}_i^m = \sum_{j\in\mathcal{N}_i^m} \text{Softmax}_{j\in\mathcal{N}_i^m}(s_{ij}^m)\boldsymbol{Y}_j^{\text{fut}}, \tag{4}$$

$$e_i^m = \text{MSE}\left(\hat{\boldsymbol{Y}}_i^m, \boldsymbol{Y}_i^{\text{fut}}\right), \quad w_i^m = \text{Softmax}_{m\in\mathcal{M}}(-e_i^m), \quad W^m = \frac{1}{N_R}\sum_{i=1}^{N_R} w_i^m. \tag{5}$$

Thus, a larger W^m indicates higher retrieval-based forecasting utility of modality m.

Modality-Independent Retrieval. As illustrated in Figure 3(b), given a query q, we perform retrieval independently for each available modality to accommodate various modality types and numbers. Each modality retrieves its top-K reference cases, which are then merged and re-ranked using the *Forecasting Contribution Weight*. Let $s_{qj}^m = s(\boldsymbol{E}_q^m, \boldsymbol{E}_j^m)$ denote the similarity between the query and reference case $\boldsymbol{r}_j$ under modality m:

$$\mathcal{C}_q^m = \mathrm{TopK}_j(s_{qj}^m), \quad \mathcal{U}_q = \bigcup_{m \in \mathcal{M}} \mathcal{C}_q^m, \quad S_{qj} = \sum_{m \in \mathcal{M}} W^m s_{qj}^m, \quad \mathcal{C}_q = \mathrm{TopK}_{j \in \mathcal{U}_q}(S_{qj}), \tag{6}$$

where $\mathcal{C}_q^m$ denotes the top-K reference cases retrieved by modality m, $\mathcal{U}_q$ denotes the union of all modality-specific candidate sets, S_{qj} denotes the contribution-weighted retrieval score of reference case $\boldsymbol{r}_j$ for query q, and $\mathcal{C}_q$ denotes the final top-K reference cases after re-ranking.

Candidate-Aware Enhancer. Given the final candidate set $\mathcal{C}_q$, we incorporate the retrieved information into the backbone representation $\boldsymbol{H}^B$ through cross-attention. For each candidate $\boldsymbol{r}_j \in \mathcal{C}_q$, we encode its representation $\boldsymbol{H}_j^{\mathrm{cand}}$ and construct a candidate identifier $\boldsymbol{I}_{qj}$ from its modality contribution W^m and retrieval similarity s_{qj}^m:

$$\boldsymbol{H}_j^{\mathrm{cand}} = \mathrm{MLP}_{\mathrm{cand}}(\boldsymbol{r}_j), \quad \boldsymbol{I}_{qj} = \mathrm{MLP}_I\left([W^m,\, s_{qj}^m]\right), \quad \boldsymbol{Q} = \mathrm{MLP}_Q(\boldsymbol{H}^B), \tag{7}$$

$$\boldsymbol{K}_j, \boldsymbol{V}_j = \mathrm{MLP}_{KV}\left(\boldsymbol{H}_j^{\mathrm{cand}} + \boldsymbol{I}_{qj}\right), \quad \boldsymbol{H}^{\mathrm{Enh}} = \mathrm{CrossAttn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}), \tag{8}$$

where $\boldsymbol{K}$ and $\boldsymbol{V}$ are obtained by stacking the candidate-wise representations $\boldsymbol{K}_j$ and $\boldsymbol{V}_j$ over all $\boldsymbol{r}_j \in \mathcal{C}_q$, respectively. $\boldsymbol{H}^{\mathrm{Enh}}$ is the enhanced representation of the input time series.

3.1.3 Endo-Multimodal Predictor

To explicitly model the evolution of endogenous multimodal information from history to the future, we jointly predict the future time series together with its Endo-Text and Endo-Image modalities. Rather than directly generating text tokens or image pixels, we predict their representations for computational efficiency and to avoid unnecessary token- and pixel-level generation noise.

After optionally incorporating exogenous information through the *Exo-Multimodal Retrieval Enhancer*, we retain only the forecast-window representations $\boldsymbol{H}_{\mathrm{pred}}^{\mathrm{Enh}} \in \mathbb{R}^{N_f \times d}$ for efficient decoding.

Time Series Decoding. To support uncertainty quantification, we employ an MLP-based multi-quantile prediction head to produce probabilistic forecasts at Q quantile levels:

$$\hat{\boldsymbol{Y}} = \mathrm{QuantileHead}\left(\boldsymbol{H}_{\mathrm{pred}}^{\mathrm{Enh}}\right) \in \mathbb{R}^{H \times Q}. \tag{9}$$

Endo-Multimodal Projector. We further project the forecast-window representations into the CLIP space to predict future Endo-Text representations $\boldsymbol{F}^{\mathrm{Text}}$ and Endo-Image representations $\boldsymbol{F}^{\mathrm{Image}}$:

$$\boldsymbol{F}^{\mathrm{Text}} = \mathrm{Project}_{\mathrm{T}}\left(\boldsymbol{H}_{\mathrm{pred}}^{\mathrm{Enh}}\right) \in \mathbb{R}^{N_f \times d_{\mathrm{Clip}}}, \ \boldsymbol{F}^{\mathrm{Image}} = \mathrm{Project}_{\mathrm{Image}}\left(\boldsymbol{H}_{\mathrm{pred}}^{\mathrm{Enh}}\right) \in \mathbb{R}^{N_f \times d_{\mathrm{Clip}}}. \tag{10}$$

Here, the projectors are MLPs that map the forecast representations to the CLIP space.

3.2 Training Strategy

To progressively equip QiYao-M with numerical forecasting, endogenous multimodal modeling, and exogenous retrieval capabilities, we adopt a three-stage training strategy, as illustrated in Figure 2: 1) *Numerical-only Training* establishes the basic temporal modeling and forecasting capability; 2) *Endo-Multimodal Training* further learns endogenous multimodal representations and their evolution from history to the future; and 3) *Retrieval Training with Endo-Modality Proxies* trains the *Exo-Multimodal Retrieval Enhancer* using readily available endogenous modalities as retrieval proxies, avoiding reliance on large-scale exo-multimodal pretraining data.

3.2.1 Numerical-only Training

In this stage, we train only the numerical forecasting pathway, including the temporal encoder, *Transformer backbone*, and multi-quantile prediction head, while all multimodal modules are excluded. Given the predicted quantile forecasts $\hat{\boldsymbol{Y}} \in \mathbb{R}^{H \times Q}$ and the ground-truth future sequence

$\boldsymbol{Y} \in \mathbb{R}^H$, we optimize these parameters using the Pinball loss (Ansari et al., 2025):

$$\mathcal{L}_{\text{TS}} = \frac{1}{HQ} \sum_{h=1}^{H} \sum_{q=1}^{Q} \max \left(\tau_q \left(Y_h - \hat{Y}_{h,q} \right), (\tau_q - 1) \left(Y_h - \hat{Y}_{h,q} \right) \right), \tag{11}$$

where $\tau_q \in (0, 1)$ denotes the q-th quantile level.

3.2.2 Endo-Multimodal Training

In this stage, we optimize the numerical pathway and endogenous multimodal modules, while excluding the *Exo-Multimodal Retrieval Enhancer* and freezing the pretrained CLIP encoder.

To provide explicit supervision for endogenous evolution, we transform the ground-truth future sequence $\boldsymbol{Y}$ into Endo-Text and Endo-Image modalities using the same procedure as in Section 3.1.1, and encode them with the frozen CLIP encoder to obtain future representation targets $\boldsymbol{T}^{\text{Text}}, \boldsymbol{T}^{\text{Image}} \in \mathbb{R}^{N_f \times d_{\text{Cli'P}}}$. The representation-level supervision is defined as follows:

$$\mathcal{L}_{\text{Text}} = \frac{\left\| \boldsymbol{F}^{\text{Text}} - \boldsymbol{T}^{\text{Text}} \right\|_F^2}{N_f d_{\text{Clip}}}, \qquad \mathcal{L}_{\text{Image}} = \frac{\left\| \boldsymbol{F}^{\text{Image}} - \boldsymbol{T}^{\text{Image}} \right\|_F^2}{N_f d_{\text{Clip}}}, \tag{12}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The overall objective is $\mathcal{L}_{\text{Multi}} = \mathcal{L}_{\text{TS}} + \lambda_{\text{Text}} \mathcal{L}_{\text{Text}} + \lambda_{\text{Image}} \mathcal{L}_{\text{Image}}$, where λ_{Text} and λ_{Image} balance the two endogenous supervision terms.

3.2.3 Retrieval Training with Endo-Modality Proxies

In this stage, we optimize only the *Exo-Multimodal Retrieval Enhancer* and forecasting head. We introduce *Endo-Modality Proxy Training*, which dynamically samples various types and numbers of endo-modalities as retrieval proxies, avoiding the need for exo-multimodal pretraining data.

Let $\mathcal{M}_{\text{endo}}$ denote the set of available endogenous modalities. For each training query q, we randomly sample a non-empty subset $\mathcal{S}_q$ and treat the selected modalities as retrieval proxies.

The sampled proxies are then processed by the same retrieval pipeline described in Section 3.1.2, to obtain the final enhanced latent $\boldsymbol{H}^{\text{Enh}}_{\text{proxy}}$. The forecasting head then generates the prediction $\hat{\boldsymbol{Y}}^{\text{proxy}}$, and only the parameters of the retrieval enhancer and head are optimized using the forecasting loss:

$$\mathcal{L}_{\text{proxy}} = \frac{1}{HQ} \sum_{h=1}^{H} \sum_{q=1}^{Q} \max \left(\tau_q \left(Y_h - \hat{Y}^{\text{proxy}}_{h,q} \right), (\tau_q - 1) \left(Y_h - \hat{Y}^{\text{proxy}}_{h,q} \right) \right). \tag{13}$$

The endogenous proxies are not intended to mimic the semantic content of exogenous modalities; instead, they expose the retrieval enhancer to various modality types, numbers, and combinations while preserving the same retrieval-to-response learning process used at inference time.

