LONG-HORIZON PREDICTION FROM MULTI-MODAL AND IRREGULAR BEHAVIOR DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we study long-horizon prediction from years of multimodal behavioral data and propose an adaptive multi-scale learning framework for effective cross-modal fusion under irregular temporal observations. Discriminative features are first extracted from modality-specific encoders and fused through scale-wise cross attention. To exploit the complementary information across modalities, we introduce a distillation mechanism based on unimodal teachers that learns shared representations through cross-modal contrastive alignment while retaining modality-specific information through correlation minimization. An error-aware adaptive scale router identifies the best-suited temporal resolution for final prediction. In the internal evaluation of an insurance company, with real data collected from $300,000$ users for critical-illness claim prediction, our method outperforms existing multimodal time-series methods. To validate its generality, we also conduct evaluation in the multi-model temporal prediction from two financial benchmarks and achieve state-of-the-art performance.

1 INTRODUCTION

Long-horizon prediction from individual activities aims to infer future outcomes from multi-modal and partially observed user records. This problem can be modeled as multi-modal time series forecasting (Xu et al., 2025a; Liu et al., 2024), where the raw inputs considered in this paper include structured tabular records, unstructured text and their associated timestamps. These records may span horizons of years, e.g., two years in our experimental setting, and are sparsely and irregularly sampled. Consequently, observations from different modalities are temporally misaligned.

Existing multimodal time-series methods have shown the benefits of incorporating textual or contextual information beyond numerical observations (Liu et al., 2024; Li et al., 2026; Nguyen et al., 2026). Prior work primarily combines heterogeneous modalities through temporal fusion, aggregation of modality-specific predictions, or frequency-domain interaction. However, two challenges remain for long-horizon behavioral prediction. First, joint optimization can lead to imbalanced learning across modalities (Ma et al., 2025; Wang et al., 2025b). In long-horizon behavioral prediction, tabular user attributes, text-based behavioral information, and numerical behavioral sequences may differ substantially in predictive strength and learning dynamics. As a result, stronger modalities can dominate training, causing weaker modalities to be insufficiently optimized (Du et al., 2023).

Second, long behavioral histories contain patterns at multiple temporal resolutions, which are not fully exploited by existing multimodal time-series methods. Fine-grained representations preserve local variations but are often sparse and noisy, whereas coarser representations capture more stable long-term patterns at the cost of local detail. Multi-scale modeling has shown clear benefits for time-series forecasting (Shabani et al., 2023; Fan et al., 2024). However, existing methods are primarily designed for numerical time series (Wang et al., 2024; Chen et al., 2024), while our setting integrates tabular, textual, and numerical information, requiring multi-scale temporal modeling to be considered jointly with cross-modal fusion. Moreover, when predictions from multiple temporal scales are combined through adaptive weighting, a scale assigned a smaller weight has less influence on the final prediction and consequently receives a weaker training signal from the task loss. This may lead to insufficient learning of that scale and degrade its predictive capability, preventing the model from fully exploiting complementary information across temporal scales.

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.

In this work, we propose a multimodal temporal framework that improves modality learning under joint optimization and adaptively combines information across temporal resolutions. To mitigate modality imbalance, we first train a teacher model for each input modality. We then introduce a decomposed unimodal teacher (DUMT) mechanism that decomposes each modality representation into shared and private components and reconstructs the corresponding teacher representation from their combination. The shared components are aligned across modalities, while shared-private correlation is minimized to retain modality-specific information. For temporal-scale modeling, we construct textual and numerical behavior sequences at multiple resolutions and perform multimodal fusion at each scale. Scale-specific supervision is applied to preserve the predictive information captured at different temporal resolutions. An error-aware router then uses scale-specific prediction errors to adaptively combine the predictions of different scales for each sample.

Extensive experiments demonstrate the effectiveness of the proposed framework. We evaluate our method on a large-scale real-world insurance dataset constructed from operational records of a leading insurance company, covering approximately 300,000 users. For critical-illness claim prediction using a two-year observation window, our method achieves a ROC-AUC of 0.6651, outperforming the strongest multimodal baseline by 0.0118 absolute ROC-AUC. According to the insurer's operational assessment, an absolute ROC-AUC improvement of 0.01 is estimated to increase the claim capture rate among the top 2% highest-risk users by approximately 0.33 percentage points. Under an assumed intervention coverage rate of 95% and intervention effectiveness rate of 90%, this corresponds to an estimated annual claim loss reduction of approximately RMB 11 million, highlighting the practical value of the performance gain. We additionally evaluate the framework on two public multimodal financial forecasting benchmarks, S&P 500 and HS300 (Xu et al., 2025a), achieving the best MAE and MSE across all input lengths and reducing MAE by 7.3% and 12.6%, respectively, at the longest 144-day input window.

2 Related Work

2.1 Multimodal Time Series Learning

Recent multimodal time-series studies have moved from loosely combining heterogeneous signals toward more structured cross-modal interaction. Time-MMD (Liu et al., 2024) separately models numerical and textual streams and combines their predictions, while TaTS (Li et al., 2026) integrates text embeddings directly as auxiliary temporal variables for joint sequence modeling. Subsequent methods explore richer interactions, including frequency-domain fusion in SpecTF (Nguyen et al., 2026), sparse expert-based integration in FTS-Text-MoE (Xu et al., 2025b), and cross-modal representation alignment with pretrained language models in CALF (Liu et al., 2025). ChatTime (Wang et al., 2025a) models numerical and textual information within a unified foundation model, while Multi-Modal Forecaster (Kim et al., 2024) jointly predicts time-series and textual outputs from aligned multimodal inputs. These studies demonstrate the benefit of incorporating contextual information beyond numerical observations. However, they primarily focus on how heterogeneous modalities should interact, while differences in modality learning dynamics and temporal-scale relevance remain less jointly explored. Our framework addresses these two dimensions by mitigating modality imbalance and adaptively modeling temporal information across multiple resolutions.

