Existence Precedes Value: Joint Modeling of Observational Existence and Evolving States in Time Series Forecasting

Yifan Hu[1,2,*], **Hongzhou Chen**[1,*], **Peiyuan Liu**[2,*], **Yiding Liu**[1], **Zewei Dong**[1,†], **Jiang-Ming Yang**[1]
[1] Ant International [2] Tsinghua University
{hyf476357, hongzhou.chz, zewei.dong}@ant-intl.com

Abstract

Real-world time series are often highly incomplete and irregular due to sensor dormancy, transmission delays, and event-driven sampling, making reliable forecasting fundamentally challenging. Existing methods have evolved from impute-then-forecast pipelines to continuous-time models such as Neural ODEs and continuous-time graph networks. While these approaches improve the modeling of historical irregularity, they still rely on an implicit oracle assumption at inference time: the timestamps of future valid observations are presumed to be known in advance. This assumption limits practical relevance, since in many real systems the more fundamental question is not only what the future value will be, but also whether a valid observation will occur at all. In this paper, we propose `Timeflies`, a unified framework that reformulates forecasting as a joint problem of future observability inference and value estimation. To explicitly model the interaction between observation dynamics and state evolution, `Timeflies` adopts an observation stream and a value stream, coupled through three dedicated modules for reliability-aware embedding, observation-guided dependency modeling, and joint prediction. We further construct `Shadow`, a benchmark that combines natural missingness from public datasets with real-world industrial data, and introduce the Observation-Value Joint Entropy (OVJE) metric to comprehensively evaluate this coupled predictability. Extensive experiments show strong joint forecasting performance and competitive value accuracy. The additional value-forecasting benefit depends on the observation mechanism and is not guaranteed under random or value-independent missingness. Code and dataset are available in https://github.com/ant-intl/Timeflies.

1 Introduction

"*Time flies over us, but leaves its shadow behind.*" — Nathaniel Hawthorne

In real-world time series systems, the generation of observations is often characterized by significant incompleteness and non-uniformity, due to factors such as sensor dormancy, transmission latencies, and event-driven selective sampling [1, 2, 3]. Despite these inherent irregularities, achieving precise forecasting remains mission-critical in these areas. To this end, time series forecasting (TSF) has evolved through two major paradigms as shown in Fig. 1. The first generation treated missing data as a defect to be eliminated, using "impute-then-forecast" strategies that forced data onto a rigid, uniform grid [4, 5]. To overcome these limitations, a second generation of models has emerged,

*Equal contribution. † Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

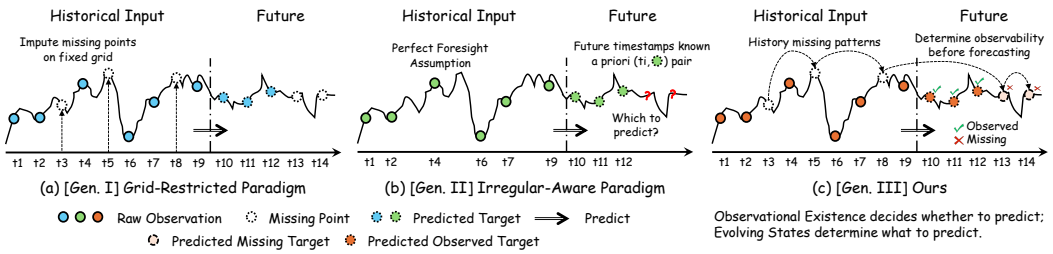

Figure 1: Evolution of Forecasting Paradigms for Time Series. (a) **Grid-Restricted Paradigm**: Treats missingness as a defect and employs a rigid impute-then-forecast approach for aligned time slots. (b) **Irregular-Aware Paradigm**: Models continuous latent trajectories from irregular historical observations. However, it requires prior knowledge of future valid timestamps during inference. (c) **Observation-State Joint Paradigm** (Ours): Elevates forecasting to a cascaded dual-track task. It explicitly decodes historical observational regularity to jointly predict future observability and value evolution. This formulation provides complementary information for downstream decisions.

including Neural Ordinary Differential Equations (Neural ODEs) and continuous-time graph networks. These frameworks treat time as a continuous variable, enabling state estimation at arbitrary timestamps and significantly improving the modeling of non-uniform historical sequences [6, 7].

However, a critical yet overlooked limitation remains: while second-generation models excel at fitting historical irregularities, they operate under the assumption of ***perfect foresight*** during inference. Specifically, these models assume that the timestamps of future valid observations are known in advance [8, 9]. In real-world systems, however, it is often unclear whether a valid observation will even occur at a specific future moment, as this depends on the underlying logic of the system [10, 11, 12]. Predicting a numerical value alone does not indicate whether a future measurement will be available to verify it; jointly predicting observability can therefore inform sensing and decision-making.

Missingness can provide a temporal signal reflecting the system's behavioral patterns. Fundamentally, real-world time series are driven by two coupled processes: ❶ ***Observational Existence***, which determine whether that state is captured as valid data, and ❷ ***Evolving States***, which dictate the evolution of physical or logical attributes. *Over time, the value of observational existence has been underestimated* [13, 14, 15, 16]. However, the physical manifestation of these dynamics, or the trajectory of historical missingness, is rarely random. Rather, it encodes structural priors such as periodic sensor cycles, load-driven sparsity, or state-triggered sampling. Consequently, historical missing patterns provide vital context for predicting the future observability of the system [17, 18, 19].

We therefore formulate forecasting as joint prediction of values and future observability. These complementary outputs address two questions:

① ***Future Observational Existence.*** Based on historical missing trajectories, the model transforms missingness from a noise factor into a structured prior for predicting future observations.

② ***Observation-Aware Evolving States.*** The model predicts future values using the available history, while the observation forecast characterizes whether those values will be measured.

By explicitly modeling the dependency between historical missingness and future observational states, the captured observational regularity can provide additional context for numerical regression when it carries state-relevant information [20, 21].

Based on the above motivation, we propose `Timeflies`, a unified forecasting framework that reformulates time series prediction as a joint process of existence inference and value estimation. To operationalize this dual-dynamic paradigm, our architecture explicitly captures these coupled processes through a symmetric design: a dedicated *value stream* and an *observation stream*. Specifically, the architecture is driven by three synergistic modules. *Reliability-Gated Patch Embedding* calculates a deterministic reliability score based on the observation ratio and missing intervals. It then explicitly attenuates sparse, noise-dominated patches, ensuring robust semantic-rich representations enter the subsequent network. *Cross-track Conditioned Transformer* introduces an observation-conditioned attention mechanism that explicitly incorporates learned historical missing patterns from the observation stream as a structural prior to guide the value stream. *Bifurcated Observation-Evolution Head*

simultaneously determines future observability probabilities and regresses expected state values under a joint optimized objective, aligning with our "existence precedes value" philosophy.

Furthermore, to rigorously validate the capability of our framework in realistic deployments, we curate `Shadow`, a benchmark combining natural missing gaps from open-source datasets with real-world industrial data. Crucially, moving beyond traditional marginal error metrics, we argue that a holistic evaluation must be grounded in a joint probability perspective. Therefore, we introduce the Observation-Value Joint Entropy (OVJE) metric, designed to explicitly quantify the synergy between accurate existence inference and precise state estimation. Evaluated under this unified criterion on the `Shadow` benchmark, `Timeflies` achieves strong aggregate joint performance, supporting future observability as an explicit forecasting target.

In a nutshell, our contributions are as follows:

- ***Joint Forecasting Formulation.*** We treat future observational existence as an explicit forecasting target alongside value prediction, relaxing the assumption that future valid timestamps are supplied. The observation forecast provides additional information; improvements in value accuracy depend on the observation mechanism.
- ***Technical advancement.*** We propose `Timeflies`, a framework that explicitly decomposes the time series into two interactive streams: *observation stream* and *value stream*. By extracting historical observation rhythms as dynamic context, the architecture enables the existence patterns to explicitly guide and refine the state predictions. The contribution is their task-specific coupling within a joint prediction formulation; attention, focal loss, and residual fusion are not individually new components.
- ***Empirical and Practical Utility.*** To bridge the evaluation gap in realistic scenarios, we construct `Shadow`, a rigorous benchmark combining open-source and real-world industrial datasets with natural missing patterns. Moreover, we propose the Observation-Value Joint Entropy (OVJE) metric to holistically evaluate the coupled capability of existence inference and value estimation. Experiments support strong aggregate joint performance, with data-dependent gains in value forecasting.

2 Related Work

For time series forecasting with missing values, one line of research follows a typical grid-restricted paradigm, which first aligns irregular or incomplete sequences onto a unified time grid and then performs forecasting on the reconstructed series [4, 22, 23]. In this setting, missing values are treated as defects to be removed (typically through interpolation or zero imputation), while the main effort is devoted to developing stronger forecasting architectures [24], including RNN-based [25, 26], CNN-based [27, 28, 29], MLP-based [30, 31, 32, 33, 34], GNN-based [35, 36, 37], and Transformer-based [4, 38, 39, 40, 41] models. While effective under regular sampling or mild missingness, these methods retain the same underlying assumption: missingness is treated only as a source of input degradation rather than as part of the forecasting target itself [4]. As a result, they still predict values for all future time steps and evaluate only on positions with available labels, effectively treating the future as a fully predictable uniform grid and leaving future observability unmodeled [42].

Beyond explicit imputation, another line of research moves away from discrete reconstruction and instead models irregular sequences directly in continuous-time or weakly aligned spaces [43], corresponding to the irregular-aware paradigm. A representative direction is based on Neural ODEs [44, 45, 46, 47], which are naturally suited to irregular sampling because they model hidden-state evolution as a continuous-time dynamical system and can propagate states over arbitrary time intervals without requiring observations to lie on a fixed grid. Other methods build on continuous-time graph networks [48, 49] and combine them with patch-based alignment [6, 7], where each patch covers the same time span but may contain a different number of valid observations, thereby relaxing strict point-wise alignment while preserving local temporal semantics. Although these methods substantially improve the modeling of historical irregularity, they still assume that the timestamps of valid future observations are known at inference time and evaluate prediction performance only on those pre-specified positions [48]. Therefore, while they move beyond the impute-then-forecast pipeline, they still do not model the generation of future observations themselves. In contrast, we explicitly treat future observability as a prediction target and jointly model whether a future observation will occur and what value it will take.

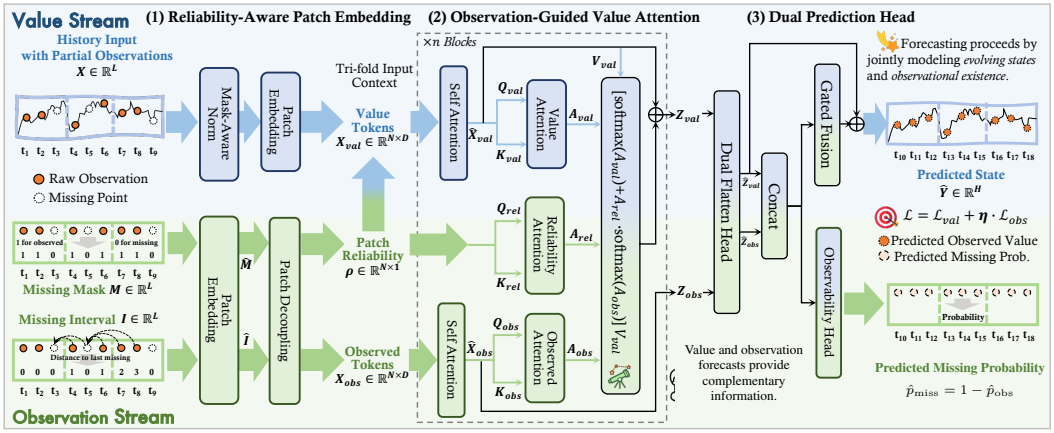

Figure 2: The overall architecture of `Timeflies`. (1) **Reliability-Aware Patch Embedding** refines value tokens by incorporating patch-level reliability based on observed patterns and missingness intervals. (2) **Observation-Guided Value Attention** integrates historical observation regularities into the attention mechanism to enhance value predictions. (3) **Dual Prediction Head** jointly predicts both future values and observation probabilities.

3 Methodology

In this section, we present `Timeflies` for jointly forecasting future values and their observability. We first formalize the problem by specifying the historical inputs, prediction targets, and probability conventions. We then introduce the dual-stream architecture and describe how its embedding, attention, and prediction modules are trained under a joint objective.

Conventional time series forecasting predicts a future sequence $Y \in \mathbb{R}^H$ from a historical input $X \in \mathbb{R}^L$, where L and H denote the input and prediction lengths, respectively. However, this formulation is incomplete for real-world settings with missing observations. Therefore, we move beyond the previous forecasting paradigm by jointly modeling future observability and future value. Specifically, in addition to the future value sequence Y, we introduce a future observation indicator sequence $O \in \{0,1\}^H$, where $O_t = 1$ indicates that a valid observation exists at time step t, and $O_t = 0$ otherwise. To infer future observability from historical missingness patterns, we further introduce two auxiliary inputs: a missing mask $M \in \{0,1\}^L$ and a missing interval sequence $I \in \mathbb{R}^L$. Here, $M_t = 0$ indicates that the value at position t is missing, and $M_t = 1$ indicates that it is observed. The missing interval is defined as $I_t = \log(1 + \delta_t)$, where δ_t denotes the distance from the current position to the most recent missing event; the logarithm is used to reduce scale variation and improve numerical stability. In summary, our model takes the triplet (X, M, I) as input and jointly predicts future values and observation probabilities $(\hat{Y}, \hat{p}_{\text{obs}})$. Throughout, $\hat{O}_t$ is an observation logit, $\hat{p}_{\text{obs},t} = \sigma(\hat{O}_t)$ is the probability of being observed, and $\hat{p}_{\text{miss},t} = 1 - \hat{p}_{\text{obs},t}$ is the probability of missingness.

3.1 Pipeline Overview

As shown in Fig. 2, `Timeflies` processes the inputs through two parallel streams. The value stream takes the historical value sequence X and learns its temporal evolution, generating the future value prediction $\hat{Y} \in \mathbb{R}^H$. In parallel, the observation stream processes the missingness mask M and missing intervals I to capture observation regularities, producing the future observation probability $\hat{p}_{\text{obs}} \in [0,1]^H$, where each element represents the probability of a valid observation at each time step. The entire framework consists of three key modules: (1) **Reliability-Aware Patch Embedding** divides (X, M, I) into patch tokens, extracting local observation patterns and generating value tokens enhanced with patch reliability information; (2) **Observation-Guided Value Attention** models long-range dependencies within each stream and modulates the attention mechanism of value stream using the patterns learned from the observation stream, ensuring that value predictions are informed by observation regularities; (3) **Dual Prediction Head** maps the high-level representations from

both streams to the future horizon, predicting both future values and observation probabilities, and incorporates observation information back into the value prediction through a gating mechanism.

3.2 Reliability-Aware Patch Embedding

In this stage, we first apply mask-aware normalization to $X \in \mathbb{R}^L$, using only the observed entries to compute the normalization statistics, which prevents distortion from missing values. We then divide X, M, and I into non-overlapping patches and map them to embeddings of the same dimensionality:

$$\hat{X}, \hat{M}, \hat{I} = \text{Patch_Embedding}(X, M, I), \tag{1}$$

where $\hat{X}, \hat{M}, \hat{I} \in \mathbb{R}^{N \times D}$, with N denoting the number of patches and D the embedding dimension.

For the observation stream, we encode missing positions and intervals using a gated fusion mechanism, which helps mitigate the challenge of directly combining features with different properties. The gated mechanism ensures that the information from missingness and intervals is appropriately modulated before integration, producing the observed token $\hat{X}_{\text{obs}} \in \mathbb{R}^{N \times D}$:

$$h_{\text{obs}} = \text{Linear}([\hat{M}; \hat{I}]), \quad \hat{X}_{\text{obs}} = h_{\text{obs}} + \sigma(h_{\text{obs}}) \odot \hat{I}, \tag{2}$$

where $[\cdot;\cdot]$ denotes concatenation, $\sigma(\cdot)$ is the Sigmoid function, and $\odot$ is element-wise product.