4 Experiments

We conduct extensive experiments to evaluate QiYao-M. Detailed experimental settings are provided in Appendix A, including the pretraining corpus (Appendix A.1), benchmarks (Appendix A.3), baselines (Appendix A.4), and other implementation details. Sections 4.1 and 4.2 evaluate forecasting performance on unimodal benchmarks (GIFT-Eval (Aksu et al., 2024) and TIME (Qiao et al., 2026)) and multimodal benchmarks with 2–4 modalities (Time-MMD (Liu et al., 2024b), MoTime (Zhou et al., 2025), and FinMultiTime (Xu et al., 2025)), respectively. Section 4.3 provides further model analyses.

As an overview, Figure 4 shows that QiYao-M achieves state-of-the-art MASE on unimodal benchmarks and MSE on multimodal benchmarks, demonstrating strong performance in scenarios both with and without exo-modalities.

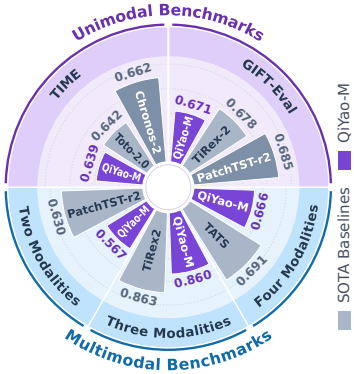
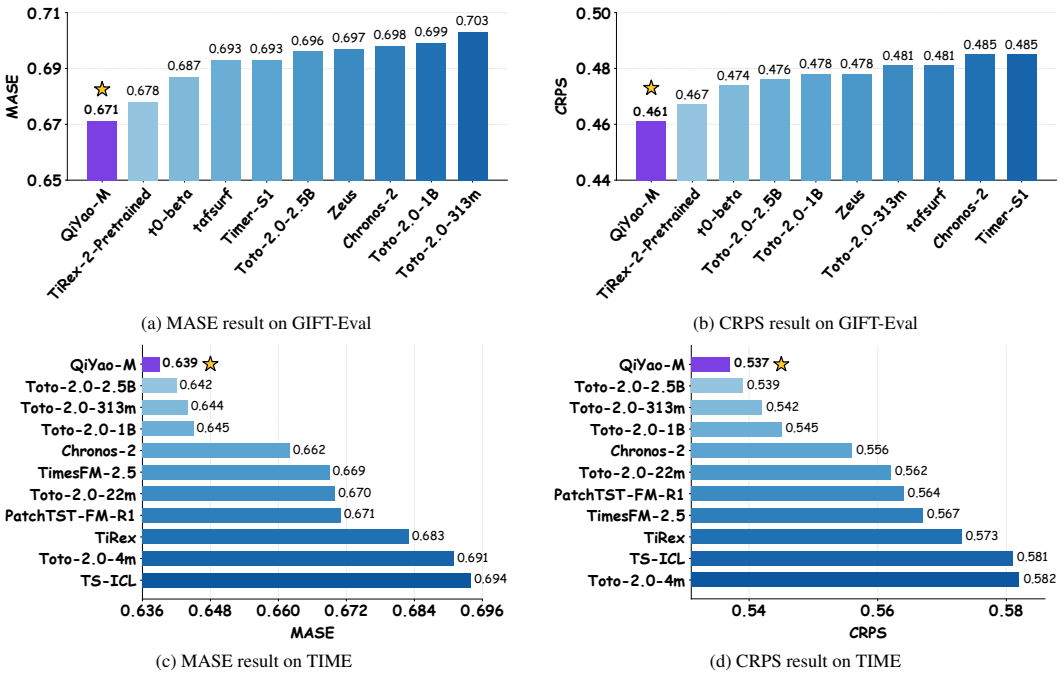


Figure 4: Evaluation summary.

4.1 Performance on Unimodal Benchmarks

To evaluate QiYao-M with endo-modalities only, we conduct experiments on the unimodal benchmarks GIFT-Eval and TIME. As shown in Figure 5, QiYao-M achieves the best MASE and CRPS on both benchmarks. On GIFT-Eval, QiYao-M reduces MASE and CRPS by 1.0% and 1.3%, respectively, compared with the strongest baseline TiRex-2-Pretrained. It also achieves consistently strong MASE and CRPS rankings across datasets (see Appendix C.1), demonstrating robust generalization across diverse forecasting scenarios. On TIME, QiYao-M achieves a MASE of 0.639 and a CRPS of 0.537, outperforming Toto-2.0-2.5B at 0.642 and 0.539. These consistent gains demonstrate the effectiveness of endo-multimodal modeling across different forecasting benchmarks.



(a) MASE result on GIFT-Eval

(b) CRPS result on GIFT-Eval

(c) MASE result on TIME

(d) CRPS result on TIME

Figure 5: Performance of QiYao-M on unimodal benchmarks.

4.2 Performance on Multimodal Benchmarks

To evaluate QiYao-M with both endo- and exo-modalities, we conduct experiments on multimodal datasets with diverse modality combinations from Time-MMD, MoTime, and FinMultiTime. As shown in Table 1, incorporating exo-modalities reduces the average MSE of QiYao-M by 5.1%. Compared with baseline foundation and end-to-end models, QiYao-M ranks first on 14/18 metrics and within the top two on 17/18 metrics. Compared with the strongest unimodal TSFM on each dataset, QiYao-M reduces the MSE by 4.3% on average. Moreover, while existing multimodal TSFMs primarily support exogenous textual information and struggle with various modality types and numbers, QiYao-M effectively handles diverse exo-modalities, achieving an average MSE reduction of 14.6%. These consistent improvements demonstrate the effectiveness of QiYao-M in utilizing exogenous information for forecasting across different domains.

4.3 Model Analysis

Ablation Studies. To analyze the contribution of each component of QiYao-M, we conduct ablation studies on four benchmarks and report MASE or MSE results in Table 2. We have the following observations: 1) Comparing Rows 1 and 2, *Endo-Multimodal Fusion (Endo. Fusion)* consistently improves performance, showing the benefit of incorporating patch-level endo-multimodal information. 2) Comparing Rows 2 and 3, *Endo-Multimodal Predictor* and *Endo-Multimodal Supervision (Endo. Sup.)* further improve performance by explicitly supervising the evolution of endo-modalities. 3) Comparing Rows 3 and 4, *Exo-Multimodal Retrieval Enhancer (Retrieval Enhancer)*

Table 1: Average forecasting results on multimodal datasets. The best and second-best results are highlighted in **purple** and blue, respectively. Full results are listed in Section C.2 of Appendix C.

Modal Type	TS + Text										TS + Text + Image				TS + Text + Image + Table			
Methods	Agriculture		Climate		Energy		Health		Social Good		TAOBAO		Tianchi		HS300		SP500	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models																		
GPT4MTS	0.225	0.298	1.182	0.889	0.262	0.380	1.464	0.799	0.920	0.450	0.534	0.257	1.320	0.147	0.780	0.535	0.767	0.642
CALF	0.250	0.315	1.286	0.922	0.244	0.365	1.491	0.775	0.906	0.401	0.512	0.250	1.320	0.119	0.674	0.502	0.743	0.617
Time-VLM	0.237	0.302	1.195	0.899	0.260	0.374	1.565	0.860	0.868	0.444	0.528	0.271	1.368	0.137	**0.662**	0.499	0.756	0.618
TATS	0.215	0.301	1.180	0.887	0.255	0.368	1.356	0.767	0.918	0.428	0.525	0.263	1.374	0.136	0.717	0.509	0.665	0.581
Unimodal Foundation Models																		
Zeus	0.203	0.295	0.852	0.738	0.239	0.337	1.590	0.845	0.999	0.434	0.425	0.172	1.378	0.083	0.755	0.524	0.858	0.628
Chronos-2	0.238	0.319	0.859	0.730	0.226	0.326	1.054	0.663	0.902	0.384	0.445	0.184	1.453	0.083	0.810	0.501	0.717	0.577
Toto-2	0.236	0.311	0.853	0.729	0.231	0.337	1.109	0.662	0.799	0.309	0.439	0.182	41.618	0.185	0.885	0.507	0.703	0.571
PatchTST-r2	0.235	0.354	0.845	0.724	0.245	0.353	0.989	0.647	0.834	0.366	0.526	0.200	1.411	0.091	0.819	0.521	1.080	0.634
TiRex2	0.255	0.318	0.847	0.724	0.224	0.330	1.097	0.638	0.749	0.325	0.427	**0.171**	**1.299**	**0.074**	0.674	0.499	0.709	0.580
Multimodal Foundation Models																		
ChatTime	0.196	0.293	1.144	0.856	0.258	0.355	2.278	1.001	1.319	0.541	0.479	0.201	1.676	0.087	1.208	0.686	5.985	1.779
Aurora	0.272	0.348	0.865	0.749	0.255	0.370	1.553	0.850	0.838	0.516	0.442	0.210	1.403	0.090	0.736	0.524	0.862	0.649
Ours (w/o Exo.)	0.199	0.288	0.828	0.725	0.237	0.338	0.915	0.612	0.760	0.296	**0.417**	**0.171**	1.423	0.081	0.690	0.496	0.694	0.569
Ours (w Exo.)	**0.174**	**0.284**	**0.827**	**0.723**	**0.211**	**0.325**	**0.901**	**0.608**	**0.722**	**0.293**	**0.417**	0.174	1.302	0.076	0.672	**0.493**	**0.660**	**0.561**

brings further improvements by incorporating exogenous information from historical reference cases. 4) Comparing Rows 4 and 5, *Exo-Multimodal Cases Analysis (Cases Analysis)* further improves performance by estimating the *Forecasting Contribution Weight* of different exo-modalities. Overall, the complete QiYao-M achieves the best performance compared to all the variants across all benchmarks.

Table 2: Ablation study of the key components.

Row	Endo-Modeling		Exo-Modeling		Unimodal Bench.		Multimodal Bench.	
	Endo. Fusion	Endo. Sup.	Retrieval Enhancer	Case Analysis	GIFT-Eval	TIME	FinMulti Time	Time -MMD
1	✗	✗	✗	✗	0.686	0.646	0.741	0.605
2	✓	✗	✗	✗	0.678	0.644	0.728	0.591
3	✓	✓	✗	✗	**0.671**	**0.639**	0.712	0.588
4	✓	✓	✓	✗	–	–	0.672	0.573
5	✓	✓	✓	✓	–	–	**0.666**	**0.567**

Exo-Modeling does not affect unimodal benchmarks; thus, the corresponding entries are marked as "–".