2.2 Modality Imbalance

Modality imbalance arises when differences in predictive strength and learning dynamics cause stronger modalities to dominate joint optimization (Wang et al., 2020; Huang et al., 2022). Existing methods rebalance modality-specific learning through gradient or optimization control, including OGM-GE (Peng et al., 2022), AGM (Li et al., 2023), CGGM (Guo et al., 2024), PMR (Fan et al., 2023), ReconBoost (Hua et al., 2024), and LFM (Yang et al., 2024). A complementary line preserves unimodal capability through unimodal supervision. UMT (Du et al., 2023) transfers representations from independently trained unimodal teachers, while later methods combine teacher guidance with curriculum learning, gradient coordination, or modality-aware optimization (Zhou et al., 2023; Wei & Hu, 2024; Rakib & Bagavathi, 2025). However, teacher-based approaches generally transfer unimodal representations without separating information shared across modalities from modality-specific information. Our decomposed unimodal teacher explicitly separates these components while preserving task-relevant unimodal knowledge.

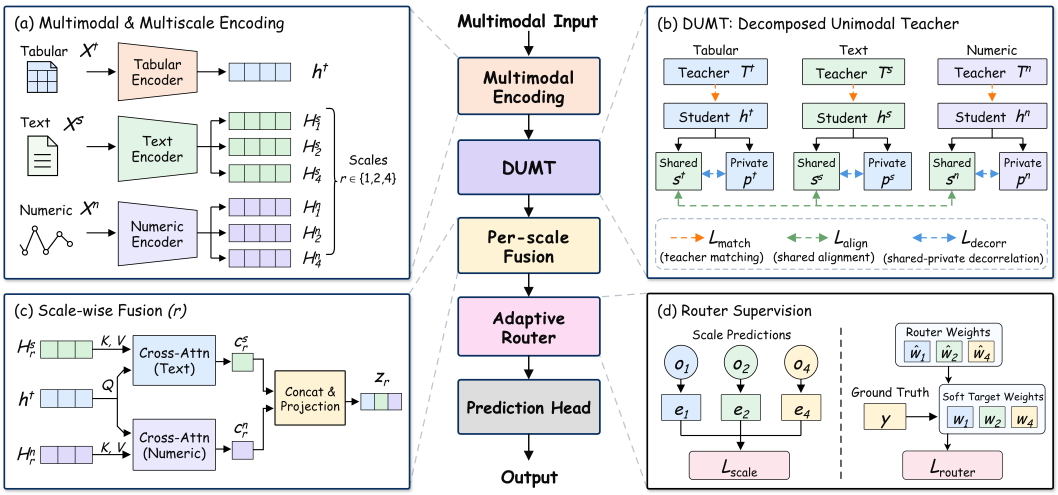


Figure 1: Overall architecture of our approach for long-horizon prediction from irregular multimodal behavior data. (a) Modality-specific encoding constructs multi-scale textual and numerical representations alongside the tabular representation. (b) DUMT regularizes modality learning through teacher matching, shared alignment, and shared–private decorrelation. (c) Tabular-guided cross-attention performs multimodal fusion independently at each temporal scale. (d) Scale-specific prediction errors provide supervision for adaptive routing.

2.3 Multi-Scale Temporal Modeling

Multi-scale temporal modeling has evolved from explicit multi-resolution modeling (Liu et al., 2022; Challu et al., 2023; Zhang & Yan, 2023; Wang et al., 2023) toward cross-scale interaction and adaptive scale selection. Scaleformer (Shabani et al., 2023) iteratively refines forecasts across multiple temporal scales with shared model parameters, while MG-TSD (Fan et al., 2024) uses coarse-grained targets at different resolutions to guide the learning process. TimeMixer (Wang et al., 2024) further models interactions between fine- and coarse-grained temporal patterns through multiscale mixing, and Pathformer (Chen et al., 2024) introduces adaptive pathways that dynamically adjust scale usage according to the input series. These methods demonstrate the benefit of multi-resolution modeling, but primarily operate on numerical time series and do not jointly address modality imbalance. In contrast, we first construct a multimodal representation within each temporal scale and then perform error-guided adaptive routing across scales, while separately regularizing modality-specific learning.

3 Methodology

Given static tabular features $\mathbf{X}^t$, a textual sequence $\mathbf{X}^s$, and a numerical sequence $\mathbf{X}^n$, we address two challenges in long-horizon multimodal prediction. First, tabular, textual, and numerical modalities can differ substantially in predictive strength and learning dynamics, causing stronger modalities to dominate joint optimization. Second, long behavioral sequences contain useful patterns at different temporal resolutions, while different samples may benefit from different resolutions. We therefore combine decomposed unimodal teacher supervision with scale-wise multimodal fusion and adaptive scale routing, together with direct supervision for each temporal scale. An overview is shown in Fig. 1.

3.1 Multi-Scale Multimodal Backbone

Modality-specific encoding. We adopt a three-tower architecture to preserve the distinct structures of tabular, textual, and numerical inputs before multimodal interaction. The tabular features are processed by an FTT-style Transformer (Gorishniy et al., 2021), and the output of its learnable

classification token is used as the tabular representation

$$\mathbf{h}^t = E_t(\mathbf{X}^t). \tag{1}$$

The tabular branch is shared across all temporal resolutions, providing a common subject-level representation for cross-modal fusion and scale routing.

For the two sequential modalities $m \in \{s, n\}$, we construct token representations at multiple temporal resolutions. Let $\mathcal{R} = \{1, 2, 4\}$ denote the set of temporal scales, and let $D_r(\cdot)$ denote temporal average downsampling at scale r. The representation at each scale is encoded as

$$\mathbf{H}_r^m = E_m\big(P_m(D_r(\mathbf{X}^m)) + \mathbf{E}_{1:L_r}^{m,\mathrm{pos}} + \mathbf{e}_r^{\mathrm{scale}}\big), \qquad m \in \{s, n\},\ r \in \mathcal{R}, \tag{2}$$

where L_r is the sequence length at scale r, P_m is the modality-specific input projection, and $\mathbf{e}_r^{\mathrm{scale}}$ is a learnable scale embedding broadcast to all tokens at the corresponding resolution. The projection P_m and encoder E_m are shared across scales within each modality, while the scale embeddings allow the model to distinguish different temporal resolutions. Each scale is encoded independently in the prediction pathway, with cross-scale aggregation deferred to the routing stage.