Directly fusing $\hat{X}_{\text{obs}}$ into $\hat{X}$ may inject noisy cues from highly sparse patches, where limited observations provide weak and unstable evidence for value modeling. To address this, we compute a patch-level reliability score $\rho \in \mathbb{R}^{N \times 1}$ based on observation sufficiency:

$$\rho = \text{Linear}([\text{Mean}(\hat{M}); \text{Mean}(\hat{I})]), \tag{3}$$

where $\text{Mean}(\hat{M})$ and $\text{Mean}(\hat{I})$ represent the missingness ratio and missing interval duration, respectively. The reliability score ρ reflects the trustworthiness of each patch's observation information. This score is then used to modulate the injection of observation information into the value embedding, resulting in the enhanced value token $\hat{X}_{\text{val}} \in \mathbb{R}^{N \times D}$:

$$\hat{X}_{\text{val}} = \hat{X} + \text{Linear}(\rho \odot \hat{X}_{\text{obs}}). \tag{4}$$

As a result, $\hat{X}_{\text{val}}$ retains the core semantics of the original value tokens while incorporating observation patterns in a reliability-aware manner.

3.3 Observation-Guided Value Attention

After obtaining $\hat{X}_{\text{val}}$ and $\hat{X}_{\text{obs}}$, we first apply self-attention within each stream to capture long-range dependencies across patches:

$$\hat{X}_{\text{val}} = \text{Attention}(\hat{X}_{\text{val}}, \hat{X}_{\text{val}}, \hat{X}_{\text{val}}), \quad \hat{X}_{\text{obs}} = \text{Attention}(\hat{X}_{\text{obs}}, \hat{X}_{\text{obs}}, \hat{X}_{\text{obs}}), \tag{5}$$

where the value stream $\hat{X}_{\text{val}}$ models temporal dependencies in the value sequence, while the observation stream $\hat{X}_{\text{obs}}$ captures long-range observation patterns.

However, modeling each stream independently is insufficient, as the reliability of future value predictions depends on historical observation patterns. Thus, the observation stream should not just be a supplementary branch but should influence the value stream's dependencies in a higher-level semantic space. To achieve this, we construct three attention maps:

$$A_{\text{val}} = \frac{Q_{\text{val}} \cdot K_{\text{val}}^{\top}}{\sqrt{D}}, \quad A_{\text{obs}} = \frac{Q_{\text{obs}} \cdot K_{\text{obs}}^{\top}}{\sqrt{D}}, \quad A_{\text{rel}} = \sqrt{\rho \cdot \rho^{\top}}, \tag{6}$$

where $Q_{\text{val}}, K_{\text{val}} \in \mathbb{R}^{N \times D}$ are derived from $\hat{X}_{\text{val}}$, and $Q_{\text{obs}}, K_{\text{obs}} \in \mathbb{R}^{N \times D}$ are derived from $\hat{X}_{\text{obs}}$. Accordingly, $A_{\text{val}} \in \mathbb{R}^{N \times N}$ captures semantic correlations between value tokens, $A_{\text{obs}} \in \mathbb{R}^{N \times N}$ captures structural correlations between observation tokens, and $A_{\text{rel}} \in \mathbb{R}^{N \times N}$ modulates the contribution of observation patterns according to patch reliability. We then combine these three maps into an observation-aware attention map:

$$A = \text{Softmax}(A_{\text{val}}) + A_{\text{rel}} \odot \text{Softmax}(A_{\text{obs}}). \tag{7}$$

This fusion retains the value dependencies as the primary structure while injecting observation patterns learned from the observation stream into the value attention. Moreover, the influence of the observation stream is amplified for patches with higher reliability and suppressed for sparse patches, thus preventing over-interference with value modeling while effectively leveraging historical missingness patterns in more reliable regions.

Finally, we update the value stream using the fused attention map:

$$Z_{\text{val}} = A \cdot V_{\text{val}} + \hat{X}_{\text{val}}, \quad Z_{\text{obs}} = \hat{X}_{\text{obs}}, \tag{8}$$

where V_{val} represents the value tokens derived from $\hat{X}_{\text{val}}$. Semantically, $Z_{\text{val}} \in \mathbb{R}^{N \times D}$ encodes value dynamics enhanced by historical observation patterns, while $Z_{\text{obs}} \in \mathbb{R}^{N \times D}$ preserves observation information for future observability prediction.

3.4 Dual Prediction Head

This module takes Z_{val} and Z_{obs} as the input and maps these representations from both streams to the prediction horizon. Specifically, we first apply a linear Flatten Head to obtain $\hat{Z}_{\text{val}}$ and $\hat{Z}_{\text{obs}} \in \mathbb{R}^H$:

$$\hat{Z}_{\text{val}} = \text{Flatten_Head}(Z_{\text{val}}), \quad \hat{Z}_{\text{obs}} = \text{Flatten_Head}(Z_{\text{obs}}). \tag{9}$$

Based on these two representations, we first produce observation logits $\hat{O} \in \mathbb{R}^H$ and probabilities:

$$\hat{O} = \text{Linear}([\hat{Z}_{\text{val}}; \hat{Z}_{\text{obs}}]), \quad \hat{p}_{\text{obs}} = \sigma(\hat{O}). \tag{10}$$

Both streams are used here because future observability is governed by not only historical observation patterns but also the underlying value dynamics. The predicted observability is then used to modulate the final value prediction $\hat{Y} \in \mathbb{R}^H$:

$$\hat{Y} = \hat{Z}_{\text{val}} + \hat{p}_{\text{obs}} \odot \hat{Z}_{\text{obs}}. \tag{11}$$

In this way, future observability serves as a soft gate on the contribution of observation features to value prediction, amplifying their effect on likely observed steps and attenuating it elsewhere.

3.5 Training Loss

The training objective consists of two components: the value prediction loss $\mathcal{L}_{\text{val}}$ and the observation prediction loss $\mathcal{L}_{\text{obs}}$. Specifically, we compute the regression loss only at positions with valid future observations and the classification loss over all prediction steps:

$$\mathcal{L}_{\text{val}} = \text{MSE}(\hat{Y}, Y), \quad \mathcal{L}_{\text{obs}} = \text{FocalBCE}(\hat{p}_{\text{obs}}, O). \tag{12}$$

Here, $\mathcal{L}_{\text{val}}$ is only evaluated on time steps $O_t = 1$, while $\mathcal{L}_{\text{obs}}$ uses Focal BCE Loss to address the class imbalance between observed and missing positions. The total loss function is expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{val}} + \eta \cdot \mathcal{L}_{\text{obs}}, \tag{13}$$

where η is a hyperparameter that balances the two tasks. By jointly optimizing these two objectives, the model learns to predict both the likelihood of future observations and, under the condition of valid observations, the corresponding future state values.

4 Experiments

4.1 `Shadow` Benchmark

Datasets Construction. To ensure comprehensive and realistic evaluation, `Shadow` contains 31 datasets collected from both public sources and real-world industrial systems. It includes 15 public datasets from GIFT-Eval [50], covering diverse domains with frequencies ranging from 5 minutes to 1 month, and 16 proprietary e-commerce datasets from a production online-retail platform, each recording hourly transaction volumes across different global regions.

Notably, the proprietary datasets contain strictly *natural missing values without artificial masking*. Their missingness is highly *non-random* and encodes structural patterns, such as regional off-peak inactivity, holiday effects, and pipeline outages. With missing ratios ranging from below 0.1% to above 89%, the full benchmark spans 6 domains and 16 geographical regions, providing a rigorous testbed for irregular time series forecasting. Table 1 summarizes all 31 dataset-frequency units; per-split missingness statistics are provided in App. A.1.

Table 1: Overview of all **Shadow** datasets. Freq.: sampling frequency; #Var: target variates; Pred Len (S/M/L): short/medium/long prediction lengths; # Obs: total observations. BB-FS/BB-rnd abbreviate Bitbrains-Fast Storage/Bitbrains-rnd; E-com. denotes E-commerce. Missing Regimes follows Table 5.

Shadow				Series Length						Missing Regimes			
Dataset	**Domain**	**Freq.**	**Series**	**Avg**	**Min**	**Max**	**# Obs**	**#Var**	**Pred Len (S / M / L)**	**No**	**Low**	**Med**	**High**
GIFT-Eval public datasets													
Electricity	Energy	1H	370	35,064	35,064	35,064	12,973,680	1	[48 / 480 / 720]	✓	✓	✓	✓
Electricity	Energy	1D	370	1,461	1,461	1,461	540,570	1	[30 / – / –]	✓	✓	✓	✓
Electricity	Energy	1W	370	208	208	208	76,960	1	[8 / – / –]	✓	✓	✓	✓
BB-FS	Cloud	5min	1,250	8,640	8,640	8,640	10,800,000	2	[48 / 480 / 720]	✓	✓	✓	✓
BB-FS	Cloud	1H	1,250	721	721	721	901,250	2	[48 / 480 / 720]	✓	✓	✓	✓
BB-rnd	Cloud	5min	500	8,640	8,640	8,640	4,320,000	2	[48 / 480 / 720]	–	✓	✓	✓
BB-rnd	Cloud	1H	500	720	720	720	360,000	2	[48 / – / –]	✓	✓	✓	✓
Car Parts	Retail	1M	2,674	51	51	51	136,374	1	[12 / – / –]	✓	–	–	✓
KDD Cup 2018	Air Qual.	1H	270	10,898	9,504	10,920	2,942,364	1	[48 / 480 / 720]	–	✓	✓	✓
KDD Cup 2018	Air Qual.	1D	270	455	396	455	122,791	1	[30 / – / –]	✓	✓	✓	✓
Restaurant	Commerce	1D	807	358	67	478	289,303	1	[30 / – / –]	✓	✓	✓	✓
Jena Weather	Weather	10min	1	52,704	52,704	52,704	52,704	21	[48 / 480 / 720]	–	✓	–	–
Jena Weather	Weather	1H	1	8,784	8,784	8,784	8,784	21	[48 / 480 / 720]	–	✓	–	–
Hierarchical Sales	Retail	1D	118	1,825	1,825	1,825	215,350	1	[30 / – / –]	–	✓	–	–
Temperature Rain	Weather	1D	32,072	725	725	725	23,252,200	1	[30 / – / –]	✓	✓	✓	–
Proprietary e-commerce datasets													
ecom_AD	E-com.	1H	26	11,838	9,769	19,968	307,793	1	[48 / 480 / 720]	–	✓	✓	–
ecom_AI	E-com.	1H	9	11,096	10,910	12,582	99,862	1	[48 / 480 / 720]	–	–	–	✓
ecom_BB	E-com.	1H	8	6,022	6,022	6,022	48,176	1	[48 / 480 / 720]	–	–	✓	–
ecom_BM	E-com.	1H	8	10,667	9,767	13,366	85,334	1	[48 / 480 / 720]	–	–	✓	–
ecom_BY	E-com.	1H	10	13,384	9,768	19,968	133,836	1	[48 / 480 / 720]	–	–	✓	–
ecom_HN	E-com.	1H	16	7,316	7,316	7,316	117,056	1	[48 / 480 / 720]	–	–	–	✓
ecom_HR	E-com.	1H	1	19,956	19,956	19,956	19,956	1	[48 / 480 / 720]	–	–	✓	–
ecom_JP	E-com.	1H	8	11,953	11,184	17,333	95,621	1	[48 / 480 / 720]	–	✓	–	–
ecom_ME	E-com.	1H	9	10,223	9,927	12,593	92,009	1	[48 / 480 / 720]	–	–	✓	✓
ecom_NG	E-com.	1H	6	12,920	12,920	12,920	77,520	1	[48 / 480 / 720]	–	–	✓	–
ecom_NO	E-com.	1H	16	6,024	6,016	6,032	96,384	1	[48 / 480 / 720]	–	✓	✓	–
ecom_NR	E-com.	1H	8	6,032	6,032	6,032	48,256	1	[48 / 480 / 720]	–	✓	–	–
ecom_PL	E-com.	1H	8	12,266	11,182	19,856	98,130	1	[48 / 480 / 720]	–	–	✓	–
ecom_SB	E-com.	1H	24	12,262	11,141	19,968	294,287	1	[48 / 480 / 720]	–	✓	–	✓
ecom_TA	E-com.	1H	8	6,032	6,032	6,032	48,256	1	[48 / 480 / 720]	–	✓	–	–
ecom_VN	E-com.	1H	16	11,842	9,769	13,370	189,474	1	[48 / 480 / 720]	–	✓	✓	–

Evaluation Protocol. To assess the performance of models under varying degrees of sparsity and predictive difficulty, we establish a multi-dimensional evaluation protocol along two main criteria:

- **Forecasting Horizon.** Aligning with GIFT-Eval [50], we evaluate models across short-, medium-, and long-term forecasting horizons, with dataset-specific prediction lengths detailed in App. A.1.
- **Sparsity Regime.** To decouple the impact of missingness from general forecasting complexity, we categorize the series into four distinct sparsity regimes based on their overall missing ratios: **no missing** (0%), **low** ($(0, 10\%]$), **medium** ($(10\%, 40\%]$), and **high** ($(40\%, 100\%]$).

Combining the dataset-frequency units and prediction horizons in Table 1 with the missingness groups in Table 5 defines a benchmark grid of 149 dataset-frequency–horizon–regime combinations. Regime labels refer to overall missingness; the missing ratios of individual splits may differ.

4.2 Experimental Setup

Baselines. We compare **Timeflies** against a comprehensive set of baselines, including Transformer-based methods: PatchTST [40], iTransformer [41], Crossformer [51] and FEDformer [38]; Linear-based methods: OLinear [32], WPMixer [52], TimeMixer [31] and DLinear [30]; CNN-based methods: FACT [29] and MICN [28]. All baselines are adapted to the missing-aware setting by receiving ***zero-filled*** inputs with accompanying observation masks.

Metrics. To evaluate the dual-stream pipeline, we report task-specific metrics and a joint criterion. For the observation stream, **ROC-AUC** measures discrimination, with observed positions ($O_t = 1$) as the positive class; it is undefined for single-class targets. For the value stream, **MSE** and **MAE** are

Table 2: Forecasting results on **Shadow** benchmark. For each missing-ratio regime, we first calculate the average of the test metrics across forecasting horizons for each dataset, and then report the arithmetic mean of these scores for all datasets within the same regime. See Tab. 7 for full results.

Shadow	High Missing				Medium Missing				Low Missing				No Missing		
Method	MSE↓	MAE↓	AUC↑	OVJE↓	MSE↓	MAE↓	AUC↑	OVJE↓	MSE↓	MAE↓	AUC↑	OVJE↓	MSE↓	MAE↓	OVJE↓
FEDformer	1.666	0.429	0.614	1.307	0.648	0.395	0.558	1.277	1.490	0.438	0.525	1.136	0.420	0.314	0.862
PatchTST	1.687	0.350	0.719	0.872	0.634	0.360	0.614	1.215	1.439	0.380	0.541	0.924	0.340	0.255	0.708
Crossformer	1.852	0.369	0.654	0.944	0.655	0.380	0.553	1.243	1.458	0.400	0.497	0.990	0.436	0.287	0.780
iTransformer	1.590	0.354	0.680	0.846	0.638	0.367	0.566	1.263	1.452	0.387	0.524	1.017	0.324	0.251	0.730
DLinear	1.714	0.383	0.638	1.069	0.611	0.367	0.528	1.328	1.433	0.387	0.507	1.245	0.371	0.286	1.020
TimeMixer	4.490	0.730	0.706	1.118	0.638	0.372	0.621	1.344	1.485	0.400	0.546	1.021	0.410	0.289	0.741
OLinear	1.622	0.328	0.740	0.762	0.608	0.352	0.623	1.475	1.438	0.376	0.525	0.958	0.317	0.247	0.695
WPMixer	1.601	0.342	0.578	0.930	0.615	0.349	0.536	1.202	1.442	0.377	0.500	1.255	0.331	0.253	0.947
MICN	8.191	0.888	0.676	1.472	7.253	1.190	0.657	1.468	13.961	1.399	0.561	1.194	17.814	1.685	0.984
FACT	1.644	0.334	0.730	0.823	0.617	0.352	0.581	1.176	1.461	0.383	0.551	0.927	0.307	0.240	0.639
Timeflies (Ours)	**1.546**	**0.303**	**0.805**	**0.611**	**0.555**	**0.285**	**0.704**	**0.914**	**1.362**	**0.349**	**0.588**	**0.762**	**0.296**	**0.229**	**0.477**

Table 3: Value forecasting performance across missingness regimes without the missingness classification head. Results are reported as two-stage arithmetic means, aggregated first over forecasting horizons within each dataset and then over datasets within each regime. See Tab. 8 for full results.