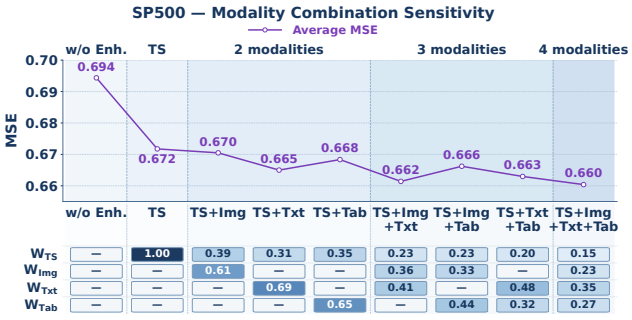


Figure 6: Modality analysis.

Modality Analysis. Figure 6 examines different modality combinations on SP500. We have the following observations: 1) Adding exogenous modalities generally improves forecasting performance, with consistent gains over TS-only retrieval across all tested combinations, suggesting robustness to various modality types and numbers. 2) *The Forecasting Contribution Weights* ($\boldsymbol{W}_{\text{TS}}, \cdots, \boldsymbol{W}_{\text{Table}}$) partially reflect the predictive utility of different modalities: when individually added to TS-only retrieval, text, table, and image receive progressively lower weights, consistent with their decreasing MSE reductions.

More analyses on the sensitivity of the retrieval mechanism for exogenous multimodal information are provided in Appendix B.1.

5 Conclusion

In this work, we present QiYao-M, a role-aware multimodal TSFM that distinguishes endo- and exo-modalities according to their forecasting roles. QiYao-M explicitly models endogenous temporal evolution and uses historical-case retrieval to accommodate diverse exo-modality types and numbers without updating TSFM parameters. Experiments on five benchmarks demonstrate strong cross-domain generalization in scenarios both with and without exo-modalities. Our code is available at `https://anonymous.4open.science/r/QiYao-M-4413/`.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing. All AI-assisted text were carefully reviewed and verified by the authors. The research questions, methodology, experimental design, analysis, and scientific claims were developed independently by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on methodological research for multimodal and covariate-aware time series forecasting. Our experiments use existing benchmark datasets and pretraining corpora, without involving human subjects or the intentional collection of personally identifiable information. The proposed method is intended for general-purpose forecasting research, and we are not aware of any direct ethical risks arising from the methodology itself.

REPRODUCIBILITY STATEMENT

The model architecture and training strategy procedure of QiYao-M are described in Section 3.1 and 3.2, with experimental settings, dataset details, and implementation provided in Appendix A. We report the configurations and evaluation protocols required to reproduce the main experiments. The source code and necessary scripts will be released.

REFERENCES

Md Atik Ahamed, Mihir Parmar, Palash Goyal, Chun-Liang Li, Qiang Cheng, Tomas Pfister, and Jinsung Yoon. Reasoning-aware training for time series forecasting, 2026.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-eval: A benchmark for general time series forecasting model evaluation. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821.

Peng Chen, Yihang Wang, Yang Shu, Yunyao Cheng, Kai Zhao, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. CC-Time: Cross-model and cross-modality time series forecasting, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Hajar Emami Gohari, Xuan-Hong Dang, Syed Yousaf Shah, and Petros Zerfos. Modality-aware transformer for financial time series forecasting. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 677–685, 2024.

Yisong Fu, Zezhi Shao, Chengqing Yu, Yujie Li, Yongjun Xu, Xueqi Cheng, and Fei Wang. Zeus: Towards tuning-free foundation model for time series analysis. In *Forty-third International Conference on Machine Learning*, 2026.

Shanghua Gao, Teddy Koker, Owen Queen, Tom Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. GPT4MTS: Prompt-based large language model for multimodal time-series forecasting. In *AAAI*, volume 38, pp. 23343–23351, 2024.

Yuhan Jia, Jianping Wu, and Ming Xu. Traffic flow prediction with rainfall impact using a deep learning method. *Journal of advanced transportation*, 2017(1):6575947, 2017.

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.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era, 2026.

Md Kowsher, Md. Shohanur Islam Sobuj, Nusrat Jahan Prottasha, E. Alejandro Alanis, Ozlem Garibay, and Niloofar Yousefi. LLM-mixer: Multiscale mixing in LLMs for time series forecasting. In *Proceedings of the 4th Table Representation Learning Workshop*, July 2025.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. In *International Conference on Learning Representations (ICLR)*, 2026.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

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 *AAAI*, volume 39, pp. 18915–18923, 2025a.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025b.

Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng, and Rui Yan. Causality-guided multi-memory interaction network for multivariate stock price movement prediction. In *ACL*, pp. 12164–12176, 2023.

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

Young-Jae Park, Doyi Kim, Minseok Seo, Hae-Gon Jeon, and Yeji Choi. Data-driven precipitation nowcasting using satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. doi: 10.1609/aaai.v39i27.35049.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex-2: Generalizing tirex to multivariate data and streaming, 2026.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's TIME: Towards the next generation of time series forecasting benchmarks. In *ICML*, 2026.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: towards comprehensive and fair benchmarking of time series forecasting methods. *Proc. VLDB Endow.*, 17(9):2363–2377, 2024.

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

Kathy Razmadze and Yoli Shavit. A universal gating framework for multi-expert fusion in heterogeneous multimodal time series forecasting. *Scientific Reports*, 2026. doi: 10.1038/s41598-026-54540-x.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Özbayoglu. Financial time series forecasting with deep learning : A systematic literature review: 2005-2019. *Appl. Soft Comput.*, 90:106181, 2020.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. TEST: text prototype aligned embedding to activate llm's ability for time series. In *International Conference on Learning Representations (ICLR)*, 2024.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12694–12702, 2025a.

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Chenjuan Guo, Bin Yang, and Yang Shu. Unlocking the value of text: Event-driven reasoning and multi-level alignment for time series forecasting. In *ICLR*, 2026.

Yihang Wang, Yuying Qiu, Peng Chen, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Lightgts: A lightweight general time series forecasting model. In *ICML*, 2025b.

Yihang Wang, Yuying Qiu, Peng Chen, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series forecasting model with unified representation and adaptive transfer. In *ICML*, 2025c.

Yunshi Wen, Wesley M. Gifford, Chandra Reddy, Lam M. Nguyen, Jayant Kalagnanam, and Anak Agung Julius. Revisiting the generic transformer: Deconstructing a strong baseline for time series foundation models, 2026. URL https://arxiv.org/abs/2602.06909.

Xingjian Wu, Xiangfei Qiu, Hongfan Gao, Jilin Hu, Bin Yang, and Chenjuan Guo. K^2VAE: A koopman-kalman enhanced variational autoencoder for probabilistic time series forecasting. In *ICML*, 2025.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *ICLR*, 2026.

Wenyan Xu, Dawei Xiang, Yue Liu, Xiyu Wang, Yanxiang Ma, Liang Zhang, Shu Hu, Chang Xu, and Jiaheng Zhang. Finmultitime: A four-modal bilingual dataset for financial time-series analysis, 2025.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *International Conference on Machine Learning (ICML)*, 2025.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One Fits All: Power general time series analysis by pretrained LM. In *NeurIPS*, 2023.

Xin Zhou, Weiqing Wang, Francisco J. Baldán, Wray Buntine, and Christoph Bergmeir. MoTime: A dataset suite for multimodal time series forecasting, 2025.

A Experimental Details

A.1 Pretraining Corpus

We construct the pretraining corpus of QiYao-M from two complementary sources: a large-scale unimodal time series corpus and a multimodal corpus tailored to the three-stage training strategy. The unimodal corpus consists of GIFT-Eval Pretrain, the GIFT-Eval training set (Aksu et al., 2024), and the Chronos corpus, providing broad temporal coverage across heterogeneous domains. Building on this corpus, we further construct 5 million patch-level endogenous multimodal samples and 10 million sequence-level samples for proxy exogenous training, supporting the multimodal objectives at different temporal granularities.

Unimodal Pre-training Data. The unimodal time series corpus combines three large-scale public data sources:

- **GIFT-Eval Pretrain.** We include the pre-training collection released with GIFT-Eval, which contains a diverse set of time series collected from multiple application domains and sampling frequencies. This corpus serves as one of the primary sources of real-world temporal patterns during pre-training.
- **GIFT-Eval Training Set.** In addition to the dedicated pre-training split, we incorporate the training portions of the datasets included in GIFT-Eval (Aksu et al., 2024). Only training data are used for pre-training, while the corresponding evaluation portions are kept separated according to the benchmark protocol to avoid test-set leakage.
- **Chronos Corpus.** We further use the large-scale time-series corpus released for Chronos (Ansari et al., 2024), which aggregates time series from heterogeneous real-world sources. The corresponding evaluation portions are kept separated according to the benchmark protocol to avoid test-set leakage.

Multimodal Pre-training Data. Building on the unimodal time-series corpus, we construct additional multimodal training samples. In total, the multimodal corpus consists of approximately **5 million patch-level endogenous multimodal samples** and **10 million sequence-level multimodal samples** used as proxies for exogenous multimodal observations.

- **Patch-level Endogenous Multimodal Data ($\sim$5M samples).** We construct multimodal observations at the local patch level, where auxiliary textual and visual information is derived from and aligned with localized temporal patterns within a time series. These samples are designed to establish fine-grained correspondence between numerical dynamics and their multimodal representations. Because the auxiliary modalities describe information intrinsic to the observed time series itself, we refer to them as *endogenous multimodal* data. This corpus is primarily used to strengthen cross-modal representation learning and to encourage the model to recover temporal structures from complementary representations of the same underlying signal.
- **Sequence-level Multimodal Data ($\sim$10M samples).** We additionally construct a substantially larger collection of multimodal samples at the sequence level. Instead of describing individual local patches, the associated textual and visual information is paired with the time series at a coarser sequence-level granularity, providing contextual signals that resemble the external information available in downstream multimodal forecasting scenarios. We therefore use these samples as *proxies for exogenous multimodal training*. This design exposes the model to heterogeneous context during pre-training and encourages it to integrate temporal observations with complementary information beyond local numerical patterns.

A.2 Multimodal Dataset Generation Method

We construct endogenous text and image modalities directly from numerical observations using deterministic templates. The resulting modalities therefore provide complementary views of the same numerical evidence. We generate representations at two granularities: (i) *patch-level modalities*, which describe fine-grained local dynamics in each patch, and (ii) *sequence-level modalities*, which summarize the global evolution of a complete input window.