Scale-wise multimodal fusion. We perform multimodal fusion separately at each temporal resolution, keeping representations from different scales separate before adaptive routing. For each scale r, the tabular representation $\mathbf{h}^t$ serves as a single query token to attend to the textual and numerical sequences at that resolution:

$$\mathbf{c}_r^m = \mathrm{CAF}_r^m\left(\mathbf{h}^t, \mathbf{H}_r^m\right), \qquad m \in \{s, n\}, \tag{3}$$

where CAF denotes a cross-attention fusion block with residual and feed-forward connections, and $\mathbf{c}_r^m \in \mathbb{R}^d$ denotes the resulting query-conditioned representation of modality m at scale r. Since cross-attention uses the single tabular representation as the query and the temporal tokens as keys and values, it summarizes the scale-specific sequence conditioned on the tabular state.

The textual and numerical representations are then concatenated within each temporal scale:

$$\mathbf{z}_r = [\mathbf{c}_r^s; \mathbf{c}_r^n]. \tag{4}$$

Thus, each $\mathbf{z}_r$ forms a multimodal representation at temporal resolution r. These scale-specific representations are retained separately and passed to the adaptive routing module.

3.2 Decomposed Unimodal Teacher

Joint multimodal optimization may leave weaker modalities insufficiently optimized when stronger modalities dominate training. We therefore use independently trained unimodal teachers to regularize modality-specific representations. Unlike direct unimodal distillation, which transfers each teacher representation as a whole, we decompose the transferred knowledge into a shared component that captures cross-modal consistency and a private component that preserves modality-specific information.

Let $\mathcal{M} = \{t, s, n\}$ denote the three modalities. For each modality $m \in \mathcal{M}$, we independently train and freeze a unimodal teacher. We denote the modality-level representations produced by the multimodal student and the corresponding unimodal teacher as $\mathbf{u}_S^m$ and $\mathbf{u}_T^m$, respectively. For the student model, the tabular representation is directly taken from the tabular encoder, while each sequential representation is obtained by mean-pooling within each temporal scale and then averaging across scales:

$$\mathbf{u}_S^m = \begin{cases} \mathbf{h}^t, & m = t, \\ \dfrac{1}{|\mathcal{R}|}\displaystyle\sum_{r\in\mathcal{R}} \frac{1}{L_r}\sum_{\ell=1}^{L_r} \mathbf{H}_{r,\ell}^m, & m \in \{s, n\}. \end{cases} \tag{5}$$

Thus, DUMT operates on modality-level representations after multi-scale encoding but before cross-modal fusion, separating modality-level regularization from scale selection.

Each student representation is projected into shared and private channels:

$$\mathbf{h}_{\mathrm{sh}}^m = f_{\mathrm{sh}}^m\left(\mathbf{u}_S^m\right), \qquad \mathbf{h}_{\mathrm{pr}}^m = f_{\mathrm{pr}}^m\left(\mathbf{u}_S^m\right), \tag{6}$$

where $\mathbf{h}_{\mathrm{sh}}^m, \mathbf{h}_{\mathrm{pr}}^m \in \mathbb{R}^d$ and d denotes the representation dimension. We adopt an additive decomposition and preserve the task-relevant knowledge of the corresponding unimodal teacher through

$$\mathcal{L}_{\mathrm{DUMT}} = \sum_{m \in \mathcal{M}} \left\| \mathbf{u}_T^m - \left(\mathbf{h}_{\mathrm{sh}}^m + \mathbf{h}_{\mathrm{pr}}^m\right) \right\|_2^2 . \tag{7}$$

This additive formulation allows task-relevant unimodal knowledge to be distributed between the two complementary channels while jointly reconstructing the teacher representation.

To encourage the shared channels to capture cross-modal consistent information, we apply

$$\mathcal{L}_{\mathrm{CL}} = \sum_{m \neq m'} \mathcal{L}_{\mathrm{InfoNCE}} \left(\mathbf{h}_{\mathrm{sh}}^m, \mathbf{h}_{\mathrm{sh}}^{m'} \right), \tag{8}$$

$$\mathcal{L}\mathrm{corr} = \frac{1}{\|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{1}{d^2} \left| \mathrm{Corr}\left(\mathbf{h}_{\mathrm{sh}}^m, \mathbf{h}_{\mathrm{pr}}^m\right) \right|_F^2, \tag{9}$$

where $\mathrm{Corr}(\mathbf{h}_{\mathrm{sh}}^m, \mathbf{h}_{\mathrm{pr}}^m)$ is a $d \times d$ matrix whose (i, j)-th entry is the Pearson correlation between the i-th shared dimension and the j-th private dimension of modality m, computed over samples in the minibatch, and $\| \cdot \|_F$ denotes the Frobenius norm.

Together, $\mathcal{L}_{\mathrm{DUMT}}$ preserves task-relevant unimodal knowledge, $\mathcal{L}_{\mathrm{CL}}$ promotes cross-modal consistency in the shared channel, and $\mathcal{L}_{\mathrm{corr}}$ discourages the private channel from redundantly encoding shared information. The decomposition heads are used only during training and introduce no additional inference-time branch.

3.3 Error-Aware Adaptive Scale Routing

Building on the scale-specific multimodal representations, we adaptively combine temporal resolutions according to their sample-dependent predictive utility. Since adaptive weighting can provide uneven task supervision across scales, we maintain an independent predictor for each scale and explicitly supervise its prediction before learning error-aware routing across scales.

Scale-aware routing. Each scale-specific representation produces an independent task prediction $\mathbf{o}_r = f_r(\mathbf{z}_r)$. To determine the contribution of each scale, the router projects the tabular state and each scale-specific representation into a common routing space and evaluates their compatibility:

$$a_r = \mathrm{Score}\big(\mathrm{TabProj}(\mathbf{h}^t) + \mathrm{ScaleProj}(\mathbf{z}_r)\big), \qquad w_r = \mathrm{Softmax}_r(a_r), \tag{10}$$

where $\mathrm{TabProj}$ and $\mathrm{ScaleProj}$ are lightweight projection networks and Score is a shared scalar scoring network across scales. The final prediction is obtained by weighting the scale-specific predictions:

$$\mathbf{o} = \sum_{r \in \mathcal{R}} w_r \mathbf{o}_r. \tag{11}$$

This formulation allows the router to compare each scale-specific multimodal representation against the tabular state rather than deriving all scale weights from an aggregated scale representation.

