IMPUTE OR FORECAST DIRECTLY? A SYSTEMATIC BENCHMARK FOR TIME SERIES FORECASTING WITH MISSING VALUES

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series forecasting with missing values (TSFMV) is commonly addressed through either impute-then-forecast pipelines or end-to-end missingness-aware models. Recent studies increasingly advocate end-to-end modeling to avoid error propagation between separately optimized imputation and forecasting stages. Whether this design consistently improves forecasting performance, however, remains unclear because existing studies differ substantially in their datasets, missingness settings, and evaluation protocols. We introduce **TSFMV-Bench**, a systematic benchmark spanning three real-world datasets, seven missingness scenarios, and three missing rates. Beyond performance rankings, we investigate when the relative advantage between the two paradigms changes and how different missingness characteristics shape forecasting difficulty. Our evaluation yields three findings. First, strong impute-then-forecast pipelines remain competitive across a majority of evaluated conditions, challenging the common concern that stage-wise error propagation necessarily makes them inferior. Second, neither paradigm dominates uniformly, and their relative advantage changes with the type and severity of missingness. Third, paradigm preference is shaped by both the recoverability of missing observations and where reconstruction errors occur. Cross-variable information helps only when supporting observations remain available, while errors near the forecasting boundary can substantially disadvantage two-stage pipelines. These findings challenge a blanket preference for end-to-end modeling and motivate evaluating both paradigms across diverse missing-data settings. Code and benchmark resources are available in our anonymous repository.

1 INTRODUCTION

Time series forecasting (TSF) predicts future values from historical observations and serves as a fundamental component of data-driven decision-making. Recent advances have substantially improved forecasting accuracy by modeling trend and periodicity (Lai et al., 2018; Wu et al., 2021), temporal multi-scale structures (Wang et al., 2024b), time-frequency characteristics (Xu et al., 2024), and cross-variable interactions (Wu et al., 2020). Most existing methods, however, are developed under a **complete-observation assumption**, where the historical context available to the forecasting model is assumed to be fully observed. This assumption is frequently violated in practice due to sensor failures, temporary disconnections, irregular measurements, privacy restrictions, and incomplete records. Forecasting models must therefore often infer future dynamics from only partially observed histories. We refer to this problem as **time series forecasting with missing values** (TSFMV).

There are two mainstream ways to tackle the TSFMV task. As illustrated in Figure 1, the *impute-then-forecast* paradigm first fills missing historical values using an imputation model Nie et al. (2024) and subsequently applies a standard forecasting model to predict. In contrast, the *end-to-end* paradigm directly infers future predictions from incomplete observations and their missingness information. Existing end-to-end approaches either incorporate missing-value recovery into the forecasting process Che et al. (2018); Cao et al. (2018); Tang et al. (2020); Yu et al. (2024) or directly construct missing-aware representations without explicitly reconstructing missing observations Chen et al. (2024); Peng et al. (2025); Wu et al. (2025). Recent work has increasingly emphasized end-to-end modeling, motivated in part by the concern that separately optimized imputation

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

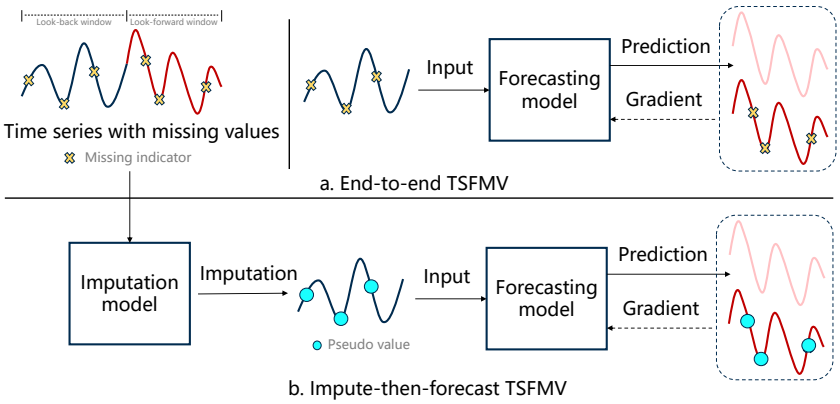


Figure 1: The two TSFMV paradigms. The end-to-end method learns a mapping from partial observations to the forecast, whereas the impute-then-forecast pipeline completes each available historical window before forecasting.

and forecasting stages may propagate reconstruction errors into downstream predictions. Together with reported improvements over conventional two-stage baselines, this trend has increasingly motivated end-to-end modeling for forecasting with incomplete histories.

Yet whether this conclusion generalizes remains unclear because existing evidence is fragmented in both **evaluation protocol** and **missingness coverage**. Each study adopts its own missingness assumptions, missingness construction strategies, evaluation datasets, and baselines (disagreement shown in Table 1). Meanwhile, existing evaluations primarily characterize missingness by structural patterns, such as point-wise and block-wise missingness, while rarely distinguishing the underlying missingness mechanisms—missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) Rubin (1976); Cai et al. (2025). Consequently, a more systematic evaluation is required to understand how the relative effectiveness of different forecasting paradigms varies across missing-data conditions.

We therefore establish **TSFMV-Bench**, a unified benchmark spanning seven missingness scenarios, comprising three mechanism-based settings and four structural patterns, at three missing rates across three real-world datasets. We evaluate five end-to-end approaches and six impute-then-forecast pipelines under a standardized protocol, enabling controlled comparison at the paradigm level. Our evaluation yields three main findings. First, **the superiority of end-to-end forecasting is not universal**. Strong impute-then-forecast pipelines remain competitive across most evaluated conditions, showing that stage-wise error propagation does not necessarily lead to inferior forecasting performance. Second, **paradigm preference changes with the type and severity of missingness**, rather than remaining fixed across evaluation settings. Third, our analyses suggest that these shifts are associated with both the **recoverability of missing observations** and the **forecasting relevance of reconstruction errors**. The benefit of cross-variable dependence diminishes when supporting observations become unavailable, while reconstruction errors near the forecasting boundary can substantially disadvantage impute-then-forecast pipelines. Together, these findings suggest that TSFMV is better viewed as a family of forecasting problems whose preferred modeling paradigm depends on what information remains observable and where missingness disrupts the historical context.

Our contributions are summarized as follows,

- We introduce **TSFMV-Bench**, a unified benchmark with standardized evaluation across seven missingness scenarios, three missing rates, and three real-world datasets.
- We conduct a controlled paradigm-level comparison of end-to-end and impute-then-forecast approaches under a shared evaluation protocol, revealing how their relative performance changes across different missing-data settings.
- We systematically investigate the factors associated with paradigm preference through cross-variable association, observation availability, and recent-history distortion, providing empirical explanations for when each paradigm becomes more effective.

Table 1: Coverage of missing patterns, mechanisms, and rates in time series analysis evaluations.

	Missing patterns				Missing mechanisms			
Method	Point	Block-T	Block-S	Block-ST	MCAR	MAR	MNAR	Missing rates (%)
GRU-D (Che et al., 2018)	–	–	–	–	–	–	✓	80
LGnet (Tang et al., 2020)	✓	✓	–	–	✓	–	–	10/30/50/70/90
BTF (Chen & Sun, 2022)	✓	–	–	✓	✓	–	–	40/60
BiTGraph (Chen et al., 2024)	✓	–	–	–	✓	–	–	10/20/40/60/80
GinAR (Yu et al., 2024)	–	–	✓	–	✓	–	–	25/50/75/90
HD-TTS (Marisca et al., 2024)	✓	✓	–	✓	✓	–	–	5/27/67
S4M (Peng et al., 2025)	–	✓	✓	–	✓	–	–	3/6/12/24
Merlin (Yu et al., 2025)	✓	–	–	–	✓	–	–	25/50/75/90
STMMoE (Wu et al., 2025)	✓	✓	✓	✓	✓	–	–	25/50/75/90
ChannelTokenFormer (Jang et al., 2026)	–	✓	–	✓	✓	–	–	20/40/60/70

2 Related Work

Time Series Forecasting Time series forecasting has been extensively studied from both temporal and cross-variable perspectives (Wang et al., 2026; Chen et al., 2023). Existing methods capture intra-variable dynamics such as trend and periodicity (Lai et al., 2018; Wu et al., 2021), multi-scale patterns (Wang et al., 2024a), and time–frequency characteristics (Xu et al., 2024), while modeling cross-variable dependencies through MLPs (Shao et al., 2022), attention mechanisms (Liu et al., 2024), and graph neural networks (Wu et al., 2020) with different inter-variable inductive biases. Beyond endogenous dependencies, incorporating exogenous covariates has become increasingly important for exploiting external factors that influence future dynamics, as explored by Wang et al. (2024b) and more recent covariate-aware forecasting models (Li et al., 2026). Meanwhile, large-scale pretraining has led to time-series foundation models with transferable and zero-shot forecasting capabilities (Das et al., 2024; Cohen et al., 2025), Despite these advances, most methods are developed under the assumption of fully observed sequences. Missing observations may disrupt spatio-temporal correlations, leaving it unclear whether modeling principles established on complete time series remain effective under different missingness conditions.

Time Series Forecasting with Missing Values TSFMV can naturally be seen as a two-stage task, i.e., applying first time series imputation (Nie et al., 2024) and then general time series forecasting. Existing methods increasingly adopt end-to-end forecasting frameworks that integrate missing-data handling into predictive modeling, avoiding potential error propagation. These approaches mainly differ in how missingness information is incorporated into the learning objective. One line adopts *reconstruction-augmented learning*, where recovering missing observations provides auxiliary supervision alongside downstream forecasting, encouraging representations to preserve information from incomplete histories (Cao et al., 2018; Zhang et al., 2023). Another line is *forecasting-oriented*, i.e., rather than explicitly supervising the reconstruction of missing values, it directly learns predictive representations from incomplete observations through local–global temporal modeling (Tang et al., 2020), missing-aware temporal and cross-variable dependency modeling (Chen et al., 2024; Marisca et al., 2024), variable-level interpolation (Yu et al., 2024), or state-space modeling of block missingness (Peng et al., 2025). Recent methods further improve robustness to varying or open-world missing conditions through representation alignment, adaptive experts, or flexible channel modeling (Yu et al., 2025; Wu et al., 2025; Jang et al., 2026). Despite these advances, existing studies employ various missing patterns, rates, and evaluation protocols, leaving it unclear when reconstruction supervision is beneficial and which forecasting strategies remain robust across different missing conditions.

3 Preliminaries

Notations Let $\mathbf{X} = [x_{t,n}] \in \mathbb{R}^{T\times N}$ denote multivariate time series with T timestamps and N variables, $\mathbf{M} = [m_{t,n}] \in \{0,1\}^{T\times N}$ denote whether the value is missing, where $m_{t,n} = 1$ if $z_{t,n}$ is observed and $m_{t,n} = 0$ otherwise. Considering this time series with missing val-

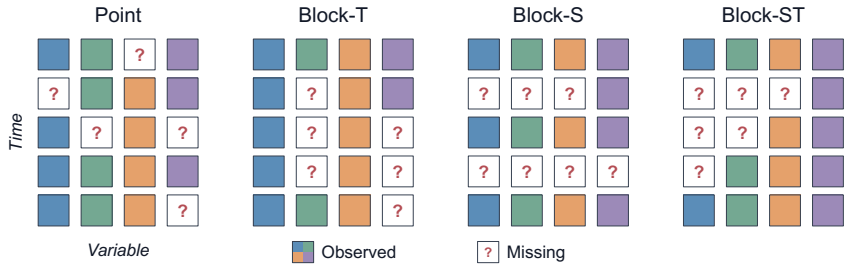


Figure 2: Missingness patterns. Rows and columns represent time steps and variables; marked cells are missing.

ues, a sliding window with a historical length H and a forecasting horizon L is utilized to construct supervised learning samples, which are denoted as $\mathbf{X}^\tau = \mathbf{X}_{\tau-H+1:\tau,:} \in \mathbb{R}^{H\times N}$ and $\mathbf{Y}^\tau = \mathbf{X}_{\tau+1:\tau+L,:} \in \mathbb{R}^{L\times N}$ with corresponding mask matrices $\mathbf{M}_x^\tau = \mathbf{M}_{\tau-H+1:\tau,:} \in \{0,1\}^{H\times N}$ and $\mathbf{M}_y^\tau = \mathbf{M}_{\tau+1:\tau+L,:} \in \{0,1\}^{L\times N}$.

