EXOST: MULTI-SYSTEM EXOGENOUS VARIABLE MODELING FOR SPATIO-TEMPORAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatio-temporal (ST) forecasting is crucial for modeling physical world systems. However, existing ST models primarily focus on target variables within a single system, often neglecting their interactions with external systems. Although early research attempts introduced exogenous variables, they merely treated the spatial dimension as an external auxiliary feature, essentially belonging to covariate modeling within a single system. This paper first conducts a systematic study on the modeling of multi-system exogenous variables in ST forecasting. We identify two core challenges in this field: the inconsistent effects of distinct system on the target variables and the imbalance effects between historical and future information. To tackle this, we propose ExoST, a simple yet effective exogenous variable modeling general framework highly compatible with existing ST backbones that follows a "select-then-balance" paradigm. ExoST can adaptively select informative signals from multi-system exogenous variables and perform context-aware balance of historical and future exogenous information. Extensive experiments on real-world datasets demonstrate that ExoST consistently enhances the predictive performance of various backbones while maintaining its simplicity, robustness, and efficiency.

1 INTRODUCTION

Spatio-temporal forecasting aims to learn the evolution of structured systems whose states vary across both space and time (Shao et al., 2024; Chen & Liang, 2025a;b). This task is fundamental to traffic management (Avila & Mezić, 2020), environmental monitoring (Liang et al., 2023), and meteorology forecasting (Bi et al., 2023). Recent advances have focused on spatio-temporal (ST) backbones (Jin et al., 2023; 2024a) that integrate graph neural operators (Yu et al., 2017; Ma et al., 2025b) for learning spatial dependencies and sequence operators (Wu et al., 2019; Guo et al., 2019; Gao et al., 2023; Han et al., 2024) for modeling temporal patterns. *However, existing models predominantly operate under a single-system paradigm, focusing on modeling the target variables themselves while largely overlooking interactions with external dynamic systems.*

Spatio-temporal exogenous variables encompass contextual data from multiple external dynamic systems (Berkani et al., 2023), including historical information and auxiliary future forecasts. These signals complement target predictions by reducing uncertainty and enhancing interpretability (Olivares et al., 2023; Wang et al., 2024; Huang et al., 2025; Arango et al., 2025). As illustrated in Figure 1, air quality forecasting relies not only on historical pollutant indicators (*e.g.*, NO_2, CO_2) but also on the environmental context. For instance, historical traffic congestion may explain sudden emission spikes, while future weather forecasts (*e.g.*, wind speed) serve as precursors for pollutant dispersion or accumulation.

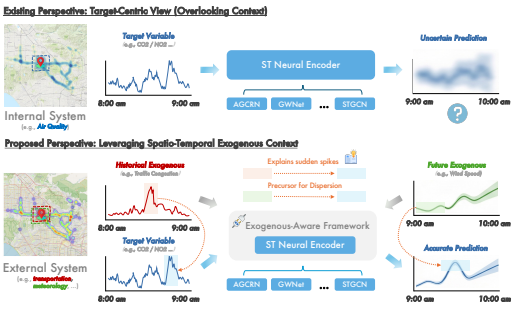


Figure 1: Exogenous-aware ST Forecasting.

Existing efforts to incorporate external information can be broadly categorized into several paradigms, as summarized in Figure 2 and Table 1. Early time-series (TS) forecasting approaches (including statistical and TS-specific models) simplify single-feature ST datasets into multivariate time series,

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.

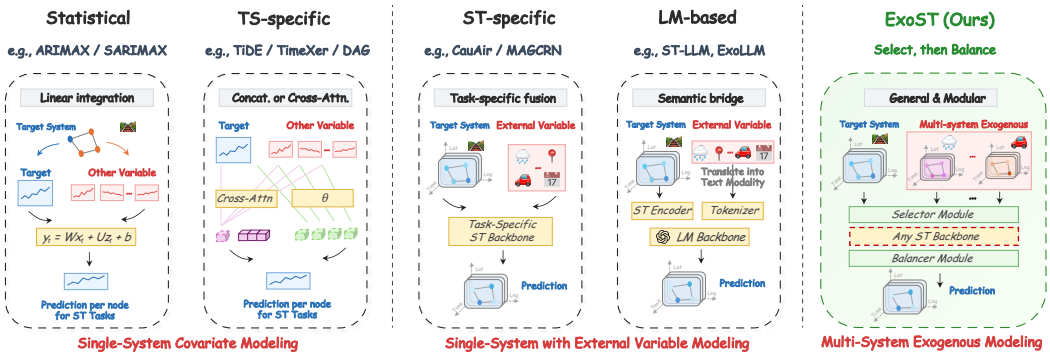


Figure 2: Paradigm comparison of exogenous modeling for ST forecasting. Existing methods either treat exogenous variables as single-system covariates, rely on task-specific fusion, or use LLM as semantic bridges. ExoST models multi-system exogenous with a general ST plug-in framework.

Table 1: The advantages and differences of our proposed ExoST compared with existing methods.

Dimension		Statistical	TS-specific	ST-specific	LM-based	ExoST (Ours)
Universality	applicable backbone	✗ strict assumption	✗ per-node models	✗ ST-specific	✓ semantic bridge	✓ any ST backbone
Simplicity	modeling complexity	✓ very simple	▲ medium	✗ ad-hoc fusion	▲ requires LLM adaption	✓ native / simple modules
Robustness	noise immunity	✗ weak	▲ limited	▲ moderate	▲ text quality dependent	✓ strong / generalizable
Efficiency	computation cost	▲ moderate	▲ moderate	✓ efficient	✗ costly	✓ efficient

thereby treating non-target spatial nodes as exogenous variables (Williams, 2001; Vagropoulos et al., 2016; Olivares et al., 2023; Wang et al., 2024; Zhou et al., 2025; Chen et al., 2026; Qiu et al., 2025b). As a result, they fundamentally remain within a *single-system covariate modeling* framework. Although such methods can adapt to ST forecasting by processing each location independently, they fail to exploit spatial dependency priors, leading to significant performance degradation (as demonstrated by comprehensive experimental evidence in Sec. 5.2 Tab. 2). More recent studies attempt to integrate external information into ST models through task-specific architectural designs (Ma et al., 2025a; Giganti et al., 2024) or by leveraging foundation models (*e.g.*, language model) as semantic bridges (Huang et al., 2025; Liu et al., 2025a;b; Arango et al., 2025). However, these approaches essentially operate on *single-system with static external variables* and are often tightly coupled with specific tasks, lacking generality or incurring substantial computational overhead. *Overall, there is still a lack of a unified framework that can enable effective native multi-system exogenous modeling while remaining seamlessly compatible with diverse spatio-temporal backbones.*

To address this gap, we propose *native multi-system exogenous modeling for ST forecasting* (ExoST), a simple and effective paradigm. ExoST enables the systematic integration of multi-system exogenous variables within a unified framework, while remaining compatible with diverse ST backbones. Our principal insight is that unified integration introduces a dual dilemma: inconsistency across multi-system variables and imbalance in temporal patterns, which fundamentally limits the ability of existing ST backbones to fully exploit cross-system exogenous information. To this end, we introduce a "select-then-balance" approach. Specifically, we augment the backbone network with selector and balancer modules before and after it, forming a sandwich-style architecture. The former aims to incorporate exogenous conditions into the representation, and leverages expert-based discretization to approximate the continuous integral of data generation, thereby explicitly modeling variable inconsistency and providing robust representations for downstream models. The latter is designed to learn a dynamic fusion function, whose context-awareness at inference time enables adaptive modeling of temporal imbalance. In summary, our contributions are as follows:

- *Problem Level:* We present the systematic formalization and exploration of native multi-system exogenous variable modeling for ST forecasting, addressing a long-neglected gap in the ST field.
- *Technical Level:* We identify the key challenges in modeling multi-system exogenous variables and propose ExoST, a simple yet effective general framework with targeted modules. ExoST demonstrates high compatibility with existing spatio-temporal backbones without any modification.
- *Evaluation Level:* We systematically evaluate ExoST across multiple aspects, verifying its efficacy, universality, robustness, design rationality, efficiency, and lightweight features. Code is available at https://anonymous.4open.science/r/ExoST and open-sourced for future research.

2 Related Work

Spatio-Temporal Forecasting. Predicting future spatio-temporal signals from historical observations, a spatially extension of multivariate time series forecasting, has evolved significantly. Early statistical (Box & Pierce, 1970; Biller & Nelson, 2003) or independent time series models (Nie et al., 2023; Wu et al., 2023; Qiu et al., 2025a) at each location proved insufficient for complex spatio-temporal dynamics (Shao et al., 2024), which involve intricate spatial correlations and heterogeneity (Chen & Liang, 2025a;b). This led to hybrid architectures (Li et al., 2018; Wu et al., 2019; Liu et al., 2023; Chen et al., 2025b; Cao et al., 2025; Ma et al., 2025c) that integrate spatial (*e.g.*, graph operators for node interactions) and temporal (*e.g.*, sequence operators for long-term trends) modeling to jointly learn rich spatio-temporal representations for accurate prediction. *However, a critical limitation persists: these models largely focus on the single-system's internal dynamics while neglecting exogenous variables, external inputs that are independent of the system's state but directly influence its evolution, and whose inclusion is vital for robust and accurate predictions.*

Exogenous Variables Modeling. Distinct from our ST forecasting setting, which models multiple systems as exogenous variables, existing literature primarily focuses on simple single-system covariate modeling within multivariate time series forecasting to capture correlations between observations and target variables. Early statistical methods, such as ARIMAX (Williams, 2001) and SARIMAX (Vagropoulos et al., 2016), addressed the relationships among exogenous series, endogenous series, and autoregressive components. Subsequent deep time series methods explored broader scenarios like leveraging future exogenous values (Das et al., 2023; Olivares et al., 2023; Qiu et al., 2025b) and handling time lags or missing data (Wang et al., 2024; Zhou et al., 2025; Chen et al., 2026). Although TS forecasting models can adapt to ST prediction by processing each location independently, as mentioned earlier, their failure to utilize spatial dependency priors leads to poor performance. Some preliminary attempts explored integrating exogenous information into ST backbones through task-specific model architectures (Lindström et al., 2014; Zagouras et al., 2015; Dong et al., 2023; Giganti et al., 2024; Ma et al., 2025a;d; Wang et al., 2025) or language-mediated proxy methods (Liu et al., 2024; 2025a;c; Huang et al., 2025), but this leads to inflexibility due to task coupling or information loss caused by language proxies. Meanwhile, they still only process single-system with static external variables. *In contrast, we introduce native multi-system exogenous variable modeling for ST forecasting, proposing a simple, effective, and general framework applicable to various spatio-temporal backbones without architectural modifications.*

3 Preliminaries

Definition 1. *(Spatio-Temporal Data).* Spatio-temporal data records the evolution of dynamic systems across space and time. Formally, we represent this as a sequence $\mathcal{X} = \{X_t\}_{t=1}^{T}$, where $X_t \in \mathbb{R}^{N \times F}$ denotes the system state at time t. The spatial domain is modeled as a network $\mathcal{G}$ with N nodes, where each node observes F features (*e.g.*, indicators such as NO_2, CO_2 *etc.* in air quality system).