For multi-horizon forecasting, each scale r produces an H-step forecast, where $\mathbf{o}_{j,r}$ denotes its prediction at step j. To allow the router to use different scale weights at different forecast steps, we introduce a learnable forecast-step embedding $\mathbf{e}_j^{\mathrm{step}}$ and compute

$$a_{j,r} = \mathrm{Score}\big(\mathrm{TabProj}\big(\mathbf{h}^t + \mathbf{e}_j^{\mathrm{step}}\big) + \mathrm{ScaleProj}(\mathbf{z}_r)\big), \qquad w_{j,r} = \mathrm{Softmax}_r(a_{j,r}), \tag{12}$$

yielding the prediction at forecast step j as

$$\mathbf{o}_j = \sum_{r \in \mathcal{R}} w_{j,r} \mathbf{o}_{j,r}, \qquad j = 1, \ldots, H. \tag{13}$$

Here, $\mathbf{e}_j^{\mathrm{step}}$ indicates the current forecast step, allowing the router to assign different scale weights across future steps. Thus, binary classification uses one scale distribution per sample, whereas multi-horizon forecasting learns a separate scale distribution for each forecast step.

Scale-specific prediction supervision. Although the final task loss jointly optimizes the scale-specific predictors and the router, the gradient propagated to each scale-specific predictor through the weighted output is modulated by its routing weight. Consequently, a low-weight scale may receive insufficient supervision and remain under-optimized. We therefore directly apply the task loss to the prediction from every scale. For each scale r, we define its task error as

$$e_r = \ell_{\text{task}}\left(\mathbf{o}_r, y\right), \tag{14}$$

and optimize

$$\mathcal{L}_{\text{scale}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} e_r. \tag{15}$$

Unlike the supervision propagated through the weighted prediction, $\mathcal{L}_{\text{scale}}$ is independent of the current routing weights and directly maintains the predictive capability of every temporal scale. For multi-horizon forecasting, the same supervision is applied independently at each forecast step and averaged over the prediction horizon.

Error-guided router supervision. Direct scale-specific supervision maintains predictive capability at each temporal resolution, but it does not explicitly tell the router which scale is more reliable for a particular sample. The final task loss only evaluates the combined prediction, so it does not directly reveal how well each scale performs on the current sample or which scales should receive more weight. We therefore use scale-specific prediction errors to guide the router, assigning larger weights to scales that perform better on the current sample.

During training, we use the scale-specific prediction errors to construct a soft reliability target:

$$p_r^{\text{err}} = \frac{\exp\left(-\text{sg}(e_r)/\tau\right)}{\sum_{r' \in \mathcal{R}} \exp\left(-\text{sg}(e_{r'})/\tau\right)}. \tag{16}$$

where τ is a temperature parameter and $\text{sg}(\cdot)$ denotes stop-gradient. Smaller prediction errors receive larger target weights, while τ controls the softness of the resulting scale distribution. The router is then aligned with this error-derived target through

$$\mathcal{L}_{\text{router}} = D_{\text{KL}}\left(\mathbf{p}^{\text{err}} \| \mathbf{w}\right). \tag{17}$$

For multi-horizon forecasting, the scale errors, error-derived targets, and routing losses are computed independently for each forecast step and then averaged over the horizon. Since scale-specific predictions are initially unstable, $\mathcal{L}_{\text{router}}$ is activated only after a short warm-up period. The error-derived target is used only during training, and inference requires neither ground-truth labels nor additional supervision.

3.4 Training Objective

The core training objective combines the task loss with modality-level regularization and scale-level supervision:

$$\mathcal{L}_{\text{core}} = \mathcal{L}_{\text{task}} + \lambda_{\text{D}}\mathcal{L}_{\text{DUMT}} + \lambda_{\text{C}}\mathcal{L}_{\text{CL}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{S}}\mathcal{L}_{\text{scale}} + \lambda_{\text{R}}\mathcal{L}_{\text{router}}. \tag{18}$$

For binary classification, $\mathcal{L}_{\text{task}}$ is the binary classification loss and no additional task-specific regularization is introduced.

For multi-horizon forecasting, we additionally encourage neighboring forecast steps to exhibit smooth scale preferences:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{H-1} \sum_{j=1}^{H-1} \left\| \mathbf{w}_{j+1} - \mathbf{w}_j \right\|_2^2. \tag{19}$$

Following prior multi-granularity forecasting objectives (Fan et al., 2024), we further employ a lightweight coarse-grained auxiliary loss $\mathcal{L}_{\text{aux}}$ by aggregating future targets over 4- and 8-step blocks, providing complementary supervision for longer-range temporal structure. The forecasting objective is

$$\mathcal{L}_{\text{forecast}} = \mathcal{L}_{\text{core}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}. \tag{20}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Tasks. We evaluate our method on a proprietary insurance dataset containing approximately 300K users and two public financial forecasting benchmarks. For health risk assessment, multimodal behavioral records are used to predict whether a user will file a critical-illness claim. Each user is represented by approximately 400 tabular features together with textual and numerical behavioral sequences derived from timestamped expenditure activities. To evaluate multi-scale modeling under increasingly long historical contexts, we use look-back windows of 26, 52, 78, and 104 weeks. We further evaluate on the S&P 500 and HS300 datasets under the FinMultiTime benchmark (Xu et al., 2025a), which similarly provides tabular, textual, and numerical modalities. The forecasting horizon is fixed at 24 trading days, with historical input lengths of 72, 96, 120, and 144 trading days.

Baselines. We compare against two groups of baselines. The first includes five multimodal time-series methods: Time-MMD (Liu et al., 2024), TaTS (Li et al., 2026), SpecTF (Nguyen et al., 2026), FTS-Text-MoE (Xu et al., 2025b), and CALF (Liu et al., 2025), covering prediction-level fusion, joint multimodal modeling, frequency-domain interaction, mixture-of-experts, and cross-modal alignment. The second includes two multi-scale methods, TimeMixer (Wang et al., 2024) and Pathformer (Chen et al., 2024), which model cross-scale interactions and adaptive multi-scale pathways, respectively. We adapt the methods to the same multimodal setting for comparison.