Shadow	High Missing		Medium Missing		Low Missing		No Missing	
Method	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓
FEDformer	1.688	0.408	0.657	0.384	1.496	0.443	0.423	0.316
PatchTST	1.766	0.337	0.650	0.354	1.446	0.372	0.351	0.259
Crossformer	1.909	0.369	0.659	0.368	1.587	0.440	0.434	0.291
iTransformer	1.611	0.348	0.634	0.353	1.462	0.379	0.326	0.251
DLinear	1.716	0.369	0.614	0.354	1.438	0.380	0.367	0.283
TimeMixer	1.935	0.452	0.636	0.354	1.456	0.381	0.396	0.283
OLinear	1.586	0.311	0.596	0.333	1.427	0.364	0.301	0.240
WPMixer	1.633	0.327	0.637	0.345	1.434	0.369	0.303	0.241
MICN	9.151	0.925	88.328	2.585	16.245	1.611	19.409	1.737
FACT	1.683	0.340	0.627	0.348	1.447	0.371	0.293	0.234
Timeflies (Ours)	**1.501**	**0.294**	**0.557**	**0.276**	**1.403**	**0.355**	**0.284**	**0.226**

computed only on ground-truth observed positions. We additionally report the Observation-Value Joint Entropy (**OVJE**) as a sample-aligned score of both outputs.

Let $o_t = O_t \in \{0, 1\}$ and $\hat{p}_t = \hat{p}_{\text{obs},t}$ denote the target observation indicator and its predicted probability. The implementation uses a symmetric relative error (sMAPE, without a factor of 100):

$$e_t = \frac{2|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t| + \epsilon}, \qquad q_t = \exp(-e_t), \qquad \epsilon = 10^{-8}. \tag{14}$$

Here q_t is a value-quality factor, not a normalized predictive density over y_t. We clip $\hat{p}_t$ to $[\epsilon, 1 - \epsilon]$ before taking logarithms. The joint score is

$$\begin{aligned} \text{OVJE} &= -\frac{1}{T}\sum_{t=1}^{T}\left[o_t \log(\hat{p}_t q_t) + (1 - o_t)\log(1 - \hat{p}_t)\right] \\ &= \underbrace{-\frac{1}{T}\sum_{t=1}^{T}\left[o_t \log \hat{p}_t + (1 - o_t)\log(1 - \hat{p}_t)\right]}_{\text{Bernoulli NLL}} + \underbrace{\frac{1}{T}\sum_{t:o_t=1} e_t}_{\text{observed-position sMAPE contribution}}. \end{aligned} \tag{15}$$

Both terms are averaged over all evaluated positions, not only observed targets. Missing targets contribute only observation NLL. Lower OVJE is better: it complements separate metrics through a sample-aligned, confidence-sensitive score, while decomposing exactly into the two terms above. An example with equal AUC/MAE but different OVJE appears in App. A.3.

4.3 Main Results

To distinguish joint prediction from value-only forecasting, we evaluate **Timeflies** under two settings: joint forecasting (with the classification head) and pure value regression (without it).

Table 4: Ablation studies of **Timeflies**. Within each missing-ratio regime, performance is summarized using the same two-stage arithmetic averaging procedure as in Tab. 2.

Shadow	High Missing				Medium Missing				Low Missing			
Variant	MSE↓	MAE↓	AUC↑	OVJE↓	MSE↓	MAE↓	AUC↑	OVJE↓	MSE↓	MAE↓	AUC↑	OVJE↓
w/o Missing Interval	1.580	0.321	0.744	0.733	0.586	0.289	0.663	0.977	1.522	0.372	0.544	0.836
w/o Focal Loss	1.509	0.322	0.760	0.691	0.584	0.291	0.686	1.042	1.522	0.375	0.557	0.777
w/o Mask-Aware Norm	1.549	0.367	0.745	0.731	0.617	0.333	0.684	0.945	1.517	0.379	0.549	0.801
w/o Obs-Conditioned Attn	1.538	0.326	0.732	0.724	0.585	0.287	0.673	0.943	1.518	0.374	0.530	0.799
w/o Patch Reliability	1.500	0.321	0.751	0.708	0.583	0.286	0.683	0.958	1.522	0.377	0.553	0.801
w/o Residual Fusion	**1.491**	0.318	0.751	0.697	0.585	**0.285**	0.672	0.967	1.526	0.372	0.542	0.808
Timeflies (Full)	1.546	**0.303**	**0.805**	**0.611**	**0.555**	**0.285**	**0.704**	**0.914**	**1.362**	**0.349**	**0.588**	**0.762**

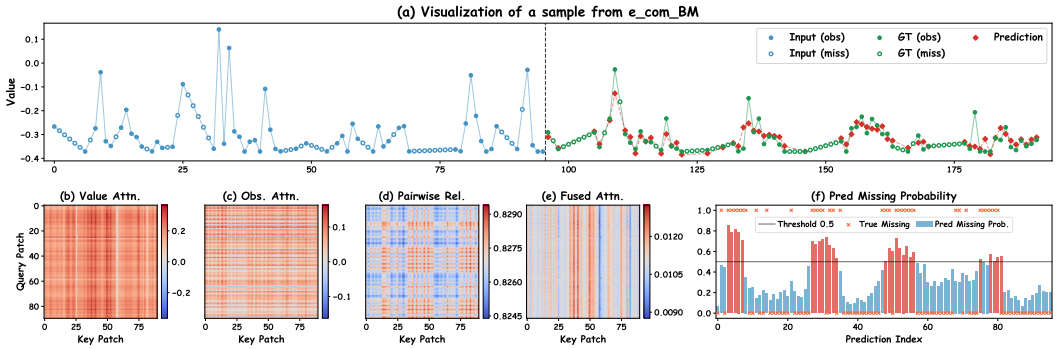

Figure 3: Visualization of **Timeflies** on a medium-irregular sample from ecom_BM. (a) **Forecasting trajectory**: Filled and hollow circles denote valid and missing observations. Despite historical sparsity, the model closely follows the ground truth (green) with predictions (red). (b)–(e) **Attention dynamics**: Construction of the observation-conditioned routing mechanism, including (b) value attention, (c) observed attention, (d) reliability attention, and (e) fused attention. (f) **Existence inference**: The complementary missingness probability $\hat{p}_{\text{miss}} = 1 - \hat{p}_{\text{obs}}$ over the forecast horizon. × denotes actual missing events, and the horizontal line denotes the 0.5 threshold, indicating accurate future observability prediction.

Joint Forecasting of Observability and Value. In Tab. 2, **Timeflies** achieves the lowest OVJE in every sparsity regime. Averaging OVJE equally across the four regimes, it improves over the strongest baseline, FACT [29], by approximately ***22.5%***. The aggregate joint-score gains are particularly pronounced in sparse regimes. They support predicting observability as an additional output, but do not establish that future-observability supervision universally improves MAE or MSE. Its incremental value benefit depends on the observation mechanism, as examined by the controlled experiments in App. C.

Historical Missing Patterns as Vital Context. To verify our claim that *historical missing patterns provide vital context for understanding the system's underlying dynamics*, Tab. 3 evaluates **Timeflies** tasked solely with value regression (without the classification head). Even without explicit observability supervision, **Timeflies** consistently yields the lowest MSE and MAE. The improvements peak under medium missingness (***17.1%*** MAE vs. OLinear), where irregular distribution shifts heavily disrupt standard attention mechanisms. This supports the use of historical observation information. Under IID missingness, such information identifies valid inputs; learning state-related observation dynamics requires a structured, state-relevant observation process.

4.4 Ablation Studies

To validate the effectiveness of **Timeflies**, we evaluate several variants in Tab. 4. The results characterize each component's contribution, whose magnitude and direction can vary across metrics and missingness regimes.

- **Mask-Aware Normalization:** Removing this component causes the largest drop, especially under High Missing, showing that excluding NaN-filled zeros is crucial for stable normalization.

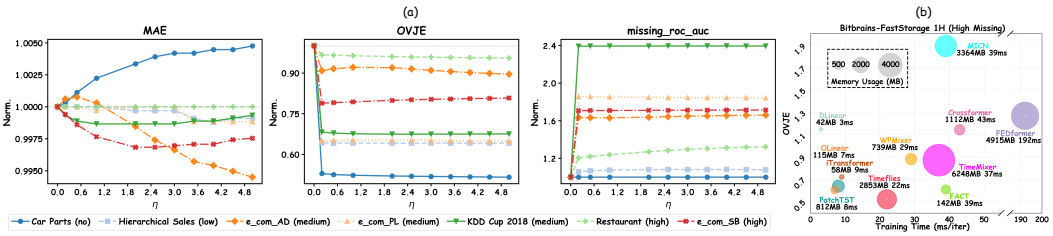

Figure 4: (a) Sensitivity of **Timeflies** to the missingness loss weight η across datasets spanning four missing-ratio regimes. Parenthesized labels denote the missingness regime. (b) Model efficiency comparison under Bitbrains-FastStorage 1H dataset with high missing.

- **Obs-Conditioned Attn & Patch Reliability:** Without these modules, the model fails to adapt its attention to observation quality. This confirms that explicitly differentiating informative patches and routing attention based on missingness priors is critical for sparse representation learning.
- **Missing Interval:** Removing the time-since-last-observation features degrades predictive accuracy, validating its role in capturing the irregular temporal gaps between valid signals.
- **Focal Loss:** Replacing the focal loss diminishes the existence inference capability (AUC), supporting its use for imbalanced observation targets, without implying better probability calibration.
- **Residual Fusion:** Removing this component worsens MAE, AUC, and OVJE under High Missing, but improves MSE from 1.546 to 1.491, indicating a metric-dependent contribution.

Controlled missingness and contribution analysis. On five complete datasets, IID masks yield near-chance observation AUC; BLOCK-50 reaches 0.930, but does not improve value accuracy over the matched backbone. Under SHIFT-10-50, a simpler mask-aware baseline performs better. M/V0–V3 controls attribute No-Missing gains to architecture and auxiliary optimization, and IID gains to input-validity awareness. Experimental settings and both result tables appear in App. C.

4.5 Model Analysis

Case study on Observation-Guided Value Attention. Figure 3(a)–(e) illustrates trend recovery from fragmented history and attention that emphasizes informative patches. The detailed interpretation is retained in App. B.4. Panel (f) shows missingness probability $\hat{p}_{\text{miss}} = 1 - \hat{p}_{\text{obs}}$; higher plotted values indicate future missing events. Training and OVJE instead use $\hat{p}_{\text{obs}} = \sigma(\hat{O})$, with observed positions as the positive class.

Influence of Loss Weight η As shown in Fig. 4(a), sensitivity to the loss weight η depends strongly on dataset missingness. In the fully observed setting (Car Parts Dataset), MAE is nearly unaffected by η. Under moderate-to-high missingness, the effect on MAE is dataset-dependent: some datasets improve as η increases, while others change little or degrade. The improvements in OVJE and ROC-AUC on several datasets when $\eta > 0$ support explicit mask supervision. ROC-AUC is unchanged when both observation labels and probabilities are complemented, as in the missingness visualization. The range $\eta \in [1.0, 3.0]$ can serve as a starting point for dataset-specific tuning.

Efficiency Analysis. We comprehensively compare the forecasting performance (OVJE), training speed, and memory footprint of **Timeflies** and other baselines, using the official model configurations and the same batch size. As shown in Fig. 4(b), the efficiency of **Timeflies** exceeds other Transformer-based and CNN-based methods.

5 Conclusion

In this work, we challenge the conventional ***perfect foresight*** assumption in time series forecasting, arguing that predicting *whether* an observation occurs is as critical as predicting its value. We propose **Timeflies** to jointly model observational existence and state evolution, the **Shadow** benchmark with natural missingness, and the Observation-Value Joint Entropy (OVJE) score to evaluate both outputs. Experiments show strong aggregate joint performance and competitive value forecasting. Additional value gains are data- and mechanism-dependent: predictable observability alone is insufficient when it contains no state-relevant information. Overall, by bridging observability and numerical regression, we open a new avenue for time series forecasting in irregular settings.

References

[1] Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

[2] Jie Yang, Yifan Hu, Kexin Zhang, Luyang Niu, Philip S Yu, and Kaize Ding. Revisiting multivariate time series forecasting with missing values. *arXiv preprint arXiv:2509.23494*, 2025.

[3] Tong Nie, Guoyang Qin, Wei Ma, Yuewen Mei, and Jian Sun. Imputeformer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2260–2271, 2024.

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

[5] Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Jigang Bao, Yong Jiang, and Shu-Tao Xia. Periodicity decoupling framework for long-term series forecasting. In *International Conference on Learning Representations*, 2024.

[6] Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *The International Conference on Machine Learning*, 2024.

[7] Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-patch: Hierarchical patch GNN for irregular multivariate time series. In *The International Conference on Machine Learning*, 2025.

[8] Timothée Hornek, Youngsub Lee, Sergio Potenciano Menci, and Ivan Pavić. The value of battery energy storage in the continuous intraday market: Forecast vs. perfect foresight strategies. In *2025 IEEE Kiel PowerTech*, pages 1–7. IEEE, 2025.

[9] Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23873–23881, 2026.

[10] Zahra Karevan and Johan AK Suykens. Transductive lstm for time-series prediction: An application to weather forecasting. *Neural Networks*, 125:1–9, 2020.

[11] Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shu-tao Xia, Dawei Cheng, and Changjun Jiang. Fintsb: A comprehensive and practical benchmark for financial time series forecasting. *arXiv preprint arXiv:2502.18834*, 2025.

[12] Wanneng Shu, Ken Cai, and Neal Naixue Xiong. A short-term traffic flow prediction model based on an improved gate recurrent unit neural network. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):16654–16665, 2021.

[13] Caizheng Liu, Zhengyu Zhu, Wanming Hao, and Gangcan Sun. Heterogeneous multivariate time series imputation by transformer model with missing position encoding. *Expert Systems with Applications*, 271:126435, 2025.

[14] Chengqing Yu, Fei Wang, Zezhi Shao, Tangwen Qian, Zhao Zhang, Wei Wei, Zhulin An, Qi Wang, and Yongjun Xu. Ginar+: A robust end-to-end framework for multivariate time series forecasting with missing values. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[15] Jie Yang, Yifan Hu, Yuante Li, Kexin Zhang, Kaize Ding, and Philip S. Yu. From observations to states: Latent time series forecasting. In *Forty-third International Conference on Machine Learning*, 2026.

[16] Yixin Wang, Yifan Hu, Peiyuan Liu, Naiqi Li, Tao Dai, and Shu-Tao Xia. Forecasting as rendering: A 2d gaussian splatting framework for time series forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26. Association for Computing Machinery, 2026.

[17] Alireza Zamanian, Henrik von Kleist, Octavia-Andreea Ciora, Marta Piperno, Gino Lancho, and Narges Ahmidi. Analysis of missingness scenarios for observational health data. *Journal of Personalized Medicine*, 14(5):514, 2024.

[18] Yiding Liu, Yifan Hu, Hongjie Xia, Peiyuan Liu, Hongzhou Chen, Xilin Dai, Zewei Dong, and Jiang-Ming Yang. Falcon-x: A time series foundation model for heterogeneous multivariate modeling. *arXiv preprint arXiv:2605.27286*, 2026.

[19] Hongjie Xia, Yiding Liu, Yifan Hu, Peiyuan Liu, and Zewei Dong. Into the orbit for time series: Training regimes for foundation models. *arXiv preprint arXiv:2608.13262*, 2026.

[20] Alex Eric Yuan and Wenying Shou. Data-driven causal analysis of observational biological time series. *Elife*, 11:e72518, 2022.

[21] Yifan Hu, Xilin Dai, Zhiyuan Qu, Yiding Liu, Zewei Dong, Jiang-ming Yang, and Qiang Xu. When tomorrow becomes today: Self-evolving policies for agentic time-series forecasting. *arXiv preprint arXiv:2609.24862*, 2026.

[22] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[23] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430, 2021.

[24] Xiangjie Kong, Zhenghao Chen, Weiyao Liu, Kaili Ning, Lechao Zhang, Syauqie Muhammad Marier, Yichen Liu, Yuhao Chen, and Feng Xia. Deep learning for time series forecasting: a survey. *International Journal of Machine Learning and Cybernetics*, 16(7):5079–5112, 2025.