Definition 2. *(Multi-System Exogenous Variables).* Exogenous variables are external factors independent of the internal dynamics of the target system and serve as auxiliary inputs. We additionally consider exogenous variables from M auxiliary systems: $\mathcal{E} = \{\mathbf{E}^{(1)}, \mathbf{E}^{(2)}, \ldots, \mathbf{E}^{(M)}\}$, where each $\mathbf{E}^{(m)}$ corresponds to a heterogeneous external source, such as traffic , meteorological , calendar, or other distinct systems. The aggregated exogenous variables across all systems are denoted as $\mathcal{E} = \{E_t\}_{t=1}^{T}$, which follow a ST data structure. $\mathcal{E}$ comprises past (*e.g.*, historical traffic flow) and future exogenous variables (*e.g.*, weather forecasts), distinguished by their temporal availability.

Problem. *(ST Forecasting with Exogenous).* We aim to predict future spatio-temporal signals given historical observations, optional exogenous inputs, and a graph prior $\mathcal{G}$. Let $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{|\mathcal{D}|} \sim \mathbb{P}$ denote the training set, where $\mathbf{X}_i \in \mathbb{R}^{N \times T_p \times F}$ and $\mathbf{Y}_i \in \mathbb{R}^{N \times T_f \times F}$ represent historical features (length T_p) and prediction targets (length T_f), respectively. Standard spatio-temporal forecasting optimizes model parameters θ^* by minimizing the expected loss:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \mathbb{P}} \left[\mathcal{L}(f_\theta(\mathbf{X}_i, \mathcal{G}), \mathbf{Y}_i) \right],$$

where f_θ is the ST backbone network and $\mathcal{L}$ denotes the error metric (typically the mean absolute error). With exogenous inputs $\mathbf{E}_i = (\mathbf{E}_i^p, \mathbf{E}_i^f)$, the objective changes to:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(\mathbf{X}_i, \mathbf{E}_i, \mathbf{Y}_i) \sim \mathbb{P}} \left[\mathcal{L}(f_\theta(\mathbf{X}_i, \mathbf{E}_i, \mathcal{G}), \mathbf{Y}_i) \right],$$

where $\mathbf{E}_i^p \in \mathbb{R}^{N \times T_p \times F_p}$ and $\mathbf{E}_i^f \in \mathbb{R}^{N \times T_f \times F_f}$ correspond to past and future exogenous features with dimensions F_p and F_f. In practice, following (Jin et al., 2023; 2024a; Shao et al., 2024), we usually focus only on the target variable to be predicted in the features, *i.e.*, F in $\mathbf{Y}_i$ is equal to 1.

4 METHODOLOGY

The ExoST framework is built upon a "select-then-balance" paradigm designed to enable native multi-system exogenous variable modeling, as shown in Fig. 3. To address the dual dilemma of inconsistency and imbalance effects, the process is decoupled into two consecutive stages forming a sandwich-style architecture. Specifically, we augment the ST backbone with a selector module before feature encode and a balancer module afterward. Both stages are fully differentiable and backbone-agnostic, meaning any modern ST encoder can be plugged in without architectural modifications.

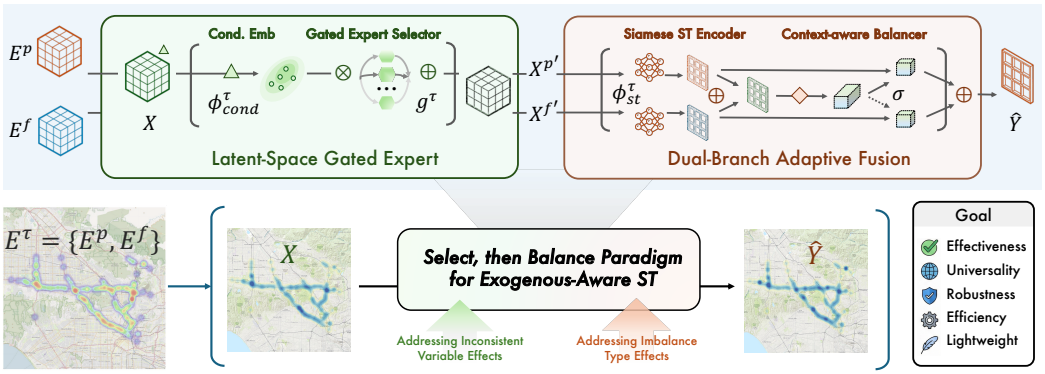


Figure 3: Overview of exogenous-aware spatio-temporal forecasting framework ExoST.

4.1 SELECT: LATENT-SPACE GATED EXPERT FOR INCONSISTENT VARIABLE EFFECTS

Motivation. The first obstacle of native multi-system exogenous modeling is *inconsistent variable effects*: exogenous variables originating from heterogeneous external systems exhibit highly uneven correlations with the target variable. As visualized in Figure 4 (Pearson correlations between exogenous variables and the target across randomly sampled stations in *AQI-19*), informativeness spans from highly predictive signals (*e.g.*, traffic flow) to weakly related or even noisy ones (*e.g.*, traffic incidents). This inconsistency is further amplified in the multi-system, because the predictive strength of each variable varies jointly across spatial nodes, time steps, and exogenous regimes (*e.g.*, rush hour vs. off-peak, sunny vs. rainy). Existing methods typically resort to uniform concatenation or projection (Giganti et al., 2024; Arango et al., 2025), which treats all variables equally, entangles informative signals with noise, and induces conflicts across heterogeneous regimes. Thus, we require a dynamic filter capable of approximating the continuous integral of latent data-generation regimes, effectively disentangling conflicting signals.

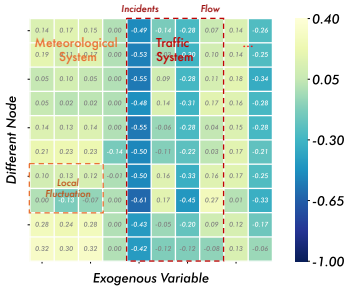


Figure 4: Inconsistent effect case.

Implementation Details. To achieve this goal, the Select stage operates in two consecutive steps:

Conditional Embedding. We first address the semantic space and dimension size alignment of endogenous variables with the past and future exogenous variables. For each type $\tau \in \{p, f\}$ (i.e., past, future), we project the raw exogenous tensor $\mathbf{E}^\tau \in \mathbb{R}^{N \times T_\tau \times F_\tau}$ and the historical endogenous input $\mathbf{X} \in \mathbb{R}^{N \times T_p \times F}$ into a shared condition latent space: $\mathbf{X}^\tau = \phi^\tau_{\text{cond}}(\mathbf{X}, \mathbf{E}^\tau)$, where ϕ^τ_{cond} is instantiated as an affine transformation block: $\phi^\tau_{\text{cond}}(\mathbf{X}, \mathbf{E}^\tau) = \text{Dropout}\Big(\text{Act}\left(W^\tau_x \mathbf{X} + W^\tau_e \mathbf{E}^\tau + \mathbf{b}^\tau\right)\Big)$, with learnable projections $W^\tau_x \in \mathbb{R}^{H \times F}$ and $W^\tau_e \in \mathbb{R}^{H \times F_\tau}$. Following standard protocols (Shao et al., 2024; Jin et al., 2024a), we assume uniform past and future sequence lengths to enable element-wise addition. For more arbitrary cases, the zero-padding preprocessing trick can also easily ensure alignment. This yields conditional representations $\mathbf{X}^p, \mathbf{X}^f \in \mathbb{R}^{N \times T \times H}$ that inject cross-system exogenous conditions into latent space.

Gated Expert Selector. To further mitigate semantic degeneration and channel entanglement caused by inconsistent variable effects, we apply a gated expert selector on each $\mathbf{X}^\tau$. Specifically, we

employ K expert projections $\{W_k^\tau \in \mathbb{R}^{H\times H}\}_{k=1}^K$ together with a gating network $\mathbf{g}^\tau : \mathbb{R}^H \to \Delta^K$. The gating scores, which determine each expert's contribution under a specific spatio-temporal context, are: $\mathbf{g}^\tau = \mathrm{Softmax}\big(W_g^\tau \mathbf{X}^\tau\big)$, with a learnable projection W_g^τ. The refined representation is then synthesized by weighted aggregation: $\mathbf{X}^{\tau'} = \sum_{k=1}^K g_k^\tau W_k^\tau \mathbf{X}^\tau$. This explicitly disentangles representations via a latent bottleneck and selective recombination, producing robust inputs $\mathbf{X}^{p'}$ and $\mathbf{X}^{f'}$ for the downstream backbone rather than relying on entangled global mappings.

Remark. By combining conditional representation with expert selection, this module instantiates an endogenous-dependent operator. The conditional embedding fuses the endogenous history and multi-system exogenous signals into a unified semantic basis, enabling the gating mechanism to act as a regime posterior approximator. Soft routing ensures that the learning signal from each contextual sample primarily updates its corresponding expert parameters. This explicitly decomposes conflicting signals across experts, dramatically reducing negative transfer and providing the downstream backbone with robust, regime-adaptive input features. *Theoretically*, we provide a more detailed analysis of the above in Appendix A.1. *Empirically*, our ablation study in Section 5.5 further validates the effectiveness of this module, and Fig. 9 demonstrates the smooth combination of gating distributions.

4.2 Balance: Dual-Branch Adaptive Fusion for Imbalanced Effect

Motivation. The second obstacle is *imbalanced type effects*. Even after the Select stage resolves inconsistency *within* each exogenous type, a distinct distributional asymmetry remains *between* historical signals and proactive future forecasts. As visualized in Figure 5, past and future exogenous variables on *AQI-19* display markedly different distribution patterns (*e.g.*, varying peak densities and tail behaviors), since historical signals reflect already-observed dynamics while future signals are noisy predictions produced by upstream forecasting systems. Their predictive importance is also highly instance-dependent: in some contexts (*e.g.*, sudden congestion) historical information dominates, whereas in others (*e.g.*, incoming wind) future forecasts are more informative. Current methods typically rely on joint modeling or simple fixed / globally-learnable weighted fusion (Das et al., 2023; Olivares et al., 2023; Wang et al., 2024; Zhou et al., 2025), which collapse onto a single global optimum on the training distribution and cannot adapt at inference time. This motivates us to *decouple* the two types during ST modeling and *dynamically balance* them through a context-aware fusion function.

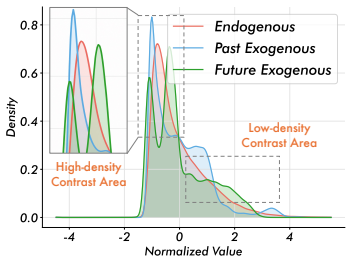


Figure 5: Imbalance effect case.

Implementation Details. The Balance stage comprises two components: a Siamese ST encoder pair for decoupled modeling and a context-aware balancer for adaptive fusion.

Siamese ST Encoders. To respect the distinct characteristics of past- and future-conditioned semantic representations, we adopt a Siamese architecture with two ST encoders that share structure but have independent parameters, while a naturally parallel design avoids inefficiencies. Each encoder ϕ_{st}^τ, instantiable by an arbitrary ST backbone (Yu et al., 2017; Li et al., 2018; Guo et al., 2019; Wu et al., 2019; Bai et al., 2020; Gao & Ribeiro, 2022; Satorras et al., 2022), captures intrinsic spatial correlations and temporal dependencies: $\mathbf{Y}^\tau = \phi_{\text{st}}^\tau(\mathbf{X}^{\tau'})$, yielding $\mathbf{Y}^p, \mathbf{Y}^f \in \mathbb{R}^{N\times T_f\times 1}$, which encode high-level spatio-temporal features derived from past and future exogenous contexts.