Implementation Details. For the internal insurance dataset, approximately 400 features are used for the tabular modality after preprocessing. For the textual modality, behavioral records within each temporal interval are aggregated into text descriptions and encoded using the pretrained BGE-small-zh-v1.5 model (Xiao et al., 2024). For the numerical modality, more than 60 behavioral statistics, including event frequencies and monetary amounts, are aggregated over temporal intervals and projected into the common hidden space. All modality representations use a hidden dimension of 512, and the temporal scales are set to $\mathcal{R} = \{1, 2, 4\}$. The sequence encoders are shared across scales, whereas the cross-attention modules and prediction heads are scale-specific.

We optimize the model using AdamW with a learning rate of 5×10^{-5} and a weight decay of 10^{-4}. The loss coefficients are set to $\lambda_{\mathrm{D}} = 10$, $\lambda_{\mathrm{C}} = 0.5$, $\lambda_{\mathrm{corr}} = 5.0$, $\lambda_{\mathrm{S}} = 0.1$, and $\lambda_{\mathrm{R}} = 0.1$. The error-guided router uses a target temperature of $\tau = 0.1$, with router supervision activated after a three-epoch warm-up. For financial forecasting, we use the same backbone configuration and additionally set $\lambda_{\mathrm{smooth}} = 0.01$ and $\lambda_{\mathrm{aux}} = 0.2$ for the forecasting-specific objectives. Additional experimental details are provided in Appendix D.

Evaluation Metrics. We use ROC-AUC as the primary metric for critical-illness claim risk prediction. ROC-AUC measures ranking quality across all thresholds, serving as a stable indicator of discriminative ability; in insurance risk control, this reflects the model's capacity to distinguish users who will file a claim from those who will not. Since claim events are inherently stochastic and available features capture only part of the risk information, AUC gains are typically modest but meaningful for risk prioritization. For financial forecasting, we follow FinMultiTime (Xu et al., 2025a) and report MAE and MSE. All results are evaluated on the held-out test sets.

4.2 MAIN RESULTS

We evaluate our method under increasing historical input lengths on two prediction tasks: binary health-risk classification on the internal insurance dataset and 24-day forecasting on public financial benchmarks. As shown in Table 1, our method consistently achieves the best performance on the internal insurance dataset across all input lengths, with ROC-AUC improving from 0.6602 at 26 weeks to 0.6651 at 104 weeks. The disease-specific results in Appendix C further show benefits from longer historical observations across the evaluated disease categories, including an absolute ROC-AUC gain of 0.0255 for autoimmune disease. These results suggest that the proposed multi-scale modeling can better exploit informative long-range behavioral patterns. A similar advantage is observed on the financial benchmarks, where our method achieves the best MAE and MSE across the reported input lengths and remains comparatively stable across different historical contexts.

Table 1: Predictive performance under different input horizons. Insurance horizons are measured in weeks, while financial horizons are measured in trading days.

Dataset	Metric	Horizon	SpecTF	TaTS	Time-MMD	FTS-Text-MoE	CALF	Path-former	Time-Mixer	**Ours**
Insurance	ROC-AUC ↑	26	0.6423	0.6112	0.6494	0.6369	0.6251	0.6218	0.6288	**0.6602**
		52	0.6424	0.6112	0.6512	0.6361	0.6269	0.6292	0.6397	**0.6610**
		78	0.6427	0.6110	0.6530	0.6374	0.6267	0.6269	0.6389	**0.6632**
		104	0.6428	0.6158	0.6533	0.6376	0.6254	0.6313	0.6373	**0.6651**
S&P 500	MAE ↓	72	0.3441	0.2815	0.3647	0.3167	0.2835	0.2899	0.2919	**0.2692**
		96	0.3502	0.2850	0.3652	0.3249	0.2921	0.3005	0.2989	**0.2697**
		120	0.3482	0.2878	0.3699	0.3144	0.3091	0.3135	0.3047	**0.2678**
		144	0.3484	0.2888	0.3710	0.3163	0.3147	0.3194	0.3123	**0.2678**
	MSE ↓	72	0.4095	0.3062	0.4241	0.3477	0.3551	0.3400	0.3411	**0.3024**
		96	0.4285	0.3097	0.4295	0.3622	0.3725	0.3564	0.3513	**0.3050**
		120	0.4259	0.3122	0.4352	0.3478	0.4132	0.3780	0.3643	**0.3030**
		144	0.4240	0.3127	0.4333	0.3505	0.4195	0.3860	0.3726	**0.3044**
HS300	MAE ↓	72	0.3071	0.2399	0.3094	0.2776	0.2278	0.2347	0.2424	**0.2211**
		96	0.3062	0.2439	0.3169	0.2780	0.2361	0.2482	0.2415	**0.2157**
		120	0.3048	0.2500	0.3103	0.2754	0.2457	0.2542	0.2523	**0.2186**
		144	0.3106	0.2517	0.3107	0.2727	0.2539	0.2605	0.2542	**0.2200**
	MSE ↓	72	0.2666	0.1619	0.2762	0.2064	0.1824	0.1808	0.1938	**0.1564**
		96	0.2632	0.1651	0.2839	0.2048	0.1973	0.2016	0.1919	**0.1517**
		120	0.2710	0.1696	0.2721	0.1987	0.2090	0.2101	0.2087	**0.1541**
		144	0.2855	0.1708	0.2729	0.1976	0.2213	0.2195	0.2108	**0.1556**

4.3 Ablation Studies

Decomposed Unimodal Teacher. On the internal insurance dataset, decomposition alone performs similarly to UMT (0.6622 vs. 0.6623), indicating limited benefit from decomposition. Adding contrastive learning improves ROC-AUC to 0.6645, while applying it without decomposition or distillation drops performance to 0.6504, showing that cross-modal alignment is most effective within the decomposed teacher-guided representation. Adding the correlation constraint further improves ROC-AUC to 0.6651. HS300 results are reported in Appendix A.

Table 2: DUMT ablation on the insurance dataset.