Problem Formulation TSFMV aims to learn a forecasting method $\mathcal{F}_\Theta$ that predicts the next L timestamps based on the look-back window:

$$\widehat{\mathbf{Y}}^\tau = \mathcal{F}_\Theta(\mathbf{X}^\tau, \mathbf{M}_x^\tau) \in \mathbb{R}^{L\times N}. \tag{1}$$

Here, $\mathcal{F}_\Theta$ denotes a general forecasting system, where an end-to-end forecasting model implements $\mathcal{F}_\Theta$ as a single jointly optimized network. An impute-then-forecast manner can instead be expressed as a composition

$$\mathcal{F}_\Theta = \mathcal{P}_\phi \circ \mathcal{G}_\psi, \tag{2}$$

where $\mathcal{G}_\psi$ fills the incomplete historical window and $\mathcal{P}_\phi$ produces the future forecast, with $\Theta = \{\psi, \phi\}$. Notably, TSFMV involves a partial-supervision problem when observations in the future window used as training labels are missing.

Missing Mechanisms and Patterns Missingness can be characterized along two complementary dimensions: the probabilistic process governing whether an observation is available and the structural distribution of missing observations across time and variables. Following established missing-data theory Rubin (1976); Little & Rubin (2019); Yi et al. (2016), we refer to these dimensions as the *missingness mechanism* and the *missingness pattern*, respectively. The former specifies how the probability of missingness depends on observed or unobserved values, whereas the latter describes where missing observations occur and how they are grouped along the temporal and variable dimensions.

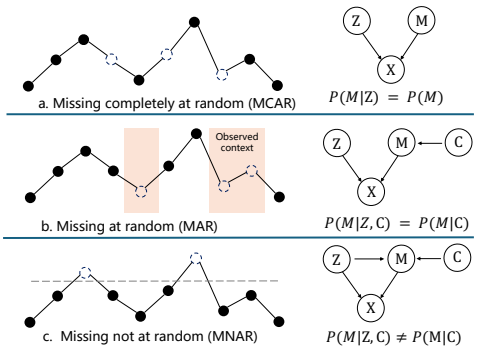


Figure 3: Missingness mechanisms. Arrows into M indicate the variables on which the observation process depends.

Missingness mechanisms. Figure 3 illustrates three standard mechanisms. Under *missing completely at random* (MCAR), missingness is independent of both observed and unobserved values. Under *missing at random* (MAR), missingness may depend on observed information but not on the missing values themselves. Under *missing not at random* (MNAR), the probability of missingness may additionally depend on the unobserved values.

Missingness patterns. Figure 2 illustrates four structural patterns. *Point missingness* consists of isolated missing observations. *Block missingness* instead forms contiguous missing regions along the temporal dimension (Block-T), the variable dimension (Block-S), or both dimensions simultaneously (Block-ST).

4 TSFMV-BENCH

We establish *STFMV-Bench*, a unified benchmark that systematically evaluates diverse forecasting methods across missing patterns, mechanisms, and ratios under a consistent experimental protocol. As shown in Figure 4, TSFMV-Bench covers both end-to-end approaches that directly learn from incomplete histories and two-stage pipelines that separate imputation from forecasting. Rather than merely comparing forecasting accuracy, we aim to understand how missingness characteristics interact with forecasting paradigms and model inductive biases, and to identify which modeling principles remain reliable under incomplete observations.

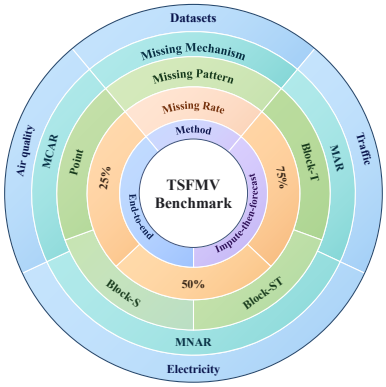


Figure 4: Overview of STFMV-Bench.

4.1 DATASETS

We construct TSFMV-Bench from three public multivariate time series datasets spanning distinct application domains. ETTh1 contains hourly measurements associated with electricity transformers (Zhou et al., 2021). PEMS-BAY contains five-minute traffic-speed measurements from the loop-detector sensors deployed in the San Francisco Bay Area (Li et al., 2018). ChinaAQI contains hourly air-quality measurements from multiple monitoring stations (Yi et al., 2016; Cini et al., 2022). Each dataset is chronologically divided into training, validation, and test sets using a 7:1:2 split. Table 2 summarizes their statistics.

Table 2: Dataset statistics. T and N denote the numbers of timestamps and variables.

Dataset	Domain	Sample rate	#Timestamps	#Variables
ETTh1	Electricity	1 hour	17,420	7
PEMS-BAY	Transportation	5 minutes	52,128	325
ChinaAQI	Air quality	1 hour	8,759	36

4.2 MISSINGNESS CONSTRUCTION

We construct incomplete multivariate time series with missing values by applying synthetic binary masks to the complete observations. We consider seven missingness scenarios in two complementary groups. The first comprises three mechanism-based settings, which control how missingness depends on the underlying observations. The second comprises four structure-based patterns, which control how missing entries are distributed across time and variables.

Under *missing completely at random* (MCAR), each observation is independently masked with the same probability, regardless of its value or other observations:

$$\Pr\left(M_{t,j} = 0 \mid \mathbf{X}\right) = r, \qquad M_{t,j} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1-r), \tag{3}$$

where r denotes the target missing rate.

Under *missing at random* (MAR), a subset of variables $\mathcal{O}$ is randomly selected and kept fully observed, while the missingness of each remaining variable $j \notin \mathcal{O}$ depends on the observed variables through a logistic model:

$$\Pr\left(M_{t,j} = 0 \mid \mathbf{X}t, \mathcal{O}\right) = \sigma\left(\mathbf{X}t, \mathcal{O}^{\top}\mathbf{w}_j + b_j\right), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The coefficients are randomly initialized and normalized, and the intercepts are calibrated to achieve the desired global missing rate. The proportions of fully observed variables are set to 0.50, 0.286, and 0.167 for target missing rates of 25%, 50%, and 75%, respectively.

Under *missing not at random* (MNAR), we adopt a self-masking mechanism in which the missing probability depends directly on the value being masked:

$$\Pr\left(M_{t,j} = 0 \mid \mathbf{X}t, j\right) = \sigma\left(a_j \widetilde{X}t, j + c_j\right), \tag{5}$$

where $\widetilde{X}_{t,j}$ denotes the standardized value of variable j. The coefficient a_j controls the direction and strength of value dependence, while the intercept c_j is calibrated to achieve the target missing rate r. Because the masking probability depends on the value that is subsequently hidden, this setting represents simulated self-masking MNAR.

Missingness patterns. We further consider four structure-based patterns. *Point missingness* uniformly samples exactly $\mathrm{round}(rTN)$ entries from the $T \times N$ observation matrix and sets their masks to zero. *Block-T* introduces contiguous missing segments along the temporal dimension, producing temporal gaps in individual variables. *Block-S* introduces structured missingness along the variable dimension. Specifically, at each time step, a starting variable is sampled uniformly and a cyclically contiguous block of $k = \mathrm{round}(rN)$ variables is masked. The block is independently resampled over time, avoiding boundary bias while keeping each variable partially observable over the full sequence. *Block-ST* introduces joint temporal–variable corruption by partitioning the observation matrix into non-overlapping $p_t \times p_n$ patches and randomly masking patches until the requested number of missing entries is reached. We use $p_t = p_n = 3$ by default. If the remaining masking budget is smaller than one patch, only the required number of entries in the final patch is removed to preserve the target global missing count.

4.3 Baselines

We organize the compared methods into two paradigms according to whether missing observations are explicitly reconstructed before forecasting.

Impute-then-forecast paradigm. Two-stage pipelines first complete the incomplete time series and then apply a forecasting model to the reconstructed sequence. We consider two imputation strategies with different levels of complexity. *Mean imputation* replaces missing observations using variable-wise means estimated from the observed training data. *ImputeFormer* learns low-rank spatio-temporal dependencies for multivariate time series imputation (Nie et al., 2024). Each imputation method is combined with three forecasting architectures representing different inductive biases. *MTGNN* learns adaptive cross-variable dependencies while modeling temporal dynamics with dilated causal convolutions (Wu et al., 2020). *PatchTST* uses a channel-independent Transformer over temporal patches (Nie et al., 2023). *STID* employs a lightweight MLP architecture augmented with spatial and temporal identity embeddings (Shao et al., 2022). Their Cartesian combination yields six impute-then-forecast pipelines. The imputation module is trained independently and remains fixed when the downstream forecasting model is optimized.

End-to-end paradigm. End-to-end methods directly map incomplete observations and their missingness information to future predictions. We include forecasting adaptations of *GRU-D*, which incorporates missingness masks and elapsed-time decay into recurrent dynamics (Che et al., 2018), and *BRITS*, which models incomplete time series through bidirectional recurrent dynamics (Cao et al., 2018). We further evaluate *BiTGraph*, which combines biased temporal convolution with adaptive graph learning (Chen et al., 2024); *S4M*, which integrates temporal prototypes and missing-aware dual-stream representations into structured state-space modeling (Peng et al., 2025); and *HiPatch*, which uses hierarchical intra- and inter-patch graph modeling to capture multi-scale temporal and cross-variable dependencies (Luo et al., 2025).

4.4 Protocol

Considering realistic missing data, end-to-end methods are trained directly on incomplete inputs and partially observed forecasting targets, with supervision restricted to available target observations. For impute-then-forecast pipelines, the imputation model is first trained using the training split and then frozen. It is subsequently used to fill both historical inputs and forecasting targets in the training samples, after which the forecasting model is trained on the completed sequences. For validation and testing, only historical inputs are imputed, while future targets are retained for evaluation. At

Table 3: Forecasting performance averaged across three missing rates (25%/50%/75%) under three missingness mechanisms (MAR, MCAR, and MNAR) and four missingness patterns (POINT, BLOCK-S, BLOCK-T, and BLOCK-ST). Best results are highlighted in **bold red**, and second-best results are underlined. IF denotes ImputeFormer. "–" indicates that the method exceeded the time limit.