Context-Aware Balancer. To overcome the limitations of static fusion, we introduce a dynamic context-aware mechanism to integrate $\mathbf{Y}^p$ and $\mathbf{Y}^f$. We first form a unified context tensor $\mathbf{Y} = \mathbf{Y}^p + \mathbf{Y}^f$, from which an instance-specific balancing weight α is derived: $\alpha = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathbf{Y})))$, where $\mathrm{AvgPool}(\cdot)$ aggregates spatial-temporal representation into a channel descriptor, which is then passed through an MLP bottleneck followed by Sigmoid activation $\sigma(\cdot)$. The final prediction $\hat{\mathbf{Y}}$ is synthesized with a residual connection to facilitate information propagation: $\hat{\mathbf{Y}} = \alpha \odot \mathbf{Y}^p + (1-\alpha)\odot \mathbf{Y}^f + \mathbf{Y}$. This adaptive calibration prioritizes past versus future contexts according to instance-specific dynamics, producing a robust and balanced spatio-temporal prediction.

Remark. In stark contrast to static fusion, our context-aware balancer formulates an input-dependent, non-linear mapping: $\Phi(x,y) = (1-w(x,y))x + w(x,y)y + (x+y)$. Moreover, by conditioning the balance strictly on the immediate data characteristics (the input past and future exogenous variables and endogenous itself), the model accounts for the shifting distributions between past and future exogenous contexts, achieving continuous sensitivity at inference time. This yields a robust

prediction that dynamically, rather than statically, prioritizes past versus future signals. Theoretically, we provide a more detailed analysis of the above in Appendix A.2. Empirically, the ablation study in Section 5.5 further validates the effectiveness of this module, and Fig. 10 demonstrates the natural weight distribution that outperforms other fusion strategies on inference samples.

5 EXPERIMENTS

In this section, we conduct extensive experiments to investigate the following core research questions:

- **RQ1: (Effectiveness)** Can ExoST outperform existing other methods across various real scenarios?
- **RQ2: (Universality)** Can ExoST achieve consistent improvements on any existing ST backbones?
- **RQ3: (Robustness)** Can ExoST maintain performance with missing or noisy exogenous inputs?
- **RQ4: (Mechanism&Rationality)** How does ExoST work? Which components or strategies are crucial, and are they sensitive to different designs or hyper-parameters? How is the interpretability?
- **RQ5: (Efficiency & Lightweight)** How does ExoST compare in efficiency and parameters?

5.1 EXPERIMENTAL SETUP

Datasets. Common ST benchmarks (*e.g.*, *METR-LA* or *PeMS-BAY*) lack real exogenous variables from heterogeneous systems (typically containing only temporal date), which also largely explains the long-standing neglect of exogenous modeling due to the scarcity of multi-system datasets. We employ the *AQI-19* dataset [1], the only publicly available dataset known to meet this requirement, for systematic evaluation. Additionally, to evaluate generalizability, we further curated and extended three different datasets for future research. Specifically, we study four forecasting tasks: (I) *AQI-19* and *AQI-22*: predicting NO_2 concentration in the air quality domain using traffic and meteorological variables as exogenous inputs; (II) *Speed-19* and *Intensity-22*: forecasting traffic speed / flow intensity by incorporating other system indicators as exogenous variables. The datasets were chronologically split into training, validation, and test sets in a 7:2:1 ratio. Date encoding was included in both past and future exogenous variables. For more dataset details, it can be found in Appendix B.1.

Baseline. We compared with three comparable types of baselines: (i) advance time-series models with exogenous modeling: TiDE (Das et al., 2023), NBEATSx (Olivares et al., 2023), TimeXer (Wang et al., 2024), CrossLinear (Zhou et al., 2025), XLinear (Chen et al., 2026), DAG (Qiu et al., 2025b); (ii) advance spatio-temporal models with exogenous integration: MAGCRN (Giganti et al., 2024); (iii) general exogenous modeling framework (for foundation model): ChronosX (Arango et al., 2025). To test ExoST's universality, we further instantiate it with various classic spatio-temporal backbones, including AGCRN (Bai et al., 2020), GWNet (Wu et al., 2019), GGNN (Satorras et al., 2022), GRUGCN (Gao & Ribeiro, 2022), STGCN (Yu et al., 2017), and DCRNN (Li et al., 2018). For more details about baseline methods and ST backbone networks, please refer to Appendix B.2.

Protocol. We use the past 24 steps to forecast the next 24 steps (*i.e.*, 1-day) and report MAE, RMSE, MAPE, and MRE. To assess long-term robustness, we additionally perform 2- and 3-day rolling forecasts without retraining. Baseline models follow their recommended settings. For ExoST, we use 4 gated experts and hidden size 64. Each experiment is run five times; we report mean ± standard deviation. Best results appear in **bold blue**, second-best in underlined green. For *Speed-19**, MAE and RMSE are scaled by 10^2 for clearer comparison. For more protocol details, it can be found in Appendix B.3.

5.2 EFFECTIVENESS STUDY (RQ1)

Table 2 presents a performance comparison of our ExoST against state-of-the-art time-series, spatio-temporal, and general foundation models with exogenous variable modeling techniques. To ensure a fair comparison, for all methods that can adopt a decoupled design, we uniformly employ the classic spatio-temporal encoder AGCRN (Bai et al., 2020) as the backbone network. Our observations are as follows: ❶ Pure time-series methods (first six) consistently yield the lowest performance, even though these are advanced modeling approaches incorporating exogenous information, which aligns with existing spatio-temporal benchmark observations (Shao et al., 2024). This deficiency stems from their inability to model spatial dependencies, resulting in suboptimal accuracy. ❷ While general method ChronosX achieves relative gains by integrating the ST backbone, it frequently suffers from non-convergence on specific tasks (*e.g.*, *Speed-19* and *Intensity-22*), despite extensive hyperparameter tuning. We attribute this instability to the inherent design of the underlying Chronos-type foundation

[1] https://zenodo.org/records/7308425

Table 2: Comparison of the overall performance of advance methods and our ExoST in real tasks.

Datasets		AQI-19			Speed-19 *			AQI-22			Intensity-22		
Method	Metric	1-day	2-day	3-day	1-day	2-day	3-day	1-day	2-day	3-day	1-day	2-day	3-day
TiDE TMLR'23	MAE	14.73±0.01	15.68±0.00	15.66±0.01	6.70±1.83	7.43±1.75	7.58±1.73	10.51±0.05	11.32±0.04	11.89±0.04	150.18±0.02	157.00±0.01	161.17±0.02
	RMSE	20.73±0.03	21.76±0.04	21.70±0.04	44.47±5.16	50.37±4.21	51.26±4.10	14.67±0.00	15.76±0.01	16.43±0.01	217.05±0.15	223.13±0.16	226.96±0.17
	MAPE (%)	93.01±0.53	98.80±0.43	98.60±0.41	16.38±4.62	18.02±4.42	18.37±4.36	87.51±0.98	95.07±0.93	99.22±0.92	146.50±0.03	147.09±0.07	150.07±0.06
	MRE (%)	56.96±0.04	60.67±0.01	60.67±0.02	17.44±4.78	19.33±4.56	19.72±4.51	56.27±0.24	60.56±0.20	63.62±0.20	54.43±0.01	56.82±0.00	58.21±0.01
NBEATSx IJoF'23	MAE	11.63±0.08	13.88±0.06	14.10±0.09	5.21±0.30	5.99±0.29	6.17±0.28	8.13±0.11	9.71±0.19	9.97±0.17	38.91±1.68	67.02±1.78	85.03±1.43
	RMSE	18.44±0.04	20.65±0.06	21.20±0.04	39.56±2.02	45.80±2.23	46.80±2.20	12.84±0.08	14.84±0.19	15.02±0.19	70.08±4.91	117.62±2.98	141.30±1.92
	MAPE (%)	56.54±1.34	69.08±1.30	69.99±1.47	13.56±0.77	15.59±0.75	16.06±0.73	55.89±1.69	67.26±2.25	71.37±2.27	25.12±0.89	38.23±0.99	48.76±0.73
	MRE (%)	44.96±0.32	53.77±0.24	54.55±0.33	13.17±0.34	23.33±0.45	30.50±0.26	43.50±0.61	51.96±1.02	53.35±0.89	14.11±0.61	24.26±0.64	30.71±0.51
TimeXer NeurIPS'24	MAE	13.11±0.02	13.92±0.29	14.08±0.18	5.07±0.09	5.61±0.11	5.71±0.11	9.65±0.59	10.13±0.08	9.88±0.46	128.05±1.05	138.34±4.13	149.21±6.48
	RMSE	20.20±0.03	20.88±0.05	21.23±0.03	38.84±2.09	41.25±2.23	41.84±2.19	14.98±0.46	15.57±0.12	15.14±0.59	191.72±1.27	199.38±8.40	210.42±8.93
	MAPE (%)	64.48±0.27	65.90±4.25	68.33±3.68	12.42±0.51	13.67±0.55	13.93±0.55	63.07±7.20	66.30±3.65	65.05±2.85	191.61±3.61	188.09±1.44	190.11±0.93
	MRE (%)	50.71±0.08	53.93±1.11	54.49±0.68	13.20±0.24	14.59±0.29	14.86±0.29	51.64±3.13	54.21±0.40	52.86±2.47	46.41±0.34	49.96±2.01	53.88±2.67
CrossLinear KDD'25	MAE	13.32±1.32	14.65±1.83	15.39±0.76	4.44±0.05	5.09±0.06	5.21±0.06	8.44±0.23	9.93±0.21	10.26±0.25	46.07±13.80	70.75±9.89	84.35±0.67
	RMSE	19.29±1.54	21.26±1.86	22.11±0.55	37.77±2.13	40.69±2.23	41.53±2.22	13.20±0.13	14.99±0.09	15.27±0.08	81.98±24.55	124.28±12.73	142.91±0.14
	MAPE (%)	70.55±0.09	77.78±0.14	81.03±0.08	10.07±0.00	11.67±0.00	12.00±0.00	53.06±0.04	63.76±0.03	68.33±0.03	25.91±0.07	37.08±0.06	43.25±0.01
	MRE (%)	51.62±0.05	56.70±0.07	59.59±0.03	11.56±0.00	13.24±0.00	13.54±0.00	45.17±0.01	53.13±0.01	54.87±0.01	16.69±0.05	25.59±0.04	30.46±0.01
XLinear AAAI'26	MAE	11.74±0.003	14.13±0.04	13.75±0.04	3.60±0.00	4.54±0.00	4.80±0.00	8.15±0.05	10.12±0.06	10.23±0.10	49.58±0.52	78.60±1.17	84.88±1.04
	RMSE	18.58±0.09	21.28±0.10	20.61±0.03	37.60±0.00	42.05±0.00	43.32±0.00	13.35±0.08	15.74±0.14	15.76±0.16	98.99±1.59	138.51±2.11	145.84±1.91
	MAPE (%)	54.71±0.13	67.34±0.02	66.03±0.34	8.42±0.00	10.73±0.00	11.42±0.00	50.36±0.32	65.53±0.28	68.46±0.81	28.30±0.81	43.76±1.21	47.18±1.17
	MRE (%)	45.40±0.007	54.70±0.15	53.28±0.14	9.37±0.00	11.80±0.00	12.49±0.00	43.63±0.27	54.13±0.35	54.73±0.52	17.97±0.19	28.45±0.42	30.66±0.38
DAG ICML'26	MAE	11.48±0.10	13.39±0.07	13.55±0.07	17.35±1.29	18.05±1.36	17.85±1.31	8.32±0.19	9.84±0.24	9.95±0.17	57.68±8.59	82.09±6.10	89.70±4.11
	RMSE	18.47±0.12	20.47±0.10	20.89±0.10	104.01±9.44	109.77±10.34	109.78±10.34	13.26±0.11	15.10±0.23	15.06±0.22	101.29±10.47	137.06±6.26	146.73±4.37
	MAPE (%)	53.40±0.01	63.18±0.00	63.59±0.00	39.64±0.03	41.46±0.04	41.11±0.03	54.93±0.04	66.86±0.04	70.26±0.03	33.59±0.06	45.07±0.05	48.86±0.03
	MRE (%)	44.38±0.00	51.88±0.00	52.43±0.00	45.16±0.03	46.97±0.04	46.43±0.03	44.53±0.01	52.65±0.01	53.25±0.01	20.91±0.03	29.71±0.02	32.39±0.01
MAGCRN IEEE'24	MAE	9.82±0.11	10.52±0.13	11.03±0.27	3.87±0.05	3.88±0.05	3.89±0.05	7.41±0.24	7.80±0.25	8.06±0.33	37.47±1.06	52.51±6.80	65.32±11.02
	RMSE	15.18±0.25	16.03±0.24	16.55±0.38	31.83±0.02	31.84±0.03	31.91±0.02	11.43±0.14	11.91±0.23	11.96±0.40	68.52±4.56	93.26±9.74	**108.66**±17.87
	MAPE (%)	38.66±0.27	**42.53**±0.41	45.08±1.15	9.41±0.07	9.43±0.08	9.44±0.08	**43.79**±4.21	**45.89**±4.35	**49.74**±4.19	20.12±0.26	27.20±2.58	34.88±4.66
	MRE (%)	37.97±0.44	40.69±0.50	42.72±1.03	10.08±0.13	10.08±0.13	10.09±0.13	39.66±1.30	41.74±1.34	43.15±1.77	13.58±0.38	19.00±2.46	23.59±3.98
ChronosX AISTATS'25	MAE	12.46±0.28	13.02±0.12	13.30±0.03	39.20±0.57	39.19±0.56	39.20±0.55	8.87±0.06	9.30±0.13	9.54±0.02	142.06±3.19	148.59±3.20	156.41±4.19
	RMSE	18.37±0.36	19.17±0.11	19.41±0.13	139.33±0.10	139.36±0.09	139.43±0.08	13.23±0.11	13.87±0.12	13.92±0.14	216.75±3.33	222.80±4.13	231.04±4.85
	MAPE (%)	55.86±0.72	59.48±1.72	60.74±2.43	99.50±0.44	99.54±0.44	99.58±0.42	58.84±2.39	61.46±2.74	65.24±2.63	58.60±3.79	64.23±1.65	69.75±1.09
	MRE (%)	48.17±1.08	50.39±0.49	51.53±0.13	102.00±1.52	101.98±1.46	101.98±1.41	47.49±0.34	49.78±0.71	51.07±0.11	51.35±0.92	53.78±1.16	56.49±1.52
ExoST	MAE	9.33±0.13	10.04±0.19	10.41±0.12	3.79±0.01	3.80±0.01	3.80±0.01	7.33±0.21	7.61±0.23	7.94±0.29	31.91±1.35	50.59±0.04	64.13±3.86
	RMSE	14.24±0.18	15.12±0.26	15.51±0.16	31.73±0.02	31.78±0.02	31.85±0.02	11.14±0.21	11.52±0.49	11.79±0.51	56.49±2.54	89.10±7.36	109.01±9.19
	MAPE (%)	38.57±0.32	42.62±0.52	44.65±0.08	9.30±0.03	9.32±0.03	9.31±0.03	44.90±0.88	46.59±0.64	50.27±1.18	17.63±0.97	26.68±1.69	34.52±2.90
	MRE (%)	36.09±0.49	38.84±0.75	40.34±0.47	9.88±0.03	9.89±0.03	9.90±0.03	39.23±1.11	40.74±1.24	42.50±1.57	11.57±0.49	18.31±1.14	23.16±1.40