Variant	S/P Decomp.	Distill.	CL	Corr.	ROC-AUC ↑
UMT		✓			0.6623
CL only			✓		0.6504
Decomp.	✓	✓			0.6622
+ CL	✓	✓	✓		0.6645
DUMT	✓	✓	✓	✓	**0.6651**

Modality Learning. We assess modality-specific representations by freezing each encoder and training a lightweight MLP probe. As shown in Table 3, multimodal training reduces the mean ROC-AUC from 0.6377 to 0.6313, while UMT largely recovers this degradation. DUMT further improves all three modalities and achieves the highest mean ROC-AUC of 0.6416, showing that decomposed teacher supervision better preserves modality-specific predictive information.

Table 3: Unimodal probe performance on the insurance dataset.

Method	Tabular	Text	Numerical	Mean
Uni-Modal Model	0.6300	0.6512	0.6318	0.6377
MM Baseline	0.6228	0.6451	0.6261	0.6313
UMT	0.6297	0.6529	0.6322	0.6383
DUMT	**0.6329**	**0.6552**	**0.6367**	**0.6416**

Decompostion Analysis. To examine the task-relevant information retained by the decomposed representations, we freeze the learned representations and train a lightweight MLP on the shared, private, and joint representations. As shown in Table 4, the joint representation achieves the best performance, with the highest mean ROC-AUC of 0.6416.

Table 4: Performance of decomposed representations on the insurance dataset.

Repr.	Tabular	Text	Numerical	Mean
Shared	0.6306	0.6490	0.6356	0.6384
Private	0.6320	0.6548	0.6348	0.6405
Joint	**0.6322**	**0.6554**	**0.6372**	**0.6416**

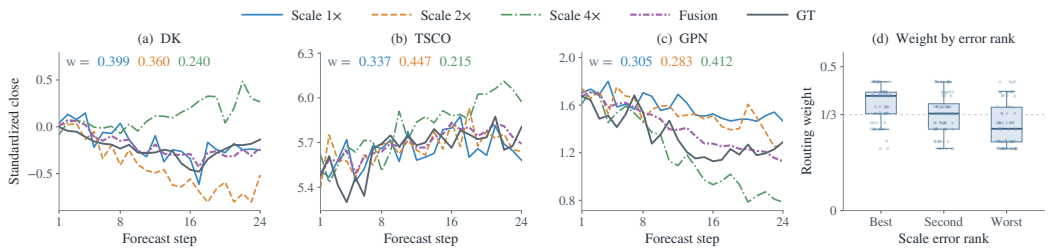


Figure 3: Visualization of adaptive scale routing on the S&P 500 dataset. Panels (a)–(c) show representative 24-step forecasts with different dominant temporal scales. Panel (d) groups routing weights by scale-specific prediction-error rank, where lower-error scales tend to receive larger weights.

Adaptive Scale Routing. We evaluate multi-scale modeling and the two scale-level supervision objectives on the internal insurance dataset. As shown in Table 5, introducing multiple temporal resolutions improves ROC-AUC from 0.6642 to 0.6645 over the single-scale setting. Applying either $\mathcal{L}_{\text{scale}}$ or $\mathcal{L}_{\text{router}}$ alone yields limited additional benefit, whereas combining both achieves the best ROC-AUC of 0.6651. This suggests that maintaining scale-specific predictive capability and using prediction errors to guide routing are complementary. Additional HS300 results are provided in Appendix B.

We further examine the contribution of different temporal resolutions on the internal insurance dataset. As shown in Fig. 2, among the single-scale variants, the $1\times$ resolution performs best, while performance decreases as the resolution becomes coarser. Combining the $1\times$, $2\times$, and $4\times$ scales achieves the best ROC-AUC of 0.6651, outperforming every individual scale. This suggests that fine-grained patterns remain important, while coarser resolutions provide complementary information when combined with the fine-grained representation.

Table 5: Ablation of multi-scale modeling and routing supervision on the insurance dataset.

Variant	Multi-scale	$\mathcal{L}_{\text{scale}}$	$\mathcal{L}_{\text{router}}$	ROC-AUC ↑
Single-scale				0.6642
Multi-scale	✓			<u>0.6645</u>
w/o $\mathcal{L}_{\text{router}}$	✓	✓		0.6644
w/o $\mathcal{L}_{\text{scale}}$	✓		✓	<u>0.6645</u>
Full model	✓	✓	✓	**0.6651**

Visualization of Scale Selection We analyze scale-selection behavior on the S&P 500 dataset using 24-step forecasting examples, as shown in Fig. 3. Panels (a)–(c) show three stocks whose largest routing weights correspond to the $1\times$, $2\times$, and $4\times$ scales, respectively, indicating that the preferred temporal resolution varies across samples rather than collapsing to a fixed scale. Panel (d) groups routing weights by the prediction-error rank of each scale. Lower-error scales tend to receive larger routing weights, while the worst-performing scales receive smaller weights. These results support the effectiveness of multi-scale modeling and error-aware supervision for adaptive scale allocation.

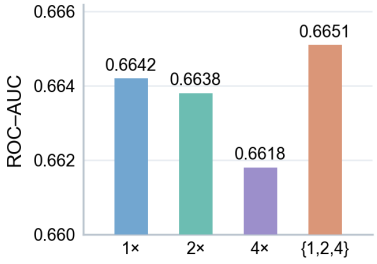


Figure 2: Performance with different temporal scales.

5 Conclusion

We study long-horizon prediction from multimodal and irregular behavioral data and propose a unified framework for jointly addressing modality imbalance and multi-scale temporal modeling. By combining decomposed unimodal teacher regularization with adaptive multi-scale modeling, our method preserves modality-specific information while dynamically exploiting temporal patterns at different resolutions. Experiments on large-scale insurance data and public financial benchmarks consistently demonstrate its effectiveness across classification and multi-horizon forecasting tasks.

Reproducibility statement

We provide detailed descriptions of the proposed framework, model architecture, training objectives, optimization procedure, evaluation protocols, and key implementation settings in the main paper and Appendix D. For the public financial experiments, we use the S&P 500 and HS300 datasets from FinMultiTime (Xu et al., 2025a) and provide the data construction, training, and evaluation settings adopted in our study to facilitate reproduction on publicly accessible data. The insurance dataset used in this study is proprietary and contains sensitive internal records, and therefore cannot be publicly released due to privacy and confidentiality restrictions. Due to institutional restrictions associated with the proprietary insurance setting, the corresponding internal implementation is also not publicly released. Nevertheless, the methodological and experimental details provided in the paper are intended to support independent implementation and reproduction of the proposed approach where the required data are available.