[25] Shengsheng Lin, Weiwei Lin, Wentai Wu, Feiyu Zhao, Ruichao Mo, and Haotong Zhang. SegRNN: Segment recurrent neural network for long-term time series forecasting. *IEEE Internet of Things Journal*, 2025.

[26] Yixuan Tan, Liyan Xie, and Xiuyuan Cheng. Neural differential recurrent neural network with adaptive time steps. *arXiv preprint arXiv:2306.01674*, 2023.

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

[28] 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*, 2022.

[29] Huiqiang Wang, Jieming Shi, and Li Qing. FACT: Fine-grained across-variable convolution for multivariate time series forecasting. In *The International Conference on Learning Representations*, 2026.

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

[31] 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 *International Conference on Learning Representations*, 2024.

[32] Wenzhen Yue, Yong Liu, Hao Wang, Haoxuan Li, Xianghua Ying, Ruohao Guo, Bowei Xing, and Ji Shi. OLinear: A linear model for time series forecasting in orthogonally transformed domain. In *Neural Information Processing Systems*, 2026.

[33] Yifan Hu, Peiyuan Liu, Peng Zhu, Dawei Cheng, and Tao Dai. Adaptive multi-scale decomposition framework for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17359–17367, 2025.

[34] Yifan Hu, Jie Yang, Tian Zhou, Peiyuan Liu, Yujin Tang, Rong Jin, and Liang Sun. Bridging past and future: Distribution-aware alignment for time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026.

[35] Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. TimeFilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *International Conference on Machine Learning*, 2025.

[36] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763, 2020.

[37] Wanlin Cai, Yuxuan Liang, Xianggen Liu, Jianshuai Feng, and Yuankai Wu. Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[38] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286. PMLR, 2022.

[39] Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-tao Xia. TimeBridge: Non-stationarity matters for long-term time series forecasting. In *International Conference on Machine Learning*, 2024.

[40] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[41] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

[42] Yifan Hu, Jie Yang, Xilin Dai, Wanxu Cai, Kuiye Ding, Yuante Li, Qinghua Liu, Enze Ma, Zhiyuan Qu, Yixin Wang, Binyan Xu, Kexin Zhang, Peiyuan Liu, Zhijian Xu, Guibin Zhang, Yujin Tang, Yanwei Yue, Kening Zheng, Chengze Li, Hanrong Zhang, Haoyan Xu, Naiqi Li, Tao Dai, Dawei Cheng, John Paparrizos, Kaize Ding, Tian Zhou, Qiang Xu, Shu-tao Xia, Shirui Pan, and Philip S. Yu. The landscape of agentic time series systems: Architectures, reliability, and frontiers. Preprint, 2026.

[43] Xuanying Li, Sha Xiang, and Cheng Dai. A survey of forecasting methods for irregular time series. In *IEEE International Conferences on Internet of Things*, pages 83–90, 2025.

[44] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

[45] Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

[46] Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *International conference on machine learning*, pages 19388–19405. PMLR, 2022.

[47] Christian Klötergens, Vijaya Krishna Yalavarthi, Randolf Scholz, Maximilian Stubbemann, Stefan Born, and Lars Schmidt-Thieme. Physiome-ode: A benchmark for irregularly sampled multivariate time series forecasting based on biological odes. *arXiv preprint arXiv:2502.07489*, 2025.

[48] Xiaodan Chen, Xiucheng Li, Bo Liu, and Zhijun Li. Biased temporal convolution graph network for time series forecasting with missing values. In *International Conference on Learning Representations*, 2024.

[49] Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. Grafiti: Graphs for forecasting irregularly sampled time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16255–16263, 2024.

[50] 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. *arXiv preprint arXiv:2410.10393*, 2024.

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

[52] Md Mahmuddun Nabi Murad, Mehmet Aktukmak, and Yasin Yilmaz. WPMixer: Efficient multi-resolution mixing for long-term time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 19581–19588, 2025.

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[54] Diederik P Kingma. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[55] Yifan Hu, Peiyuan Liu, Yuante Li, Dawei Cheng, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. Finmamba: Market-aware graph enhanced multi-level mamba for stock movement prediction. *arXiv preprint arXiv:2502.06707*, 2025.

A Additional Experimental Details

A.1 Dataset Details

Tab. 1 provides a detailed characterization of the `Shadow` benchmark, which integrates 15 public datasets from *GIFT-Eval* with 16 proprietary e-commerce datasets spanning domains such as energy, cloud computing, and retail. For each dataset, we report key temporal statistics—including sampling frequency, series count, and sequence length distributions—alongside the number of target variates and multi-horizon prediction lengths ($S/M/L$).

To evaluate model robustness under varying data sparsity, we stratify the series into four missingness regimes: *None* (0%), *Low* ($(0, 10\%]$), *Medium* ($(10\%, 40\%]$), and *High* ($(40\%, 100\%]$). While public datasets exhibit heterogeneous patterns—with some, like *Electricity*, spanning all regimes and others localized to specific subsets—the proprietary datasets are predominantly hourly and characterized by higher missing ratios. This distribution reflects the inherent sparsity of real-world transactional data.

Collectively, `Shadow` offers a broad spectrum of missingness severities, facilitating a rigorous evaluation of forecasting models across controlled and realistic scenarios.

Table 5 details the per-split missingness ratios across all dataset–frequency–regime configurations under the GIFT-Eval 80/10/10 protocol. Car Parts in the High regime has no observed validation or test targets; observed-target MAE/MSE are therefore undefined for this group. While most datasets exhibit "front-loaded" missingness (higher sparsity in training than in evaluation), the e-commerce datasets (`ecom_NG` and `ecom_VN`) present a more challenging inverse pattern. Specifically, in the Medium regime, `ecom_NG`'s missing ratio escalates from 0.15% in training to 94.43% in testing; `ecom_VN` follows a similar trajectory, rising from 1.13% to 93.03%. High missingness is already present during validation (54.33% for `ecom_NG` and 83.48% for `ecom_VN`). These natural regime changes therefore differ from the previously unseen test-only shift in App. C; they are stress tests rather than general robustness guarantees.

Table 5: Missing-ratio statistics for each dataset under different frequencies and missing regimes. We report the overall missing ratio, together with the missing ratios in the training, validation, and test splits. Missing regimes are abbreviated as **High**, **Medium**, **Low**, and **No**.

Dataset	Freq.	Missing Ratio	Overall	Train	Val	Test
GIFT-Eval public datasets						
`Bitbrains-Fast Storage`	5T	High	77.14%	83.57%	54.60%	48.25%
		Low	0.26%	0.16%	1.00%	0.27%
		Medium	19.51%	22.74%	7.25%	5.89%
		No	0.00%	0.00%	0.00%	0.00%
	H	High	77.00%	83.50%	54.12%	48.26%
		Low	3.81%	2.70%	0.00%	16.35%
		Medium	19.25%	22.62%	5.88%	5.88%
		No	0.00%	0.00%	0.00%	0.00%
`Bitbrains-rnd`	5T	High	69.45%	81.75%	26.21%	14.29%
		Low	4.15%	5.15%	0.30%	0.03%
		Medium	33.98%	42.44%	0.23%	0.02%
	H	High	69.68%	82.04%	26.15%	14.32%
		Low	4.13%	5.16%	0.00%	0.00%
		Medium	34.52%	43.15%	0.00%	0.00%
		No	0.00%	0.00%	0.00%	0.00%
`Car Parts`	M	High	72.75%	65.26%	100.00%	100.00%
		No	0.00%	0.00%	0.00%	0.00%
`Electricity`	D	High	67.03%	83.59%	2.05%	0.00%
		Low	5.07%	6.34%	0.00%	0.00%
		Medium	25.44%	31.82%	0.00%	0.00%
		No	0.00%	0.00%	0.00%	0.00%
	H	High	67.04%	83.55%	2.04%	0.00%
		Low	5.08%	6.35%	0.00%	0.00%
		Medium	25.44%	31.80%	0.00%	0.00%
		No	0.00%	0.00%	0.00%	0.00%
	1W	High	66.99%	83.68%	2.14%	0.00%
		Low	4.81%	6.02%	0.00%	0.00%
		Medium	25.45%	31.89%	0.00%	0.00%

Continued on next page

Table 5 – continued from previous page

Dataset	Freq.	Missing Ratio	Overall	Train	Val	Test
		No	0.00%	0.00%	0.00%	0.00%
Hierarchical Sales	D	Low	1.48%	1.16%	2.75%	2.73%
Jena Weather	10T	Low	0.02%	0.02%	0.00%	0.00%
	H	Low	0.01%	0.01%	0.00%	0.00%
KDD Cup 2018	D	High	64.79%	60.93%	80.00%	80.43%
		Low	4.59%	5.50%	0.67%	1.15%
		Medium	19.48%	19.80%	16.44%	19.93%
		No	0.00%	0.00%	0.00%	0.00%
	H	High	56.68%	53.35%	67.20%	72.83%
		Low	6.91%	7.87%	3.14%	3.03%
		Medium	18.18%	20.42%	7.93%	10.45%
Restaurant	D	High	44.19%	45.08%	43.07%	38.35%
		Low	4.03%	4.27%	2.91%	3.24%
		Medium	19.50%	20.05%	17.71%	17.02%
		No	0.00%	0.00%	0.00%	0.00%
Temperature Rain	D	Low	2.67%	3.32%	0.04%	0.03%
		Medium	17.74%	22.17%	0.00%	0.00%
		No	0.00%	0.00%	0.00%	0.00%
Proprietary e-commerce datasets						
ecom_AD	1H	Low	0.10%	0.09%	0.00%	0.29%
		Medium	17.32%	10.85%	27.79%	58.53%
ecom_AI	1H	High	56.01%	58.55%	30.08%	61.62%
ecom_BB	1H	Medium	22.52%	20.24%	34.05%	29.19%
ecom_BM	1H	Medium	27.21%	22.34%	47.84%	45.48%
ecom_BY	1H	Medium	24.43%	16.52%	49.27%	62.91%
ecom_HN	1H	High	83.50%	86.74%	72.23%	68.83%
ecom_HR	1H	Medium	24.40%	30.50%	0.00%	0.00%
ecom_JP	1H	Low	0.24%	0.30%	0.00%	0.00%
ecom_ME	1H	High	63.33%	67.23%	48.69%	46.78%
		Medium	24.40%	29.63%	4.53%	2.46%
ecom_NG	1H	Medium	14.99%	0.15%	54.33%	94.43%
ecom_NO	1H	Low	5.40%	6.42%	1.16%	1.49%
		Medium	36.32%	39.34%	25.62%	22.89%
ecom_NR	1H	Low	6.68%	5.76%	9.95%	10.76%
ecom_PL	1H	Medium	17.02%	19.39%	11.06%	3.99%
ecom_SB	1H	High	79.16%	81.14%	73.67%	68.82%
		Low	0.08%	0.10%	0.00%	0.00%
ecom_TA	1H	Low	0.13%	0.17%	0.00%	0.00%
ecom_VN	1H	Low	0.13%	0.06%	0.05%	0.80%
		Medium	18.56%	1.13%	83.48%	93.03%

A.2 Experiment Details

All models are implemented in PyTorch [53] and trained on a single NVIDIA B200 (180GB) GPU. We use the Adam optimizer [54], with the learning rate selected from $\{10^{-3}, 10^{-4}, 5 \times 10^{-5}\}$ via grid search. Across all experiments, the dropout rate is set to 0.2, training runs for at most 20 epochs with an early stopping patience of 3, η is fixed to 1.0 for $\mathcal{L}_{\text{val}}$, and focal loss is used for $\mathcal{L}_{\text{obs}}$.

Table 6 summarizes the dataset-specific hyperparameter settings of **Timeflies**. We set the input context length (L) and prediction horizon (H) according to the temporal granularity of each dataset. High-frequency datasets (`5T`, `10T`, `1H`) use a longer window ($L = 720, H = 336$), whereas lower-frequency datasets (`1D`, `1W`, `1M`) use shorter ranges ($L \in [12, 180], H \in [4, 30]$). The default architecture uses $d_{\text{model}} = 128$, $d_{\text{ff}} = 256$, and $N_{\text{layers}} = 2$. For fully observed settings, we increase the model capacity up to $d_{\text{model}} = 256$, $d_{\text{ff}} = 1024$, and $N_{\text{layers}} = 4$ to better capture dense temporal dependencies.

Table 6: Dataset-specific hyperparameter settings of `Timeflies` for missing-aware forecasting. Datasets and frequencies are presented hierarchically, and rows are merged when multiple missingness settings share the same configuration. Shared hyperparameters are given in the note.

Dataset	Freq.	Missingness	d_{model}	n_{heads}	n_{layers}	d_{ff}	L	H	Batch
Bitbrains-FastStorage	5T	low / medium / high / none	128	8	2	256	720	336	1024
	1H	low / medium / high	128	8	2	256	336	48	256
		none	256	8	4	1024	336	48	256
Bitbrains-Rnd	5T	low / high	256	16	4	1024	720	336	1024
		medium	128	8	2	256	720	336	1024
	1H	low / medium / high / none	128	8	2	256	336	48	256
Car Parts	1M	high / none	128	8	2	256	12	12	256
Electricity	1D	low / medium / none	128	8	2	256	180	8	256
		high	256	8	2	1024	96	8	256
	1H	low	128	8	2	256	720	336	256
		medium / high	128	8	2	256	720	336	1024
	1W	low	128	8	2	256	96	4	64
		medium / high / none	128	8	2	256	96	4	256
Hierarchical Sales	1D	low	256	8	2	1024	96	96	256
Jena Weather	10T	low	128	8	2	256	720	336	256
	1H	low	128	8	2	256	720	336	256
KDD Cup 2018	1H	low / medium / high	128	8	2	256	720	336	256
	1D	low / medium / none	128	8	2	256	96	30	256
		high	64	4	1	128	96	30	32
Restaurant	1D	low / medium / high / none	128	8	2	256	48	7	256
Temperature Rain	1D	low	256	8	4	1024	336	30	256
		medium	64	8	1	256	336	30	256
		none	128	8	2	256	336	30	256
e_com_SB	1H	low / high	128	8	2	256	720	336	256
e_com_HN	1H	high	128	8	2	256	720	336	256
e_com_AI	1H	high	128	8	2	256	720	336	256
e_com_BY	1H	medium	128	8	2	256	720	336	256
e_com_BB	1H	medium	128	8	2	256	720	336	256
e_com_NR	1H	low	128	8	2	256	720	336	256
e_com_NO	1H	low / medium	128	8	2	256	720	336	256
e_com_VN	1H	low / medium	128	8	2	256	720	336	256
e_com_JP	1H	low	128	8	2	256	720	336	256
e_com_HR	1H	medium	128	8	2	256	720	336	256
e_com_TA	1H	low	128	8	2	256	720	336	256
e_com_ME	1H	medium / high	128	8	2	256	720	336	256
e_com_AD	1H	low / medium	128	8	2	256	720	336	256
e_com_PL	1H	medium	128	8	2	256	720	336	256
e_com_NG	1H	medium	128	8	2	256	720	336	256
e_com_BM	1H	medium	128	8	2	256	720	336	256

Note. Shared hyperparameters across all experiments: dropout $= 0.2$, learning rate $= 10^{-4}$, epochs $= 20$, patience $= 3$, $\eta = 1$, and $\mathcal{L}_{\text{obs}}$ = focal.