models (Ansari et al., 2024). ❸ MAGCRN often produces suboptimal results as it is tailored for specific scenarios yet neglects the inconsistent and unbalanced effects of exogenous variables. ❹ Conversely, our ExoST achieves superior performance across all metrics on the four datasets. By effectively selecting and balancing exogenous information, ExoST adaptively handles diverse sources

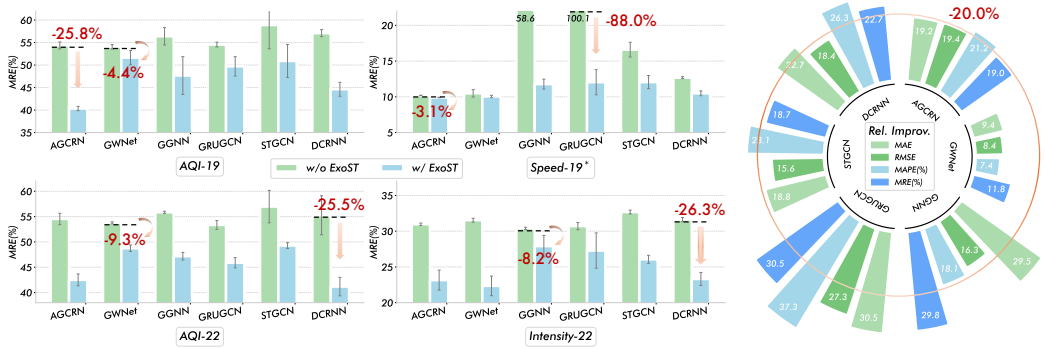


Figure 6: Left: Performance gains of different models on different datasets w/ and w/o ExoST. Right: Average relative improvement of metrics for each model across all datasets (Full results in Table. 6).

5.3 Universality Study (RQ2)

We then evaluate the universality of the "select, then balance" paradigm in ExoST across diverse ST backbones. Figure 6 illustrates the absolute and average relative performance improvements of ExoST across four tasks (left) and six models (right) for 3-day forecasting. ❶ ExoST yields consistent performance boosts across varying architectures and datasets, validating the efficacy of our exogenous modeling paradigm. ❷ From the task perspective, air quality forecasting achieves substantial gains due to the strong influence of exogenous factors (*e.g.*, traffic and meteorology), whereas traffic speed forecasting sees moderate improvements, reflecting a limited dependency on exogenous variables. ❸ From the model perspective, while gains vary across backbones, most exceed 20%; notably, even the vanilla GWNet achieves a significant improvement of nearly 10%. ❹ Supplementary results for 1-day and 2-day horizons (Tables 7 and 8 in Appendix C.1) further reveal that performance gains amplify as the forecasting horizon extends, underscoring the robust generalization capability of our approach.

5.4 Robustness Study (RQ3)

In real-world scenarios, spatio-temporal systems frequently encounter unreliable exogenous data collection (*e.g.*, signal loss from sensor anomalies), which compromises prediction accuracy and robustness. To assess the ExoST under such conditions, we designed a systematic experiment simu-

Table 3: Performance under abnormal conditions with missing exogenous signals (3-day scenario).

Datasets		AQI-19				Speed-19*				Intensity-22			
Masking Strategy		MAE	RMSE	MAPE(%)	MRE(%)	MAE	RMSE	MAPE(%)	MRE(%)	MAE	RMSE	MAPE(%)	MRE(%)
Zero	20%	10.60±0.15	15.87±0.16	44.83±1.39	41.07±0.60	3.87±0.03	31.94±0.06	9.47±1.20	10.08±0.90	64.31±9.49	**108.23**±18.00	**34.19**±3.45	23.23±3.43
	40%	11.33±0.02	17.07±0.09	46.13±1.45	43.91±0.09	3.89±0.07	31.98±0.08	9.53±1.80	10.12±1.60	67.80±1.82	115.25±3.73	38.21±0.21	24.49±0.65
	60%	11.44±0.52	17.30±0.84	48.30±2.02	44.34±2.01	3.87±0.05	31.96±0.09	9.48±1.50	10.08±1.40	68.50±10.78	118.41±17.47	36.37±6.38	24.74±3.89
	80%	12.33±0.00	18.94±0.00	52.03±0.00	47.79±0.00	3.83±0.02	31.89±0.06	9.40±0.90	9.95±0.50	69.70±10.99	118.51±18.88	38.18±6.62	25.17±3.97
Random	20%	10.67±0.16	16.17±0.19	44.23±2.37	41.32±0.62	3.86±0.03	31.94±0.08	9.44±1.40	10.05±1.00	67.32±8.42	114.89±16.03	34.88±3.82	24.32±3.04
	40%	11.30±0.08	17.09±0.23	46.57±1.19	43.78±0.32	3.84±0.02	31.92±0.04	9.42±0.60	9.99±0.50	64.43±10.78	110.41±17.43	34.34±7.21	23.27±3.89
	60%	11.37±0.05	17.26±0.22	47.43±1.15	44.04±0.18	3.87±0.05	31.95±0.09	9.49±1.40	10.06±1.40	65.27±6.06	109.81±11.90	35.34±3.27	23.57±2.19
	80%	12.33±0.00	18.94±0.00	52.03±0.00	47.79±0.00	3.88±0.05	31.93±0.09	9.48±1.60	10.08±1.20	68.13±10.61	115.96±18.89	36.61±4.61	24.60±3.83
No Masking		**10.41**±0.12	**15.51**±0.16	**42.62**±0.52	**38.84**±0.75	**3.80**±0.01	**31.85**±0.02	**9.31**±0.03	**9.90**±0.03	**64.13**±3.86	109.01±9.19	34.52±2.90	**23.16**±1.40

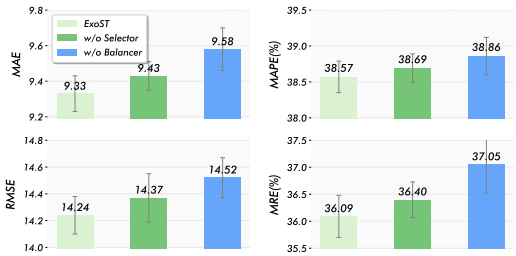


Figure 7: Ablation study on *AQI-19* dataset comparing ExoST, *w/o Selector*, and *w/o Balancer* variants in 1-day forecasting horizons.

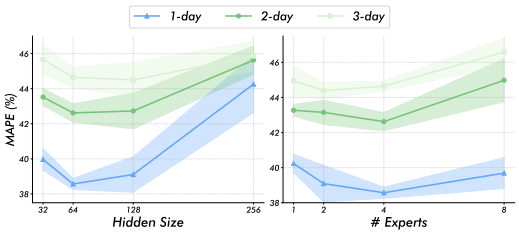


Figure 8: Performance comparison on *AQI-19* for different hyperparameter: *hidden size* and *the number of experts* across various horizons.

lating data corruption. Specifically, we randomly replaced exogenous variables at varying degradation levels (20%, 40%, 60%, and 80%) with zeros or noise from a standard normal distribution $\mathcal{N}(0, 1)$. The 3-day results in Table 3 reveal that: ❶ While performance gradually declines as the missing data ratio increases, the absolute drop remains minimal, indicating our framework maintains reasonable efficacy even with limited auxiliary information. ❷ Surprisingly, for certain tasks, appropriate levels of missing signals improve performance. This suggests that random corruption acts as a data augmentation strategy, forcing the model to predict under incomplete or noisy conditions, thereby enhancing generalization. Similar trends for 1- and 2-day scenarios are reported in Appendix C.2.