Mask-aware preprocessing. Let a numerical input be $\mathbf{x} = (x_1, \ldots, x_L)$, with an observation mask $\mathbf{m}^{\text{obs}}$ and a padding mask $\mathbf{m}^{\text{pad}}$. We define

$$v_t = m_t^{\text{obs}}(1 - m_t^{\text{pad}}), \tag{14}$$

where $v_t = 1$ indicates an actually observed value. Left padding and naturally missing observations are recorded separately: padding describes positions outside the available context, whereas missingness describes unobserved positions inside the context. Neither is interpreted as an observed zero. For a patch size $P = 32$, the input is divided chronologically into $K = \lceil L/P \rceil$ non-overlapping patches. If the first patch is incomplete, it is left-padded, and its padded positions remain masked throughout text and image generation. During training, historical modalities are generated from historical values and used as model inputs. Modalities generated from ground-truth future patches may only be used as auxiliary prediction targets for the *Endo-Multimodal Predictor*. At inference time, only modalities derived from the available history are constructed; future values or future-generated modalities are never supplied to the model.

A.2.1 Patch-Level Endogenous Modalities

Patch-level text generation. For patch k, let $\mathcal{V}_k = \{r : v_{k,r} = 1\}$ denote its valid positions. We compute a fixed set of mask-aware statistics:

$$\rho_k = \frac{|\mathcal{V}_k|}{P}, \qquad \text{(valid-data coverage)} \tag{15}$$

$$\Delta_k = x_{k,r_{\text{last}}} - x_{k,r_{\text{first}}}, \qquad \text{(net change)} \tag{16}$$

$$\beta_k = \frac{\sum_{r \in \mathcal{V}_k} (r - \bar{r})(x_{k,r} - \bar{x}_k)}{\sum_{r \in \mathcal{V}_k} (r - \bar{r})^2}, \qquad \text{(least-squares trend)} \tag{17}$$

$$R_k = Q_{0.95}(\mathbf{x}_k) - Q_{0.05}(\mathbf{x}_k), \qquad \text{(robust range)} \tag{18}$$

$$V_k = \text{median}_r \left| d_{k,r}^{(1)} - \text{median}(\mathbf{d}_k^{(1)}) \right|, \qquad \text{(local volatility)} \tag{19}$$

where quantiles are computed only over valid values.

The normalized slope, robust range, volatility, lag-one dependence, number of turning points, and largest standardized jump are mapped to predefined linguistic bins. These bins determine the fields `[trend]`, `[shape]`, `[variation]`, `[volatility]`, `[event]`, and `[regime]`. The coverage and the two masks determine `[padding]`, `[missingness]`, and `[coverage]`. All thresholds are fixed before downstream evaluation; when data-dependent thresholds are needed, they are estimated from the training split only.

The categorical fields are inserted into the following deterministic template:

> *"This patch contains [padding] and [missingness], leaving [coverage] valid data coverage. Within the valid observations, the series shows [trend] and follows [shape]. Its local variation is [variation], with [volatility] volatility. The patch contains [event]. Overall, its local behavior is characterized as [regime]."*

This constrained vocabulary makes every statement traceable to a numerical statistic and prevents unsupported domain-specific interpretations.

Figure 7 shows the intended presentation of a patch-level example. The middle line explicitly displays the numerical patch from which the text is generated.

Patch-level three-channel image generation. First- and second-order differences are defined as

$$d_{k,r}^{(1)} = x_{k,r} - x_{k,r-1}, \tag{20}$$

$$d_{k,r}^{(2)} = d_{k,r}^{(1)} - d_{k,r-1}^{(1)} = x_{k,r} - 2x_{k,r-1} + x_{k,r-2}. \tag{21}$$

The corresponding difference is valid only when all values required by the operation are valid. Hence, differencing never crosses a missing or padded position. Each numerical patch is also rendered as a 224×224 RGB image. The three channels encode the value, velocity, and acceleration-like views of the same patch:

$$\mathcal{I}_k = \text{Stack}_{\text{RGB}} \left(\mathcal{R}(\widetilde{\mathbf{x}}_k), \mathcal{R}(\widetilde{\mathbf{d}}_k^{(1)}), \mathcal{R}(\widetilde{\mathbf{d}}_k^{(2)}) \right), \tag{22}$$

Template: This patch contains `[padding]` and `[missingness]`, leaving `[coverage]` valid data coverage. Within the valid observations, the series shows `[trend]` and follows `[shape]`. Its local variation is `[variation]`, with `[volatility]` volatility. The patch contains `[event]`. Overall, its local behavior is characterized as `[regime]`.

Example Series: `[76, 41, 27, 26, 15, 6, 6, 4, 11, 8, 11, 13, 4, 8, 12, 8, 10, 16, 26, 17, 18, 24, 23, 21, 22, 27, 29, 17, 25, 21, 23, 23]`. This is the first 32-point patch of a real Taobao-Fashion series. It contains no padding or missing observations.

Generated Text: This patch contains no padding and no missing observations, leaving full valid-data coverage. The series starts at 76, drops rapidly to 4 within the first eight positions, and then fluctuates at a lower level before partially recovering. Its local variation and volatility are high. The patch contains a prominent early decline and several subsequent turning points. Overall, its local behavior is characterized as an irregular downward regime.

Figure 7: Deterministic patch-level text construction for a real 32-point patch from Taobao-Fashion item `948150`. The template, numerical patch, and generated description are shown together to make the supervision auditable.

where $\mathcal{R}(\cdot)$ denotes mask-aware curve rasterization. The original value curve is placed in the red channel, the first-order difference curve in the green channel, and the second-order difference curve in the blue channel. Their roles are complementary:

- **Value channel (red):** preserves the local level, direction, turning points, and overall shape of the observed patch.
- **First-difference channel (green):** represents local increments $d^{(1)}_{k,r}$ and exposes the direction and magnitude of short-term changes. A nearly horizontal trace indicates approximately constant local increments.
- **Second-difference channel (blue):** represents changes in the first differences. It highlights curvature, acceleration, deceleration, and abrupt changes that may be visually subtle in the raw-value curve.

To reduce sensitivity to isolated outliers, the value channel is robustly normalized using valid patch quantiles:

$$\widetilde{x}_{k,r} = \mathrm{clip}\left(\frac{x_{k,r} - Q_{0.50}(\mathbf{x}_k)}{Q_{0.95}(\mathbf{x}_k) - Q_{0.05}(\mathbf{x}_k) + \epsilon}, -1, 1\right). \tag{23}$$

The two difference channels share a common robust scale

$$s_{\Delta,k} = Q_{0.95}\left(\left\{|d^{(1)}_{k,r}|\right\} \cup \left\{|d^{(2)}_{k,r}|\right\}\right) + \epsilon, \tag{24}$$

and are normalized by $\widetilde{d}^{(j)}_{k,r} = \mathrm{clip}(d^{(j)}_{k,r}/s_{\Delta,k}, -1, 1)$. Sharing the scale preserves the relative magnitudes of the first- and second-order changes. Horizontal coordinates are determined by the original patch positions, and normalized amplitudes determine vertical coordinates. Curves are drawn only between consecutive valid positions. Missing or padded positions remain white and interrupt the curve, rather than being filled or connected across.

A.2.2 Sequence-Level Endogenous Modalities

Patch-level modalities emphasize local behavior but do not explicitly capture how local regimes evolve over a long context. We therefore additionally construct sequence-level views from the complete input window. These views use the same masks, robust statistics, and deterministic vocabulary as their patch-level counterparts. Unlike patch-level modalities, sequence-level modalities are not directly fused into the numerical backbone as endogenous inputs. Instead, they serve as retrieval proxies during the training of the exo-multimodal retrieval enhancer. Specifically, the generated sequence-level text and images simulate the heterogeneous exogenous information associated with a historical time series, allowing the retrieval module to learn how different exogenous modalities

Example Series: `[76, 41, 27, 26, 15, 6, 6, 4, 11, 8, 11, 13, 4, 8, 12, 8, 10, 16, 26, 17, 18, 24, 23, 21, 22, 27, 29, 17, 25, 21, 23, 23]`. The red channel encodes the normalized value curve, the green channel encodes $d_t^{(1)}$, and the blue channel encodes $d_t^{(2)}$.

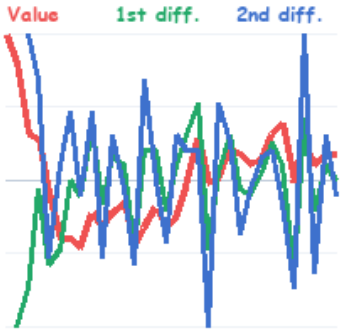


Figure 8: Actual 224×224 patch-level RGB image generated from the listed real 32-point Taobao-Fashion patch. The three colored curves encode the value, first-difference, and second-difference views, respectively.

should identify and aggregate relevant reference cases. This proxy-based training provides scalable supervision for the exogenous module without requiring every pretraining sequence to be paired with naturally occurring exogenous multimodal data.

Sequence-level text generation. For a complete input window, we compute global coverage, robust range, global slope, volatility, dominant periodicity, and salient jumps. We also arrange the patch-level regimes chronologically and compare adjacent patches. A regime transition is recorded when the trend category, variation category, or turning behavior changes between neighboring valid patches. These statistics populate the sequence template:

> *"This sequence contains [number] patches with [coverage] valid-data coverage. Across the observed history, it shows [global trend] with [global variation] variation and [global volatility] volatility. The local dynamics progress from [early regime] to [late regime], with [transition] near [location]. The sequence contains [global event] and exhibits [periodicity]. Overall, its temporal evolution is characterized as [global pattern]."*

Sequence-level line image. For the complete sequence, we draw the valid value, first-difference, and second-difference curves in chronological order. The construction follows the same definitions as the patch-level RGB image, but uses the full historical window and a sequence-level robust scale. This view makes long-term trend, regime changes, repeated oscillations, and change points visible in a single image. Missing intervals remain blank, and curves are not connected across them.