AI Use Statement

In this work, generative AI tools were used only for language editing and improving the clarity and readability of the manuscript. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of the paper.

References

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramírez, Max Mergenthaler Canseco, and Artur Dubrawski. NHITS: Neural hierarchical interpolation for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6989–6997, 2023. doi: 10.1609/aaai.v37i6.25854.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=lJkOCMP2aW.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8632–8656. PMLR, 2023. URL https://proceedings.mlr.press/v202/du23e.html.

Xinyao Fan, Yueying Wu, Chang Xu, Yuhao Huang, Weiqing Liu, and Jiang Bian. MG-TSD: Multi-granularity time series diffusion models with guided learning process. In *The Twelfth International Conference on Learning Representations*, 2024.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038. IEEE, 2023.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 133328–133344, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f0c68d99827dc09ed28aa073455efcbe-Abstract-Conference.html.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19573–19597. PMLR, 2024. URL https://proceedings.mlr.press/v235/hua24a.html.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259. PMLR, 2022.

Kai Kim, Howard Tsai, Rajat Sen, Abhimanyu Das, Zihao Zhou, Abhishek Tanpure, Mathew Luo, and Rose Yu. Multi-Modal Forecaster: Jointly predicting time series and textual data. *CoRR*, abs/2411.06735, 2024. URL https://arxiv.org/abs/2411.06735.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22157–22167. IEEE, 2023.

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*, volume 2026, pp. 24437–24484, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/29de5722d7d2a6946c08d0c0162a1c71-Abstract-Conference.html.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 77888–77933, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8e7768122f3eeec6d77cd2b424b72413-Abstract-Datasets_and_Benchmarks_Track.html.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. SCINet: Time series modeling and forecasting with sample convolution and interaction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5816–5828, 2022.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):18915–18923, 2025. doi: 10.1609/aaai.v39i18.34082.

Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41827–41836. PMLR, 2025.

Huu Hiep Nguyen, Minh Hoang Nguyen, Dung Nguyen, and Hung Le. Spectral text fusion: A frequency-aware approach to multimodal time-series forecasting. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4213–4221. PMLR, 2026. URL https://proceedings.mlr.press/v300/nguyen26d.html.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8237. IEEE, 2022.

Mohammed Rakib and Arunkumar Bagavathi. G2D: Boosting multimodal learning with gradient-guided distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4059–4068. IEEE, 2025.

Mohammad Amin Shabani, Amir H. Abdi, Lili Meng, and Tristan Sylvain. Scaleformer: Iterative multi-scale refining transformers for time series forecasting. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=sCrnllCtjoE.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):12694–12702, 2025a. doi: 10.1609/aaai.v39i12.33384.

Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. MICN: Multi-scale local and global context modeling for long-term series forecasting. In *International Conference on Learning Representations*, 2023.

Jun Wang, Fuyuan Cao, Zhixin Xue, Xingwang Zhao, and Jiye Liang. CMoB: Modality valuation via causal effect for balanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 91756–91778, 2025b.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52559–52572. PMLR, 2024. URL https://proceedings.mlr.press/v235/wei24d.html.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878.

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. *arXiv preprint arXiv:2506.05019*, 2025a. URL https://arxiv.org/abs/2506.05019.

Wenyan Xu, Dawei Xiang, Rundong Wang, Yonghong Hu, Liang Zhang, Jiayu Chen, and Zhonghua Lu. Learning explainable stock predictions with tweets using mixture of experts. *arXiv preprint arXiv:2507.20535*, 2025b. URL https://arxiv.org/abs/2507.20535.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62108–62122, 2024.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, and Wenwu Zhu. Intra- and inter-modal curriculum for multimodal learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3724–3735, 2023. doi: 10.1145/3581783.3612468.

A Ablation Study of Decomposed Unimodal Teacher

To further assess whether the benefit of DUMT generalizes beyond the internal insurance classification task, we conduct an additional ablation study on HS300 using a 96-day input window and a 24-day forecasting horizon.

Table 6: DUMT ablation on HS300 with a 96-day input window and a 24-day forecasting horizon.

Variant	MAE ↓	MSE ↓
UMT	<u>0.2164</u>	<u>0.1549</u>
CL only	0.2248	0.1559
Decomp.	0.2172	<u>0.1549</u>
Decomp. + CL	0.2199	0.1566
DUMT	**0.2157**	**0.1517**

As shown in Table 6, the full DUMT achieves the best performance on both MAE and MSE, indicating that its benefit is not limited to the insurance classification task. UMT already provides a strong baseline, suggesting that preserving unimodal knowledge remains useful for financial forecasting. In contrast, contrastive alignment alone performs worse, while decomposition without the complete objective provides only limited improvement. The full combination nevertheless achieves the lowest forecasting errors, suggesting that the complete DUMT objective is more effective than applying its individual components in isolation.

B Ablation Study of Adaptive Scale Routing

To further evaluate whether adaptive scale routing generalizes beyond the internal insurance task, we conduct an additional ablation study on HS300 using a 96-day input window and a 24-day forecasting horizon, examining the effects of multi-scale modeling, scale-specific supervision, and error-aware routing.

Table 7: Adaptive scale routing ablation on HS300 with a 96-day input window and a 24-day forecasting horizon.

Variant	MAE ↓	MSE ↓
Single-scale	<u>0.2181</u>	<u>0.1545</u>
Multi-scale	0.2203	0.1564
w/o $\mathcal{L}_{\text{router}}$	0.2205	0.1574
w/o $\mathcal{L}_{\text{scale}}$	0.2207	0.1568
Full model	**0.2157**	**0.1517**

As shown in Table 7, simply introducing multiple temporal resolutions does not improve over the single-scale model, indicating that multi-scale representations alone are insufficient for effective forecasting. Removing either $\mathcal{L}_{\text{scale}}$ or $\mathcal{L}_{\text{router}}$ also degrades performance, whereas the full model achieves the lowest MAE and MSE. These results suggest that maintaining the predictive capability of individual scales and using their prediction errors to guide scale weighting are both important for effectively exploiting multi-scale information.