5.5 Mechanism & Rationality Study (RQ4)

Ablation Study. We conducted systematic ablation studies to comprehensively evaluate the impact of different components on ExoST's performance from both data and model perspectives.

From the data perspective, we investigated the auxiliary role of distinct exogenous variable types. Given the universality of date information, we treat it as a separate category. As shown in Table 4: ❶ Future exogenous variables outperform past variables, indicating that forward-looking information provides more direct predictive signals; however, combining both yields superior performance, suggesting they offer complementary patterns. ❷ Date exogenous variables perform poorly in isolation but drive improvements when combined with other types. ❸ The best performance is achieved by integrating all exogenous variables, demonstrating that ExoST effectively extracts and balances contributions from multi-system diverse exogenous contexts.

Table 4: Ablation study from the data perspective on *AQI-19* for 1-day forecasting, where P, F, and D denote Past, Future, and Data exogenous variables.

P	F	D	MAE	RMSE	MAPE (%)	MRE (%)
✓	-	-	13.09±0.65	19.35±0.52	65.54±5.87	50.64±2.50
-	✓	-	9.97±0.36	15.09±0.52	41.20±1.97	38.56±1.38
-	-	✓	12.47±0.15	19.15±0.20	57.42±5.11	48.21±0.58
✓	-	✓	12.20±0.09	18.72±0.08	59.06±1.72	47.16±0.33
-	✓	✓	9.63±0.00	14.69±0.00	39.68±0.03	37.24±0.02
✓	✓	-	9.69±0.33	14.93±0.59	39.42±0.69	37.46±1.28
✓	✓	✓	**9.33**±0.13	**14.24**±0.18	**38.57**±0.32	**36.09**±0.49

From the model perspective, we examined the contribution of specific modules by constructing variants without the Selector or the Balancer module (replaced by direct element-wise fusion). As Figure 7 illustrates: ❶ The *w/o Selector* variant suffers significant degradation across all metrics, proving the importance of addressing inconsistent variable effects. Without semantic alignment and adaptive selection, the model struggles to disentangle heterogeneous exogenous variables from endogenous ones, leading to feature space competition. ❷ The *w/o Balancer* variant exhibits even more pronounced degradation, highlighting that temporal distribution shifts between past and future exogenous information pose a fundamental challenge. Sophisticated balancing mechanisms are key to manage noise-contaminated data and uncertainty-laden future forecasts. More results in Appx. C.3.1.

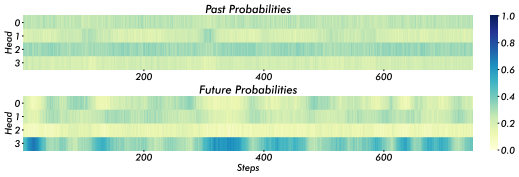


Figure 9: Visualization of gated expert selector.

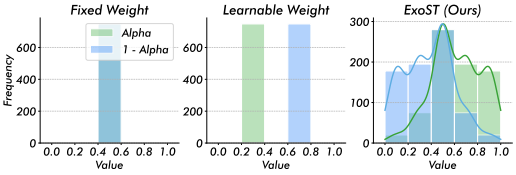


Figure 10: Visualization of weight distribution.

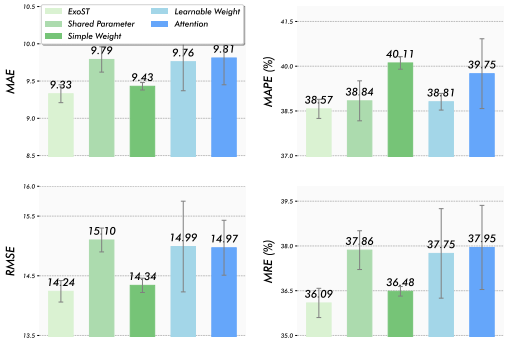


Figure 11: Different strategy performance comparison on *AQI-19* comparing ExoST, share parameter, simple weight, learned parameter and attention variants on 1-day forecasting scenario.

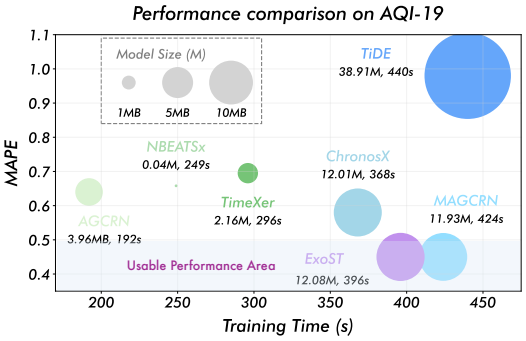


Figure 12: Efficiency & Lightweight study across different models. To ensure fair comparison given varying convergence rates, we report the average training time for the first 100 epochs.

Parameter Study. We then performed a parameter sensitivity analysis to assess the impact of key hyperparameters on ExoST. As shown in Figure 8: ❶ A hidden size of *64* achieves optimal performance across all prediction horizons, with larger sizes causing degradation due to overfitting. ❷ We found that four experts maximize performance without excessive overhead. While one or two experts lack sufficient representational power, eight experts lead to redundancy. This confirms that our latent space gating expert module can be lightweight while maintaining effectiveness.

Strategy Study. We compared several alternative strategies: (i) *Shared Parameter*: Siamese branches sharing parameters.(ii) *Simple Weight*: A static balancing weight ($\alpha = 0.5$) fusing outputs.(iii) *Learnable Weight*: Dynamically adjusting importance via learnable parameters.(iv) *Attention*: Using cross-attention to balance contributions. As shown in Figure 11: ❶ ExoST outperforms all alternative strategies, confirming the effectiveness of our design.❷ *Simple Weight* degrades performance significantly due to its inability to adapt to varying input contexts.❸ *Attention* and *Learnable Weight* strategies underperform while increasing computational cost; attention mechanisms introduce excessive parameter overhead, and learnable weights lack the inference-adaptive modeling required to handle the nuanced differences between past and future exogenous. More results in Appendix. C.3.2.

Visualization Study. Fig. 9 and 10 visualize the gating and weight distributions in the AQI-19 experiment, showing that ❶ the selector's smooth probability distribution enables dynamic routing of local contexts to exogenous experts via latent-space discretization. ❷ Unlike the irregular patterns of alternative strategies, ExoST's uniform distribution underscores its adaptability during inference.

5.6 Efficiency & Lightweight Study (RQ5)

We evaluated the efficiency of ExoST against other TS and ST models incorporating exogenous variables. As illustrated in Figure 12, ExoST strikes a favorable balance between performance, efficiency, and parameter count. Specifically: ❶ Compared to time series models, purely lightweight designs often lead to inferior performance. ❷ Compared to spatio-temporal and general exogenous variable modeling methods using the same backbone (*i.e.*, MAGCRN, ChronosX), ExoST achieves superior performance while maintaining comparable parameter and training times.

6 Conclusion and Future Work

This paper presents the systematic exploration of native multi-system exogenous variable modeling for spatio-temporal forecasting. To address inconsistent and imbalanced effects, we propose ExoST, a general framework that leverages a latent-space gated expert and a context-aware balancer to dynamically route and fuse multi-system exogenous signals. Extensive experiments demonstrate its effectiveness, universality, and robustness. In future work, we aim to explore how to expand the modality availability of exogenous variables and their feasibility in extremely data-scarce scenarios.

AI USE STATEMENT

Following the conference guidelines regarding generative artificial intelligence (GenAI) models, we declare that all aspects of this paper—including conceptual development, experimental methodology, data analysis, and the core content of the manuscript—were completed independently by the authors without any assistance from GenAI.

ETHICS STATEMENT

This paper introduces an algorithm designed to advance the modeling of multi-source spatiotemporal models. Our work has many potential societal impacts, but we believe none require specific highlighting here.

REPRODUCIBILITY STATEMENT

We used publicly available datasets for training and evaluation, and plan to open-source the code. The numerical tables in the manuscript were manually double-checked and cross-verified against the records. The authors take responsibility for the text, claims, and research findings of this study.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Henri Joseph Senetaire, Ali Caner Turkmen, Oleksandr Shchur, Danielle C Maddix, et al. Chronosx: Adapting pretrained time series models with exogenous variables. In *International Conference on Artificial Intelligence and Statistics*, pp. 2242–2250. PMLR, 2025.

Allan M Avila and Igor Mezić. Data-driven analysis and forecasting of highway traffic dynamics. *Nature communications*, 11(1):2090, 2020.

Lei Bai, Lina Yao, Can Li, Xianjie Wang, and Chunrong Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *NeurIPS*, 2020.

Safaa Berkani, Bassma Guermah, Mehdi Zakroum, and Mounir Ghogho. Spatio-temporal forecasting: A survey of data-driven models using exogenous data. *IEEE Access*, 11:75191–75214, 2023.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Bahar Biller and Barry L Nelson. Modeling and generating multivariate time-series input processes using a vector autoregressive technique. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(3):211–237, 2003.

George EP Box and David A Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65 (332):1509–1526, 1970.

Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11463–11471, 2025.

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

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Wei Chen, Bing Zhu, Junwei Liang, and Zixuan Yuan. Climateiqa: A new dataset and benchmark to advance vision-language models in meteorology anomalies analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5322–5333, 2025a.

Min Chen, Guansong Pang, Wenjun Wang, and Cheng Yan. Information bottleneck-guided mlps for robust spatial-temporal forecasting. In *Forty-second International Conference on Machine Learning*, 2025b.

Wei Chen and Yuxuan Liang. Expand and compress: Exploring tuning principles for continual spatio-temporal graph forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Wei Chen and Yuxuan Liang. Learning with calibration: Exploring test-time computing of spatio-temporal forecasting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Wei Chen, Xixuan Hao, Yuankai Wu, and Yuxuan Liang. Terra: A multimodal spatio-temporal dataset spanning the earth. *NeurIPS*, 37:66329–66356, 2024.

Xinyang Chen, Huidong Jin, Yu Huang, and Zaiwen Feng. Xlinear: A lightweight and accurate mlp-based model for long-term time series forecasting with exogenous inputs. *AAAI*, 2026.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023.

Chengxiang Dong, Xiaoliang Feng, Yongchao Wang, and Xin Wei. Spatiotemporal exogenous variables enhanced model for traffic flow prediction. *IEEE Access*, 11:95958–95973, 2023.

Haotian Gao, Renhe Jiang, Zheng Dong, Jinliang Deng, and Xuan Song. Spatio-temporal-decoupled masked pre-training for traffic forecasting. *arXiv preprint arXiv:2312.00516*, 2023.

Jiaxin Gao and Bruno Ribeiro. On the equivalence between temporal and static equivariant graph representations. In *International Conference on Machine Learning*, pp. 7052–7076. PMLR, 2022.

Antonio Giganti, Sara Mandelli, Paolo Bestagini, Umberto Giuriato, Alessandro D'Ausilio, Marco Marcon, and Stefano Tubaro. Back to the future: Gnn-based no 2 forecasting via future covariates. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3872–3876. IEEE, 2024.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 922–929, 2019.

Jindong Han, Weijia Zhang, Hao Liu, Tao Tao, Naiqiang Tan, and Hui Xiong. Bigst: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proceedings of the VLDB Endowment*, 17(5):1081–1090, 2024.

Jindong Han, Yansong Ning, Zirui Yuan, Hang Ni, Fan Liu, Tengfei Lyu, and Hao Liu. Large language model powered intelligent urban agents: Concepts, capabilities, and applications. *arXiv preprint arXiv:2507.00914*, 2025.