Sequence-level heatmap. To expose relationships across patches, the sequence is reshaped into a $K \times P$ patch matrix. For valid position (k, r), we set

$$H_{k,r} = \mathrm{clip}\left(\frac{x_{k,r} - Q_{0.50}(\mathbf{x})}{Q_{0.75}(\mathbf{x}) - Q_{0.25}(\mathbf{x}) + \epsilon}, -3, 3\right). \tag{25}$$

Rows correspond to chronological patches and columns correspond to within-patch positions. A fixed diverging color map represents negative and positive deviations from the sequence median; masked entries are rendered white. Because the normalization is shared across the whole sequence rather than performed independently for every row, differences in level and amplitude remain comparable across patches.

Template: This sequence contains `[number]` patches with `[coverage]` valid-data coverage. Across the observed history, it shows `[global trend]` with `[global variation]` variation and `[global volatility]` volatility. The local dynamics progress from `[early regime]` to `[late regime]`, with `[transition]` near `[location]`. The sequence contains `[global event]` and exhibits `[periodicity]`. Overall, its temporal evolution is characterized as `[global pattern]`.

Example Series: `[76, 41, 27, 26, 15, 6, 6, 4, 11, 8, 11, 13, 4, 8, 12, 8, 10, 16, 26, 17, 18, 24, 23, 21, 22, 27, 29, 17, 25, 21, 23, 23, 18, 19, 32, 25, 31, 34, 33, 45, 54, 61, 80, 57, 63, 56, 44, 51, 37, 55, 27, 31, 36, 42, 48, 41, 51, 47, 57, 60, 66, 65, 57, 44, 65, 78, 65, 45, 68, 49, 39, 47, 64, 41, 56, 43, 53, 50, 37, 51, 41, 39, 45, 65, 52, 49, 71, 69, 46, 41, 48, 66, 60, 62, 49, 172, 57, 54, 41, 59, 54, 60, 59, 43, 49, 58, 36, 80, 64, 58, 36, 40, 33, 47, 73, 58, 85, 54, 41, 52, 52, 62, 78, 61, 51, 60, 109, 43, 48, 74, 35, 37, 51, 48, 41, 47, 46, 51, 22, 34, 25, 44, 37, 37, 22, 18, 18, 33, 35, 40, 39, 27, 41, 39, 30, 30, 26, 35, 26, 33, 25, 46, 29, 39, 40, 44, 31, 33, 38, 54, 41, 51, 49, 50, 39, 38, 36, 57, 37, 34, 47, 34, 54, 54, 47, 37, 16, 15, 6, 5, 5, 3, 2, 1, 4, 7, 12, 11, 14, 23, 30, 53, 69, 76, 85, 92, 95, 75, 61, 73, 81, 83, 155, 150, 129, 149, 133, 30, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 10, 11, 9, 9, 10, 5, 7, 10, 8, 8, 9, 8, 4, 13, 8, 6, 6, 11, 10, 7, 13, 9, 7]`. This is a real 256-day window from Taobao-Fashion item `948150`, covering 2014-08-08 to 2015-04-20. It contains eight fully observed 32-point patches; zero-valued observations are retained as valid values.

Generated Text: This sequence contains eight patches with full valid-data coverage. Across the observed history, it shows substantial variation and high volatility, with several distinct local regimes. It begins with a sharp decline, moves into a sustained medium-level fluctuating regime, and contains a prominent spike of 172 near the end of the third patch. The middle patches gradually return to lower values. A later rebound reaches 155, followed by an abrupt collapse into a near-zero regime and a small recovery near the end. The repeated changes in level dominate any stable periodic pattern. Overall, the sequence is characterized as a non-stationary multi-regime trajectory with multiple abrupt transitions.

Figure 9: Deterministic sequence-level text construction from a real 256-point Taobao-Fashion input window. The item identifier and date range are retained to make the example traceable to the source data.

Sequence-level spectrum image. The frequency-domain view describes periodic and multi-scale behavior that may be difficult to identify in the time-domain curves. For a contiguous valid block $\mathcal{B}$, we remove its robust center, apply a Hann window w_t, and compute the periodogram

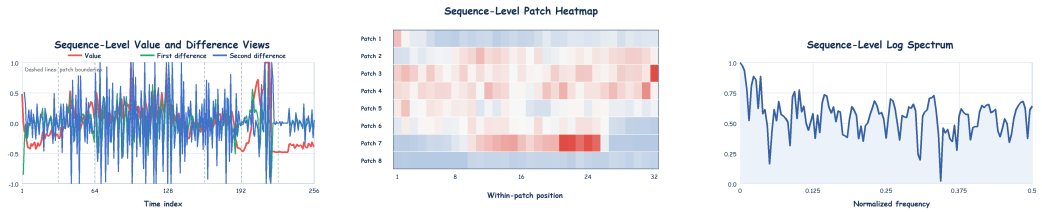
$$S_{\mathcal{B}}(f) = \frac{\left|\sum_{t\in\mathcal{B}} w_t(x_t - \bar{x}_{\mathcal{B}})e^{-\mathrm{i}2\pi ft}\right|^2}{\sum_{t\in\mathcal{B}} w_t^2 + \epsilon}. \tag{26}$$

When missing intervals exist, periodograms are computed separately for valid contiguous blocks and combined on a common normalized-frequency grid, weighted by block length. We never interpolate across a missing interval solely to make the Fourier transform visually continuous. The plotted vertical coordinate is $\log(1 + S(f))$, which compresses extreme spectral peaks while retaining the dominant frequencies.

Dataset construction and reproducibility. For every training window, we store the numerical values, observation mask, padding mask, patch boundaries, patch-level text, patch-level RGB images, sequence-level text, and the three sequence-level image views. Each artifact is associated with the source identifier, window endpoint, and generation-version identifier. Identical numerical inputs and masks therefore produce identical multimodal examples. Dataset splitting is performed chronologically before generation statistics are estimated, and any threshold or normalization summary shared across examples is obtained from the training partition only. This construction preserves the forecasting information boundary while providing aligned local and global multimodal supervision.

A.3 Benchmarks

We evaluate QiYao-M on both unimodal and multimodal forecasting benchmarks. For unimodal forecasting, we adopt GIFT-Eval (Aksu et al., 2024) and TIME (Qiao et al., 2026). For fore-



(a) Value, first-difference, and second-difference curves.

(b) Mask-aware patch-by-position heatmap.

(c) Log-scaled frequency spectrum.

Figure 10: The three sequence-level image views generated deterministically from the same real 256-point Taobao-Fashion historical sequence.

casting with exogenous modalities, we consider representative multimodal benchmarks covering increasingly diverse modality combinations: Time-MMD (Liu et al., 2024a) (time series + text), MoTime (Zhou et al., 2025) (time series + text + image), and FinMultiTime (Xu et al., 2025) (time series + text + image + table).

Endogenous Multimodal Forecasting. We evaluate forecasting without externally provided modalities on GIFT-Eval and TIME. In this setting, multimodal representations are constructed solely from the observed time series, allowing us to examine whether endogenous multimodal modeling improves forecasting even in the absence of external information. GIFT-Eval contains 23 datasets spanning diverse domains and sampling frequencies, while TIME consists of 50 datasets, resulting in a total of **73 evaluation datasets**. We follow the official evaluation protocols of the two benchmarks, including their predefined forecasting horizons, sampling frequencies, and evaluation windows.

Exogenous Multimodal Forecasting. To evaluate the ability of QiYao-M to exploit externally provided information, we further consider 9 multimodal datasets from Time-MMD, MoTime, and FinMultiTime. Specifically, we use Agriculture, Climate, Energy, Health, and Social Good from Time-MMD; Taobao-Fashion and Tianchi from MoTime; and HS300 and SP500 from FinMultiTime. These datasets cover different combinations of time series, text, images, and tabular information, enabling evaluation under heterogeneous multimodal settings. For Agriculture, Climate, and Social Good, we use prediction lengths $F \in \{6, 8, 10, 12\}$. For Energy and Health, the prediction lengths are $F \in \{12, 24, 36, 48\}$. For Taobao-Fashion and Tianchi, we evaluate $F \in \{1, 7, 14, 21, 28\}$. For HS300 and SP500, the prediction lengths are $F \in \{24, 48, 96\}$.

A.4 Baselines

We compare QiYao-M against representative forecasting models from three categories. For GIFT-Eval and TIME, we use baseline results reported on the official leaderboards under the corresponding benchmark protocols, which ensures consistent and standardized evaluation across methods. For multimodal forecasting, we consider four end-to-end supervised models, including GPT4MTS (Jia et al., 2024), CALF (Liu et al., 2025a), Time-VLM (Zhong et al., 2025), and TATS (Li et al., 2026). We further include five unimodal time-series foundation models, namely Zeus (Fu et al., 2026), Chronos-2 (Ansari et al., 2025), Toto-2 (Khwaja et al., 2026), PatchTST-FM-r2 (Wen et al., 2026), and TiRex-2 (Podest et al., 2026), as well as two multimodal foundation models, ChatTime (Wang et al., 2025a) and Aurora (Wu et al., 2026). The codebases and implementation details of all baselines are provided in Table 6.

A.5 Pretraining Settings

Stage I: Numerical Pretraining. We first train the time series backbone using numerical data only. Historical series are normalized and divided into patches of size $p = 32$, with a maximum input length of 8192 observations. The backbone and forecasting head learn general temporal representations and probabilistic forecasting before any multimodal components are introduced. This stage provides the numerical initialization for the subsequent stages.

Table 3: Dataset statistics and forecasting configurations of GIFT-Eval. Multiple frequencies of the same underlying dataset are grouped into one row.