A.3 Metric Details

The proposed Observation-Value Joint Entropy (OVJE) metric represents a mathematical formalization of our "existence precedes value" philosophy, evaluating observation forecasts together with value estimates at observed positions. To project the regression accuracy into a value-quality scale that can be jointly evaluated with classification, we map the symmetric relative estimation error (sMAPE) e_t to a bounded state quality score $q_t \in (0, 1]$:

$$e_t = \frac{2|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t| + \epsilon}, \quad q_t = \exp(-e_t), \tag{16}$$

where $\epsilon = 10^{-8}$ and $\hat{p}_t = \hat{p}_{\text{obs},t}$ is clipped to $[\epsilon, 1 - \epsilon]$. Here $o_t = O_t$ and q_t is a quality factor rather than a normalized density. The joint quality score P_t is formulated piecewise:

$$P_t = \begin{cases} \hat{p}_t \cdot q_t, & o_t = 1 \\ 1 - \hat{p}_t, & o_t = 0 \end{cases} \tag{17}$$

This explicitly unifies the two streams: if the point naturally exists ($o_t = 1$), the score accounts for the probability assigned to its existence ($\hat{p}_t$) and the accuracy of its estimated value (q_t). Conversely, if the point is missing ($o_t = 0$), correctness relies solely on predicting its absence ($1 - \hat{p}_t$). This logic compactly reduces to:

$$P_t = (\hat{p}_t \cdot q_t)^{o_t} \cdot (1 - \hat{p}_t)^{1-o_t}. \tag{18}$$

Taking the negative logarithm of this joint quality score over all T evaluated positions yields the final OVJE metric:

$$\text{OVJE} = -\frac{1}{T}\sum_{t=1}^{T} \log P_t = -\frac{1}{T}\sum_{t=1}^{T} \left[o_t \cdot \log(\hat{p}_t \cdot q_t) + (1 - o_t) \cdot \log(1 - \hat{p}_t) \right]. \tag{19}$$

As shown in Eq. (15), this score is exactly the sum of Bernoulli NLL and the observed-position sMAPE contribution, both normalized by the total number of evaluated positions. It cannot replace reporting these components separately, and its value term depends on the observed fraction. Comparisons should therefore use the same evaluation positions and missingness regime.

Unlike AUC, OVJE is sensitive to probability confidence. For example, take $o = (1, 1, 0, 0)$ and exact value predictions at the two observed positions. Probability vectors $(0.9, 0.8, 0.2, 0.1)$ and $(0.02, 0.01, 0.009, 0.008)$ both yield AUC $= 1$ and MAE $= 0$, but their OVJE scores are 0.164 and 2.134, respectively: the latter assigns very low probability to observations that actually occur. This example illustrates the NLL contribution; the score still decomposes exactly as in Eq. (15).

B Full Results

B.1 Full Missingness Prediction Results

Tab. 7 summarizes the dataset-level results, using the arithmetic mean to aggregate forecasting (`MAE`, `MSE`), classification (`AUC`), and joint (`OVJE`) metrics. **`Timeflies`** consistently outperforms baselines, particularly in high-sparsity environments. Notably, on `ecom_NG` and `ecom_VN`, our model maintains performance despite a massive missingness shift (e.g., from 0.15% during training to 94.43% at test time), whereas competing models prove brittle. These cases suggest sensitivity to observation density in some baselines. High missingness is already visible during validation, however, so they do not establish robustness to an entirely unseen observation process.

Table 7: Full missingness prediction Results. Abbreviations: BB-FS = Bitbrains Fast Storage, BB-rnd = Bitbrains Random, Hier. Sales = Hierarchical Sales, Jena Wea. = Jena Weather, KDD Cup = KDD Cup 2018, Temp. Rain = Temperature Rain. Datasets with the same name appearing in multiple row groups differ in sampling frequency. Single-class evaluations use the implementation fallback AUC $= 0.500$, for which AUC is mathematically undefined.

Dataset	Missing Ratio	Metric	**`Timeflies`**	Crossformer	DLinear	FACT	FEDformer	MICN	OLinear	PatchTST	TimeMixer	WPMixer	iTransformer
BB-FS	High	MAE↓	0.092	0.255	0.190	0.105	0.147	0.184	0.104	0.102	0.113	0.099	0.095
		MSE↓	0.702	2.904	1.202	0.773	0.738	0.860	0.717	0.747	0.721	0.727	0.709
		AUC↑	0.995	0.860	0.853	0.954	0.830	0.943	0.956	0.954	0.951	0.809	0.957
		OVJE↓	0.378	0.803	0.919	0.487	1.179	0.835	0.462	0.485	0.495	0.710	0.505
	Medium	MAE↓	0.554	0.561	0.624	0.560	0.694	0.644	0.589	0.555	0.652	0.590	0.650
		MSE↓	0.668	0.769	0.647	0.766	0.788	0.996	0.730	0.729	0.945	0.819	0.962
		AUC↑	0.995	0.820	0.783	0.932	0.384	0.826	0.906	0.939	0.973	0.440	0.931
		OVJE↓	0.796	1.037	1.324	0.823	1.309	0.970	0.919	1.039	0.907	1.349	0.981
	Low	MAE↓	0.120	0.134	0.135	0.128	0.166	0.207	0.122	0.134	0.205	0.137	0.129
		MSE↓	0.223	0.234	0.217	0.229	0.226	0.372	0.221	0.247	0.753	0.252	0.249
		AUC↑	0.625	0.483	0.589	0.559	0.515	0.409	0.577	0.469	0.407	0.482	0.520
		OVJE↓	0.182	0.271	0.287	0.184	0.294	0.363	0.173	0.249	0.225	0.580	0.222
	No	MAE↓	0.077	0.081	0.079	0.085	0.126	0.088	0.080	0.079	0.086	0.094	0.081
		MSE↓	0.062	0.069	0.063	0.063	0.109	0.062	0.063	0.063	0.066	0.065	0.057
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.162	0.171	0.172	0.192	0.277	0.186	0.170	0.170	0.188	0.534	0.193
BB-FS	High	MAE↓	0.097	0.353	0.306	0.125	0.237	0.409	0.107	0.110	0.148	0.152	0.151
		MSE↓	0.674	2.781	1.771	0.771	1.120	1.671	0.706	0.696	0.855	0.830	0.865
		AUC↑	0.921	0.629	0.563	0.985	0.524	0.568	0.973	0.958	0.826	0.534	0.910
		OVJE↓	0.523	1.143	1.146	0.611	1.267	1.890	0.607	0.640	0.872	0.882	0.723
	Medium	MAE↓	0.596	0.580	0.570	0.592	0.590	0.645	0.644	0.630	0.618	0.595	0.637
		MSE↓	0.723	0.609	0.563	0.785	0.586	0.689	0.813	0.858	0.760	0.693	0.807
		AUC↑	1.000	0.596	0.328	0.539	0.484	0.799	0.951	0.631	0.762	0.477	0.598
		OVJE↓	0.954	1.161	1.632	1.070	1.634	1.159	1.165	1.124	1.126	1.722	1.566

Continued on next page

Table 7 – Continued from previous page

Dataset	Missing Ratio	Metric	Timeflies	Crossformer	DLinear	FACT	FEDformer	MICN	OLinear	PatchTST	TimeMixer	WPMixer	iTransformer
BB-FS	Low	MAE↓	0.910	1.008	0.968	0.952	0.993	1.853	0.942	0.907	0.907	0.928	0.930
		MSE↓	17.040	17.974	18.117	18.225	18.063	18.818	18.173	17.869	17.773	18.101	17.953
		AUC↑	0.810	0.546	0.472	0.578	0.491	0.462	0.269	0.217	0.760	0.500	0.349
		OVJE↓	0.788	0.843	1.098	0.980	1.114	1.441	0.937	0.815	0.675	1.145	0.936
	No	MAE↓	0.156	0.296	0.237	0.168	0.289	0.206	0.175	0.202	0.190	0.208	0.159
		MSE↓	0.329	1.074	0.549	0.312	0.710	0.431	0.319	0.485	0.425	0.484	0.256
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.241	0.278	0.416	0.252	0.405	0.300	0.255	0.247	0.256	0.838	0.305
BB-rnd	High	MAE↓	0.472	0.480	0.533	0.460	0.541	0.515	0.493	0.482	0.499	0.453	0.465
		MSE↓	8.445	7.836	8.121	8.690	8.022	8.030	8.910	9.384	8.678	8.310	8.136
		AUC↑	0.988	0.962	0.865	0.978	0.957	0.953	0.969	0.971	0.955	0.676	0.986
		OVJE↓	0.585	0.619	1.011	0.603	1.346	1.153	0.642	0.681	0.908	0.976	0.617
	Medium	MAE↓	0.163	0.209	0.195	0.174	0.203	0.261	0.178	0.182	0.198	0.161	0.170
		MSE↓	0.149	0.172	0.161	0.156	0.166	0.193	0.156	0.171	0.187	0.151	0.153
		AUC↑	0.700	0.359	0.492	0.228	0.583	0.656	0.488	0.665	0.455	0.477	0.570
		OVJE↓	0.598	0.838	1.105	0.755	1.068	1.060	0.829	0.792	0.481	1.031	0.756
	Low	MAE↓	0.190	0.191	0.200	0.220	0.217	0.241	0.198	0.202	0.239	0.217	0.202
		MSE↓	0.483	0.577	0.473	0.791	0.493	0.526	0.565	0.570	0.810	0.612	0.717
		AUC↑	0.680	0.493	0.442	0.665	0.524	0.515	0.514	0.554	0.625	0.464	0.646
		OVJE↓	0.430	0.504	0.649	0.452	0.495	0.684	0.452	0.516	1.096	0.822	0.521
BB-rnd	High	MAE↓	0.546	0.709	0.744	0.686	0.939	2.539	0.557	0.722	0.821	0.684	0.637
		MSE↓	6.366	6.528	6.671	6.816	6.823	13.335	6.445	6.581	6.826	6.521	6.391
		AUC↑	0.978	0.901	0.801	0.801	0.571	0.566	0.961	0.902	0.688	0.593	0.664
		OVJE↓	0.784	1.242	1.599	1.612	1.704	2.121	0.945	1.193	1.442	1.199	1.218
	Medium	MAE↓	0.138	0.340	0.154	0.185	0.391	3.533	0.143	0.195	0.215	0.140	0.231
		MSE↓	0.087	0.194	0.087	0.116	0.256	19.761	0.087	0.118	0.125	0.085	0.152
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.417	1.178	1.020	1.051	1.416	1.716	0.950	0.723	1.051	0.943	1.126
	Low	MAE↓	0.214	0.222	0.255	0.236	0.242	0.265	0.225	0.262	0.243	0.264	0.246
		MSE↓	0.521	0.701	0.594	0.609	0.515	0.679	0.547	0.707	0.646	0.661	0.599
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.505	0.650	1.101	0.562	0.658	0.617	0.723	0.721	0.599	0.972	0.665
	No	MAE↓	0.080	0.112	0.104	0.095	0.146	1.491	0.090	0.081	0.108	0.091	0.094
		MSE↓	0.046	0.078	0.054	0.068	0.083	5.492	0.069	0.047	0.069	0.048	0.062
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.182	0.447	0.933	0.407	0.961	1.152	0.422	0.210	0.275	0.839	0.596
Car Parts	No	MAE↓	0.358	0.440	0.493	0.363	0.426	0.441	0.361	0.426	0.467	0.397	0.366
		MSE↓	0.665	0.683	0.707	0.671	0.682	0.682	0.669	0.671	0.820	0.670	0.688
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.835	1.520	1.883	0.915	1.636	1.073	0.940	1.655	1.450	1.518	1.008
Electricity	High	MAE↓	0.020	0.038	0.060	0.020	0.084	0.675	0.019	0.022	0.024	0.024	0.019
		MSE↓	0.016	0.015	0.023	0.016	0.035	22.089	0.014	0.018	0.019	0.020	0.013
		AUC↑	0.833	0.658	0.601	0.832	0.549	0.663	0.833	0.824	0.833	0.563	0.829
		OVJE↓	0.078	0.854	0.932	0.523	2.043	2.560	0.143	1.315	0.245	0.612	0.431
	Medium	MAE↓	0.243	0.338	0.318	0.314	0.409	6.612	0.320	0.268	0.329	0.299	0.324
		MSE↓	0.104	0.168	0.143	0.144	0.257	11.675	0.147	0.118	0.156	0.135	0.152
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.318	0.838	1.065	0.960	1.102	1.149	0.926	0.502	1.035	0.872	1.034
	Low	MAE↓	0.243	0.338	0.318	0.314	0.409	6.612	0.320	0.268	0.329	0.299	0.324
		MSE↓	0.104	0.168	0.143	0.144	0.257	131.675	0.147	0.118	0.156	0.135	0.152
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.318	0.838	1.065	0.960	1.102	1.149	0.926	0.502	1.035	0.872	1.034
	No	MAE↓	0.025	0.034	0.034	0.025	0.046	0.072	0.025	0.026	0.026	0.029	0.023
		MSE↓	0.011	0.016	0.018	0.009	0.022	0.026	0.011	0.011	0.013	0.014	0.012
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.124	0.190	0.343	0.129	0.207	0.372	0.127	0.133	0.141	0.503	0.169
Electricity	High	MAE↓	0.022	0.029	0.028	0.023	0.039	0.041	0.024	0.023	0.025	0.029	0.023
		MSE↓	0.020	0.022	0.024	0.021	0.025	0.032	0.023	0.021	0.024	0.023	0.020
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.045	0.155	0.577	0.052	1.000	0.540	0.049	0.140	0.058	0.599	0.101
	Medium	MAE↓	0.019	0.026	0.022	0.021	0.037	0.034	0.021	0.024	0.024	0.030	0.021
		MSE↓	0.034	0.026	0.031	0.063	0.097	0.051	0.036	0.038	0.027	0.034	0.029
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.123	0.349	0.261	0.148	0.347	0.522	0.156	0.170	0.154	0.568	0.210
	Low	MAE↓	0.188	0.216	0.189	0.218	0.355	0.193	0.218	0.196	0.219	0.175	0.221
		MSE↓	0.073	0.100	0.076	0.096	0.207	0.079	0.086	0.076	0.094	0.072	0.096
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.409	0.508	0.684	0.442	0.891	0.412	0.482	0.484	0.459	1.130	0.537
Electricity	Medium	MAE↓	0.017	0.053	0.071	0.018	0.056	1.290	0.025	0.019	0.027	0.035	0.025
		MSE↓	0.029	0.331	0.155	0.035	0.058	79.755	0.073	0.037	0.093	0.155	0.065
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.096	0.291	1.115	0.144	0.411	1.210	0.278	0.163	0.136	0.468	0.744
	Low	MAE↓	0.216	0.514	0.218	0.222	0.381	10.743	0.220	0.219	0.220	0.222	0.222
		MSE↓	0.070	0.392	0.063	0.071	0.205	117.859	0.065	0.072	0.070	0.064	0.066
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.546	1.389	0.985	0.865	1.143	2.162	0.862	0.875	0.977	0.666	0.962
	No	MAE↓	0.019	0.047	0.068	0.021	0.037	2.454	0.033	0.022	0.027	0.026	0.019
		MSE↓	0.006	0.040	0.031	0.008	0.008	75.830	0.017	0.007	0.012	0.010	0.006
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.114	0.329	0.999	0.110	0.194	1.296	0.492	0.206	0.121	0.408	0.230