Kethmi Hirushini Hettige, Jiahao Ji, Cheng Long, Shili Xiang, Gao Cong, and Jingyuan Wang. A modular multitask reasoning framework integrating spatio-temporal models and llms. *arXiv preprint arXiv:2506.20073*, 2025.

Qihe Huang, Zhengyang Zhou, Kuo Yang, and Yang Wang. Exploiting language power for time series forecasting with exogenous variables. In *Proceedings of the ACM on Web Conference 2025*, pp. 4043–4052, 2025.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23343–23351, 2024.

Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5388–5408, 2023.

Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I Webb, Irwin King, and Shirui Pan. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *ICLR*, 2024b.

Ming Jin, YiFan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position: What can large language models tell us about time series analysis. In *41st International Conference on Machine Learning*. MLResearchPress, 2024c.

Wenbin Li, Di Yao, Ruibo Zhao, Wenjie Chen, Zijie Xu, Chengxue Luo, Chang Gong, Quanliang Jing, Haining Tan, and Jingping Bi. Stbench: Assessing the ability of large language models in spatio-temporal analysis. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 749–752, 2025.

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

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5351–5362, 2024.

Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14329–14337, 2023.

Johan Lindström, Adam A Szpiro, Paul D Sampson, Assaf P Oron, Mark Richards, Tim V Larson, and Lianne Sheppard. A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and ecological statistics*, 21:411–433, 2014.

Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. Spatial-temporal large language model for traffic prediction. In *2024 25th IEEE International Conference on Mobile Data Management (MDM)*, pp. 31–40. IEEE, 2024.

Chenxi Liu, Kethmi Hirushini Hettige, Qianxiong Xu, Cheng Long, Shili Xiang, Gao Cong, Ziyue Li, and Rui Zhao. St-llm+: Graph enhanced spatio-temporal large language models for traffic prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2025a.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18780–18788, 2025b.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 4125–4129, 2023.

Yuhang Liu, Yingxue Zhang, Xin Zhang, Ling Tian, Yanhua Li, and Jun Luo. Urbanmind: Urban dynamics prediction with multifaceted spatial-temporal large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 1951–1962, 2025c.

Kang Luo, Yuanshao Zhu, Wei Chen, Kun Wang, Zhengyang Zhou, Sijie Ruan, and Yuxuan Liang. Towards robust trajectory representations: isolating environmental confounders with causal learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 2243–2251, 2024.

Jiaming Ma, Zhiqing Cui, Binwu Wang, Pengkun Wang, Zhengyang Zhou, Zhe Zhao, and Yang Wang. Causal learning meet covariates: Empowering lightweight and effective nationwide air quality forecasting. In *International Joint Conference on Artificial Intelligence*, 2025a.

Jiaming Ma, Binwu Wang, Guanjun Wang, Kuo Yang, Zhengyang Zhou, Pengkun Wang, Xu Wang, and Yang Wang. Less but more: Linear adaptive graph learning empowering spatiotemporal forecasting. In *NeurIPS*, 2025b.

Jiaming Ma, Binwu Wang, Pengkun Wang, Zhengyang Zhou, Xu Wang, and Yang Wang. Robust spatio-temporal centralized interaction for ood learning. In *Forty-second International Conference on Machine Learning*, 2025c.

Jiaming Ma, Guanjun Wang, Sheng Huang, Kuo Yang, Binwu Wang, Pengkun Wang, and Yang Wang. Spatiotemporal causal decoupling model for air quality forecasting. *arXiv preprint arXiv:2505.20119*, 2025d.

Juntong Ni, Shiyu Wang, Ming Jin, Qi He, and Wei Jin. Streasoner: Empowering llms for spatio-temporal reasoning in time series via spatial-aware reinforcement learning. *arXiv preprint arXiv:2601.03248*, 2026.

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

Kin G Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2):884–900, 2023.

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

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1185–1196, 2025a.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Hanyin Cheng, Xingjian Wu, Chenjuan Guo, Bin Yang, and Jilin Hu. Dag: A dual causal network for time series forecasting with exogenous variables. *arXiv preprint arXiv:2509.14933*, 2025b.

Victor G Satorras, Syama Sundar Rangapuram, and Tim Januschowski. Multivariate time series forecasting with latent graph inference. *arXiv preprint arXiv:2203.03423*, 2022.

Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, et al. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. Test: Text prototype aligned embedding to activate llm's ability for time series. In *ICLR*, 2024.

Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. Are language models actually useful for time series forecasting? *NeurIPS*, 2024.

Stylianos I Vagropoulos, GI Chouliaras, Evaggelos G Kardakos, Christos K Simoglou, and Anastasios G Bakirtzis. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. In *2016 IEEE international energy conference (ENERGYCON)*, pp. 1–6. IEEE, 2016.

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. *NeurIPS*, 2024.

Zhigaoyuan Wang, Ying Sun, and Hengshu Zhu. Unifying knowledge from diverse datasets to enhance spatial-temporal modeling: A granularity-adaptive geographical embedding approach. In *Forty-second International Conference on Machine Learning*, 2025.

Billy M Williams. Multivariate vehicular traffic flow prediction: evaluation of arimax modeling. *Transportation Research Record*, 1776(1):194–200, 2001.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*, 2023.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

Yiwu Xu and Mengchi Liu. Gpt4tfp: Spatio-temporal fusion large language model for traffic flow prediction. *Neurocomputing*, 625:129562, 2025.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

Athanassios Zagouras, Hugo TC Pedro, and Carlos FM Coimbra. On the role of lagged exogenous variables and spatio–temporal correlations in improving the accuracy of solar forecasting methods. *Renewable Energy*, 78:203–218, 2015.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4120–4131, 2025.

SUPPLEMENTARY MATERIAL

SELECT, THEN BALANCE: A GENERAL FRAMEWORK OF EXOGENOUS VARIABLE MODELING FOR SPATIO-TEMPORAL FORECASTING

TABLE OF CONTENTS

A More Analysis **16**

A.1 More Analysis on Select Module 16

A.2 More Analysis on Balance Module 17

B Experimental Details **17**

B.1 Datasets Details 17

B.2 Baseline Details 18

B.3 Protocol Details 20

C More Results **20**

C.1 More Result on Universality Study 20

C.2 More Result on Robustness Study 20

C.3 More Result on Mechanism Study 22

D More Discussion **24**

D.1 Impact Statement 24

D.2 Current Limitation 24

D.3 Future Direction 25

A MORE ANALYSIS

A.1 MORE ANALYSIS ON SELECT MODULE

We provide theoretical insights into why the proposed select stage (conditional embedding + latent-space gated experts) improves exogenous-aware spatio-temporal forecasting, addressing the challenge of *inconsistent variable effects*. Our analysis is backbone-agnostic and focuses on representation, approximation, and optimization perspectives.

✦ *Input-dependent linear operator and regime-adaptive processing.* The gated expert selector in equation $\mathbf{X}^{\tau'}$ can be reformulated as an input-dependent linear operator:

$$\mathbf{X}^{\tau'} = W(\mathbf{X}^{\tau})\,\mathbf{X}^{\tau}, \quad W(\mathbf{X}^{\tau}) = \sum_{k=1}^{K} g_k^{\tau}(\mathbf{X}^{\tau})\,W_k^{\tau},$$

where $g^{\tau}(\cdot)$ is a simplex-valued gate. Unlike a static projection that enforces a global transformation, it implements a *dynamic filter* whose parameters evolve with local spatio-temporal contexts encoded in $\mathbf{X}^{\tau}$. This explicitly models the inconsistent effects, where the contribution of endogenous channels varies across different weather conditions or temporal regimes.

✦ *Latent regime approximation.* We postulate that the data-generating process involves a continuous latent regime variable $z \in \mathcal{Z}$ (*e.g.*, a continuous spectrum from clear to extreme weather). The predictive distribution can be viewed as an integral over these regimes:

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{E}) = \int_{\mathcal{Z}} p(\mathbf{Y} \mid \mathbf{X}, z)\,p(z \mid \mathbf{E})\,dz.$$

Our module approximates this integral via discretization. The conditional embedding $\phi_{\text{cond}}^{\tau}$ aligns heterogeneous sources, enabling the gate $g^{\tau}(\mathbf{X}^{\tau})$ to approximate the regime posterior $p(z \mid \mathbf{E})$. The K experts W_k^{τ} then serve as support points in the function space, effectively decomposing the complex multimodal mapping into a mixture of local experts:

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{E}) \approx \sum_{k=1}^{K} g_k^{\tau}(\mathbf{X}^{\tau})\,p_k(\mathbf{Y} \mid \mathbf{X}).$$

✦ *Reduced gradient interference via soft routing.* Inconsistent effects also manifest as *gradient conflicts*: samples from different contexts induce updates in opposing directions when sharing parameters. For an expert parameter θ_k (e.g., W_k^{τ}), the per-sample gradient satisfies

$$\nabla_{\theta_k}\mathcal{L}_i \approx g_{i,k}^{\tau}\,\nabla_{\theta_k}\widetilde{\mathcal{L}}_i,$$

where $g_{i,k}^{\tau}$ is the gate weight for sample i. Thus, samples primarily update the experts they activate, which *decomposes* conflicting gradients across experts and reduces negative transfer. Empirically, this is consistent with smoother gate trajectories and clearer expert specialization (Fig. 9), improving optimization stability and downstream forecasting accuracy.

Overall, the key advantage of the Select Module is that it explicitly incorporates exogenous-condition-induced variations in variable effects into both the representation learning and routing mechanisms. Specifically, the conditional embedding first fuses endogenous histories with heterogeneous, multi-source exogenous signals into context-dependent latent representations, thereby establishing a unified semantic basis for subsequent selection. Building on this latent space, the gated experts further instantiate an input-dependent dynamic operator, enabling the model to automatically activate more appropriate processing pathways across different time steps, spatial nodes, and exogenous contexts. Meanwhile, soft routing assigns the learning signal from each contextual sample primarily to its corresponding expert parameters, effectively alleviating gradient interference and negative transfer that commonly arise under shared-parameter training, which in turn promotes context-specialized expert division of labor and improves optimization efficiency. Consequently, by combining *conditional representation* with *expert selection*, the Select stage realizes adaptive modeling of inconsistent variable effects and provides the downstream backbone with more robust input features.

A.2 More Analysis on Balance Module

Although existing methods (Wang et al., 2024; Giganti et al., 2024; Arango et al., 2025) have explored various strategies for balancing past and future exogenous variables, our approach introduces a fundamental distinction in its design philosophy. We theoretically scrutinize the discrepancies among these methods through the lens of functional complexity and the representational capacity of their respective function spaces:

- *Fixed Weighting Strategy:* This approach operates as a rudimentary static linear mapping, formalized as:
$$f(x, y) = (1 - \alpha)x + \alpha y$$
The representational capacity of this method is severely constrained, as the dimension of the function space is 1, governed exclusively by the scalar hyperparameter α. It imposes a rigid interpolation that is agnostic to data distribution.

- *Learnable Weighting Strategy:* While this method introduces optimization, it fundamentally remains a linear mapping within the family of affine transformations:
$$f(x, y) = (1 + \tilde{w}_1)x + (1 + \tilde{w}_2)y$$
Although the parameters are learnable, the objective is to converge upon a single, global optimal weighted balance across the entire training set. Consequently, the learned weights remain static during the inference phase and cannot adapt to the specific variations of individual samples.

- *Context-Aware Balancing (Ours):* In stark contrast, our proposed mechanism functions as an input-dependent, non-linear mapping:
$$f(x, y) = (1 - w(x, y))x + w(x, y)y + (x + y) = \Phi(x, y)$$
Theoretically, this formulation is capable of approximating arbitrary fusion functions. By conditioning the weights $w(x, y)$ on the input itself, the model achieves context sensitivity at inference time, allowing for the dynamic modulation of the fusion strategy based on immediate data characteristics.