Dataset	Domain	Frequency	Variates	Prediction Length
Jena Weather	Nature	10T / H / D	21	{48, 480, 720} (10T/H); 30 (D)
BizITObs–Application	Web/CloudOps	10S	2	{60, 600, 900}
BizITObs–Service	Web/CloudOps	10S	2	{60, 600, 900}
BizITObs–L2C	Web/CloudOps	5T / H	7	{48, 480, 720}
Bitbrains–Fast Storage	Web/CloudOps	5T / H	2	{48, 480, 720} (5T); 48 (H)
Bitbrains–rnd	Web/CloudOps	5T / H	2	{48, 480, 720} (5T); 48 (H)
Restaurant	Sales	D	1	30
ETT1	Energy	15T / H / D / W	7	{48, 480, 720} (15T/H); 30 (D); 8 (W)
ETT2	Energy	15T / H / D / W	7	{48, 480, 720} (15T/H); 30 (D); 8 (W)
Loop Seattle	Transport	5T / H / D	1	{48, 480, 720} (5T/H); 30 (D)
SZ-Taxi	Transport	15T / H	1	{48, 480, 720} (15T); 48 (H)
M_DENSE	Transport	H / D	1	{48, 480, 720} (H); 30 (D)
Solar	Energy	10T / H / D / W	1	{48, 480, 720} (10T/H); 30 (D); 8 (W)
Hierarchical Sales	Sales	D / W	1	30 (D); 8 (W)
M4	Econ/Fin	A / Q / M / W / D / H	1	6 / 8 / 18 / 13 / 14 / 48
Hospital	Healthcare	M	1	12
COVID Deaths	Healthcare	D	1	30
US Births	Healthcare	D / W / M	1	30 / 8 / 12
Saugeen	Nature	D / W / M	1	30 / 8 / 12
Temperature Rain	Nature	D	1	30
KDD Cup 2018	Nature	H / D	1	{48, 480, 720} (H); 30 (D)
Car Parts	Sales	M	1	12
Electricity	Energy	15T / H / D / W	1	{48, 480, 720} (15T/H); 30 (D); 8 (W)

Stage II: Endogenous Multimodal Training. We then construct patch-level text and images from the time series using the deterministic generation procedure in Section A.2. The first 16 Transformer layers are frozen, while the remaining 8 layers and the endogenous multimodal modules are trained. The model jointly predicts future numerical values and the latent representations of future endogenous text and images. This design encourages the shared representation to capture temporal evolution from complementary numerical, textual, and visual perspectives, while preserving the general forecasting ability learned in Stage I.

Stage III: Exogenous Retrieval Training. Finally, we freeze the numerical backbone and the endogenous multimodal modules. Only the exogenous retrieval head and its candidate-aware cross-attention components are optimized. Reference cases are retrieved independently through the available modalities, and their known future values provide retrieval-based context for the forecast. Retrieval banks and modality weights are constructed or calibrated using the permitted training data; test targets are never used for training or calibration. A null key–value candidate allows the cross-attention module to reduce its reliance on unhelpful retrieved cases.

Downstream Forecasting. At inference time, endogenous text and images are generated solely from the observed history. For the exogenous setting, reference cases are selected from the pre-constructed bank, and the trained retrieval module enhances the forecast without updating model parameters. We report metrics over all valid forecast points and set drop_last=False during evaluation so that incomplete final batches are not discarded.

A.6 Evaluation Metrics

We follow the official evaluation protocols of the respective benchmarks. For point forecasting, we report Mean Squared Error (MSE) and Mean Absolute Error (MAE). For probabilistic forecasting, we report Mean Absolute Scaled Error (MASE) and Continuous Ranked Probability Score (CRPS) where required. All metrics are computed over valid forecast points only; padded or missing targets are excluded. Let $y_{i,t}$ and $\hat{y}_{i,t}$ denote the target and point forecast for series i at future step t, and let $m_{i,t} \in \{0, 1\}$ indicate whether that target is valid. Define $N = \sum_{i,t} m_{i,t}$.

Table 4: Dataset statistics and forecasting configurations of the TIME benchmark. Different sampling frequencies of the same underlying source are treated as separate evaluation datasets.

Domain	Dataset	Frequency	Prediction Length
Nature	Water Quality Darwin	15T	$\{16, 96, 288\}$
	Current Velocity	5T	$\{36, 288, 864\}$
	Current Velocity	10T	$\{18, 144, 432\}$
	Current Velocity	15T	$\{12, 96, 288\}$
	Current Velocity	20T	$\{9, 72, 216\}$
	Current Velocity	H	$\{24, 168, 336\}$
	CPHL	15T	$\{12, 96, 288\}$
	CPHL	30T	$\{12, 48, 144\}$
	CPHL	H	$\{24, 168, 336\}$
	Coastal T&S	5T	$\{36, 288, 864\}$
	Coastal T&S	15T	$\{12, 96, 288\}$
	Coastal T&S	20T	$\{9, 72, 216\}$
	Coastal T&S	H	$\{24, 168, 336\}$
	SG Weather	D	$\{3, 7, 14\}$
	SG PM2.5	H	$\{24, 72, 168\}$
	NE China Wind	H	$\{24, 72, 168\}$
Energy	Australia Solar	H	$\{24, 72, 168\}$
	EPF Electricity Price	H	$\{24, 72, 168\}$
	OpenElectricity NEM	5T	$\{24, 96, 288\}$
	EWELD Load	15T	$\{24, 96, 672\}$
Transport	SG Carpark	15T	$\{16, 96, 672\}$
	Finland Traffic	15T	$\{16, 96, 672\}$
	Port Activity	D	30
	Port Activity	W	13
Healthcare	ECDC COVID	D	30
	ECDC COVID	W	13
	Global Influenza	W	13
Finance	Crypto	D	30
	US Term Structure	B	20
	Oil Price	B	20
Economics	Job Claims	W	13
	Uncertainty 1M	M	6
	Housing Inventory	M	12
	JOLTS	M	12
	US Labor	M	12
	Vehicle Supply	M	12
	Auto Production SF	M	12
	Commodity Production	M	12
	Commodity Import	M	12
	WUI Global	Q	4
	Global Price	Q	4
Sales	Vehicle Sales	M	12
	Online Retail II UCI	D	30
	Supply Chain Customer	D	30
	Supply Chain Location	D	30
CloudOps	Azure2019 D	5T	288
	Azure2019 I	5T	288
	Azure2019 U	5T	48
	Smart Manufacturing	H	$\{24, 168, 336\}$
	MetroPT-3	5T	$\{48, 288, 576\}$

Point forecasting metrics. The MSE penalizes large prediction errors more strongly, whereas the MAE measures their average absolute magnitude:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i,t} m_{i,t}(y_{i,t} - \hat{y}_{i,t})^2, \qquad \mathrm{MAE} = \frac{1}{N}\sum_{i,t} m_{i,t}|y_{i,t} - \hat{y}_{i,t}|. \tag{27}$$

For models producing predictive quantiles, we use the median (0.5 quantile) as the point forecast.

Scale-normalized metrics. When reporting normalized errors, we calculate the standard deviation σ_{train} from the training interval only and apply the same scale to every model evaluated on that dataset:

$$\text{Norm-MSE} = \frac{\mathrm{MSE}}{\sigma_{\mathrm{train}}^2}, \qquad \text{Norm-MAE} = \frac{\mathrm{MAE}}{\sigma_{\mathrm{train}}}. \tag{28}$$

Normalization changes only the reported metric; it does not use test-set statistics or alter the model forecasts.

Table 5: Statistics of the exogenous multimodal forecasting datasets used in our experiments. The modality column reports the external information available to each dataset.

Benchmark	Dataset	Modalities	Frequency	Scale	Prediction Length
TimeMMD	Agriculture	TS + Text	Monthly	496×1	$\{6, 8, 10, 12\}$
	Climate	TS + Text	Monthly	496×5	$\{6, 8, 10, 12\}$
	Energy	TS + Text	Weekly	$1{,}479 \times 9$	$\{12, 24, 36, 48\}$
	Health	TS + Text	Weekly	$1{,}389 \times 11$	$\{12, 24, 36, 48\}$
	Social Good	TS + Text	Monthly	900×1	$\{6, 8, 10, 12\}$
MoTIME	TaobaoFashion	TS + Text + Image	Daily	365×890	$\{1, 7, 14, 21, 28\}$
	Tianchi	TS + Text + Image	Daily	$184 \times 36{,}397$	$\{1, 7, 14, 21, 28\}$
FinMultiTime	HS300	TS + Text + Image + Table	Trading Day	35 stocks	$\{24, 48, 96\}$
	SP500	TS + Text + Image + Table	Trading Day	35 stocks	$\{24, 48, 96\}$

Table 6: Code and model repositories of the baseline methods used in our experiments.

Model Type	Model	Code / Model Repository
Multimodal End-to-End	GPT4MTS	https://github.com/Flora-jia-jfr/GPT4MTS-Prompt-based-Large-Language-Model-for-Multimodal-Time-series-Forecasting
	CALF	https://github.com/Hank0626/CALF
	Time-VLM	https://github.com/CityMind-Lab/ICML25-TimeVLM
	TATS	https://github.com/iDEA-iSAIL-Lab-UIUC/TaTS
Unimodal Foundation	Zeus	https://github.com/GestaltCogTeam/Zeus
	Chronos-2	https://github.com/amazon-science/chronos-forecasting
	Toto-2.0	https://github.com/DataDog/toto
	PatchTST-FM-r2	https://huggingface.co/ibm-granite/granite-timeseries-patchtst-fm-r2
	TiRex-2	https://github.com/NX-AI/tirex-2
Multimodal Foundation	ChatTime	https://github.com/ForestsKing/ChatTime
	Aurora	https://github.com/decisionintelligence/Aurora

Mean Absolute Scaled Error. MASE scales the absolute forecast error by the in-sample seasonal naive error. For series i with training observations $x_{i,1:L_i}$ and seasonality s_i, the scaling factor is

$$d_i = \frac{1}{L_i - s_i} \sum_{u=s_i+1}^{L_i} |x_{i,u} - x_{i,u-s_i}|, \qquad \text{MASE} = \frac{1}{N} \sum_{i,t} m_{i,t} \frac{|y_{i,t} - \hat{y}_{i,t}|}{d_i}. \tag{29}$$

The benchmark's official evaluator determines the seasonality and handles degenerate denominators.

Continuous Ranked Probability Score. For a predictive cumulative distribution function $F_{i,t}$, CRPS compares the full forecast distribution with the realized target:

$$\text{CRPS} = \frac{1}{N} \sum_{i,t} m_{i,t} \int_{-\infty}^{\infty} \left(F_{i,t}(z) - \mathbb{I}\{y_{i,t} \le z\}\right)^2 \, dz. \tag{30}$$

We use each benchmark's official evaluator to compute its reported CRPS from the model's probabilistic forecasts.

A.7 Model Configurations

The numerical backbone contains 312.68M parameters. The endogenous image and text projectors, predictors, and mask tokens add 36.73M parameters. The exogenous target encoder and five

Table 7: Architecture and parameter counts of QiYao-M. The main-model count excludes the frozen CLIP feature encoder and does not double-count the shared forecasting head.