	MAR		MCAR		MNAR		POINT		BLOCK-S		BLOCK-T		BLOCK-ST	
Method	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
Dataset: ETTh1														
ONE-STAGE FORECASTING														
BRITS	1.9589	3.3214	1.8574	3.2737	2.1274	3.7334	1.8688	3.2654	1.8273	3.2287	1.8884	3.3603	2.0642	3.7614
GRU-D	1.8216	3.1814	1.8636	3.3118	2.0560	3.7451	1.8143	3.2757	1.7952	3.2355	1.8243	3.3178	2.0786	3.9134
BiTGraph	1.7143	3.0988	1.6967	3.2048	1.9496	3.8523	1.7224	3.2698	1.6952	3.1959	1.7329	3.2744	1.9346	3.6947
Hi-Patch	1.9291	3.4568	1.8876	3.4946	2.1344	4.0219	1.9172	3.5617	1.8875	3.5657	1.9608	3.7172	2.2096	4.2773
S4M	1.8276	3.2276	2.0155	3.6096	2.2578	4.0230	2.0622	3.7663	2.0094	3.6268	2.1527	3.8793	2.2239	4.1519
IMPUTE-THEN-FORECAST														
MTGNN+Mean	3.1288	4.9736	3.0319	5.0687	3.9850	6.4755	3.0765	5.1015	3.0694	5.1181	3.0734	5.1075	3.0929	5.2410
STID+Mean	3.1491	5.0168	3.0379	5.1028	3.8795	6.2967	3.0528	5.1263	3.0577	5.1212	3.0408	5.0832	3.0389	5.1421
PatchTST+Mean	2.7866	4.4482	2.5262	4.3610	3.4989	5.8005	2.5579	4.4273	2.6172	4.5002	2.5499	4.3740	2.6446	4.6089
MTGNN+IF	1.7202	3.0287	1.6937	3.0306	1.9257	3.5191	1.6754	3.0269	1.7102	3.1403	1.7827	3.2876	1.9734	3.7126
STID+IF	**1.5628**	**2.9437**	**1.5188**	**2.9123**	1.6690	**3.2351**	**1.5233**	**2.9174**	**1.5299**	**2.9434**	**1.6046**	**3.0804**	**1.7489**	**3.4688**
PatchTST+IF	1.5860	3.0067	1.5390	2.9824	**1.6661**	3.2950	1.5443	2.9973	1.5589	3.0295	1.6047	3.1123	1.7512	3.5037
Dataset: AQI														
ONE-STAGE FORECASTING														
BRITS	42.1663	62.1568	42.5341	63.3346	43.9772	65.1420	43.8412	65.4989	43.2259	64.6112	43.7156	63.6983	41.8992	60.9137
GRU-D	43.3069	63.1255	41.3277	60.2445	46.1894	68.6049	41.9795	60.9646	42.7551	62.7519	42.5674	61.5445	42.7534	61.3426
BiTGraph	40.7312	60.3611	40.2884	59.3436	43.4425	64.9175	40.5105	59.8681	40.8137	60.5393	42.5300	61.8617	41.1184	59.7696
Hi-Patch	44.4165	64.6143	41.5526	61.3585	45.3635	66.7373	42.8829	63.4583	41.9399	61.6843	**41.5326**	**60.6218**	43.8410	63.2776
S4M	51.6207	69.4647	49.2381	65.5092	53.1396	71.9466	49.7706	66.0073	49.6136	65.6143	50.0835	66.2586	50.2223	66.5476
IMPUTE-THEN-FORECAST														
MTGNN+Mean	55.4390	73.8090	50.9884	67.8703	56.6464	76.0145	50.9820	67.8446	52.5732	69.2076	52.0665	68.6984	51.4110	68.3031
STID+Mean	53.4317	72.4956	49.9644	66.8206	55.5831	75.3697	50.0548	66.9968	50.1303	67.1246	50.2283	66.9355	49.7512	66.8183
PatchTST+Mean	52.9034	72.0135	47.8341	64.4664	53.6138	72.7312	47.9110	64.5264	47.8576	64.5252	47.7113	64.3055	48.3876	65.4729
MTGNN+IF	47.1112	65.0308	47.8907	65.5042	49.3640	67.2968	48.3646	65.8973	46.7952	64.4603	46.8137	64.6101	47.7636	65.4775
STID+IF	**40.4438**	**58.3631**	**40.0876**	**58.0889**	**40.4685**	**58.6411**	**40.2086**	**58.2343**	**40.2362**	**58.1393**	43.0584	61.1645	**40.5549**	**58.5177**
PatchTST+IF	44.8569	64.7082	44.9326	65.0005	44.8814	64.7882	45.0122	65.1586	44.7777	64.8459	46.8546	66.5131	45.1454	65.2315
Dataset: PEMS-BAY														
ONE-STAGE FORECASTING														
BRITS	2.4710	5.1200	2.4456	5.0934	2.6490	5.7112	2.4297	5.0804	2.4573	5.1270	2.4435	5.1314	2.4731	5.1110
GRU-D	2.6712	5.4495	2.6824	5.4918	2.8546	6.0127	2.6764	5.4771	2.6488	5.4079	2.7552	5.5204	2.7921	5.6035
BiTGraph	**2.1303**	**4.7106**	2.0321	4.5486	**2.2702**	**5.2780**	2.0241	4.5427	2.0369	4.5750	**2.0752**	**4.6488**	2.1208	4.7130
Hi-Patch	—	—	—	—	—	—	—	—	—	—	—	—	—	—
S4M	3.1522	5.8187	3.0285	5.6580	3.2317	6.2849	3.0181	5.6171	3.0313	5.6736	3.1018	5.8460	3.0764	5.7722
IMPUTE-THEN-FORECAST														
MTGNN+Mean	4.4858	7.9571	4.3022	7.8034	4.8218	8.7149	4.3067	7.7944	4.3044	7.7834	4.3204	7.7958	4.2330	7.6761
STID+Mean	4.5268	7.9923	4.3142	7.8077	4.8345	8.7032	4.3160	7.8095	4.3072	7.7751	4.3624	7.9055	4.2833	7.8002
PatchTST+Mean	4.5271	8.3687	3.9688	7.7403	4.7245	8.8262	3.9602	7.7472	3.9615	7.7379	3.9352	7.7700	3.9777	7.9444
MTGNN+IF	2.2300	4.9946	1.9818	4.5379	2.3106	5.5444	1.9863	4.5595	2.0108	4.5799	2.1165	4.7132	2.1431	4.7988
STID+IF	2.2291	4.9707	**1.9686**	**4.4244**	2.3221	5.5709	**1.9819**	**4.4534**	**1.9990**	**4.4799**	2.1374	4.6685	**2.1113**	**4.6511**
PatchTST+IF	3.2685	7.2998	3.0666	7.0218	3.1917	7.4051	3.0678	7.0330	3.0870	7.0771	3.2610	7.3209	3.2281	7.2525

every forecasting origin, imputing the historical window is strictly restricted to information available before the forecast horizon and cannot access future observations.

4.5 Metrics

We use mean absolute error (MAE) and root mean squared error (RMSE) as the primary evaluation metrics. Both metrics are computed between the forecasts and ground-truth values, with lower values indicating better forecasting accuracy.

5 Benchmark Evaluation and Analysis

5.1 RQ1: Paradigm Comparison

Table 3 compares end-to-end (E2E) and impute-then-forecast (ITF) methods across three datasets and seven missingness scenarios, with results averaged over the three missing rates. The results provide no evidence that E2E forecasting is uniformly superior. Across the 42 dataset–scenario–metric comparisons, the overall best-performing method belongs to the ITF paradigm in 34 cases. On ETTh1, ITF achieves the best performance under every evaluated scenario. A similar trend is observed on AQI, where ITF leads in all scenarios except Block-T. PEMS-BAY exhibits a more

mixed pattern. E2E, represented by BiTGraph, performs best under MAR, MNAR, and Block-T, whereas ITF leads under MCAR, Point, Block-S, and Block-ST. Overall, strong ITF pipelines remain highly competitive, while the preferred paradigm varies across datasets and missingness scenarios.

5.2 RQ2: Conditions for Paradigm Preference

Figure 5 resolves the paradigm comparison by missing rate. Across the 21 dataset–scenario combinations, nine change their preferred paradigm as the missing rate increases. These transitions follow neither a common threshold nor a consistent direction, indicating that missing severity alone is insufficient for paradigm selection.

The type of missingness further modulates these transitions. MCAR consistently favors ITF across all datasets and missing rates, whereas each of the other missingness scenarios exhibits a preference change on at least one dataset as missingness increases. Dataset characteristics also substantially affect these behaviors. ETTh1 consistently favors ITF, while AQI and PEMS-BAY exhibit rate-dependent and sometimes opposite transitions under the same missingness scenario. Paradigm preference therefore reflects an interaction among missing severity, missingness characteristics, and the underlying data rather than any single factor in isolation.

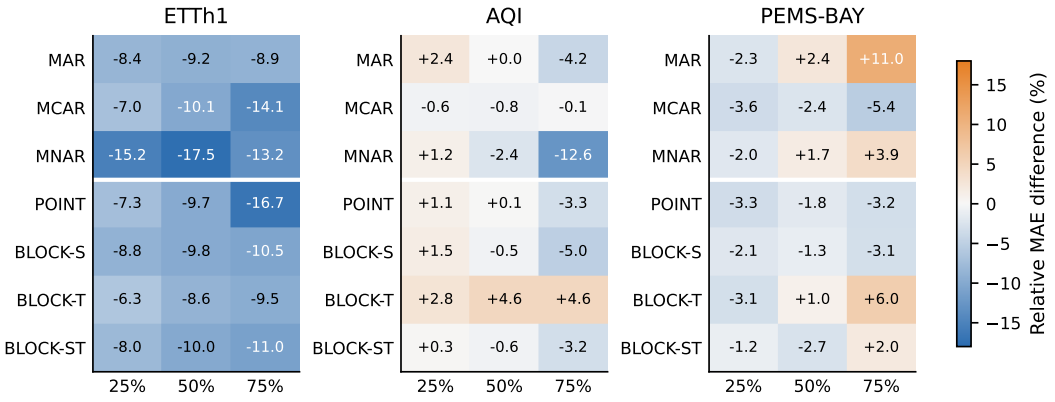
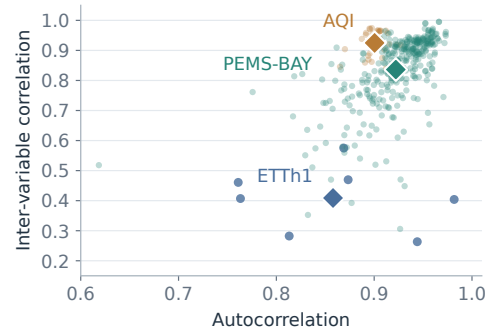


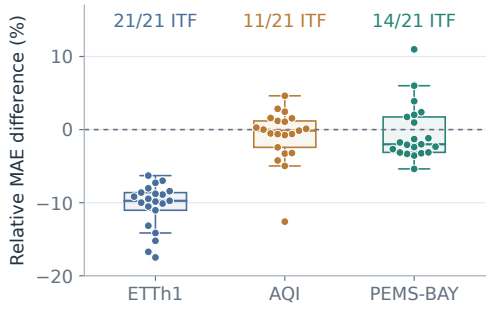
Figure 5: Relative MAE difference between ITF and E2E across missing-data scenarios. Each cell reports $100 \times (\mathrm{MAE}_{\mathrm{ITF}} - \mathrm{MAE}_{\mathrm{E2E}})/\mathrm{MAE}_{\mathrm{E2E}}$. Negative (blue) values favor ITF, while positive (orange) values favor E2E. The three panels cover 63 scenarios in total.

5.3 RQ3: Explaining Paradigm Differences

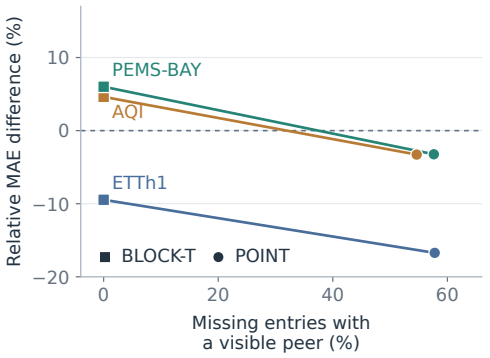
We next examine three factors that may explain paradigm preference: cross-variable association, observation availability, and recent-history distortion. **(i) Cross-variable association.** ETTh1 has substantially weaker cross-variable association than AQI and PEMS-BAY, with an average value of 0.409, while ITF wins all 21 conditions. This suggests that explicit reconstruction is particularly beneficial when cross-variable information provides limited support for direct forecasting. **(ii) Observation availability.** The usefulness of cross-variable dependence also depends on whether peer observations remain accessible. Block-T preserves synchronized observations from other variables while interrupting each variable's temporal history, and E2E wins all four 50%/75% Block-T conditions on AQI and PEMS-BAY. In contrast, Block-S removes multiple variables simultaneously and favors ITF in the corresponding high-missing-rate conditions. This contrast indicates that preserving cross-variable evidence can shift the advantage toward E2E. **(iii) Recent-history distortion.** Reconstruction errors are also more consequential when they occur near the forecasting boundary. On PEMS-BAY, repairing only the final six historical steps recovers 98.4% of the improvement obtained by repairing the full input window, indicating that recent imputation errors largely explain the ITF deficit in this setting. Together, these results suggest that paradigm preference depends not only on how much data are missing, but also on whether cross-variable evidence remains available and where missingness-induced distortion occurs relative to the forecasting boundary.



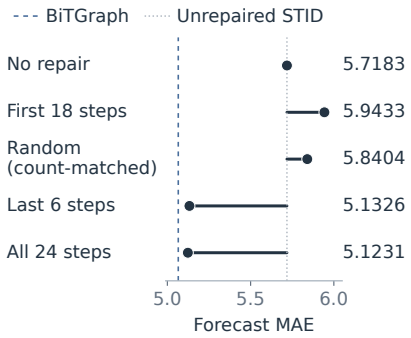
(a) Training-period dependency structure.



(b) Performance across the 63 missing conditions.



(c) Observation availability under 75% missingness.



(d) Recent-history repair on PEMS-BAY.

Figure 6: Analysis of ITF–E2E performance gap. (a) Training-only Pearson associations across 368 variables; diamonds denote dataset means. Cross-variable association averages each variable's three strongest absolute correlations, while temporal association uses weighted lags within the 24-step history. (b) Relative MAE difference (%), $(E_{\mathrm{ITF}} - E_{\mathrm{E2E}})/E_{\mathrm{E2E}}$, over all 63 core conditions; counts denote ITF wins. (c) Observation availability and MAE gaps for paired POINT and BLOCK-T masks at 75% missingness. (d) Fixed ImputeFormer–STID repairs on 2,604 highest-change PEMS-BAY windows under 75% MNAR. Random repair is count-matched to the last-six-step repair; dashed blue denotes archived BiTGraph and dotted gray unrepaired STID. The 98.4% denotes the recovered fraction of the full-repair improvement.