Therefore, our methodology diverges significantly from prior static strategies by offering superior dynamic adaptability. To validate this theoretical advantage empirically, Section 5.5 presents a comprehensive ablation study comparing our approach against degenerate fixed-weight baselines, learnable-weight strategies, and attention-based mechanisms. The empirical results robustly demonstrate the efficacy and necessity of our context-aware balancing strategy.

B Experimental Details

B.1 Datasets Details

The statistics of the real-world multi-system exogenous ST datasets in different years (*i.e.*, *AQI-19*, *Speed-19* / *AQI-22*, *Intensity-22*) used in this paper are summarized in Table 5.

Table 5: Summary of datasets used for our experiments. M: million (10^6).

Time range	01/01/2019 – 30/06/2019	01/01/2022 – 30/06/2022
Sampling Rate	1 hour	1 hour
Data points	2.03 M	2.03 M
Variables	20	20
System Domain	Air quality, Traffic, Meteorological, Date	Air quality, Traffic, Meteorological, Date

Graph Construction for GNN-based Models. We classify the GNN models into *adaptive graph construction* and *predefined graph structure* based on their graph construction approaches and perform graph construction on datasets accordingly. For *adaptive graph construction* models including AGCRN, GraphWaveNet, GGNN, and MAGCRN, the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ is computed by

embedding the learnable node. Specifically, given node embeddings $\mathbf{E} \in \mathbb{R}^{N \times d_e}$ where d_e is the embedding dimension, AGCRN computes the adjacency matrix as $\mathbf{A} = \text{softmax}(\text{ReLU}(\mathbf{E}\mathbf{E}^\top))$, where $\mathbf{E}\mathbf{E}^\top$ computes the similarity between all pairs of nodes, ReLU ensures non-negative similarities, and softmax normalizes the weights to create a valid adjacency matrix. This approach allows the model to automatically discover the latent spatial dependencies between the monitoring stations without requiring prior knowledge of their geographical relationships. GWNet uses separate source $\mathbf{E}_s$ and target embeddings $\mathbf{E}_t$ to compute the adjacency matrix as $\mathbf{A} = \text{softmax}(\text{ReLU}(\mathbf{E}_s\mathbf{E}_t^\top))$, allowing more flexible spatial relationship modeling. GGNN employs a fully connected graph structure where $\mathbf{A}_{ij} = 1$ for all $i, j \in \{1, 2, \ldots, N\}$, enabling dense spatial attention across all pairs of node. MAGCRN inherits the adaptive graph construction capability from AGCRN, automatically learning the graph structure through node embeddings. For *predefined graph structure* models that include STGCN, DCRNN, and GRUGCN, due to the lack of geographic location, we follow the mainstream practice [2] and construct the adjacency matrix based on the Pearson correlation coefficient of the target time series. Given the target variable $\mathbf{X}_t \in \mathbb{R}^{N \times F}$ at time t, we compute the correlation matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ where $\mathbf{C}_{ij} = \rho_{ij}$ represents the Pearson correlation coefficient between nodes i and j. The adjacency matrix is then constructed as $\mathbf{A}_{ij} = \begin{cases} \rho ij, & \text{if } j \in \text{TopK}(\rho_{i,:}) \\ 0, & \text{otherwise} \end{cases}$, where the $TopK$ operation retains the k most correlated connections for each node, specifically choosing $k = 8$.

Variable Scales and Units. For the *AQI-19* and *AQI-22*, the target variable NO_2 ranges from 0 to 328 μg/m^3 in Madrid-19 and 0 to 625 μg/m^3 in Madrid-22, with mean values of 36.6 and 28.0 μg/m^3 respectively. For the *Speed-19* and *Intensity-22*, the target variables have various scales: the average traffic speed in Madrid-19 ranges from 0 to 3.57 km/h with a mean of approximately 1.2 km/h, while the traffic intensity in Madrid-22 ranges from 56.5 to 289.88 vehicles/hour with a mean of approximately 170 vehicles/hour. Traffic variables show substantial scale differences: traffic intensity ranges from 56 to 429 $vehicles/hour$, traffic occupation time varies from 0.5% to 5.76%, traffic load ranges from 6.5% to 28.12%, and traffic average speed ranges from 0 to 3.57km/h. Exogenous variables maintain consistent scales across all tasks, with meteorological variables serving as future covariates and traffic variables (excluding the target) serving as past covariates. To handle these diverse scales effectively, we apply StandardScaler normalization to all variables, ensuring that each variable has zero mean and unit variance.

Temporal Variable Encoding. We employ consistent temporal encoding in four forecasting tasks using sinusoidal transformations for periodic features and one-hot encoding for categorical information. For continuous periodic features, given a timestamp t, the hour encoding is calculated as $\sin(2\pi \cdot \text{hour}(t)/24)$ and $\cos(2\pi \cdot \text{hour}(t)/24)$, and the month encoding as $\sin(2\pi \cdot (\text{month}(t) - 1)/12)$ and $\cos(2\pi \cdot (\text{month}(t) - 1)/12)$. For categorical temporal information, each weekday is represented as a binary vector $\mathbf{w} \in \{0, 1\}^7$, where $\mathbf{w}_i = 1$ if the day corresponds to the i-th day of the week (Monday=0, Sunday=6), and $\mathbf{w}_i = 0$ otherwise. These temporal features are added as *universal exogenous* variables to four tasks.

B.2 Baseline Details

In this appendix, we provide detailed descriptions of the advanced exogenous variable modeling methods and different spatio-temporal backbone models used in our default evaluation.

B.2.1 Exogenous Variable Modeling Methods

These methods are grouped into three main categories based on their approach to handling exogenous variables.

TS Models w/ Exogenous Variable Integration.

- *TiDE* (Das et al., 2023): *TiDE* is an encoder–decoder model built on multilayer perceptrons (MLPs) that fuses historical time-series data and covariates via dense MLP residual blocks to map them into low-dimensional representations, then generates forecast sequences by incorporating

[2] Torch-Spatiotemporal

future covariates. https://github.com/google-research/google-research/tree/master/tide

- *TimeXer* (Wang et al., 2024): *TimeXer* is a Transformer model tailored for time series forecasting with exogenous variables: it captures temporal dependencies in the endogenous series via patch-level self-attention and effectively integrates exogenous information through variable-level cross attention and learnable global endogenous tokens to produce high-quality predictions of the target series. https://github.com/thuml/TimeXer
- *NBEATSx* (Olivares et al., 2023): *NBEATSx* extends the original N-BEATS by incorporating convolutional encoding substructures for time-varying and static exogenous variables within a hierarchical stacked residual block architecture, enabling interpretable decomposition of trend, seasonality, and exogenous factors for high-quality target-series forecasting. https://github.com/cchallu/nbeatsx
- *CrossLinear* (Zhou et al., 2025):*CrossLinear* is a Linear-based model that introduces a plug-and-play cross-correlation embedding module to capture time-invariant dependencies between endogenous and exogenous variables, utilizing patch-wise processing and a global linear head to effectively model both short-term and long-term temporal patterns for precise forecasting. https://github.com/mumiao2000/CrossLinear
- *DAG* (Qiu et al., 2025b): *DAG* is a general framework that utilizes a Dual Causal Network along temporal and channel dimensions, it explicitly models how historical exogenous variables affect future exogenous variables (temporal) and historical endogenous variables (channel), and then injects these discovered causal relationships to improve the prediction of future endogenous variables. https://github.com/decisionintelligence/DAG

ST Models w/ Exogenous Variable Integration.

- *MAGCRN* (Giganti et al., 2024): *MAGCRN* is a spatio-temporal model that extends AGCRN into a dual-path architecture, using separate AGCRN blocks to encode historical observations and future exogenous covariates. https://github.com/polimi-ispl/MAGCRN

General Exogenous Variable Integration Framework.

- *ChronosX* (Arango et al., 2025): *ChronosX* extends MAGCRN by integrating two exogenous adaptation modules: Instance-wise Interpolation Block (IIB) and Offset Integration Block (OIB). IIB adaptively interpolates exogenous inputs across time steps, while OIB injects time-dependent bias offsets into the output layer, enabling more flexible integration of future covariates. Due to the limited public availability of the code, we implemented the IIB and OIB modules based on the formulas provided in the referenced paper. https://github.com/amazon-science/chronos-forecasting/tree/chronosx

B.2.2 Spatio-Temporal Backbone Models

To ensure fair use of these models, we use a popular spatio-temporal forecasting benchmark library: https://github.com/TorchSpatiotemporal/tsl.

- *AGCRN* (Bai et al., 2020): *AGCRN* is an adaptive graph-based recurrent model that captures fine-grained spatio-temporal dependencies by jointly learning node-specific patterns and data-driven graph structures, eliminating the need for pre-defined adjacency matrices.
- *GWNet* (Wu et al., 2019): *GWNet* is a spatial-temporal model that combines diffusion-based graph convolutions with dilated causal temporal convolutions, and incorporates a learnable self-adaptive adjacency matrix to capture latent spatial relationships from data.

- *GGNN* (Satorras et al., 2022): *GGNN* is a recurrent graph-based model that applies gated update mechanisms to iteratively refine node representations through message passing, enabling effective modeling of long-range structural dependencies.
- *GRUGCN* (Gao & Ribeiro, 2022): *GRUGCN* is a spatio-temporal neural architecture that integrates graph convolutions for spatial feature extraction and gated recurrent units for sequential modeling, effectively capturing dynamic dependencies in traffic forecasting scenarios.
- *STGCN* (Yu et al., 2017): *STGCN* is a graph-based temporal forecasting model that integrates graph convolutional layers with temporal convolution modules to jointly learn spatial and temporal dependencies in structured time series data.
- *DCRNN* (Li et al., 2018): *DCRNN* is a graph-based recurrent forecasting model that captures spatial dependencies via diffusion convolutions on directed graphs and models temporal dynamics through a sequence-to-sequence recurrent architecture.

B.3 Protocol Details

B.3.1 Metrics Detail.

We use different metrics such as MAE, RMSE, MRE, and MAPE. Formally, these metrics are formulated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

$$\text{MRE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i|}, \qquad \text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|,$$

where n represents the indices of all observed samples, y_i denotes the i-th actual sample and $\hat{y}_i$ is the corresponding prediction. The mean relative error is defined as the MAE divided by the ℓ_1 norm of the target window.

B.3.2 Parameter Detail.

The model is trained for 500 epochs using the AdamW optimizer with a cosine annealing learning rate scheduler. The learning rate is initialized at 10^{-2} and decays to 10^{-7} over the training period. A batch size of 512 is used, incorporating mixed-precision arithmetic and Z-score normalization for target variables. Early stopping is implemented to prevent overfitting, halting training if the validation MAE does not improve for 30 consecutive epochs. All experiments are carried out on a Linux server equipped with a 1 $\times$ AMD EPYC 7763 128-Core Processor CPU with 256GB memory and 4 $\times$ NVIDIA RTX A6000 GPUs, each with 48GB memory. The source code, data and experimental results can be accessed in the anonymous repository at: `https://anonymous.4open.science/r/ExoST`.

C More Results

C.1 More Result on Universality Study

We provide more experimental results on universality (*i.e.*, for 1-day and 2-day scenarios), as shown in Tables 7 and 8. We observe that while the `ExoST` exhibits some negative gains in the 1-day scenario, it exhibits improved performance as the prediction length increases, demonstrating the superior generalization of our method. The remaining observations show similar analytical results to those in the main text.