Model	Backbone Layers	Fusion After	Model Dim.	FFN Dim.	Heads	Patch Size	CLIP Dim.	Retrieval Layers	Main Parameters
QiYao-M	24	16	1024	2736	16	32	512	5	375.63M

candidate-aware cross-attention layers add 26.21M parameters, yielding 375.63M unique parameters in the main model. The frozen CLIP ViT-B/32 image and text encoder contains an additional 151.28M parameters; including it gives 526.90M parameters.

B More Analysis

B.1 Retrieval Analysis

We study the robustness of our retrieval mechanism by exploring its sensitivity to the retrieval bank size and final top-K, as shown in Figure 11.

For the retrieval bank, QiYao-M already outperforms the internal baseline and strong foundation models using only 5% of the available training samples. As the bank size increases, forecasting performance quickly stabilizes with only minor fluctuations, suggesting that an excessively large retrieval bank is unnecessary. In practice, retaining around 20% of the available samples provides a favorable trade-off between forecasting performance and retrieval efficiency. We further fix the bank size and vary the final top-k from 4 to 24. The forecasting errors remain within a narrow range across different values of k, indicating that QiYao-M is insensitive to the exact number of retrieved samples. These results demonstrate strong robustness to both retrieval bank size and top-k, substantially reducing the need for careful hyperparameter tuning.

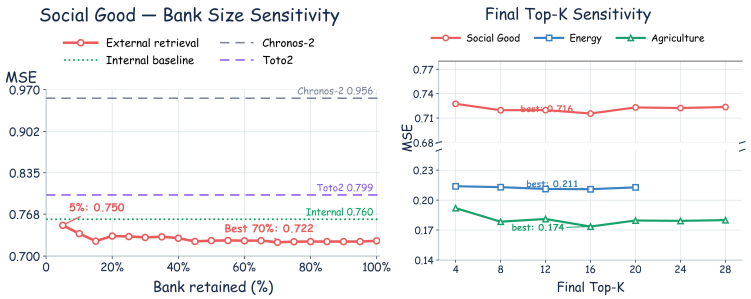


Figure 11: Sensitivity analysis of the retrieval module.

C More Results

C.1 GIFT-Eval Rankings

To provide a more comprehensive comparison on GIFT-Eval, we further report the ranking results in terms of MASE and CRPS. As shown in Figures 12 and 13, our model consistently achieves favorable rankings across both metrics, demonstrating clear improvements over the baseline forecasting models. These results further validate the effectiveness and robustness of our model under diverse time-series forecasting scenarios.

C.2 Full Results

We further provide the complete step-wise forecasting results on multimodal datasets. Specifically, we report the performance of different methods at each forecasting horizon, together with their average performance, enabling a more fine-grained comparison across short- and long-term forecasting settings. The detailed results further demonstrate the consistent effectiveness of our model across different prediction horizons and multimodal forecasting benchmarks.

Table 8: Step-wise forecasting results on **Agriculture**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Methods	H=6		H=8		H=10		H=12		Avg.	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models										
GPT4MTS	0.161	0.257	0.207	0.288	0.230	0.305	0.301	0.342	0.225	0.298
CALF	0.142	0.250	0.195	0.285	0.350	0.370	0.314	0.355	0.250	0.315
Time-VLM	0.143	0.245	0.215	0.287	0.271	0.320	0.322	0.359	0.237	0.302
TATS	0.140	0.251	0.187	0.282	0.244	0.320	0.290	0.350	0.215	0.301
Unimodal Foundation Models										
Zeus	0.122	0.237	0.173	0.279	0.228	0.315	0.289	0.348	0.203	0.295
Chronos-2	0.147	0.258	0.208	0.303	0.267	0.340	0.330	0.376	0.238	0.319
Toto-2	0.144	0.252	0.203	0.296	0.264	0.330	0.331	0.366	0.236	0.311
PatchTST-r2	0.167	0.302	0.216	0.341	0.255	0.371	0.302	0.401	0.235	0.354
TiRex2	0.148	0.254	0.216	0.300	0.288	0.338	0.369	0.380	0.255	0.318
Multimodal Foundation Models										
ChatTime	0.129	0.245	0.170	0.278	0.215	0.306	0.271	0.343	0.196	0.293
Aurora	0.184	0.295	0.242	0.335	0.297	0.365	0.365	0.398	0.272	0.348
Ours (w/o Exo.)	0.121	0.233	0.173	0.275	0.224	0.307	0.281	0.340	0.199	0.288
Ours (w Exo.)	**0.108**	**0.232**	**0.150**	**0.268**	**0.192**	**0.300**	**0.244**	**0.336**	**0.174**	**0.284**

Table 9: Step-wise forecasting results on **Climate**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Horizons	H=6		H=8		H=10		H=12		Avg.	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models										
GPT4MTS	1.199	0.895	1.205	0.899	1.173	0.885	1.152	0.876	1.182	0.889
CALF	1.231	0.910	1.227	0.911	1.508	0.989	1.177	0.883	1.286	0.922
Time-VLM	1.218	0.907	1.181	0.914	1.179	0.880	1.203	0.896	1.195	0.899
TATS	1.194	0.897	1.178	0.886	1.170	0.881	1.179	0.885	1.180	0.887
Unimodal Foundation Models										
Zeus	0.858	0.741	0.854	0.738	0.847	0.736	0.846	0.737	0.852	0.738
Chronos-2	0.858	0.730	0.857	0.728	0.859	0.729	0.863	0.732	0.859	0.730
Toto-2	0.844	0.725	0.850	0.728	0.858	0.731	0.861	0.733	0.853	0.729
PatchTST-r2	0.842	0.724	0.841	0.723	0.846	0.724	0.850	**0.726**	0.845	0.724
TiRex2	0.839	**0.723**	0.841	0.722	0.852	0.725	0.856	0.727	0.847	0.724
Multimodal Foundation Models										
ChatTime	1.100	0.842	1.177	0.863	1.158	0.866	1.143	0.854	1.144	0.856
Aurora	0.859	0.747	0.858	0.746	0.868	0.748	0.875	0.753	0.865	0.749
Ours (w/o Exo.)	0.828	0.724	**0.822**	0.721	0.830	0.725	**0.832**	0.728	0.828	0.725
Ours (w Exo.)	**0.826**	**0.723**	**0.822**	**0.719**	**0.829**	**0.722**	**0.832**	0.727	**0.827**	**0.723**

Table 10: Step-wise forecasting results on **Energy**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Horizons	H=12		H=24		H=36		H=48		Avg.	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models										
GPT4MTS	0.111	0.244	0.232	0.362	0.308	0.418	0.398	0.496	0.262	0.380
CALF	0.102	0.224	0.210	0.346	0.300	0.420	0.365	0.470	0.244	0.365
Time-VLM	0.114	0.253	0.227	0.359	0.309	0.410	0.390	0.475	0.260	0.374
TATS	0.105	0.232	0.216	0.344	0.309	0.418	0.391	0.480	0.255	0.368
Unimodal Foundation Models										
Zeus	0.094	0.207	0.193	0.312	0.288	0.383	0.380	0.444	0.239	0.337
Chronos-2	0.089	0.200	0.182	**0.301**	0.271	**0.368**	0.362	0.434	0.226	0.326
Toto-2	0.091	0.203	0.190	0.315	0.275	0.383	0.367	0.445	0.231	0.337
PatchTST-r2	0.087	0.210	0.182	0.319	0.284	0.399	0.426	0.484	0.245	0.353
TiRex2	0.088	0.202	0.181	0.304	0.266	0.373	0.360	0.441	0.224	0.330
Multimodal Foundation Models										
ChatTime	0.104	0.223	0.213	0.331	0.315	0.403	0.399	0.463	0.258	0.355
Aurora	0.117	0.245	0.226	0.354	0.292	0.409	0.383	0.472	0.255	0.370
Ours (w/o Exo.)	0.085	0.200	0.186	0.310	0.287	0.387	0.390	0.454	0.237	0.338
Ours (w Exo.)	**0.081**	**0.196**	**0.178**	0.302	**0.262**	0.374	**0.322**	**0.428**	**0.211**	**0.325**

Table 11: Step-wise forecasting results on **Health**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Horizons	H=12		H=24		H=36		H=48		Avg.	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models										
GPT4MTS	0.985	0.658	1.513	0.802	1.601	0.846	1.757	0.889	1.464	0.799
CALF	0.964	0.609	1.451	0.749	1.713	0.851	1.836	0.889	1.491	0.775
Time-VLM	1.198	0.727	1.491	0.839	1.867	0.967	1.702	0.907	1.565	0.860
TATS	0.899	0.612	1.307	0.759	1.523	0.827	1.693	0.872	1.356	0.767
Unimodal Foundation Models										
Zeus	1.230	0.668	1.610	0.847	1.728	0.917	1.791	0.949	1.590	0.845
Chronos-2	0.650	0.513	0.954	0.632	1.209	0.723	1.402	0.784	1.054	0.663
Toto-2	0.702	0.501	1.038	0.637	1.260	0.723	1.437	0.787	1.109	0.662
PatchTST-r2	0.643	0.512	0.960	0.636	1.119	0.697	1.233	0.743	0.989	0.647
TiRex2	0.754	0.518	1.116	0.635	1.236	0.685	1.280	0.712	1.097	0.638
Multimodal Foundation Models										
ChatTime	1.266	0.716	2.087	0.952	2.622	1.097	3.137	1.240	2.278	1.001
Aurora	1.093	0.668	1.572	0.849	1.688	0.920	1.857	0.963	1.553	0.850
Ours (w/o Exo.)	0.585	0.469	0.870	0.602	1.034	0.664	1.170	0.712	0.915	0.612
Ours (w Exo.)	**0.574**	**0.468**	**0.851**	**0.598**	**1.021**	**0.659**	**1.156**	**0.706**	**0.901**	**0.608**

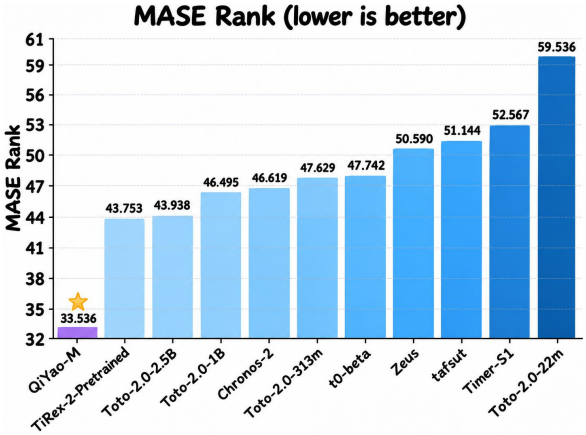


Figure 12: GIFT-Eval MASE Ranking.