6 Discussion and Limitations

Our results suggest that paradigm preference depends not only on how much data are missing, but also on what predictive information remains accessible. TSFMV-Bench studies these effects through controlled missingness scenarios, while real-world observation processes may combine multiple mechanisms and structural patterns. The identified factors are therefore intended as diagnostic evidence rather than a prescriptive selection rule, leaving data-driven paradigm selection under more complex missingness as an interesting direction for future work.

7 Conclusion

We introduced TSFMV-Bench for systematic comparison of end-to-end and impute-then-forecast approaches under diverse missing-data settings. Our results show that neither paradigm is uniformly superior and that their relative advantage depends on how missingness affects the information available for forecasting. We hope the benchmark provides a common basis for more rigorous evaluation and model development for forecasting with incomplete observations.

AI USE STATEMENT

Generative AI tools were used only for literature discovery, manuscript organization, and language refinement. All research design and analyses were conducted by the authors, and AI-assisted outputs were manually reviewed and verified against the original sources and implementations.

ETHICS STATEMENT

This work uses existing publicly available datasets and does not involve human participants or the collection of new personal data. The introduced missingness settings are constructed for controlled methodological evaluation and do not introduce additional sensitive information. The proposed benchmark is intended to facilitate systematic research on time series forecasting with missing observations. Its results should not be interpreted as evidence of reliability in safety-critical or high-stakes applications without appropriate domain-specific validation.

REPRODUCIBILITY STATEMENT

The code and data for TSFMV-Bench are available at `https://anonymous.4open.science/r/TSFMV-Bench-BE81`. Detailed dataset construction, missingness generation, experimental protocols, and implementation settings are provided in the appendix and the accompanying repository.

REFERENCES

Ruichu Cai, Kaitao Zheng, Junxian Huang, Zijian Li, Zhengming Chen, Boyan Xu, and Zhifeng Hao. Causal view of time series imputation: Some identification results on missing mechanism. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4779–4787, 2025.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

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.

Xinyu Chen and Lijun Sun. Bayesian temporal factorization for multidimensional time series prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4659–4673, 2022.

Zonglei Chen, Minbo Ma, Tianrui Li, Hongjun Wang, and Chongshou Li. Long sequence time-series forecasting with deep learning: A survey. *Information Fusion*, 97:101819, 2023.

Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the gaps: Multivariate time series imputation by graph neural networks. In *International Conference on Learning Representations*, 2022.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, et al. This time is different: An observability perspective on time series foundation models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50907–50951, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 10148–10167, 2024.

Jinkwan Jang, Hyungjin Park, Jinmyeong Choi, and Taesup Kim. Towards robust real-world multivariate time series forecasting: A unified framework for dependency, asynchrony, and missingness. In *International Conference on Learning Representations*, 2026.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104, 2018.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Chenjuan Guo, and Bin Yang. Gcgnet: Graph-consistent generative network for time series forecasting with exogenous variables. In *International Conference on Learning Representations*, volume 2026, pp. 27904–27926, 2026.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

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.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-patch: Hierarchical patch gnn for irregular multivariate time series. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Ivan Marisca, Cesare Alippi, and Filippo Maria Bianchi. Graph-based forecasting with missing data through spatiotemporal downsampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34846–34865. PMLR, 2024.

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*, pp. 2260–2271, 2024.

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.

Jing Peng, Meiqi Yang, Qiong Zhang, and Xiaoxiao Li. S4M: S4 for multivariate time series forecasting with missing values. In *International Conference on Learning Representations*, 2025.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 4454–4458, 2022.

Xianfeng Tang, Huaxiu Yao, Yiwei Sun, Charu Aggarwal, Prasenjit Mitra, and Suhang Wang. Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5956–5963, 2020.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 38626–38652, 2024a.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, pp. 469–498, 2024b.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Chenyu Wu, Zhipeng Ma, Junbo Zhang, Songyu Ke, and Yu Zheng. Spatio-temporal forecasting under open-world missingness with adaptive mixture-of-experts. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 3366–3375, 2025.

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, pp. 22419–22430, 2021.

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*, pp. 753–763, 2020.

Zhijian Xu, Ailing Zeng, and Qiang Xu. Fits: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, volume 2024, pp. 26295–26318, 2024.

Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. ST-MVL: Filling missing values in geo-sensory time series data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 2704–2710, 2016.

Chengqing Yu, Fei Wang, Zezhi Shao, Tangwen Qian, Zhao Zhang, Wei Wei, and Yongjun Xu. GinAR: An end-to-end multivariate time series forecasting model suitable for variable missing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3989–4000, 2024.

Chengqing Yu, Fei Wang, Chuanguang Yang, Zezhi Shao, Tao Sun, Tangwen Qian, Wei Wei, Zhulin An, and Yongjun Xu. Merlin: Multi-view representation learning for robust multivariate time series forecasting with unfixed missing rates. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 3633–3644, 2025.

Kai Zhang, Chao Li, and Qinmin Yang. Trid-mae: A generic pre-trained model for multivariate time series with missing values. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3164–3173, 2023.

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, pp. 11106–11115, 2021.

A Full Forecasting Results

Tables 4–6 report forecasting results across three datasets, seven missingness scenarios, three missing rates, and eleven methods. MAE, RMSE, and R^2 are reported to four decimal places; MAPE is reported as a percentage to two decimal places. Lower MAE, RMSE, and MAPE and higher R^2 indicate better performance.

Red boldface and underlining indicate the best and second-best distinct values within the one-stage and IF groups separately, for each scenario–rate–metric combination. Ties at the displayed precision share the same marking. Mean-based pipelines are included as reference baselines without highlighting. IF denotes ImputeFormer. Dashes indicate unavailable results for Hi-Patch on PEMS-BAY. Metric definitions and target handling are specified in Appendix D.

Table 4: Full forecast results on ETTh1. The best value is in red bold, and the second-best is underlined.

Dataset: ETTh1												
	Missing rate 25%				Missing rate 50%				Missing rate 75%			
Method	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑
MAR												
ONE-STAGE FORECASTING												
BRITS	1.8035	3.1428	63.66%	0.6731	1.9123	3.2798	66.25%	0.6440	2.1610	3.5417	79.83%	0.5849
GRU-D	1.7268	3.0888	61.09%	0.6843	1.7578	3.1162	61.82%	0.6787	1.9801	3.3391	70.92%	0.6310
BiTGraph	1.6017	2.9927	57.04%	0.7036	1.6710	3.0544	59.47%	0.6913	1.8701	3.2494	67.20%	0.6506
Hi-Patch	1.7529	3.2125	62.62%	0.6585	1.8298	3.2867	62.51%	0.6425	2.2047	3.8714	77.83%	0.5040
S4M	1.7182	3.1419	61.05%	0.6733	1.7893	3.1903	63.86%	0.6632	1.9754	3.3507	73.69%	0.6285
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	2.2713	3.8840	73.35%	0.5008	3.3792	5.2628	89.06%	0.0835	3.7360	5.7739	100.53%	-0.1032
STID+Mean	2.4037	4.0222	73.50%	0.4646	3.3284	5.2410	89.17%	0.0910	3.7151	5.7871	100.86%	-0.1083
PatchTST+Mean	2.0868	3.5801	64.49%	0.5758	2.8224	4.4163	76.96%	0.3546	3.4507	5.3481	95.01%	0.0535
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	1.6224	2.9130	55.67%	0.7192	1.6483	2.9703	58.83%	0.7080	1.8899	3.2028	68.92%	0.6605
STID+IF	**1.4668**	**2.8533**	54.25%	**0.7306**	**1.5178**	**2.8950**	55.45%	**0.7226**	**1.7038**	**3.0829**	65.28%	**0.6855**
PatchTST+IF	1.5106	2.9392	**53.63%**	0.7141	1.5314	2.9601	**54.69%**	0.7100	1.7161	3.1209	**61.98%**	0.6777
MCAR												
ONE-STAGE FORECASTING												
BRITS	1.7506	3.1198	59.24%	0.6779	1.8299	3.1934	62.09%	0.6625	1.9916	3.5079	69.57%	0.5928
GRU-D	1.7407	3.1029	59.89%	0.6814	1.8507	3.2888	60.95%	0.6421	1.9994	3.5437	66.25%	0.5844
BiTGraph	1.5731	2.9973	56.68%	0.7027	1.6649	3.1019	58.04%	0.6816	1.8522	3.5151	62.57%	0.5911
Hi-Patch	1.7527	3.2667	60.27%	0.6469	1.8479	3.4046	63.64%	0.6164	2.0621	3.8123	70.32%	0.5190
S4M	1.8413	3.3285	62.21%	0.6334	1.9845	3.5330	64.23%	0.5869	2.2208	3.9673	69.82%	0.4791
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	1.7513	3.2091	57.41%	0.6592	3.6886	6.0240	103.03%	-0.2009	3.6557	5.9729	102.33%	-0.1806
STID+Mean	1.7669	3.3091	56.63%	0.6376	3.6911	6.0263	103.12%	-0.2018	3.6557	5.9729	102.33%	-0.1806
PatchTST+Mean	1.8965	3.4456	58.45%	0.6071	2.4796	4.2898	70.29%	0.3910	3.2024	5.3475	88.30%	0.0537
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	1.6117	2.9501	56.27%	0.7120	1.6848	3.0384	58.71%	0.6945	1.7845	3.1034	63.69%	0.6813
STID+IF	**1.4631**	**2.8586**	54.62%	**0.7296**	**1.4964**	**2.8977**	54.88%	**0.7221**	1.5967	**2.9805**	57.88%	**0.7060**
PatchTST+IF	1.5013	2.9346	**53.35%**	0.7150	1.5254	2.9692	**54.12%**	0.7083	**1.5902**	3.0432	**57.12%**	0.6935
MNAR												
ONE-STAGE FORECASTING												
BRITS	1.9099	3.4165	63.69%	0.6137	2.0829	3.6560	68.24%	0.5577	2.3894	4.1277	79.63%	0.4362
GRU-D	1.7999	3.2872	58.45%	0.6424	1.9828	3.6387	63.93%	0.5618	2.3853	4.3093	79.69%	0.3855
BiTGraph	1.7399	3.4899	57.76%	0.5969	1.9429	3.8885	65.20%	0.4996	2.1661	4.1784	75.71%	0.4222
Hi-Patch	1.7755	3.4151	59.45%	0.6141	2.1786	4.0767	71.45%	0.4500	2.4491	4.5740	82.79%	0.3076
S4M	1.8315	3.4039	62.28%	0.6166	2.1485	3.8403	73.63%	0.5120	2.7933	4.8248	90.61%	0.2296
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	2.8248	5.0727	86.11%	0.1484	4.1470	6.6353	120.87%	-0.4570	4.9831	7.7185	141.62%	-0.9715
STID+Mean	2.5169	4.5669	79.84%	0.3098	4.1050	6.5620	120.27%	-0.4250	5.0167	7.7611	141.42%	-0.9933
PatchTST+Mean	2.4546	4.4376	75.53%	0.3483	3.4391	5.7280	101.72%	-0.0858	4.6030	7.2358	134.10%	-0.7326
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	1.6192	2.9605	53.59%	0.7100	2.0486	3.8159	64.25%	0.5181	2.1092	3.7808	70.28%	0.5270
STID+IF	**1.4752**	**2.8536**	53.07%	**0.7305**	1.6174	**3.1669**	**54.50%**	**0.6681**	1.9144	**3.6849**	**64.73%**	**0.5507**
PatchTST+IF	1.5141	2.9614	**52.80%**	0.7098	**1.6031**	3.1883	54.82%	0.6636	**1.8812**	3.7352	65.38%	0.5383
POINT												
ONE-STAGE FORECASTING												
BRITS	1.7366	3.0722	58.25%	0.6877	1.8528	3.2041	63.47%	0.6603	2.0169	3.5200	67.02%	0.5900
GRU-D	1.7166	3.0867	61.00%	0.6847	1.7822	3.1904	58.92%	0.6632	1.9440	3.5499	62.28%	0.5830
BiTGraph	1.5904	3.0187	56.32%	0.6984	1.6699	3.1643	58.15%	0.6687	1.9069	3.6265	62.08%	0.5648
Hi-Patch	1.7326	3.2208	58.14%	0.6567	1.8970	3.5631	63.29%	0.5799	2.1219	3.9011	73.16%	0.4964
S4M	1.8500	3.3908	62.19%	0.6195	2.0669	3.7594	65.71%	0.5323	2.2697	4.1486	69.10%	0.4305
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	1.8543	3.2730	59.89%	0.6455	3.6882	6.0172	103.02%	-0.1982	3.6871	6.0142	103.11%	-0.1970
STID+Mean	1.7824	3.3476	56.23%	0.6291	3.6888	6.0173	103.02%	-0.1982	3.6871	6.0141	103.10%	-0.1969
PatchTST+Mean	1.9214	3.5063	58.85%	0.5932	2.5134	4.3779	70.79%	0.3658	3.2388	5.3978	89.45%	0.0358
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	1.6050	2.9572	59.70%	0.7106	1.6288	2.9640	58.32%	0.7093	1.7924	3.1596	63.36%	0.6696
STID+IF	**1.4746**	**2.8589**	**53.45%**	**0.7295**	**1.5073**	**2.8986**	54.06%	**0.7220**	**1.5880**	**2.9948**	58.11%	**0.7032**
PatchTST+IF	1.5066	2.9455	53.83%	0.7129	1.5244	2.9831	**54.04%**	0.7055	1.6019	3.0632	**57.01%**	0.6895