Continued on next page

Table 7 – Continued from previous page

Dataset	Missing Ratio	Metric	Timeflies	Crossformer	DLinear	FACT	FEDformer	MICN	OLinear	PatchTST	TimeMixer	WPMixer	iTransformer
Electricity	High	MAE↓	0.015	0.127	0.084	0.029	0.181	1.822	0.034	0.031	0.037	0.025	0.033
		MSE↓	0.005	0.145	0.036	0.026	0.074	54.368	0.028	0.027	0.043	0.017	0.029
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.177	0.919	0.971	0.514	1.933	2.404	0.540	0.468	0.815	0.252	0.648
Hier. Sales	Low	MAE↓	0.331	0.341	0.366	0.338	0.368	0.636	0.355	0.371	0.372	0.366	0.376
		MSE↓	0.532	0.518	0.562	0.537	0.556	0.994	0.572	0.588	0.581	0.581	0.593
		AUC↑	0.507	0.528	0.471	0.573	0.547	0.514	0.399	0.528	0.507	0.510	0.591
		OVJE↓	1.043	1.109	1.659	1.046	1.167	1.507	1.089	1.113	1.104	1.302	1.141
Jena Wea.	Low	MAE↓	0.308	0.329	0.361	0.326	0.406	0.365	0.328	0.322	0.314	0.321	0.330
		MSE↓	0.309	0.295	0.349	0.332	0.417	0.353	0.337	0.327	0.316	0.333	0.336
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.615	0.872	1.116	0.662	1.205	0.809	0.663	0.694	0.611	1.258	0.879
Jena Wea.	Low	MAE↓	0.380	0.488	0.416	0.407	0.544	1.526	0.385	0.374	0.398	0.386	0.416
		MSE↓	0.386	0.497	0.400	0.412	0.584	6.486	0.389	0.383	0.405	0.386	0.429
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.770	1.212	1.573	1.269	1.728	1.590	0.788	0.760	1.214	1.473	1.356
KDD Cup	High	MAE↓	0.457	0.494	0.469	0.503	0.500	1.962	0.470	0.533	0.520	0.468	0.467
		MSE↓	0.551	0.636	0.586	0.702	0.670	5.292	0.608	0.748	0.642	0.596	0.696
		AUC↑	0.997	0.468	0.546	0.860	0.669	0.601	0.823	0.703	0.515	0.663	0.605
		OVJE↓	0.424	0.873	0.875	0.762	0.726	0.787	0.713	0.770	0.875	1.149	0.787
	Medium	MAE↓	0.435	0.450	0.450	0.448	0.458	2.570	0.429	0.450	0.439	0.432	0.452
		MSE↓	1.061	1.153	1.197	1.163	1.184	13.987	1.131	1.168	1.130	1.105	1.186
		AUC↑	0.924	0.776	0.522	0.878	0.484	0.630	0.595	0.739	0.789	0.739	0.451
		OVJE↓	0.822	1.046	1.235	0.905	1.066	1.690	1.016	0.972	0.946	0.995	1.095
	Low	MAE↓	0.438	0.438	0.442	0.456	0.485	1.646	0.432	0.446	0.440	0.436	0.463
		MSE↓	1.367	1.446	1.461	1.474	1.540	5.897	1.447	1.452	1.442	1.430	1.513
		AUC↑	0.420	0.342	0.490	0.612	0.470	0.473	0.552	0.450	0.588	0.460	0.455
		OVJE↓	0.759	0.935	1.348	0.814	0.905	1.718	0.785	0.865	0.806	1.035	0.910
	No	MAE↓	0.556	0.627	0.558	0.568	0.700	8.103	0.559	0.610	0.557	0.562	0.631
		MSE↓	0.537	0.667	0.547	0.579	0.751	71.199	0.562	0.645	0.530	0.568	0.678
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	1.065	1.910	1.587	1.604	1.783	2.467	1.622	1.604	1.721	1.249	1.809
KDD Cup	High	MAE↓	0.612	0.631	0.667	0.679	0.646	0.688	0.660	0.682	0.779	0.695	0.686
		MSE↓	2.578	2.629	2.791	2.769	2.749	2.815	2.754	2.850	3.158	2.804	2.768
		AUC↑	0.958	0.891	0.906	0.933	0.815	0.954	0.922	0.933	0.950	0.626	0.850
		OVJE↓	0.533	0.880	0.958	0.547	0.798	0.535	0.545	0.556	0.550	0.915	0.742
	Medium	MAE↓	0.442	0.445	0.451	0.478	0.484	0.473	0.447	0.446	0.468	0.482	0.484
		MSE↓	1.628	1.643	1.636	1.709	1.652	1.645	1.636	1.655	1.656	1.704	1.710
		AUC↑	0.576	0.564	0.503	0.626	0.506	0.537	0.540	0.556	0.563	0.482	0.567
		OVJE↓	1.069	1.138	1.142	1.072	1.116	1.137	1.085	1.083	1.081	1.390	1.128
	Low	MAE↓	0.487	0.482	0.489	0.515	0.522	0.518	0.507	0.506	0.531	0.521	0.520
		MSE↓	0.709	0.721	0.704	0.716	0.737	0.754	0.713	0.720	0.790	0.732	0.740
		AUC↑	0.566	0.502	0.545	0.481	0.451	0.490	0.524	0.540	0.495	0.556	0.498
		OVJE↓	1.050	1.141	1.156	1.069	1.112	1.087	1.080	1.134	1.077	1.492	1.192
Restaurant	High	MAE↓	0.466	0.546	0.577	0.502	0.559	1.406	0.518	0.495	0.515	0.516	0.507
		MSE↓	0.460	0.581	0.758	0.532	0.592	3.061	0.606	0.502	0.557	0.592	0.539
		AUC↑	0.696	0.444	0.552	0.608	0.492	0.623	0.668	0.639	0.581	0.533	0.517
		OVJE↓	1.049	1.331	1.378	1.280	1.389	1.514	1.257	1.241	1.326	1.294	1.318
	Medium	MAE↓	0.384	0.391	0.396	0.391	0.399	0.447	0.402	0.392	0.394	0.406	0.394
		MSE↓	0.302	0.303	0.313	0.302	0.311	0.364	0.322	0.308	0.304	0.331	0.306
		AUC↑	0.916	0.886	0.603	0.909	0.891	0.903	0.904	0.895	0.897	0.886	0.884
		OVJE↓	0.843	1.026	1.278	0.876	0.894	0.955	0.886	0.969	0.886	1.117	0.982
Restaurant	Low	MAE↓	0.422	0.450	0.462	0.454	0.463	0.849	0.458	0.452	0.460	0.458	0.449
		MSE↓	0.900	0.890	0.902	0.885	0.905	1.585	0.897	0.893	0.913	0.896	0.883
		AUC↑	0.795	0.725	0.617	0.770	0.756	0.762	0.772	0.708	0.637	0.776	0.698
		OVJE↓	0.987	1.054	1.446	0.963	0.990	1.301	0.973	1.016	0.977	1.346	1.034
	No	MAE↓	0.528	0.684	0.723	0.559	0.666	2.016	0.623	0.586	0.852	0.606	0.605
		MSE↓	0.451	0.728	0.804	0.492	0.775	6.032	0.585	0.562	1.199	0.563	0.594
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	1.082	1.691	2.049	1.655	1.600	1.487	1.733	1.651	2.005	1.721	1.709
Temp. Rain	Medium	MAE↓	0.302	0.362	0.359	0.305	0.463	3.358	0.337	0.319	0.317	0.306	0.319
		MSE↓	0.283	0.328	0.315	0.298	0.445	12.191	0.299	0.321	0.299	0.300	0.300
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.827	0.986	1.417	1.095	1.278	2.364	1.189	0.763	1.250	1.308	1.314
	Low	MAE↓	0.440	0.448	0.473	0.462	0.522	0.482	0.472	0.489	0.519	0.477	0.468
		MSE↓	0.740	0.805	0.747	0.809	0.828	0.799	0.775	0.783	0.882	0.775	0.809
		AUC↑	0.674	0.362	0.353	0.540	0.468	0.492	0.422	0.667	0.455	0.457	0.544
		OVJE↓	0.798	0.853	0.972	0.778	0.911	0.848	0.821	0.920	0.875	1.213	0.881
	No	MAE↓	0.260	0.264	0.280	0.273	0.389	0.294	0.271	0.267	0.283	0.261	0.279
		MSE↓	0.556	0.570	0.566	0.564	0.640	0.572	0.560	0.570	0.560	0.560	0.566
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.485	0.490	0.796	0.490	0.692	0.518	0.494	0.492	0.509	0.910	0.550
e_com_SB	High	MAE↓	0.431	0.460	0.497	0.464	0.559	0.658	0.468	0.483	0.559	0.471	0.509
		MSE↓	0.606	0.620	0.656	0.624	0.749	1.120	0.612	0.638	0.818	0.618	0.681
		AUC↑	0.822	0.675	0.674	0.591	0.476	0.684	0.555	0.524	0.720	0.525	0.565
		OVJE↓	0.764	0.915	0.979	0.915	0.958	1.041	0.959	0.953	0.854	0.936	0.943
	Low	MAE↓	0.225	0.232	0.253	0.254	0.325	0.366	0.230	0.230	0.239	0.228	0.237
		MSE↓	0.233	0.247	0.254	0.246	0.311	0.640	0.233	0.237	0.239	0.244	0.240
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.774	0.829	1.159	0.826	1.093	1.227	0.774	0.800	0.824	1.503	0.824

Continued on next page

Table 7 – Continued from previous page

Dataset	Missing Ratio	Metric	Timeflies	Crossformer	DLinear	FACT	FEDformer	MICN	OLinear	PatchTST	TimeMixer	WPMixer	iTransformer
e_com_HN	High	MAE↓	0.382	0.398	0.500	0.427	0.531	0.532	0.428	0.511	4.017	0.466	0.432
		MSE↓	0.299	0.298	0.410	0.342	0.606	0.513	0.329	0.417	34.571	0.385	0.338
		AUC↑	0.756	0.615	0.657	0.710	0.603	0.678	0.687	0.698	0.680	0.588	0.615
		OVJE↓	0.811	0.914	0.963	0.949	0.891	1.566	0.975	0.964	2.852	0.961	0.909
e_com_AI	High	MAE↓	0.309	0.325	0.374	0.335	0.518	0.505	0.372	0.330	0.341	0.360	0.510
		MSE↓	0.337	0.338	0.345	0.336	0.453	0.594	0.357	0.341	0.343	0.356	0.466
		AUC↑	0.728	0.559	0.429	0.463	0.562	0.647	0.493	0.433	0.654	0.479	0.526
		OVJE↓	1.135	1.266	1.312	1.339	1.395	1.863	1.450	1.453	1.671	1.195	1.376
e_com_BY	Medium	MAE↓	0.394	0.432	0.422	0.439	0.482	0.467	0.428	0.429	0.421	0.428	0.447
		MSE↓	0.593	0.653	0.606	0.617	0.672	0.669	0.604	0.638	0.609	0.619	0.624
		AUC↑	0.918	0.566	0.467	0.874	0.687	0.899	0.713	0.747	0.879	0.555	0.594
		OVJE↓	0.911	1.408	1.314	1.104	1.228	0.864	1.442	1.108	0.848	1.108	1.344
e_com_BB	Medium	MAE↓	0.323	0.345	0.347	0.323	0.326	0.448	0.318	0.338	0.348	0.323	0.344
		MSE↓	0.298	0.302	0.302	0.303	0.303	0.451	0.299	0.321	0.321	0.305	0.320
		AUC↑	0.761	0.518	0.524	0.531	0.614	0.721	0.616	0.608	0.661	0.500	0.531
		OVJE↓	1.504	1.835	1.836	1.620	1.609	1.749	1.666	1.521	1.548	1.679	1.784
e_com_NR	Low	MAE↓	0.550	0.581	0.621	0.557	0.576	0.719	0.563	0.586	0.595	0.551	0.562
		MSE↓	0.911	0.925	0.929	0.920	0.948	1.212	0.916	0.974	0.936	0.925	0.906
		AUC↑	0.866	0.562	0.715	0.822	0.639	0.860	0.824	0.782	0.760	0.396	0.767
		OVJE↓	1.331	1.519	1.987	1.365	1.517	1.532	1.367	1.344	1.453	1.764	1.493
e_com_NO	Medium	MAE↓	0.487	0.489	0.496	0.493	0.495	0.682	0.489	0.539	0.670	0.496	0.516
		MSE↓	3.235	3.235	3.235	3.240	3.220	3.532	3.236	3.281	3.308	3.259	3.275
		AUC↑	0.767	0.483	0.479	0.544	0.631	0.707	0.702	0.686	0.601	0.508	0.503
		OVJE↓	1.573	1.915	1.966	1.751	1.821	1.843	1.707	1.612	1.852	1.913	1.718
	Low	MAE↓	0.525	0.564	0.533	0.540	0.602	0.662	0.519	0.565	0.746	0.532	0.536
		MSE↓	1.171	1.175	1.155	1.172	1.264	1.346	1.170	1.229	1.364	1.174	1.163
		AUC↑	0.662	0.502	0.458	0.517	0.583	0.891	0.589	0.634	0.569	0.353	0.605
		OVJE↓	1.329	1.357	1.742	1.243	1.642	1.211	1.229	1.284	1.462	1.899	1.408
e_com_VN	Medium	MAE↓	0.002	0.223	0.228	0.247	0.255	0.167	0.245	0.233	0.188	0.240	0.241
		MSE↓	0.0004	0.051	0.053	0.061	0.070	0.038	0.060	0.055	0.039	0.058	0.059
		AUC↑	0.560	0.477	0.542	0.462	0.505	0.646	0.482	0.526	0.472	0.676	0.517
		OVJE↓	0.404	1.704	1.626	1.933	1.752	1.911	3.012	2.230	2.885	0.757	1.649
	Low	MAE↓	0.071	0.070	0.090	0.077	0.100	0.092	0.075	0.075	0.074	0.075	0.081
		MSE↓	0.158	0.158	0.161	0.159	0.167	0.161	0.158	0.158	0.158	0.159	0.160
		AUC↑	0.764	0.510	0.668	0.693	0.618	0.499	0.646	0.730	0.468	0.222	0.401
		OVJE↓	0.573	0.596	0.825	0.605	0.749	0.839	0.607	0.622	0.609	1.310	0.662
e_com_JP	Low	MAE↓	0.438	0.440	0.453	0.456	0.473	0.499	0.444	0.470	0.453	0.445	0.460
		MSE↓	0.975	0.983	0.986	1.030	1.104	1.038	1.001	0.978	1.025	0.981	1.030
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	1.167	1.295	1.635	1.231	1.507	1.452	1.150	1.196	1.248	1.861	1.376
e_com_HR	Medium	MAE↓	0.346	0.417	0.364	0.441	0.522	0.399	0.363	0.358	0.418	0.338	0.436
		MSE↓	0.573	0.729	0.634	0.701	0.979	0.586	0.595	0.595	0.663	0.584	0.697
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	1.030	1.377	1.560	1.371	1.956	1.160	1.241	1.293	1.370	1.616	1.482
e_com_TA	Low	MAE↓	0.443	0.488	0.469	0.453	0.532	0.492	0.451	0.492	0.474	0.447	0.455
		MSE↓	0.378	0.437	0.401	0.398	0.491	0.448	0.383	0.437	0.415	0.387	0.393
		AUC↑	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
		OVJE↓	0.979	1.085	1.646	0.956	1.347	0.975	0.981	1.029	0.978	1.727	1.368
e_com_ME	High	MAE↓	0.320	0.318	0.332	0.320	0.524	0.500	0.341	0.380	1.819	0.341	0.417
		MSE↓	0.588	0.597	0.598	0.596	0.664	0.892	0.595	0.648	5.602	0.611	0.613
		AUC↑	0.603	0.494	0.480	0.505	0.551	0.589	0.518	0.522	0.528	0.504	0.501
		OVJE↓	1.273	1.304	1.348	1.329	1.662	1.803	1.384	1.348	2.687	1.339	1.532
	Medium	MAE↓	0.494	0.578	0.529	0.536	0.554	0.666	0.500	0.511	0.502	0.498	0.512
		MSE↓	0.737	0.807	0.747	0.795	0.808	1.028	0.747	0.763	0.740	0.739	0.755
		AUC↑	0.831	0.486	0.689	0.508	0.616	0.795	0.834	0.769	0.614	0.364	0.582
		OVJE↓	1.425	1.810	1.947	1.451	1.952	1.454	1.460	1.746	1.477	2.019	1.746
e_com_AD	Medium	MAE↓	0.324	0.420	0.442	0.401	0.466	0.494	0.393	0.371	0.394	0.378	0.398
		MSE↓	0.953	1.011	0.995	1.007	1.069	1.050	0.978	0.967	0.992	0.967	1.018
		AUC↑	0.799	0.579	0.589	0.361	0.595	0.672	0.340	0.386	0.514	0.528	0.428
		OVJE↓	1.030	1.448	1.411	3.532	1.643	1.639	3.519	2.571	2.320	1.121	1.418
	Low	MAE↓	0.263	0.264	0.285	0.290	0.361	0.344	0.267	0.264	0.302	0.272	0.330
		MSE↓	1.477	1.488	1.506	1.514	1.572	1.571	1.505	1.504	1.530	1.490	1.568
		AUC↑	0.684	0.412	0.453	0.491	0.575	0.764	0.609	0.795	0.690	0.377	0.314
		OVJE↓	0.814	0.829	1.210	0.876	1.282	1.071	0.862	0.865	0.865	1.546	0.963
e_com_PL	Medium	MAE↓	0.173	0.190	0.182	0.184	0.229	0.354	0.178	0.179	0.184	0.170	0.182
		MSE↓	0.136	0.139	0.129	0.135	0.164	0.601	0.131	0.132	0.130	0.127	0.135
		AUC↑	0.804	0.638	0.763	0.716	0.568	0.714	0.685	0.690	0.678	0.448	0.789
		OVJE↓	0.996	1.161	1.212	1.110	1.539	1.826	1.130	1.153	1.183	1.498	1.137
e_com_NG	Medium	MAE↓	0.026	1.021	0.895	0.694	0.638	1.279	0.783	0.968	0.875	0.828	0.777
		MSE↓	0.001	1.069	0.809	0.501	0.460	2.998	0.629	0.975	0.855	0.696	0.636
		AUC↑	0.482	0.431	0.436	0.430	0.486	0.559	0.591	0.435	0.440	0.492	0.545
		OVJE↓	2.596	2.609	1.176	0.936	0.713	3.512	5.435	2.926	4.763	0.755	2.179
e_com_BM	Medium	MAE↓	0.115	0.119	0.190	0.148	0.137	0.175	0.147	0.153	0.129	0.142	0.152
		MSE↓	0.054	0.057	0.068	0.055	0.060	0.059	0.056	0.061	0.057	0.057	0.056
		AUC↑	0.754	0.443	0.360	0.659	0.688	0.739	0.727	0.617	0.733	0.693	0.387
		OVJE↓	0.866	0.957	1.236	0.991	0.964	0.944	0.960	1.066	0.930	1.022	1.132

B.2 Full Value Prediction Results

Table 8 presents dataset-level forecasting performance in the absence of the classification head, with MAE and MSE aggregated via the arithmetic mean. **Timeflies** consistently achieves state-of-the-art or competitive results across all regimes. Notably, the catastrophic failure of baseline models on ecom_NG and ecom_VN (Medium) persists in this isolated forecasting setting. This recurrence indicates that the value-only architecture contributes to these results, independently of the auxiliary classification objective; it does not imply uniform robustness to arbitrary observation shifts.