C Disease-specific Analysis

To further investigate the effect of longer historical observations across disease categories, we conduct a disease-specific analysis on the insurance dataset. We examine eight disease categories covering diverse clinical characteristics and temporal patterns, and report the performance of all methods under four input horizons (26, 52, 78, and 104 weeks). The results are presented in Table 8.

Table 8: Disease-specific ROC-AUC performance under different input horizons on the insurance dataset.

Disease Category	Horizon	SpecTF	TaTS	Time-MMD	FTS-Text-MoE	CALF	Path-former	Time-Mixer	**Ours**
Severe Infection	26	0.6689	0.6271	0.6903	0.6648	0.6273	0.6113	0.6174	**0.6941**
	52	0.6571	0.6018	0.6738	0.6671	0.6252	0.6248	0.6452	**0.6968**
	78	0.6447	0.6122	0.6795	0.6547	0.6260	0.6102	0.6323	**0.6918**
	104	0.6471	0.6102	0.6854	0.6630	0.6094	0.6184	0.6386	**0.7091**
Autoimmune Disease	26	0.7133	0.6703	0.7301	0.6820	0.7001	0.6914	0.7213	**0.7331**
	52	0.7100	0.6723	**0.7311**	0.6864	0.7030	0.7064	0.7135	0.7188
	78	0.7096	0.6277	**0.7464**	0.6883	0.6938	0.6982	0.7183	0.7363
	104	0.7142	0.6434	0.7445	0.6813	0.6978	0.7232	0.7103	**0.7586**
Digestive	26	0.6795	0.6480	0.6797	0.6707	0.6646	0.6596	0.6616	**0.6868**
	52	**0.6819**	0.6391	0.6800	0.6676	0.6607	0.6634	0.6674	0.6742
	78	0.6759	0.6419	0.6810	0.6738	0.6607	0.6647	0.6673	**0.6861**
	104	0.6792	0.6501	0.6801	0.6689	0.6587	0.6618	0.6637	**0.6892**
Respiratory	26	0.6734	0.6317	0.6832	0.6765	0.6773	0.6625	0.6558	**0.6974**
	52	0.6712	0.6221	**0.6848**	0.6704	0.6698	0.6699	0.6790	0.6787
	78	0.6734	0.6255	0.6816	0.6732	0.6741	0.6664	0.6696	**0.6928**
	104	0.6726	0.6441	0.6875	0.6727	0.6649	0.6686	0.6740	**0.6990**
Musculoskeletal	26	0.6938	0.6238	0.6945	0.6895	0.6802	0.6783	0.6750	**0.7045**
	52	0.6957	0.6298	**0.7038**	0.6926	0.6697	0.6789	0.6905	0.6994
	78	0.6928	0.6454	0.7036	0.6926	0.6726	0.6804	0.6904	**0.7123**
	104	0.6962	0.6379	0.6970	0.6917	0.6637	0.6819	0.6840	**0.7113**
Dermatology & ENT	26	0.6487	0.6348	0.6617	0.6423	0.6214	0.6165	0.6271	**0.6657**
	52	0.6486	0.6269	0.6636	0.6408	0.6303	0.6248	0.6303	**0.6755**
	78	0.6527	0.6308	0.6633	0.6491	0.6279	0.6217	0.6420	**0.6724**
	104	0.6479	0.6394	0.6695	0.6446	0.6258	0.6251	0.6429	**0.6746**
Obstetrics & Neonatal	26	0.6936	0.6620	0.6959	0.6762	0.6959	0.6869	0.6979	**0.7105**
	52	0.6932	0.6413	0.7038	0.6812	0.6975	0.6902	0.6955	**0.7101**
	78	0.6911	0.6454	0.7014	0.6804	0.6918	0.6820	0.6987	**0.7085**
	104	0.6967	0.6525	0.7019	0.6764	0.6941	0.6879	0.6920	**0.7134**
Mild Malignancy	26	0.7386	0.6727	0.7331	0.7434	0.7395	0.7465	0.7536	**0.7560**
	52	0.7391	0.6785	0.7439	0.7337	0.7351	0.7498	**0.7589**	0.7405
	78	0.7486	0.6839	0.7495	0.7383	0.7416	0.7400	**0.7552**	0.7447
	104	0.7444	0.6776	0.7462	0.7404	0.7325	**0.7638**	0.7513	0.7564

Overall, our method achieves the best performance in most disease and horizon settings and shows a general benefit from incorporating longer historical observations. The 104-week setting consistently outperforms the 26-week setting across all eight disease categories, with the average ROC-AUC increasing from 0.7060 to 0.7140. The improvement is particularly pronounced for autoimmune disease, where the ROC-AUC increases from 0.7331 at 26 weeks to 0.7586 at 104 weeks, corresponding to an absolute gain of 0.0255. These results further demonstrate the benefit of longer historical observations across different disease categories.

D EXPERIMENTAL SETUP

For the internal insurance dataset, we use approximately 300,000 policyholders for the claim prediction task. Behavioral records from the two-year observation period preceding policy enrollment are organized into 104 weekly intervals and used to predict the subsequent binary claim outcome. The dataset is chronologically split according to enrollment time into training and test sets with a ratio of 7:3. All models are trained for 20 epochs using AdamW with a learning rate of 5×10^{-5}, with cross-entropy loss used for binary claim classification.

For the FinMultiTime experiments, we conduct forecasting independently on the S&P 500 and HS300 datasets. For each stock sequence, we consider input windows of 72, 96, 120, and 144 trading days and construct samples by sliding each window along the temporal sequence with a stride of 1. For all input settings, the prediction horizon is fixed at 24 trading days. After sample construction, the data of each stock are chronologically divided into training, validation, and test sets with a ratio of 7:1:2, thereby preserving the temporal ordering of the observations.

A separate model is trained for each stock. During evaluation, MSE and MAE are first computed on the test set of each stock, and the corresponding metrics are then averaged across all evaluated stocks within each dataset to obtain the final results for the S&P 500 and HS300 datasets, respectively. We use the same backbone configuration as in the main experiments. Each model is trained for up to 20 epochs using AdamW with a learning rate of 5×10^{-5} and a batch size of 128, with mean squared error (MSE) used as the forecasting loss. Early stopping is applied with a patience of five epochs based on validation performance.