Continued on the next page

Table 4: Complete forecast results on ETTh1. (continued)

Dataset: ETTh1												
	Missing rate 25%				Missing rate 50%				Missing rate 75%			
Method	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑
BLOCK-S												
ONE-STAGE FORECASTING												
BRITS	1.7227	3.0818	58.81%	0.6857	1.8222	3.2484	62.30%	0.6508	1.9371	3.3558	64.93%	0.6273
GRU-D	1.7127	3.0807	60.92%	0.6859	1.7792	3.1887	60.41%	0.6635	1.8938	3.4371	61.99%	0.6091
BiTGraph	1.5947	3.0288	55.60%	0.6964	1.6896	3.1520	57.74%	0.6712	1.8013	3.4070	59.07%	0.6159
Hi-Patch	1.8490	3.5591	57.17%	0.5808	1.8594	3.4891	63.17%	0.5971	1.9542	3.6490	65.44%	0.5594
S4M	1.8377	3.3305	61.98%	0.6329	2.0463	3.6963	65.87%	0.5479	2.1443	3.8535	65.58%	0.5086
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	1.8672	3.3868	57.10%	0.6204	3.6706	5.9889	102.81%	-0.1869	3.6704	5.9787	102.43%	-0.1829
STID+Mean	1.8321	3.3959	57.93%	0.6184	3.6707	5.9889	102.81%	-0.1869	3.6704	5.9787	102.43%	-0.1829
PatchTST+Mean	1.9851	3.5856	59.50%	0.5746	2.7306	4.6815	75.21%	0.2747	3.1360	5.2334	85.46%	0.0936
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	1.5994	2.9940	57.02%	0.7034	1.7122	3.1446	56.44%	0.6728	1.8190	3.2823	60.56%	0.6435
STID+IF	**1.4547**	**2.8437**	54.15%	**0.7324**	**1.5233**	**2.9201**	54.47%	**0.7178**	**1.6117**	**3.0663**	56.40%	**0.6889**
PatchTST+IF	1.4995	2.9316	**53.46%**	0.7156	1.5439	2.9976	**54.21%**	0.7026	1.6332	3.1592	**54.41%**	0.6697
BLOCK-T												
ONE-STAGE FORECASTING												
BRITS	1.7647	3.1230	62.49%	0.6773	1.8587	3.2285	65.63%	0.6551	2.0418	3.7293	66.05%	0.5398
GRU-D	1.6885	3.0355	58.44%	0.6951	1.7888	3.2265	61.60%	0.6555	1.9957	3.6914	63.85%	0.5491
BiTGraph	1.5836	2.9729	55.76%	0.7075	1.6926	3.1669	59.31%	0.6681	1.9225	3.6835	63.87%	0.5510
Hi-Patch	1.8798	3.6381	57.72%	0.5620	1.8548	3.4740	63.14%	0.6006	2.1477	4.0395	70.20%	0.4600
S4M	2.0132	3.5716	66.89%	0.5779	2.1000	3.7559	64.91%	0.5332	2.3449	4.3106	67.60%	0.3851
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	1.8498	3.3283	58.80%	0.6334	3.6749	5.9846	102.73%	-0.1852	3.6956	6.0096	102.32%	-0.1951
STID+Mean	1.7416	3.2286	54.66%	0.6550	3.6852	6.0114	102.94%	-0.1958	3.6956	6.0096	102.32%	-0.1951
PatchTST+Mean	1.9117	3.4593	58.91%	0.6040	2.4886	4.2775	70.91%	0.3945	3.2494	5.3852	88.86%	0.0403
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	1.6381	2.9982	56.34%	0.7025	1.7169	3.1081	57.13%	0.6803	1.9932	3.7566	67.20%	0.5330
STID+IF	**1.4839**	**2.8729**	54.14%	**0.7269**	**1.5467**	**2.9591**	56.29%	**0.7102**	1.7832	3.4091	60.80%	0.6154
PatchTST+IF	1.5126	2.9549	**53.36%**	0.7110	1.5608	3.0246	**54.64%**	0.6973	**1.7406**	**3.3574**	**58.32%**	**0.6270**
BLOCK-ST												
ONE-STAGE FORECASTING												
BRITS	1.8601	3.3435	64.59%	0.6301	1.9597	3.6183	64.69%	0.5668	2.3728	4.3223	75.78%	0.3818
GRU-D	1.8151	3.3362	59.55%	0.6317	1.9985	3.7644	63.99%	0.5311	2.4223	4.6397	76.14%	0.2876
BiTGraph	1.6538	3.1366	56.55%	0.6744	1.8455	3.5498	61.67%	0.5830	2.3044	4.3977	75.12%	0.3600
Hi-Patch	1.8840	3.5145	58.51%	0.5913	2.0990	4.1390	70.71%	0.4331	2.6460	5.1785	84.46%	0.1126
S4M	1.9727	3.6446	64.38%	0.5604	2.1562	4.0603	66.47%	0.4544	2.5428	4.7509	79.94%	0.2531
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	1.9976	3.7371	60.22%	0.5378	3.6041	5.9376	101.49%	-0.1667	3.6769	6.0482	103.58%	-0.2105
STID+Mean	1.8361	3.4449	58.35%	0.6073	3.6029	5.9337	101.46%	-0.1651	3.6776	6.0476	103.60%	-0.2103
PatchTST+Mean	2.0190	3.6890	60.65%	0.5497	2.6457	4.6222	74.87%	0.2930	3.2690	5.5155	92.23%	-0.0067
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	1.6589	3.0396	55.04%	0.6942	1.8602	3.4974	57.93%	0.5952	2.4012	4.6007	75.07%	0.2995
STID+IF	**1.5210**	**2.9527**	**53.29%**	**0.7115**	1.6665	**3.2968**	55.81%	**0.6403**	2.0593	4.1568	**66.11%**	0.4282
PatchTST+IF	1.5424	3.0419	53.95%	0.6938	**1.6611**	3.3203	**55.44%**	0.6352	**2.0501**	**4.1490**	67.06%	**0.4303**

Table 5: Full forecast results on AQI. The best value is in red bold, and the second-best is underlined.

Dataset: AQI												
	Missing rate 25%				Missing rate 50%				Missing rate 75%			
Method	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑
MAR												
ONE-STAGE FORECASTING												
BRITS	40.6285	60.0219	104.68%	0.2815	42.0225	62.3107	113.28%	0.2256	43.8478	64.1377	**106.47%**	0.1796
GRU-D	40.7925	60.5638	**104.48%**	0.2684	42.2969	61.7868	113.84%	0.2386	46.8313	67.0258	114.93%	0.1040
BiTGraph	**39.0012**	**58.5685**	112.10%	**0.3159**	**40.1542**	59.4848	**107.40%**	0.2943	43.0382	63.0301	111.27%	0.2076
Hi-Patch	40.5183	59.1682	120.28%	0.3018	43.6835	63.8519	119.01%	0.1869	49.0476	70.8228	116.77%	-0.0004
S4M	49.2939	66.0092	171.78%	0.1310	50.0965	67.7237	162.21%	0.0852	55.4716	74.6613	180.71%	-0.1118
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	49.9671	68.0563	164.18%	0.0762	54.3244	72.0164	187.46%	-0.0344	62.0254	81.3542	207.85%	-0.3200
STID+Mean	46.9583	66.3541	142.56%	0.1219	52.2269	70.3483	174.40%	0.0130	61.1100	80.7844	201.89%	-0.3016
PatchTST+Mean	47.3756	66.4602	143.69%	0.1191	50.8945	69.4688	165.72%	0.0375	60.4401	80.1115	197.63%	-0.2800
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	45.9829	63.9527	151.64%	0.1843	45.4322	63.4986	145.99%	0.1958	49.9186	67.6412	164.13%	0.0875
STID+IF	39.9566	58.8936	113.06%	0.3082	40.1548	**57.6401**	123.22%	**0.3374**	**41.2199**	**58.5555**	126.70%	**0.3162**
PatchTST+IF	44.7623	64.7835	128.81%	0.1629	44.7323	64.6151	128.43%	0.1673	45.0762	64.7261	127.00%	0.1644
MCAR												
ONE-STAGE FORECASTING												
BRITS	43.5236	64.9983	129.71%	0.1574	42.9147	64.9437	118.00%	0.1588	41.1640	60.0619	110.51%	0.2805
GRU-D	40.2952	59.6637	**102.18%**	0.2900	40.7802	59.7376	110.54%	0.2883	42.9077	61.3323	126.99%	0.2498
BiTGraph	40.1629	60.4676	110.87%	0.2708	40.2569	60.1612	**107.17%**	0.2781	40.4453	**57.4020**	126.19%	**0.3428**
Hi-Patch	40.3391	60.2276	109.76%	0.2765	40.6931	60.1119	108.88%	0.2793	43.6255	63.7358	**110.24%**	0.1898
S4M	49.3373	65.4603	180.66%	0.1454	49.3857	65.3327	180.55%	0.1487	48.9911	65.7346	163.80%	0.1382
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	44.9892	62.9235	147.21%	0.2103	53.9005	70.2817	186.88%	0.0148	54.0754	70.4057	188.52%	0.0114
STID+Mean	41.7461	59.6215	122.05%	0.2910	54.0679	70.4352	188.22%	0.0105	54.0792	70.4051	188.57%	0.0114
PatchTST+Mean	45.6410	63.2597	139.96%	0.2019	47.0583	63.4734	153.94%	0.1965	50.8031	66.6660	176.92%	0.1136
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	45.9271	63.8153	151.50%	0.1878	48.1034	65.7288	162.36%	0.1383	49.6416	66.9685	168.45%	0.1055
STID+IF	**39.9220**	**58.8962**	112.94%	**0.3082**	**39.9548**	**57.5976**	120.60%	**0.3383**	**40.3860**	57.7728	123.99%	0.3343
PatchTST+IF	44.8367	64.9365	128.63%	0.1590	44.7969	64.8595	127.96%	0.1610	45.1641	65.2055	128.60%	0.1520
MNAR												
ONE-STAGE FORECASTING												
BRITS	43.6023	64.0459	118.74%	0.1819	41.2308	60.1306	116.63%	0.2789	47.0984	71.2495	**112.36%**	-0.0125
GRU-D	41.7468	62.7411	**108.04%**	0.2149	46.0615	68.1516	121.03%	0.0737	50.7600	74.9221	134.22%	-0.1195
BiTGraph	**39.5396**	**58.5327**	112.49%	**0.3167**	42.5934	64.1996	**108.39%**	0.1780	48.1946	72.0201	128.18%	-0.0345
Hi-Patch	40.7495	60.0691	112.47%	0.2803	44.2731	65.4013	117.25%	0.1469	51.0678	74.7417	137.67%	-0.1142
S4M	50.3212	67.3727	173.37%	0.0947	52.5022	71.2851	177.25%	-0.0135	56.5955	77.1821	189.86%	-0.1881
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	48.4198	68.4412	145.97%	0.0658	56.4076	75.3068	181.50%	-0.1311	65.1119	84.2954	220.92%	-0.4172
STID+Mean	45.4499	66.3354	122.16%	0.1224	55.9881	75.2963	180.45%	-0.1308	65.3114	84.4775	221.64%	-0.4233
PatchTST+Mean	46.8808	65.7866	140.02%	0.1368	51.7000	70.7885	163.15%	0.0006	62.2605	81.6185	209.98%	-0.3286
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	49.1435	66.8086	166.25%	0.1098	48.0282	66.0931	157.83%	0.1288	50.9203	68.9888	164.95%	0.0508
STID+IF	40.0077	59.0233	112.72%	0.3052	**40.2294**	**58.0107**	121.74%	**0.3288**	**41.1683**	**58.8893**	124.58%	**0.3083**
PatchTST+IF	44.5771	64.5223	128.14%	0.1697	44.7984	64.7666	128.08%	0.1634	45.2686	65.0756	128.79%	0.1554
POINT												
ONE-STAGE FORECASTING												
BRITS	44.3602	66.1113	135.92%	0.1283	41.2563	60.7649	**107.76%**	0.2636	45.9070	69.6206	135.19%	0.0333
GRU-D	39.9171	**58.6740**	108.31%	**0.3134**	41.8632	60.8659	117.07%	0.2611	44.1582	63.3539	127.05%	0.1995
BiTGraph	**39.5511**	58.9891	**107.03%**	0.3060	**40.0000**	58.8331	130.31%	0.3097	41.9804	61.7822	**123.22%**	0.2387
Hi-Patch	41.1726	61.2589	112.62%	0.2516	43.2407	64.6987	114.62%	0.1651	44.2354	64.4174	138.63%	0.1724
S4M	49.9118	66.0728	181.52%	0.1293	49.4218	65.7065	181.48%	0.1389	49.9781	66.2426	187.08%	0.1248
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	44.7858	62.6332	144.44%	0.2176	54.0771	70.4397	188.36%	0.0104	54.0830	70.4608	188.26%	0.0098
STID+Mean	41.9827	60.0863	121.22%	0.2799	54.0914	70.4448	188.53%	0.0103	54.0902	70.4594	188.37%	0.0099
PatchTST+Mean	45.7706	63.3804	141.12%	0.1988	47.0917	63.4145	154.31%	0.1980	50.8707	66.7843	177.65%	0.1104
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	45.9221	63.8678	151.69%	0.1864	48.1562	65.7539	162.47%	0.1377	51.0156	68.0702	174.86%	0.0759
STID+IF	39.9779	58.9609	112.77%	0.3067	40.0433	**57.6286**	121.80%	**0.3376**	**40.6046**	**58.1135**	125.63%	**0.3264**
PatchTST+IF	44.8258	64.9308	128.69%	0.1591	44.9268	65.0115	128.29%	0.1570	45.2841	65.5334	129.75%	0.1435