Table 8: Value Prediction Results. Abbreviations: BB-FS = Bitbrains Fast Storage, BB-rnd = Bitbrains Random, Hier. Sales = Hierarchical Sales, Jena Wea. = Jena Weather, Temp. Rain = Temperature Rain. Datasets with the same name appearing in multiple row groups differ in sampling frequency.

Dataset	Frequency	Missing Ratio	Metric	**Timeflies**	Crossformer	DLinear	FACT	FEDformer	MICN	OLinear	PatchTST	TimeMixer	WPMixer	iTransformer
BB-FS	5T	High	MAE↓	0.100	0.288	0.183	0.097	0.158	0.206	0.100	0.102	0.102	0.096	0.098
			MSE↓	0.700	3.165	1.082	0.720	0.751	1.156	0.734	0.741	0.725	0.729	0.731
		Medium	MAE↓	0.535	0.563	0.619	0.561	0.664	0.607	0.574	0.574	0.663	0.601	0.602
			MSE↓	0.611	0.705	0.650	0.731	0.754	0.830	0.699	0.777	0.951	0.872	0.847
		Low	MAE↓	0.128	0.138	0.143	0.137	0.227	0.160	0.141	0.139	0.131	0.147	0.138
			MSE↓	0.281	0.323	0.245	0.295	0.308	0.339	0.292	0.277	0.263	0.322	0.300
		No	MAE↓	0.077	0.087	0.082	0.082	0.129	0.088	0.082	0.079	0.081	0.081	0.083
			MSE↓	0.060	0.075	0.063	0.058	0.108	0.071	0.063	0.065	0.057	0.064	0.059
BB-FS	1H	High	MAE↓	0.100	0.406	0.340	0.149	0.337	0.864	0.095	0.124	0.129	0.152	0.199
			MSE↓	0.675	3.189	2.077	0.825	1.927	2.864	0.683	0.714	0.738	0.837	1.096
		Medium	MAE↓	0.636	0.634	0.584	0.689	0.644	1.647	0.583	0.734	0.682	0.743	0.658
			MSE↓	0.844	0.768	0.630	1.021	0.730	4.246	0.688	1.263	0.990	1.157	0.927
		Low	MAE↓	0.966	1.105	1.002	0.947	1.100	4.165	0.931	0.940	0.953	0.943	0.956
			MSE↓	17.631	17.898	18.183	17.985	17.881	32.882	17.883	18.104	17.813	17.574	18.100
		No	MAE↓	0.167	0.283	0.236	0.145	0.293	0.286	0.164	0.218	0.151	0.167	0.163
			MSE↓	0.270	1.034	0.547	0.255	0.748	0.523	0.293	0.580	0.275	0.312	0.275
BB-rnd	5T	High	MAE↓	0.428	0.418	0.535	0.430	0.524	0.555	0.403	0.533	0.463	0.452	0.387
			MSE↓	7.813	7.832	8.077	9.025	8.019	8.060	8.529	10.642	8.686	8.677	8.147
		Medium	MAE↓	0.167	0.213	0.189	0.175	0.200	0.453	0.175	0.179	0.170	0.161	0.177
			MSE↓	0.149	0.168	0.159	0.157	0.163	0.394	0.154	0.164	0.155	0.151	0.158
		Low	MAE↓	0.185	0.185	0.198	0.213	0.248	0.193	0.200	0.192	0.233	0.242	0.213
			MSE↓	0.499	0.594	0.470	0.681	0.512	0.537	0.561	0.512	0.905	0.922	0.734
BB-rnd	1H	High	MAE↓	0.626	0.740	0.771	0.890	0.743	2.759	0.590	0.736	0.774	0.699	0.664
			MSE↓	6.346	6.623	6.712	6.984	6.547	16.075	6.398	6.667	6.708	6.657	6.375
		Medium	MAE↓	0.152	0.347	0.148	0.184	0.393	3.542	0.143	0.203	0.223	0.142	0.231
			MSE↓	0.085	0.202	0.085	0.114	0.258	19.916	0.088	0.124	0.130	0.086	0.152
		Low	MAE↓	0.211	0.215	0.251	0.250	0.252	0.322	0.223	0.267	0.258	0.243	0.248
			MSE↓	0.548	0.456	0.589	0.659	0.535	0.622	0.530	0.715	0.651	0.624	0.599
		No	MAE↓	0.080	0.134	0.107	0.091	0.158	1.321	0.093	0.082	0.096	0.083	0.097
			MSE↓	0.046	0.081	0.055	0.054	0.087	7.745	0.053	0.046	0.057	0.046	0.057
Car Parts	1M	No	MAE↓	0.357	0.450	0.493	0.367	0.426	0.519	0.369	0.438	0.486	0.393	0.357
			MSE↓	0.666	0.660	0.707	0.667	0.683	0.722	0.658	0.678	0.851	0.657	0.687
Electricity	1D	High	MAE↓	0.021	0.039	0.058	0.020	0.087	0.827	0.020	0.023	0.021	0.022	0.020
			MSE↓	0.015	0.017	0.024	0.015	0.037	31.692	0.016	0.020	0.017	0.018	0.014
		Medium	MAE↓	0.018	0.029	0.038	0.017	0.086	0.236	0.029	0.025	0.020	0.027	0.022
			MSE↓	0.020	0.066	0.141	0.019	0.245	1.817	0.107	0.056	0.030	0.093	0.047
		Low	MAE↓	0.245	0.333	0.319	0.303	0.407	6.578	0.308	0.282	0.328	0.300	0.323
			MSE↓	0.105	0.165	0.146	0.136	0.258	131.788	0.140	0.125	0.156	0.136	0.152
		No	MAE↓	0.024	0.034	0.034	0.023	0.046	0.218	0.026	0.024	0.024	0.026	0.023
			MSE↓	0.008	0.015	0.018	0.009	0.022	0.553	0.013	0.010	0.010	0.012	0.009
Electricity	1H	High	MAE↓	0.022	0.029	0.027	0.022	0.047	0.032	0.023	0.024	0.023	0.024	0.023
			MSE↓	0.020	0.022	0.025	0.022	0.028	0.029	0.024	0.021	0.021	0.024	0.020
		Medium	MAE↓	0.018	0.020	0.017	0.021	0.040	0.030	0.021	0.020	0.028	0.029	0.016
			MSE↓	0.032	0.038	0.036	0.059	0.060	0.041	0.042	0.035	0.055	0.049	0.023
		Low	MAE↓	0.191	0.229	0.190	0.209	0.330	0.428	0.218	0.196	0.210	0.185	0.224
			MSE↓	0.074	0.108	0.078	0.091	0.187	0.393	0.085	0.075	0.092	0.077	0.098
Electricity	1W	High	MAE↓	0.017	0.226	0.084	0.028	0.178	1.675	0.029	0.032	0.037	0.027	0.032
			MSE↓	0.009	0.253	0.037	0.024	0.071	53.552	0.021	0.026	0.041	0.017	0.029
		Medium	MAE↓	0.015	0.073	0.069	0.021	0.082	1.857	0.019	0.025	0.027	0.034	0.022
			MSE↓	0.018	0.484	0.144	0.047	0.086	95.215	0.038	0.064	0.088	0.142	0.042
		Low	MAE↓	0.208	1.448	0.220	0.225	0.482	10.794	0.214	0.256	0.227	0.220	0.248
			MSE↓	0.060	3.273	0.063	0.068	0.333	120.225	0.058	0.089	0.074	0.060	0.081
		No	MAE↓	0.018	0.061	0.064	0.017	0.055	2.679	0.021	0.021	0.022	0.024	0.021
			MSE↓	0.005	0.092	0.026	0.004	0.012	86.802	0.007	0.007	0.008	0.009	0.008
Hier. Sales	1D	Low	MAE↓	0.359	0.340	0.376	0.364	0.382	2.066	0.344	0.371	0.369	0.374	0.373
			MSE↓	0.552	0.517	0.581	0.583	0.574	29.797	0.561	0.578	0.574	0.592	0.599
Jena Wea.	10T	Low	MAE↓	0.254	0.265	0.287	0.248	0.380	0.307	0.251	0.251	0.272	0.257	0.263
			MSE↓	0.227	0.223	0.251	0.225	0.387	0.273	0.230	0.226	0.255	0.233	0.242
Jena Wea.	1H	Low	MAE↓	0.378	0.498	0.411	0.399	0.546	1.723	0.382	0.378	0.393	0.382	0.416
			MSE↓	0.384	0.511	0.396	0.400	0.586	8.300	0.386	0.388	0.401	0.385	0.428

Continued on next page

Table 8 – Continued from previous page

Dataset	Frequency	Missing Ratio	Metric	**Timeflies**	Crossformer	DLinear	FACT	FEDformer	MICN	OLinear	PatchTST	TimeMixer	WPMixer	iTransformer
KDD Cup 2018	1D	High	MAE↓	0.462	0.490	0.468	0.500	0.562	1.958	0.469	0.529	0.521	0.467	0.466
			MSE↓	0.641	0.664	0.586	0.691	0.725	5.238	0.606	0.682	0.644	0.608	0.695
		Medium	MAE↓	0.436	0.445	0.444	0.451	0.463	2.709	0.445	0.450	0.441	0.434	0.450
			MSE↓	1.140	1.166	1.188	1.181	1.200	15.020	1.177	1.158	1.126	1.106	1.187
		Low	MAE↓	0.440	0.447	0.440	0.453	0.479	1.497	0.446	0.444	0.443	0.434	0.459
			MSE↓	1.474	1.423	1.461	1.473	1.555	4.945	1.456	1.449	1.457	1.414	1.509
		No	MAE↓	0.525	0.595	0.529	0.567	0.688	8.134	0.535	0.611	0.558	0.523	0.628
			MSE↓	0.489	0.598	0.520	0.577	0.726	71.655	0.521	0.648	0.532	0.508	0.675
KDD Cup 2018	1H	High	MAE↓	0.616	0.633	0.668	0.683	0.651	0.668	0.673	0.685	0.794	0.703	0.696
			MSE↓	2.577	2.614	2.801	2.857	2.797	2.837	2.774	2.847	3.275	2.866	2.792
		Medium	MAE↓	0.442	0.436	0.451	0.477	0.470	0.467	0.456	0.446	0.465	0.494	0.488
			MSE↓	1.605	1.631	1.634	1.719	1.646	1.651	1.658	1.653	1.654	1.731	1.723
		Low	MAE↓	0.498	0.484	0.491	0.511	0.515	0.527	0.507	0.509	0.523	0.530	0.514
			MSE↓	0.697	0.697	0.679	0.694	0.728	0.788	0.710	0.725	0.733	0.754	0.713
Restaurant	1D	High	MAE↓	0.465	0.556	0.577	0.503	0.564	1.423	0.511	0.495	0.523	0.516	0.510
			MSE↓	0.462	0.583	0.759	0.531	0.601	3.132	0.590	0.497	0.580	0.590	0.552
		Medium	MAE↓	0.386	0.402	0.396	0.392	0.399	0.759	0.399	0.391	0.403	0.404	0.394
			MSE↓	0.294	0.325	0.314	0.301	0.311	1.045	0.319	0.307	0.321	0.328	0.307
		Low	MAE↓	0.449	0.454	0.461	0.455	0.463	1.125	0.461	0.452	0.458	0.457	0.446
			MSE↓	0.886	0.899	0.901	0.889	0.905	2.504	0.902	0.890	0.909	0.898	0.880
		No	MAE↓	0.526	0.713	0.723	0.535	0.672	1.969	0.601	0.589	0.858	0.605	0.608
			MSE↓	0.457	0.800	0.803	0.453	0.787	5.937	0.549	0.570	1.212	0.565	0.599
Temp. Rain	1D	Medium	MAE↓	0.294	0.390	0.372	0.305	0.464	36.365	0.314	0.316	0.316	0.300	0.317
			MSE↓	0.274	0.347	0.332	0.256	0.446	1702.501	0.275	0.314	0.297	0.281	0.299
		Low	MAE↓	0.454	0.447	0.471	0.494	0.524	0.483	0.477	0.488	0.463	0.478	0.469
			MSE↓	0.761	0.800	0.747	0.822	0.849	0.824	0.773	0.767	0.830	0.775	0.812
		No	MAE↓	0.262	0.263	0.282	0.276	0.375	0.418	0.265	0.266	0.275	0.268	0.279
			MSE↓	0.556	0.549	0.567	0.560	0.638	0.677	0.554	0.554	0.559	0.556	0.563
e_com_SB	1H	High	MAE↓	0.415	0.452	0.499	0.489	0.559	0.657	0.470	0.478	0.544	0.479	0.506
			MSE↓	0.584	0.597	0.661	0.643	0.758	1.202	0.615	0.630	0.763	0.627	0.681
		Low	MAE↓	0.225	0.233	0.251	0.231	0.324	0.361	0.231	0.229	0.241	0.227	0.237
			MSE↓	0.234	0.247	0.255	0.239	0.314	0.894	0.233	0.237	0.242	0.244	0.241
e_com_HN	1H	High	MAE↓	0.225	0.233	0.251	0.231	0.324	0.361	0.231	0.229	0.241	0.227	0.237
			MSE↓	0.234	0.247	0.255	0.239	0.314	0.894	0.233	0.237	0.242	0.244	0.241
e_com_AI	1H	High	MAE↓	0.310	0.345	0.377	0.407	0.502	0.478	0.392	0.348	0.498	0.364	0.503
			MSE↓	0.341	0.333	0.347	0.386	0.436	0.512	0.377	0.348	0.466	0.357	0.457
e_com_BY	1H	Medium	MAE↓	0.404	0.430	0.423	0.427	0.480	0.473	0.427	0.431	0.423	0.429	0.451
			MSE↓	0.597	0.650	0.607	0.601	0.667	0.686	0.604	0.647	0.604	0.623	0.626
e_com_BB	1H	Medium	MAE↓	0.315	0.351	0.348	0.322	0.328	0.446	0.323	0.331	0.332	0.321	0.343
			MSE↓	0.296	0.304	0.302	0.302	0.303	0.488	0.302	0.317	0.300	0.304	0.320
e_com_NR	1H	Low	MAE↓	0.549	0.580	0.632	0.564	0.573	0.750	0.558	0.587	0.593	0.552	0.564
			MSE↓	0.895	0.929	0.938	0.921	0.944	1.262	0.919	0.973	0.922	0.914	0.902
e_com_NO	1H	Medium	MAE↓	0.492	0.483	0.496	0.501	0.493	0.678	0.492	0.541	0.512	0.504	0.519
			MSE↓	3.272	3.228	3.235	3.249	3.222	3.545	3.238	3.305	3.238	3.254	3.283
		Low	MAE↓	0.520	0.562	0.534	0.535	0.592	0.785	0.517	0.561	0.580	0.535	0.533
			MSE↓	1.176	1.171	1.156	1.172	1.247	1.539	1.170	1.228	1.189	1.173	1.160
e_com_VN	1H	Medium	MAE↓	0.003	0.216	0.225	0.245	0.253	0.328	0.243	0.237	0.222	0.240	0.243
			MSE↓	0.000	0.047	0.051	0.060	0.067	0.226	0.059	0.057	0.050	0.058	0.060
		Low	MAE↓	0.071	0.068	0.088	0.074	0.099	0.133	0.073	0.074	0.072	0.074	0.080
			MSE↓	0.158	0.157	0.160	0.158	0.165	0.162	0.158	0.158	0.156	0.157	0.160
e_com_JP	1H	Low	MAE↓	0.437	0.463	0.452	0.452	0.464	0.504	0.450	0.461	0.494	0.443	0.468
			MSE↓	0.972	1.013	0.984	1.004	1.078	1.254	1.022	0.971	1.022	0.978	1.030
e_com_HR	1H	Medium	MAE↓	0.346	0.422	0.365	0.441	0.525	0.496	0.358	0.357	0.391	0.341	0.439
			MSE↓	0.575	0.737	0.632	0.688	0.973	0.720	0.584	0.560	0.598	0.548	0.697
e_com_TA	1H	Low	MAE↓	0.439	0.490	0.473	0.460	0.531	0.630	0.449	0.463	0.480	0.451	0.460
			MSE↓	0.375	0.439	0.406	0.406	0.491	0.648	0.385	0.398	0.423	0.389	0.396
e_com_ME	1H	High	MAE↓	0.315	0.314	0.332	0.317	0.472	0.483	0.343	0.380	1.657	0.345	0.533
			MSE↓	0.592	0.579	0.585	0.595	0.626	0.875	0.597	0.649	4.186	0.613	0.719
		Medium	MAE↓	0.496	0.591	0.529	0.520	0.555	0.657	0.492	0.515	0.490	0.489	0.510
			MSE↓	0.737	0.824	0.737	0.766	0.815	0.996	0.725	0.766	0.721	0.724	0.754
e_com_AD	1H	Medium	MAE↓	0.317	0.415	0.442	0.406	0.465	0.521	0.401	0.373	0.415	0.385	0.394
			MSE↓	0.956	1.006	0.994	1.009	1.065	1.087	0.976	0.968	1.009	0.965	1.000
		Low	MAE↓	0.258	0.254	0.285	0.275	0.376	0.292	0.270	0.262	0.282	0.277	0.332
			MSE↓	1.482	1.484	1.506	1.495	1.582	1.173	1.507	1.472	1.507	1.484	1.570
e_com_PL	1H	Medium	MAE↓	0.170	0.189	0.181	0.183	0.225	0.324	0.178	0.176	0.184	0.171	0.181
			MSE↓	0.133	0.141	0.129	0.132	0.163	0.564	0.131	0.131	0.129	0.128	0.134
e_com_NG	1H	Medium	MAE↓	0.029	0.966	0.901	0.827	0.708	1.522	0.762	0.947	0.883	0.842	0.790
			MSE↓	0.002	0.943	0.819	0.699	0.556	3.839	0.594	0.928	0.854	0.719	0.668
e_com_BM	1H	Medium	MAE↓	0.115	0.124	0.189	0.151	0.134	0.176	0.154	0.154	0.145	0.152	0.152
			MSE↓	0.055	0.057	0.067	0.057	0.060	0.063	0.058	0.061	0.061	0.058	0.056