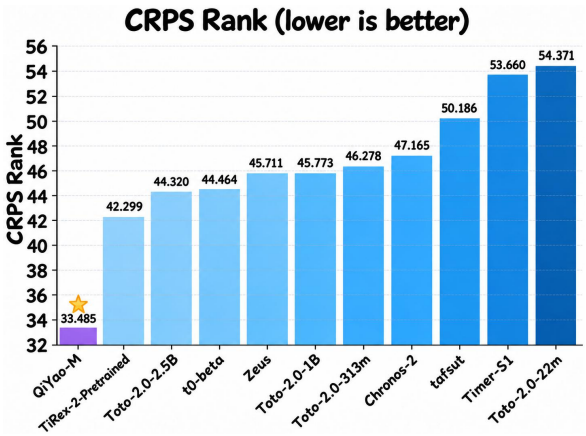


Figure 13: GIFT-Eval CRPS Ranking.

Table 12: Step-wise forecasting results on **Social Good**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Horizons	**H=6**		**H=8**		**H=10**		**H=12**		**Avg.**	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models										
GPT4MTS	0.718	0.378	0.942	0.505	0.929	0.446	1.093	0.470	0.920	0.450
CALF	0.782	0.360	0.874	0.386	0.976	0.420	0.991	0.439	0.906	0.401
Time-VLM	0.732	0.379	0.822	0.427	0.916	0.465	1.005	0.505	0.868	0.444
TATS	0.753	0.370	0.875	0.409	0.991	0.459	1.053	0.474	0.918	0.428
Unimodal Foundation Models										
Zeus	0.820	0.363	0.926	0.410	1.058	0.461	1.193	0.501	0.999	0.434
Chronos-2	0.734	0.312	0.855	0.362	0.959	0.408	1.060	0.453	0.902	0.384
Toto-2	0.683	0.261	0.773	0.296	0.841	0.326	0.900	0.355	0.799	0.309
PatchTST-r2	0.747	0.319	0.816	0.354	0.865	0.382	0.907	0.408	0.834	0.366
TiRex2	0.650	0.277	0.727	0.312	0.785	0.342	0.832	0.369	0.749	0.325
Multimodal Foundation Models										
ChatTime	1.200	0.488	1.244	0.506	1.281	0.547	1.552	0.621	1.319	0.541
Aurora	0.701	0.442	0.804	0.493	0.886	0.543	0.960	0.587	0.838	0.516
Ours (w/o Exo.)	0.675	0.257	0.739	0.284	0.789	0.309	0.835	0.332	0.760	0.296
Ours (w Exo.)	**0.634**	**0.253**	**0.702**	**0.283**	**0.753**	**0.305**	**0.800**	**0.331**	**0.722**	**0.293**

Table 13: Step-wise forecasting results on **TAOBAO-fashion**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Horizons	H=1		H=7		H=14		H=21		H=28		Avg.	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models												
GPT4MTS	0.384	0.189	0.473	0.232	0.613	0.297	0.640	0.293	0.559	0.272	0.534	0.257
CALF	0.420	0.230	0.488	0.240	0.544	0.263	0.536	0.254	0.570	0.263	0.512	0.250
Time-VLM	0.419	0.226	0.489	0.257	0.555	0.290	0.568	0.290	0.609	0.289	0.528	0.271
TATS	0.443	0.242	0.515	0.258	0.538	0.273	0.548	0.267	0.579	0.274	0.525	0.263
Unimodal Foundation Models												
Zeus	0.376	0.153	0.394	0.161	0.426	**0.172**	0.450	**0.181**	0.477	**0.191**	0.425	0.172
Chronos-2	0.378	0.153	0.411	0.169	0.449	0.186	0.477	0.199	0.509	0.212	0.445	0.184
Toto-2	0.371	0.151	0.406	0.166	0.443	0.182	0.471	0.197	0.505	0.213	0.439	0.182
PatchTST-r2	0.371	0.163	0.503	0.186	0.551	0.201	0.585	0.216	0.621	0.232	0.526	0.200
TiRex2	0.358	0.150	0.400	**0.160**	0.434	**0.172**	0.459	0.182	0.487	0.193	0.427	**0.171**
Multimodal Foundation Models												
ChatTime	0.544	0.167	0.432	0.185	0.357	0.200	0.527	0.219	0.537	0.231	0.479	0.201
Aurora	0.365	0.169	0.408	0.192	0.448	0.214	0.475	0.229	0.512	0.247	0.442	0.210
Ours (w/o Exo.)	**0.351**	**0.149**	0.391	0.161	0.424	**0.172**	**0.447**	0.182	0.473	0.193	**0.417**	**0.171**
Ours (w Exo.)	0.469	0.152	**0.359**	0.160	0.324	0.173	0.460	0.185	**0.472**	0.200	**0.417**	0.174

Table 14: Step-wise forecasting results on **Tianchi**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Horizons	H=1		H=7		H=14		H=21		H=28		Avg.	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models												
GPT4MTS	**1.931**	0.156	1.384	0.138	1.112	0.154	1.013	0.141	1.162	0.145	1.320	0.147
CALF	2.017	0.128	1.346	0.108	1.065	0.111	0.996	0.114	1.176	0.134	1.320	0.119
Time-VLM	1.982	0.124	1.416	0.135	1.217	0.155	1.027	0.129	1.200	0.140	1.368	0.137
TATS	2.035	0.135	1.411	0.134	1.165	0.138	1.052	0.133	1.206	0.139	1.374	0.136
Unimodal Foundation Models												
Zeus	2.143	0.085	1.339	0.074	1.188	0.083	1.065	0.083	1.154	0.091	1.378	0.083
Chronos-2	2.707	0.089	1.337	0.075	1.129	0.082	0.975	0.082	1.119	0.089	1.453	0.083
Toto-2	2.545	0.096	2.882	0.107	119.296	0.303	54.950	0.224	28.415	0.194	41.618	0.185
PatchTST-r2	2.150	0.087	1.337	0.083	1.396	0.097	0.974	0.091	1.197	0.099	1.411	0.091
TiRex2	2.122	**0.078**	1.324	**0.069**	**1.033**	**0.073**	**0.936**	**0.072**	**1.079**	**0.079**	**1.299**	**0.074**
Multimodal Foundation Models												
ChatTime	2.478	0.090	1.372	0.078	2.045	0.094	1.306	0.087	1.179	0.088	1.676	0.087
Aurora	2.219	0.086	1.321	0.082	1.204	0.088	0.998	0.092	1.272	0.102	1.403	0.090
Ours (w/o Exo.)	2.563	0.089	1.328	0.072	1.053	0.078	1.052	0.081	1.121	0.086	1.423	0.081
Ours (w Exo.)	2.132	0.081	**1.310**	0.070	**1.021**	0.074	0.963	0.075	1.084	0.081	1.302	0.076

Table 15: Step-wise forecasting results on **HS300**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Horizons	H=24		H=48		H=96		Avg.	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models								
GPT4MTS	0.321	0.352	0.626	0.493	1.392	0.759	0.780	0.535
CALF	0.283	0.333	0.574	0.475	**1.165**	0.696	0.674	0.502
Time-VLM	0.287	0.347	**0.500**	**0.456**	1.200	**0.693**	**0.662**	0.499
TATS	0.281	0.330	0.561	0.461	1.309	0.737	0.717	0.509
Unimodal Foundation Models								
Zeus	0.306	0.341	0.625	0.493	1.333	0.739	0.755	0.524
Chronos-2	0.397	0.329	0.728	0.475	1.306	0.699	0.810	0.501
Toto-2	0.501	0.331	0.773	0.475	1.382	0.716	0.885	0.507
PatchTST-r2	0.292	0.323	0.635	0.481	1.529	0.759	0.819	0.521
TiRex2	0.277	0.322	0.565	0.473	1.179	0.701	0.674	0.499
Multimodal Foundation Models								
ChatTime	0.628	0.492	1.115	0.669	1.880	0.898	1.208	0.686
Aurora	0.332	0.359	0.625	0.496	1.251	0.716	0.736	0.524
Ours (w/o Exo.)	**0.273**	0.319	0.567	0.466	1.232	0.703	0.690	0.496
Ours (w Exo.)	**0.273**	**0.318**	0.549	0.463	1.194	0.698	0.672	**0.493**

Table 16: Step-wise forecasting results on **SP500**. Lower MSE and MAE indicate better performance. The best and second-best results are highlighted in **purple** and blue, respectively.

Horizons	H=24		H=48		H=96		Avg.	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Multimodal End-to-End Models								
GPT4MTS	0.484	0.503	0.723	0.635	1.095	0.786	0.767	0.642
CALF	0.368	0.434	0.598	0.564	1.262	0.853	0.743	0.617
Time-VLM	0.395	0.449	0.760	0.629	1.114	0.776	0.756	0.618
TATS	0.365	0.430	0.617	0.570	**1.014**	**0.742**	0.665	0.581
Unimodal Foundation Models								
Zeus	0.385	0.427	0.728	0.595	1.461	0.862	0.858	0.628
Chronos-2	0.354	0.408	0.651	0.559	1.147	0.765	0.717	0.577
Toto-2	0.332	**0.395**	0.620	0.544	1.157	0.772	0.703	0.571
PatchTST-r2	0.471	0.434	0.948	0.609	1.822	0.858	1.080	0.634
TiRex2	0.344	0.405	0.624	0.557	1.159	0.779	0.709	0.580
Multimodal Foundation Models								
ChatTime	5.214	1.647	5.719	1.747	7.022	1.944	5.985	1.779
Aurora	0.449	0.476	0.761	0.622	1.376	0.851	0.862	0.649
Ours (w/o Exo.)	**0.329**	0.396	0.612	0.545	1.142	0.767	0.694	0.569
Ours (w Exo.)	**0.329**	0.398	**0.591**	**0.542**	1.061	0.744	**0.660**	**0.561**