Continued on the next page

Table 5: Complete forecast results on AQI. (continued)

Dataset: AQI												
	Missing rate 25%				Missing rate 50%				Missing rate 75%			
Method	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑
BLOCK-S												
ONE-STAGE FORECASTING												
BRITS	42.0148	62.6717	111.23%	0.2166	42.8440	64.0808	127.47%	0.1810	44.8189	67.0811	131.53%	0.1025
GRU-D	40.0698	59.0220	**108.45%**	0.3052	44.8092	66.9496	118.45%	0.1060	43.3864	62.2840	121.26%	0.2263
BiTGraph	**39.3460**	**58.3878**	111.92%	**0.3201**	40.2806	60.5513	**114.46%**	0.2687	42.8145	62.6787	140.08%	0.2165
Hi-Patch	40.6941	60.0992	111.73%	0.2796	41.6941	61.7244	120.57%	0.2401	43.4314	63.2292	**119.91%**	0.2026
S4M	49.7933	65.8215	183.11%	0.1359	49.1302	65.0724	179.61%	0.1555	49.9171	65.9491	180.28%	0.1326
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	49.4795	66.7673	169.50%	0.1109	54.0732	70.3668	188.82%	0.0125	54.1668	70.4886	189.06%	0.0090
STID+Mean	42.0990	60.4575	120.58%	0.2710	54.1235	70.4289	189.15%	0.0107	54.1684	70.4873	189.09%	0.0091
PatchTST+Mean	45.6674	63.2993	140.40%	0.2009	47.0856	63.5452	154.31%	0.1946	50.8198	66.7311	177.02%	0.1119
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	41.5433	59.9635	121.32%	0.2829	46.3631	64.1820	153.41%	0.1784	52.4792	69.2354	181.05%	0.0440
STID+IF	39.9545	58.8407	113.23%	0.3095	**40.0693**	**57.6401**	121.96%	**0.3374**	**40.6847**	**57.9371**	124.95%	**0.3305**
PatchTST+IF	44.8692	64.8645	128.88%	0.1609	44.6785	64.8600	128.51%	0.1610	44.7853	64.8132	129.16%	0.1622
BLOCK-T												
ONE-STAGE FORECASTING												
BRITS	44.7090	66.2382	121.09%	0.1249	41.9101	60.7370	115.67%	0.2643	44.5277	64.1198	134.33%	0.1800
GRU-D	40.3156	**58.1457**	122.81%	**0.3257**	44.2621	65.2324	118.73%	0.1513	**43.1244**	**61.2555**	**118.32%**	**0.2516**
BiTGraph	41.9032	62.5716	123.18%	0.2191	42.1589	61.1325	129.87%	0.2546	43.5278	61.8810	121.90%	0.2363
Hi-Patch	**39.8369**	59.1999	**116.90%**	0.3010	**41.2130**	**59.8865**	**111.60%**	**0.2847**	43.5478	62.7791	141.47%	0.2139
S4M	51.0629	67.3003	189.34%	0.0966	50.1994	66.4290	178.58%	0.1199	48.9882	65.0464	171.78%	0.1561
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	47.7994	65.3087	158.36%	0.1493	54.4318	70.3031	194.48%	0.0142	53.9683	70.4835	186.56%	0.0092
STID+Mean	42.3042	60.0225	123.91%	0.2815	54.4150	70.2967	194.44%	0.0144	53.9656	70.4872	186.48%	0.0091
PatchTST+Mean	45.2782	62.6600	136.85%	0.2169	47.4573	63.7766	157.57%	0.1888	50.3984	66.4800	172.90%	0.1185
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	44.2573	62.3082	140.45%	0.2257	44.3148	62.8146	137.92%	0.2131	51.8691	68.7075	176.08%	0.0585
STID+IF	40.9561	58.4233	126.56%	0.3192	43.1021	62.5272	128.12%	0.2202	45.1171	62.5431	137.40%	0.2198
PatchTST+IF	45.4928	65.2919	132.02%	0.1498	46.7999	66.9646	136.82%	0.1056	48.2711	67.2828	140.19%	0.0971
BLOCK-ST												
ONE-STAGE FORECASTING												
BRITS	41.8493	61.6057	111.63%	0.2431	41.1351	59.6764	**114.46%**	0.2897	42.7131	61.4590	**124.10%**	0.2467
GRU-D	41.6823	61.1248	117.20%	0.2548	42.1844	60.2341	123.92%	0.2764	44.3935	62.6689	135.35%	0.2167
BiTGraph	**39.9746**	59.6567	116.96%	0.2902	40.4850	58.4865	129.95%	0.3178	42.8955	61.1656	136.64%	0.2538
Hi-Patch	40.9408	60.3967	**110.05%**	0.2725	43.4217	61.9697	126.82%	0.2341	47.1604	67.4663	127.24%	0.0922
S4M	50.0243	66.3763	181.95%	0.1213	49.4330	65.7487	177.18%	0.1378	51.2098	67.5177	184.69%	0.0908
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	47.6966	65.2483	160.22%	0.1509	52.8529	69.4614	183.28%	0.0377	53.6835	70.1995	186.71%	0.0171
STID+Mean	42.9063	60.9787	125.59%	0.2584	52.6691	69.2910	182.45%	0.0424	53.6781	70.1853	186.74%	0.0175
PatchTST+Mean	46.2306	63.8209	142.62%	0.1876	47.9688	65.0212	155.78%	0.1568	50.9635	67.5765	174.82%	0.0892
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	45.1082	63.0836	147.31%	0.2063	49.4415	66.9215	167.68%	0.1068	48.7412	66.4275	161.21%	0.1199
STID+IF	40.0848	**59.0659**	114.29%	**0.3042**	**40.2359**	**57.8334**	122.05%	**0.3329**	**41.3439**	**58.6539**	128.65%	**0.3139**
PatchTST+IF	44.7526	64.6979	128.52%	0.1652	44.9772	65.0137	128.04%	0.1570	45.7065	65.9829	128.21%	0.1317

Table 6: Complete forecast results on PEMS-BAY. The best value is in red bold, and the second-best is underlined.

Dataset: PEMS-BAY												
	Missing rate 25%				Missing rate 50%				Missing rate 75%			
Method	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑
MAR												
ONE-STAGE FORECASTING												
BRITS	2.4091	5.0575	5.81%	0.7262	2.4552	5.0992	5.94%	0.7217	2.5487	5.2033	6.17%	0.7102
GRU-D	2.6330	5.3627	6.34%	0.6922	2.6669	5.4572	6.46%	0.6813	2.7137	5.5285	6.64%	0.6729
BiTGraph	1.9996	4.5134	4.72%	0.7820	**2.0894**	**4.6430**	**4.91%**	**0.7693**	**2.3020**	**4.9753**	**5.46%**	**0.7351**
Hi-Patch	–	–	–	–	–	–	–	–	–	–	–	–
S4M	3.0908	5.7884	7.33%	0.6414	3.1079	5.7113	7.25%	0.6509	3.2577	5.9564	7.73%	0.6203
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	3.1709	6.6492	7.93%	0.5268	4.5360	7.8722	10.52%	0.3367	5.7504	9.3498	13.26%	0.0644
STID+Mean	3.2045	6.6353	8.03%	0.5288	4.5917	7.9530	10.74%	0.3231	5.7841	9.3885	13.41%	0.0566
PatchTST+Mean	3.4890	7.5763	8.97%	0.3857	4.2982	8.0636	10.53%	0.3041	5.7941	9.4663	13.50%	0.0409
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	**1.9529**	4.5067	**4.71%**	0.7826	2.1392	4.8480	5.28%	0.7484	2.5979	5.6290	6.49%	0.6609
STID+IF	1.9934	**4.4991**	4.88%	**0.7834**	2.1392	4.8125	5.40%	0.7521	2.5548	5.6005	6.62%	0.6643
PatchTST+IF	3.1041	7.1353	7.53%	0.4551	3.2122	7.2319	7.96%	0.4402	3.4893	7.5321	8.76%	0.3928
MCAR												
ONE-STAGE FORECASTING												
BRITS	2.4108	5.0517	5.84%	0.7269	2.4536	5.0823	5.87%	0.7236	2.4724	5.1463	5.98%	0.7165
GRU-D	2.6503	5.4676	6.34%	0.6800	2.5985	5.3538	6.27%	0.6932	2.7983	5.6540	6.64%	0.6578
BiTGraph	1.9594	4.4224	4.55%	0.7907	2.0117	4.5160	**4.70%**	0.7817	2.1252	4.7074	4.96%	0.7628
Hi-Patch	–	–	–	–	–	–	–	–	–	–	–	–
S4M	3.0187	5.6493	7.17%	0.6584	3.0088	5.6317	7.13%	0.6605	3.0581	5.6929	7.21%	0.6531
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	2.0849	4.6953	5.13%	0.7640	5.4096	9.3565	13.90%	0.0630	5.4120	9.3583	13.91%	0.0627
STID+Mean	2.1211	4.7087	5.39%	0.7627	5.4095	9.3562	13.90%	0.0631	5.4119	9.3582	13.91%	0.0627
PatchTST+Mean	3.4330	6.9864	9.13%	0.4776	3.7150	7.6571	10.28%	0.3725	4.7585	8.5774	12.47%	0.2126
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	**1.8898**	4.4084	**4.40%**	0.7920	1.9747	4.5497	**4.70%**	0.7785	2.0810	4.6557	**4.89%**	0.7680
STID+IF	1.9312	**4.3770**	4.59%	**0.7950**	**1.9636**	**4.4196**	4.71%	**0.7909**	**2.0109**	**4.4765**	4.94%	**0.7855**
PatchTST+IF	3.0365	7.0052	7.22%	0.4748	3.0671	7.0462	7.37%	0.4686	3.0961	7.0140	7.51%	0.4735
MNAR												
ONE-STAGE FORECASTING												
BRITS	2.5099	5.4307	6.45%	0.6843	2.6428	5.7357	6.92%	0.6479	2.7944	5.9672	7.12%	0.6189
GRU-D	2.6723	5.7230	6.85%	0.6495	2.8489	5.9639	7.30%	0.6193	3.0425	6.3511	7.74%	0.5683
BiTGraph	2.0909	4.8894	5.41%	0.7441	**2.2286**	**5.2104**	**5.81%**	**0.7094**	**2.4910**	**5.7343**	**6.49%**	**0.6481**
Hi-Patch	–	–	–	–	–	–	–	–	–	–	–	–
S4M	3.0361	5.9397	7.62%	0.6224	3.1801	6.2419	8.03%	0.5830	3.4790	6.6731	8.58%	0.5234
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	2.9247	7.3997	8.49%	0.4140	4.9858	8.6424	11.70%	0.2006	6.5550	10.1026	14.19%	-0.0924
STID+Mean	2.9489	7.3803	8.56%	0.4170	5.0139	8.6475	11.79%	0.1996	6.5408	10.0818	14.17%	-0.0879
PatchTST+Mean	3.5331	7.9192	9.57%	0.3288	4.3789	8.5502	11.13%	0.2176	6.2615	10.0093	14.05%	-0.0723
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	**2.0489**	**4.7895**	**5.23%**	**0.7545**	2.2672	5.4462	6.13%	0.6825	2.6158	6.3976	7.33%	0.5619
STID+IF	2.0855	4.8576	5.49%	0.7475	2.2930	5.4948	6.38%	0.6768	2.5878	6.3603	7.44%	0.5670
PatchTST+IF	3.0675	7.1315	7.77%	0.4557	3.1783	7.3809	8.38%	0.4169	3.3292	7.7029	9.00%	0.3649
POINT												
ONE-STAGE FORECASTING												
BRITS	2.4136	5.0544	5.81%	0.7266	2.4201	5.0520	5.79%	0.7268	2.4554	5.1347	5.91%	0.7178
GRU-D	2.6080	5.4137	6.27%	0.6863	2.6612	5.4303	6.24%	0.6844	2.7600	5.5873	6.59%	0.6659
BiTGraph	1.9510	4.4258	4.55%	0.7904	1.9985	4.4851	4.65%	0.7847	2.1229	4.7172	**4.90%**	0.7618
Hi-Patch	–	–	–	–	–	–	–	–	–	–	–	–
S4M	3.0403	5.6334	7.14%	0.6603	3.0174	5.6210	7.10%	0.6618	2.9965	5.5970	7.05%	0.6647
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	2.0949	4.6691	5.22%	0.7667	5.4122	9.3562	13.91%	0.0631	5.4131	9.3578	13.91%	0.0628
STID+Mean	2.1223	4.7149	5.40%	0.7621	5.4126	9.3559	13.91%	0.0632	5.4131	9.3577	13.91%	0.0628
PatchTST+Mean	3.3994	7.0163	9.10%	0.4731	3.7226	7.6568	10.28%	0.3725	4.7585	8.5685	12.45%	0.2142
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	**1.8861**	4.4053	**4.48%**	0.7923	1.9695	4.5052	**4.63%**	0.7828	2.1033	4.7680	5.08%	0.7567
STID+IF	1.9282	**4.3735**	4.60%	**0.7953**	**1.9634**	**4.4179**	4.72%	**0.7911**	**2.0542**	**4.5688**	5.04%	**0.7766**
PatchTST+IF	3.0310	6.9939	7.21%	0.4765	3.0675	7.0595	7.37%	0.4666	3.1048	7.0456	7.53%	0.4687