B.3 Ablation Analysis of Mask-Aware Normalization

Tab. 9 examines the impact of integrating Mask-Aware Normalization (MAN) into observation-aware forecasting baselines. We evaluate three representative models across four missingness regimes. For iTransformer and DLinear, MAN yields negligible performance shifts (e.g., iTransformer stays near 0.325/0.250 under zero missingness). This consistency aligns with our theoretical expectation: when the observation mask is all-ones, MAN reduces to standard Instance Normalization, rendering the forward pass mathematically equivalent. Conversely, OLinear suffers consistent degradation (e.g., MSE increases from 0.301 to 0.367). This is because MAN replaces OLinear's default RevIN module—which utilizes learnable affine parameters—with a purely statistical normalization that lacks sufficient representational capacity. These results underscore that simply swapping normalization layers is insufficient; effective exploitation of missingness patterns requires the dedicated architectural innovations proposed in our method.

Table 9: Effect of Mask Aware Normalization on value forecasting performance. We compare models without and with mask aware normalization across different missingness regimes. Lower MSE and MAE are better.

Method	High Missing		Medium Missing		Low Missing		No Missing	
	w/o MAN	w/ MAN	w/o MAN	w/ MAN	w/o MAN	w/ MAN	w/o MAN	w/ MAN
DLinear	1.716 / 0.369	1.724 / 0.387	0.614 / 0.354	0.612 / 0.354	1.438 / 0.380	1.439 / 0.380	0.367 / 0.283	0.367 / 0.282
OLinear	1.586 / 0.311	1.959 / 0.406	0.596 / 0.333	0.684 / 0.378	1.427 / 0.364	1.500 / 0.440	0.301 / 0.240	0.367 / 0.311
iTransformer	1.611 / 0.348	1.614 / 0.358	0.634 / 0.353	0.638 / 0.354	1.462 / 0.379	1.465 / 0.380	0.326 / 0.251	0.325 / 0.250

Tab. 10 presents an ablation study of Mask-Aware Normalization (MAN) within our missing-aware framework. We observe several key insights. First, for iTransformer under the zero-missingness regime, MAN and standard instance normalization yield nearly identical results (MSE: 0.324 vs. 0.326), validating our theoretical assertion that MAN reduces to standard normalization when the observation mask is all-ones. Second, under non-trivial missingness, MAN consistently enhances MAE and OVJE for iTransformer (e.g., MAE decreases from 0.354 to 0.338 under high missingness). This suggests that restricting statistics to observed entries provides a more robust, unbiased normalization signal. Third, OLinear reveals a trade-off between learnability and missingness-awareness: while MAN improves MAE under high missingness (0.328 vs. 0.315), it degrades MSE in the no-missing case (0.317 vs. 0.377) by replacing RevIN's learnable affine parameters with purely statistical normalization. Finally, DLinear shows mixed, metric-dependent effects: under high missingness, MSE improves from 1.714 to 1.614 while OVJE worsens from 1.069 to 1.124; under medium missingness, MAE improves from 0.367 to 0.346 while MSE worsens from 0.611 to 0.634. Under no missingness, changes are small. Thus, MAN does not improve every metric or regime.

B.4 Additional interpretation of observation-guided attention

To analyze how `Timeflies` handles irregular missingness, we visualize its internal attention dynamics on a real-world sample. As shown in Fig. 3(a), the model accurately recovers and extrapolates the underlying trend despite severe discontinuity in the historical context, demonstrating its ability

Table 10: Effect of Mask-Aware Normalization (MAN) on missing-aware forecasting. We compare three baselines with and without MAN across four missingness regimes. Metrics: MSE↓ and MAE↓ evaluate value prediction accuracy; AUC↑ measures observation classification quality (observed=1); OVJE↓ (Obs-Value Joint sMAPE Log Loss) captures joint forecasting performance. "–" indicates AUC is undefined when no missing values exist. Results are averaged across all datasets and prediction horizons.

Method	Setting	High Missing				Medium Missing				Low Missing				No Missing			
		MSE↓	MAE↓	AUC↑	OVJE↓	MSE↓	MAE↓	AUC↑	OVJE↓	MSE↓	MAE↓	AUC↑	OVJE↓	MSE↓	MAE↓	AUC↑	OVJE↓
DLinear	w/o MAN	1.714	0.383	0.638	1.069	0.611	0.367	0.528	1.328	1.433	0.387	0.507	1.245	0.371	0.286	–	1.020
	w/ MAN	1.614	0.324	0.527	1.124	0.634	0.346	0.539	1.271	1.443	0.368	0.505	1.082	0.372	0.289	-	1.014
OLinear	w/o MAN	1.622	0.328	0.740	0.762	0.608	0.352	0.623	1.475	1.438	0.376	0.525	0.958	0.317	0.247	–	0.695
	w/ MAN	1.609	0.315	0.760	0.751	0.633	0.350	0.614	1.329	1.445	0.366	0.537	0.827	0.377	0.267	-	0.664
iTransformer	w/o MAN	1.590	0.354	0.680	0.846	0.638	0.367	0.566	1.263	1.452	0.387	0.524	1.017	0.324	0.251	–	0.730
	w/ MAN	1.655	0.338	0.742	0.816	0.652	0.356	0.591	1.088	1.458	0.379	0.494	0.880	0.326	0.251	-	0.737

to extract coherent temporal structure from highly fragmented inputs. Fig. 3(b)–(e) further reveal how the conditioned attention is formed. The value attention in Fig. 3(b) shows broad activation over the timeline, capturing general temporal dependencies but remaining vulnerable to noisy missing patches. In contrast, the observation-aware bias in Fig. 3(c) and the pairwise reliability attention in Fig. 3(d) encode structured patterns aligned with the missingness distribution. Their fusion in Fig. 3(e) yields selective routing, suppressing unreliable sparse keys while emphasizing informative patches. This reliability-aware filtering limits the propagation of missing-induced noise and explains the stable trend recovery in Fig. 3(a). Fig. 3(f) visualizes the complementary missingness probabilities $\hat{p}_{\text{miss}} = 1 - \hat{p}_{\text{obs}}$. Higher plotted probabilities correspond to future missing events; $\hat{p}_{\text{obs}} = \sigma(\hat{O})$ instead denotes observation probability. This distinction is used consistently in the loss and OVJE, with observed positions as the positive class.

C Controlled Missingness and Contribution Analysis

C.1 Random, structured, and shifted missingness

We use five complete datasets: Electricity H, Bitbrains FS, Restaurant, US Births, and Saugeen Day. Identical masks are shared across variants within each dataset, with the same splits and training protocol; underlying complete evaluation targets are retained for value-error evaluation. IID-10 and IID-50 independently mask 10% and 50% of positions in all splits. BLOCK-50 uses periodic missing blocks at a 50% rate, constructed independently of the underlying values. SHIFT-10-50 uses 10% IID missingness in training/validation and 50% at test time. Each result is an equal-weight arithmetic average of dataset-level metrics; these controlled results use a different evaluation protocol from naturally missing targets in `Shadow`.

Variant A is value-only PatchTST; B adds mask-aware normalization; D incorporates historical-observation modeling and residual fusion; E additionally uses future-observation supervision. Table 11 reports their MAE/MSE and the full model's observation AUC/Brier.

Table 11: Controlled missingness on five complete datasets. Value cells report MAE/MSE; the final column reports E's observation ROC-AUC/Brier. Lower MAE, MSE, and Brier are better; higher AUC is better.

Setting	Train/Val/Test	A: PatchTST	B: Mask-aware	D: Hist-Obs	E: Full	E AUC/Brier
IID-10	10%/10%/10%	0.542/1.165	0.543/1.169	0.524/1.144	0.521/1.142	0.494/0.113
IID-50	50%/50%/50%	0.570/1.156	0.562/1.146	0.539/1.111	0.538/1.110	0.507/0.266
BLOCK-50	50% periodic	0.812/2.517	0.705/2.316	0.706/2.317	0.704/2.323	0.930/0.079
SHIFT-10-50	10%/10%/50%	0.632/1.282	0.612/1.252	0.650/1.327	0.652/1.331	0.504/0.335

Under IID masks, E improves average value accuracy over A while observation AUC remains near chance. The historical mask identifies valid inputs rather than encoding predictable observation dynamics. BLOCK-50 makes observation prediction learnable (AUC 0.930), but E's MSE is slightly worse than B's (2.323 versus 2.316). Under SHIFT-10-50, B performs best and E is worse than both A and B in MAE/MSE. Thus, the in-distribution results do not imply robustness to a previously unseen observation process.

C.2 Separating architecture, historical inputs, and future supervision

Using the same five datasets and controlled protocol, we add No-Missing as a negative control. All variants share masks, initialization of shared components, and training budgets. We compare:

- **M:** a parameter-matched value-only backbone with no observation stream or future-observation supervision.
- **V0 (Neutral-O):** the generic Timeflies architecture with a constant neutral observation representation. The true historical mask is retained only for observed-entry normalization; future-observation supervision is disabled.
- **V1 (Hist-O only):** V0 with true historical masks, intervals, and patch reliability, still without future-observation supervision.

- **V2 (Shuffled Aux-O):** V1 with the full auxiliary head, trained on shuffled future-observation labels preserving the marginal observation rate.
- **V3 (Full):** the same architecture and auxiliary head with correctly aligned future-observation labels.

The comparisons distinguish generic architecture (M to V0), historical input-validity information (V0 to V1), non-informative auxiliary supervision (V1 to V2), and label alignment (V2 to V3).

Table 12: Controlled contribution analysis. M and V0–V3 cells report MAE/MSE, averaged arithmetically with equal weight across the five datasets. AUC uses observed=1; it is undefined (N/A) under No-Missing.

Regime	M	V0	V1	V2	V3	V2 AUC/Brier	V3 AUC/Brier
No-Missing	0.535/1.140	0.520/1.115	0.520/1.115	0.519/1.113	0.519/1.113	N/A/0.000	N/A/0.000
IID-10	0.543/1.169	0.532/1.156	0.524/1.144	0.523/1.143	0.521/1.142	0.501/0.114	0.494/0.113
IID-50	0.562/1.146	0.550/1.129	0.539/1.111	0.539/1.111	0.538/1.110	0.499/0.267	0.507/0.266
BLOCK-50	0.705/2.316	0.709/2.324	0.706/2.317	0.711/2.340	0.704/2.323	0.501/0.249	0.930/0.079

Under No-Missing, V0=V1 and V2=V3: constant masks contain no nontrivial observability information. The M-to-V0 improvement therefore reflects generic architecture, with a small auxiliary optimization effect thereafter. Under IID-10/50, V0-to-V1 gains reflect input-validity awareness under MCAR, not state-related observation dynamics. Under BLOCK-50, correct instead of shuffled labels raise AUC from 0.501 to 0.930 and lower Brier from 0.249 to 0.079, establishing that the model learns the structured observation process. Yet V3 and M have nearly identical value accuracy ($0.704/2.323$ versus $0.705/2.316$ MAE/MSE). Predictable observability alone is therefore insufficient to improve state estimation when missingness is value-independent. Joint modeling supplies an additional observation forecast; its incremental value-forecasting benefit is data- and mechanism-dependent, rather than guaranteed under MCAR or value-independent masks.

D Limitations

1. **Constant overhead:** The observation-aware modules incur computational overhead even when all data is observed. An adaptive gating mechanism that bypasses these modules when missingness is negligible could improve efficiency.
2. **Channel independence assumption:** Following PatchTST, `Timeflies` processes each channel independently. Cross-channel missing patterns (e.g., correlated sensor failures) are not explicitly modeled, though the observation mask implicitly captures some of this information.
3. **Observation-process transferability:** Historical observation information must remain useful at test time. IID masks can provide input-validity cues without predictable observation dynamics; predictable but value-independent masks need not improve value accuracy. Previously unseen shifts in the marginal missing rate or conditional observation mechanism can degrade performance (App. C), so the natural-shift results are not general robustness guarantees.
4. **Evaluation scope:** We focus on `Shadow` with natural missingness, supplemented by controlled masks on five complete datasets. Evaluation on additional domains (e.g., clinical time series, financial time series [55]) would strengthen generalizability claims.

E Broader Impact

Missing-aware time series forecasting has broad applications in domains where sensor failures, communication dropouts, or irregular sampling create incomplete observations—including energy grid management, environmental monitoring, healthcare, and cloud infrastructure management. By providing both value forecasts and observation-probability estimates, `Timeflies` enables downstream decision-making systems to incorporate observation uncertainty. We do not foresee negative societal impacts specific to this work beyond those inherent to time series forecasting in general.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are clearly stated and match the method, benchmark, and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: See Section "Limitations" in the Appendix.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results; all findings are empirical.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The appendix provides the hyperparameter settings. The implementation code and datasets are included in the supplementary material.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: See implementation details and supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See "Experimental Setup".

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper provides probabilistic predictions for future observability, which serve as uncertainty estimates for the classification branch alongside the reported forecasting metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See "Experiment Details".

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper complies with standard ethical guidelines for academic research.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section "Broader Impacts" in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used datasets and codebases are properly cited, and their licenses and terms of use are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new benchmark and code are documented and provided in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.