Continued on the next page

Table 6: Complete forecast results on PEMS-BAY. (continued)

Dataset: PEMS-BAY												
	Missing rate 25%				Missing rate 50%				Missing rate 75%			
Method	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑	MAE↓	RMSE↓	MAPE↓	R^2↑
BLOCK-S												
ONE-STAGE FORECASTING												
BRITS	2.4088	5.0439	5.86%	0.7277	2.4643	5.1195	5.93%	0.7195	2.4987	5.2175	6.07%	0.7086
GRU-D	2.6248	5.4245	6.23%	0.6851	2.6093	5.2988	6.22%	0.6995	2.7122	5.5003	6.54%	0.6762
BiTGraph	1.9605	4.4424	4.58%	0.7888	2.0114	4.5136	**4.67%**	0.7820	2.1387	4.7691	5.04%	0.7566
Hi-Patch	–	–	–	–	–	–	–	–	–	–	–	–
S4M	3.0263	5.6785	7.20%	0.6549	3.0126	5.6576	7.17%	0.6574	3.0551	5.6846	7.20%	0.6541
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	2.1193	4.7063	5.35%	0.7629	5.3859	9.2861	13.80%	0.0771	5.4081	9.3579	13.90%	0.0628
STID+Mean	2.1038	4.6117	5.29%	0.7724	5.4100	9.3557	13.90%	0.0632	5.4079	9.3578	13.90%	0.0628
PatchTST+Mean	3.4095	6.9991	9.08%	0.4757	3.7259	7.6402	10.25%	0.3753	4.7490	8.5744	12.45%	0.2131
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	**1.9198**	4.4344	**4.54%**	0.7895	1.9995	4.5767	4.75%	0.7758	2.1132	4.7285	**4.95%**	0.7607
STID+IF	1.9401	**4.3907**	4.64%	**0.7937**	**1.9853**	**4.4628**	4.80%	**0.7868**	**2.0717**	**4.5861**	5.04%	**0.7749**
PatchTST+IF	3.0470	7.0310	7.26%	0.4709	3.0857	7.0926	7.45%	0.4616	3.1283	7.1077	7.64%	0.4593
BLOCK-T												
ONE-STAGE FORECASTING												
BRITS	2.3945	5.0129	5.71%	0.7311	2.4355	5.1102	5.81%	0.7205	2.5004	5.2712	5.99%	0.7026
GRU-D	2.6110	5.3067	6.23%	0.6986	2.7186	5.4314	6.42%	0.6843	2.9359	5.8232	6.95%	0.6371
BiTGraph	1.9750	4.4335	4.59%	0.7896	**2.0301**	4.5810	**4.80%**	0.7754	**2.2204**	**4.9319**	**5.15%**	**0.7397**
Hi-Patch	–	–	–	–	–	–	–	–	–	–	–	–
S4M	3.0723	5.8033	7.39%	0.6395	3.0917	5.8316	7.39%	0.6360	3.1415	5.9031	7.42%	0.6270
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	2.1449	4.7214	5.38%	0.7614	5.4095	9.3089	13.85%	0.0725	5.4067	9.3572	13.90%	0.0629
STID+Mean	2.2579	5.0064	5.88%	0.7317	5.4223	9.3527	13.91%	0.0638	5.4069	9.3573	13.90%	0.0629
PatchTST+Mean	3.3339	7.0543	8.90%	0.4674	3.7229	7.6623	10.28%	0.3716	4.7488	8.5933	12.47%	0.2097
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	**1.9135**	**4.4025**	**4.51%**	**0.7926**	2.0824	4.6960	4.92%	0.7640	2.3537	5.0410	5.66%	0.7280
STID+IF	1.9676	4.4231	4.74%	0.7906	2.0498	**4.5476**	4.95%	**0.7787**	2.3948	5.0347	5.98%	0.7287
PatchTST+IF	3.1025	7.0803	7.55%	0.4635	3.2437	7.3454	7.94%	0.4225	3.4367	7.5370	8.71%	0.3920
BLOCK-ST												
ONE-STAGE FORECASTING												
BRITS	2.4149	5.0340	5.81%	0.7288	2.4606	5.1010	5.92%	0.7215	2.5437	5.1980	6.06%	0.7108
GRU-D	2.6692	5.4268	6.23%	0.6848	2.7504	5.5257	6.42%	0.6732	2.9567	5.8580	6.89%	0.6327
BiTGraph	1.9757	4.4581	4.64%	0.7873	2.0975	4.6854	**4.91%**	0.7650	**2.2893**	**4.9954**	**5.45%**	**0.7329**
Hi-Patch	–	–	–	–	–	–	–	–	–	–	–	–
S4M	3.0370	5.6982	7.20%	0.6525	3.0619	5.7671	7.30%	0.6440	3.1303	5.8513	7.47%	0.6336
IMPUTE-THEN-FORECAST: Mean												
MTGNN+Mean	2.1249	4.7847	5.25%	0.7550	5.1949	8.9313	13.15%	0.1463	5.3793	9.3124	13.80%	0.0718
STID+Mean	2.2240	4.9145	5.71%	0.7415	5.2503	9.1779	13.46%	0.0985	5.3755	9.3081	13.79%	0.0727
PatchTST+Mean	3.4330	7.1938	9.30%	0.4461	3.7145	7.8839	10.51%	0.3348	4.7855	8.7554	12.62%	0.1796
IMPUTE-THEN-FORECAST: Imputerformer												
MTGNN+IF	**1.9524**	4.4700	**4.52%**	0.7861	2.0934	4.7348	5.01%	0.7601	2.3834	5.1915	6.01%	0.7115
STID+IF	1.9566	**4.4176**	4.70%	**0.7911**	**2.0416**	**4.5157**	5.02%	**0.7817**	2.3356	5.0199	5.90%	0.7303
PatchTST+IF	3.0804	7.0784	7.38%	0.4638	3.1968	7.2483	7.85%	0.4377	3.4071	7.4309	8.55%	0.4090

B Full Imputation Results

Tables 7–9 compare Mean and ImputeFormer across all 63 benchmark conditions. Mean uses its observed mean within the 24-step history, falling back to training-global-mean when no observations are available.

Errors are computed on artificially hidden entries against the preprocessed reference and averaged over histories. Lower values are better. Red boldface and underlining mark the best and second-best distinct values among the three imputers.

Table 7: Complete history-imputation results on ETTh1. MAE and RMSE are averaged over test histories; lower values are better.

Dataset: ETTh1						
	Missing rate 25%		Missing rate 50%		Missing rate 75%	
Method	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
MAR						
Mean	2.3362	3.7492	2.4516	4.5987	2.6911	5.2976
ImputeFormer	**0.4864**	**0.6750**	**0.5685**	**0.9114**	**1.0038**	**1.5799**
MCAR						
Mean	2.4473	4.2333	2.5293	4.4322	2.6164	4.6958
ImputeFormer	**0.4405**	**0.7204**	**0.6577**	**1.1486**	**0.9768**	**1.6470**
MNAR						
Mean	3.2727	5.8805	3.0286	5.5699	3.5546	6.4160
ImputeFormer	**0.6601**	**1.2321**	**0.9911**	**2.0259**	**1.5504**	**3.0766**
POINT						
Mean	2.4959	4.3722	2.5544	4.5249	2.6432	4.7428
ImputeFormer	**0.4480**	**0.7340**	**0.6457**	**1.1218**	**1.0415**	**1.7719**
BLOCK-S						
Mean	2.5478	4.4738	2.5583	4.5258	2.5887	4.6077
ImputeFormer	**0.4028**	**0.6080**	**0.7332**	**1.2595**	**1.0829**	**1.9210**
BLOCK-T						
Mean	2.4693	4.2479	2.5164	4.4327	2.6131	4.6754
ImputeFormer	**0.8143**	**1.3939**	**1.0110**	**1.7347**	**1.3669**	**2.3921**
BLOCK-ST						
Mean	2.7911	4.5015	2.8023	4.8626	3.1084	5.4723
ImputeFormer	**0.8970**	**1.5044**	**1.2408**	**2.2478**	**1.8379**	**3.3501**

Table 8: Complete history-imputation results on AQI. MAE and RMSE are averaged over test histories; lower values are better.

Dataset: AQI						
	Missing rate 25%		Missing rate 50%		Missing rate 75%	
Method	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
MAR						
Mean	31.4926	40.2094	34.1548	44.4918	45.6108	63.8101
ImputeFormer	**10.4574**	**17.2575**	**12.6651**	**21.1418**	**18.1017**	**27.5019**
MCAR						
Mean	29.3904	37.2889	29.7999	37.9804	30.9948	40.2678
ImputeFormer	**8.0518**	**13.7286**	**9.3116**	**15.7549**	**12.2158**	**19.9515**
MNAR						
Mean	34.5908	45.2258	34.6490	45.2414	38.0921	51.2377
ImputeFormer	**10.7428**	**20.3863**	**12.4129**	**22.0606**	**17.1458**	**27.4490**
POINT						
Mean	29.2581	36.9515	29.7514	37.9432	30.8180	39.9269
ImputeFormer	**7.9323**	**13.4771**	**9.4641**	**16.1060**	**12.5583**	**20.2634**
BLOCK-S						
Mean	29.3284	37.3094	29.5856	37.6598	30.8554	39.8999
ImputeFormer	**8.3759**	**14.3614**	**9.9719**	**16.7072**	**13.0368**	**21.1474**
BLOCK-T						
Mean	28.7768	36.4615	30.3237	38.3474	30.7483	39.3375
ImputeFormer	**11.9493**	**18.4396**	**17.6144**	**24.8586**	**25.5230**	**33.5465**
BLOCK-ST						
Mean	31.8828	40.0871	33.5602	43.1904	38.0699	50.1209
ImputeFormer	**9.8212**	**16.2783**	**12.3062**	**20.6450**	**15.9565**	**25.5559**

Table 9: Complete history-imputation results on PEMS-BAY. MAE and RMSE are averaged over test histories; lower values are better.

Dataset: PEMS-BAY						
	Missing rate 25%		Missing rate 50%		Missing rate 75%	
Method	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
MAR						
Mean	2.2140	4.1543	2.2742	4.3753	3.4525	6.4466
ImputeFormer	**1.3104**	**2.6432**	**1.5711**	**3.1520**	**2.2379**	**4.2002**
MCAR						
Mean	1.9148	3.5259	1.9385	3.5985	2.0155	3.7935
ImputeFormer	**0.7002**	**1.2767**	**0.8524**	**1.6400**	**1.1374**	**2.2393**
MNAR						
Mean	2.8265	5.5462	2.6747	5.4056	2.9289	5.9496
ImputeFormer	**1.9121**	**4.2099**	**2.0240**	**4.4401**	**2.2542**	**4.7979**
POINT						
Mean	1.9168	3.5358	1.9389	3.5969	2.0132	3.7871
ImputeFormer	**0.7038**	**1.2890**	**0.8510**	**1.6319**	**1.1439**	**2.2358**
BLOCK-S						
Mean	1.9111	3.5146	1.9381	3.5887	2.0139	3.7882
ImputeFormer	**0.8087**	**1.4539**	**0.9900**	**1.8554**	**1.2879**	**2.5030**
BLOCK-T						
Mean	1.9279	3.5310	1.9388	3.5893	2.0074	3.7598
ImputeFormer	**1.1065**	**1.9676**	**1.3227**	**2.4673**	**1.7550**	**3.1368**
BLOCK-ST						
Mean	2.0611	3.8281	2.1448	4.0422	2.6356	4.9976
ImputeFormer	**1.0363**	**2.0335**	**1.4218**	**2.7651**	**1.9034**	**3.6417**

C Missingness Settings and Realized Rates

The scenarios in the result tables use the following generators. Let $Z \in \mathbb{R}^{T\times N}$ be the preprocessed reference, r the nominal missing rate, and $M_{t,j} = 1$ indicate a retained entry.

Probabilistic mechanisms. MCAR samples independent $M_{t,j} \sim \text{Bernoulli}(1-r)$. MAR keeps a randomly selected subset of variables $\mathcal{O}$ fully observed and masks the remaining variables with probability

$$\Pr(M_{t,j} = 0 \mid Z_{t,\mathcal{O}}) = \sigma(Z_{t,\mathcal{O}}^{\top} w_j + b_j), \qquad j \notin \mathcal{O}. \tag{6}$$

Here σ is the logistic function. Random coefficients are scaled to unit logit standard deviation, and intercepts are calibrated on Z to the specified deletion probability. The nominal fully observed fractions are 0.5, $1 - 0.5/0.7$, and $1 - 0.75/0.9$ at 25%, 50%, and 75% missingness, respectively, subject to integer column counts and at least one conditioning variable. Self-masking MNAR uses

$$\Pr(M_{t,j} = 0 \mid Z_{t,j}) = \sigma(a_j \widetilde{Z}_{t,j} + b_j), \tag{7}$$

where $\widetilde{Z}$ is standardized using the training time partition. Intercepts are calibrated on that partition. The sign of a_j determines whether high or low values are more likely to be missing.

Structural patterns. POINT uniformly selects $\text{round}(rTN)$ distinct entries. Block-T independently removes a contiguous temporal block of length $k = \text{round}(rT)$ from each variable, with the starting timestamp sampled uniformly for each variable. Its realized rate is k/T. Block-S removes $k = \text{round}(rN)$ consecutive variables from a random starting column at each timestamp, wrapping at the last column. Its realized rate is k/N. Block-ST randomly orders non-overlapping 3×3 time–variable patches and removes entries until the required count is reached; boundary patches may be smaller, and the last patch may be partial.

Realized rates. Table 10 reports the percentage of masked entries across the S test histories, each of length 24:

$$r_{\text{hist}} = 100 \left(1 - \frac{1}{24 * S * N} \sum_{s=1}^{S} \sum_{t=1}^{24} \sum_{j=1}^{N} M_{t,j}^{(s)} \right). \tag{8}$$

Table 10: Realized missing percentages across test histories. Column headers give nominal rates; entries count repeated timestamps once per history.

	ETTh1			AQI			PEMS-BAY		
Scenario	25%	50%	75%	25%	50%	75%	25%	50%	75%
MAR	23.91	49.18	73.11	24.98	49.94	75.48	24.90	50.01	74.83
MCAR	25.28	50.12	75.13	24.94	49.97	75.29	25.01	50.02	75.04
MNAR	28.13	53.02	76.39	25.14	50.36	75.27	25.08	49.85	74.70
POINT	25.13	50.34	75.25	25.13	50.25	75.07	24.98	50.00	74.98
BLOCK-S	28.57	57.14	71.43	25.00	50.00	75.00	24.92	49.85	75.08
BLOCK-T	25.00	50.69	76.15	24.73	51.35	75.49	24.47	49.58	75.25
BLOCK-ST	24.68	49.93	74.62	24.81	49.93	75.38	24.90	49.94	75.01

The calculation uses the synthetic masks and counts repeated timestamps once per history in which they occur.

D Metric Definitions and Evaluation Protocol

Forecasting metrics. Let $y_{m,l,n}$ and $\hat{y}_{m,l,n}$ denote the ground-truth and predicted values of the n-th variable at the l-th forecasting step for the m-th test sample, respectively, where $m = 1, \ldots, M$, $l = 1, \ldots, L$, and $n = 1, \ldots, N$. We define the prediction error as

$$e_{m,l,n} = \hat{y}_{m,l,n} - y_{m,l,n}. \tag{9}$$

All metrics are computed after inverse normalization. Specifically,

$$\mathrm{MAE} = \frac{1}{MLN} \sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{n=1}^{N} |e_{m,l,n}|, \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{MLN} \sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{n=1}^{N} e_{m,l,n}^2}, \tag{11}$$

$$\mathrm{MAPE} = \frac{100}{|\Omega_{0.1}|} \sum_{(m,l,n)\in\Omega_{0.1}} \frac{|e_{m,l,n}|}{|y_{m,l,n}| + 10^{-12}}, \tag{12}$$

$$R^2 = 1 - \frac{\sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{n=1}^{N} e_{m,l,n}^2}{\sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{n=1}^{N} \left(y_{m,l,n} - \bar{y}\right)^2}, \tag{13}$$

where

$$\bar{y} = \frac{1}{MLN} \sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{n=1}^{N} y_{m,l,n}, \tag{14}$$

and

$$\Omega_{0.1} = \{(m, l, n) \mid |y_{m,l,n}| > 0.1\}. \tag{15}$$

Here, the metrics are pooled over all test samples, forecasting steps, and variables. R^2 is computed using the pooled target mean $\bar{y}$ and may therefore take negative values. For an identical set of targets, R^2 and RMSE induce the same model ordering before rounding.

E Discussion: Relating Imputation to Forecasting Performance in Two-Stage Models

Controlled input substitutions. We freeze forecasting weights and normalization, replacing only imputed history entries. Each substitution covers 189 backbone–condition pairs. TrainingMean-trained predictors receive ImputeFormer instead of TrainingMean (TM→IF); IF-trained predictors

receive SampleMean instead of IF (IF→SM). TM uses training-observation means; SM uses per-variable observed means within each 24-step history, with TM fallback when empty. The two substitutions are therefore not strictly symmetric. Observations, time features, and targets remain fixed; effects use paired same-runtime baselines and synchronized MTGNN graph randomness.

Overall association. On natively observed, artificially masked history entries, IF outperforms both mean strategies in all 63 conditions. TM→IF improves forecast MAE in 160/189 pairs (84.7%). Excluding changes within $\pm 2\%$ leaves 84 improvements versus 19 declines; at $\pm 5\%$, the counts are 54 versus 4. IF→SM worsens all 189 pairs, including 177 declines exceeding 2% (Figure 7).

Interpretation. The 29 declines have median magnitude 2.39%; 15 concern AQI–PatchTST. Native-target scoring retains 84 improvements exceeding 2%, while reducing such declines to 13. Three inference seeds retain declines in 13/14 MTGNN cases; the remaining approximately 0.002% decline changes sign. Thus, better imputation generally benefits fixed predictors, with localized exceptions. Input-distribution changes remain entangled with reconstruction accuracy; these single-training-seed results do not establish a universal monotonic relationship or statistical significance.

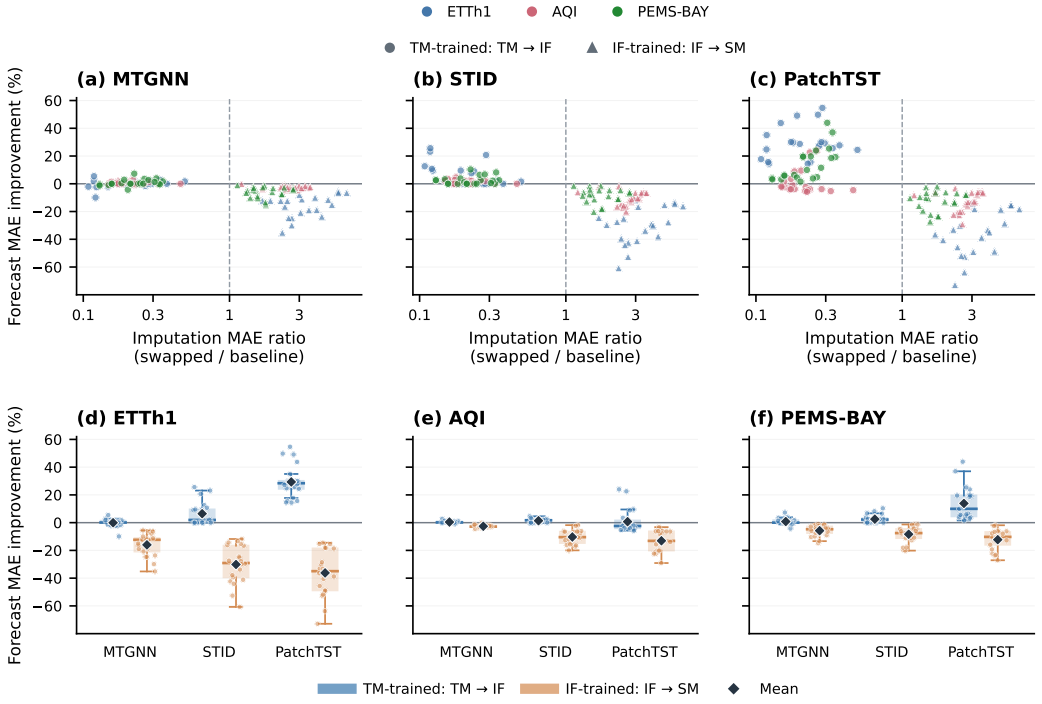


Figure 7: **Imputation quality and forecasting performance.** (a–c) Backbone-wise scatter: imputation MAE ratio $E_{\text{swapped}}/E_{\text{baseline}}$ (log scale; below one is better) versus forecast improvement $100 \times (F_{\text{baseline}} - F_{\text{swapped}})/F_{\text{baseline}}$. Positive forecast values improve; negative values deteriorate. (d–f) Dataset-wise boxes: quartiles, median lines, mean diamonds, and $1.5 \times$ IQR whiskers. All 21 conditions per backbone and substitution are shown. Forecast scores use complete reference targets; native-target sensitivity is summarized above.