C.2 More Result on Robustness Study

We provide more experimental results on robustness (*i.e.*, for 1-day and 2-day scenarios) in Tables 9 and 10. We observe similar analysis as in the main text.

Table 6: Performance comparison of different models w/ and w/o ExoST (3-day scenario).

	Model	AGCRN			GWNet			GGNN			GRUGCN			STGCN			DCRNN		
	w/ ExoST	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)
AQI-19	MAE	14.05±0.20	**10.41**±0.12	↓ **25.91**	12.67±2.25	13.33±0.41	↑ **5.21**	14.39±0.22	12.30±1.08	↓ **14.52**	14.10±0.09	12.83±0.55	↓ **9.01**	15.17±1.31	13.14±0.94	↓ **13.38**	14.75±0.20	11.52±0.40	↓ **21.84**
	RMSE	20.92±0.47	**15.51**±0.16	↓ **25.86**	19.00±3.52	19.81±0.28	↑ **4.26**	20.88±0.07	18.57±1.62	↓ **11.06**	21.20±0.04	19.58±0.54	↓ **7.64**	21.72±0.61	19.90±1.08	↓ **8.47**	21.76±0.32	17.51±0.86	↓ **19.53**
	MAPE(%)	63.79±1.32	**44.65**±0.08	↓ **30.00**	63.50±2.08	62.43±3.94	↓ **1.68**	73.36±1.23	53.55±8.41	↓ **27.00**	69.99±1.47	50.96±2.82	↓ **27.20**	82.69±20.81	57.21±7.64	↓ **30.70**	73.64±6.33	46.56±2.39	↓ **36.81**
	MRE(%)	54.36±0.77	**40.34**±0.47	↓ **25.79**	54.05±0.50	51.65±1.60	↓ **4.44**	56.39±1.93	47.67±4.20	↓ **15.46**	54.55±0.33	49.71±2.15	↓ **8.87**	58.89±5.27	50.91±3.66	↓ **13.67**	57.09±0.78	44.62±1.55	↓ **21.84**
*Speed-19**	MAE	3.94±0.01	**3.80**±0.01	↓ **3.55**	4.03±0.06	3.86±0.06	↓ **4.22**	22.53±3.92	4.53±0.38	↓ **79.89**	38.48±0.02	4.63±0.63	↓ **87.97**	6.45±0.48	4.64±0.45	↓ **28.06**	4.87±0.05	4.03±0.20	↓ **17.25**
	RMSE	31.88±0.02	**31.85**±0.02	↓ **0.09**	32.00±0.06	32.05±0.05	↑ **0.16**	45.89±3.35	32.89±0.12	↓ **28.32**	139.12±0.01	32.63±0.45	↓ **76.54**	44.46±0.50	33.84±0.60	↓ **23.90**	34.49±0.38	32.26±0.09	↓ **6.50**
	MAPE(%)	9.42±0.03	**9.31**±0.03	↓ **1.17**	9.54±0.23	9.65±0.23	↑ **1.14**	14.15±0.27	10.63±0.18	↓ **24.90**	99.86±0.01	10.67±0.50	↓ **89.33**	14.02±0.46	11.20±0.81	↓ **20.11**	11.70±0.14	9.73±0.20	↓ **16.84**
	MRE(%)	10.21±0.02	**9.90**±0.03	↓ **3.04**	10.49±0.52	10.05±0.16	↓ **4.20**	58.60±10.19	11.78±0.69	↓ **79.90**	100.09±0.05	12.04±1.78	↓ **87.97**	16.60±1.03	12.06±0.92	↓ **27.35**	12.68±0.13	10.49±0.34	↓ **17.27**
AQI-22	MAE	10.19±0.21	**7.94**±0.29	↓ **22.08**	10.05±0.04	9.11±0.12	↓ **9.35**	10.44±0.04	8.82±0.14	↓ **15.51**	9.97±0.17	8.58±0.19	↓ **13.96**	10.65±0.60	9.21±0.11	↓ **13.51**	10.33±0.72	7.70±0.34	↓ **25.47**
	RMSE	15.34±0.02	**11.79**±0.51	↓ **23.14**	15.11±0.27	13.66±0.08	↓ **9.60**	15.13±0.04	13.01±0.25	↓ **14.03**	15.02±0.19	13.29±0.43	↓ **11.51**	15.65±0.50	14.16±0.17	↓ **9.53**	15.25±0.83	11.81±0.61	↓ **22.49**
	MAPE(%)	56.91±2.06	**50.27**±1.18	↓ **11.67**	60.72±1.66	60.70±1.57	↓ **0.03**	71.58±1.28	59.79±2.29	↓ **16.47**	71.37±2.27	53.48±1.16	↓ **25.07**	73.76±10.10	59.47±0.81	↓ **19.41**	69.19±6.00	51.26±3.68	↓ **25.93**
	MRE(%)	54.54±1.12	42.50±1.57	↓ **22.08**	53.77±0.20	48.75±0.65	↓ **9.34**	55.85±0.21	47.18±0.74	↓ **15.52**	53.35±0.89	45.89±1.02	↓ **13.99**	56.98±3.20	49.28±0.58	↓ **13.51**	55.27±3.85	**41.18**±1.83	↓ **25.50**
Intensity-22	MAE	85.76±0.47	64.13±3.86	↓ **25.22**	87.34±0.75	**61.87**±3.80	↓ **29.14**	84.17±0.43	77.29±1.21	↓ **8.17**	85.03±1.43	75.55±6.83	↓ **11.16**	90.49±0.75	72.17±1.54	↓ **20.23**	87.57±0.79	64.55±2.61	↓ **26.29**
	RMSE	144.77±0.94	100.01±9.19	↓ **28.34**	147.13±3.61	**105.44**±5.53	↓ **28.34**	141.73±1.35	125.02±2.74	↓ **11.83**	141.30±1.92	122.13±11.30	↓ **13.62**	148.09±0.05	117.47±1.89	↓ **20.68**	145.99±1.77	109.83±2.77	↓ **24.89**
	MAPE(%)	59.46±3.18	34.52±2.90	↓ **41.94**	44.80±0.89	**31.74**±1.69	↓ **29.15**	47.16±0.66	45.19±1.92	↓ **4.18**	48.76±0.73	44.98±5.12	↓ **7.75**	50.02±2.67	39.01±0.67	↓ **22.01**	46.25±1.65	34.48±2.83	↓ **25.45**
	MRE(%)	30.97±0.17	23.16±1.40	↓ **25.22**	31.54±0.27	**22.35**±1.37	↓ **29.14**	30.40±0.16	27.91±1.50	↓ **8.19**	30.71±0.51	27.29±2.47	↓ **11.14**	32.68±0.27	26.06±0.56	↓ **20.39**	31.62±0.29	23.31±0.94	↓ **26.27**

Table 7: Performance comparison of different models w/ and w/o ExoST (2-day scenario).

	Model	AGCRN			GWNet			GGNN			GRUGCN			STGCN			DCRNN		
	w/ ExoST	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)
AQI-19	MAE	13.77±0.47	**10.04**±0.19	↓ **27.09**	12.29±2.45	12.75±0.21	↑ **3.74**	14.28±0.11	12.14±1.08	↓ **14.99**	13.88±0.06	12.58±0.55	↓ **9.37**	14.73±0.66	12.78±0.64	↓ **13.24**	14.27±0.31	11.31±0.01	↓ **20.74**
	RMSE	20.54±0.58	**15.12**±0.26	↓ **26.39**	18.87±3.73	19.18±0.13	↑ **1.64**	20.81±0.38	18.45±1.58	↓ **11.34**	20.65±0.06	19.47±0.50	↓ **5.71**	21.80±0.56	19.66±0.67	↓ **9.82**	20.97±0.36	17.55±0.26	↓ **16.31**
	MAPE(%)	62.14±0.54	**42.62**±0.52	↓ **31.41**	60.03±4.32	58.83±2.95	↓ **2.00**	71.16±1.37	52.19±0.62	↓ **26.66**	69.08±1.30	49.74±2.81	↓ **27.99**	75.40±9.03	54.59±6.32	↓ **27.61**	70.18±6.64	43.76±0.03	↓ **37.65**
	MRE(%)	53.33±1.79	**38.84**±0.75	↓ **27.17**	52.24±1.39	49.36±0.82	↓ **5.51**	55.63±0.78	46.97±4.17	↓ **15.58**	53.77±0.24	48.68±2.11	↓ **9.47**	57.02±2.55	49.45±2.46	↓ **13.28**	55.27±1.22	43.76±0.03	↓ **20.82**
*Speed-19**	MAE	3.93±0.01	3.80±0.01	↓ **3.31**	3.96±0.06	**3.77**±0.06	↓ **4.80**	22.33±3.90	4.28±0.34	↓ **80.83**	38.45±0.02	4.52±0.62	↓ **88.27**	6.27±0.53	4.51±0.41	↓ **28.07**	4.73±0.06	3.96±0.18	↓ **16.28**
	RMSE	31.82±0.02	31.78±0.02	↓ **0.13**	**31.38**±0.02	31.66±0.02	↑ **0.88**	45.19±3.38	32.18±0.34	↓ **28.79**	139.01±0.01	32.26±0.55	↓ **76.78**	43.66±0.77	33.12±0.96	↓ **31.82**	33.64±0.41	31.92±0.25	↓ **5.11**
	MAPE(%)	9.43±0.02	9.32±0.03	↓ **1.17**	**9.05**±0.23	9.45±0.22	↑ **4.76**	13.82±0.29	10.03±0.35	↓ **27.42**	99.85±0.01	10.44±0.70	↓ **89.55**	13.59±0.58	10.94±0.91	↓ **19.49**	11.41±0.14	9.58±0.30	↓ **16.04**
	MRE(%)	10.21±0.02	9.89±0.03	↓ **3.13**	10.29±0.46	**9.83**±0.15	↓ **4.68**	58.11±10.16	11.13±0.88	↓ **80.84**	100.08±0.05	11.77±1.61	↓ **88.24**	16.09±1.11	11.75±1.05	↓ **26.97**	12.32±0.16	10.29±0.46	↓ **16.48**
AQI-22	MAE	9.91±0.05	7.61±0.23	↓ **23.21**	9.39±0.20	8.66±0.17	↓ **7.78**	10.10±0.03	8.49±0.17	↓ **15.94**	9.71±0.19	8.10±0.07	↓ **16.57**	10.00±0.76	8.71±0.14	↓ **12.90**	9.93±0.85	**7.35**±0.21	↓ **25.97**
	RMSE	15.08±0.12	11.52±0.49	↓ **23.61**	14.51±0.34	13.26±0.04	↓ **8.60**	14.93±0.12	12.76±0.18	↓ **14.52**	14.84±0.19	13.07±0.23	↓ **11.91**	14.98±0.65	13.77±0.21	↓ **8.08**	14.85±0.89	**11.47**±0.39	↓ **22.79**
	MAPE(%)	54.78±1.05	**46.59**±0.64	↓ **14.95**	54.79±0.59	56.45±2.25	↑ **3.03**	66.35±1.23	55.59±2.78	↓ **16.19**	67.26±2.25	48.14±1.33	↓ **28.40**	68.35±11.56	54.35±1.09	↓ **20.43**	66.35±8.27	47.63±3.59	↓ **28.21**
	MRE(%)	53.01±0.29	40.74±1.24	↓ **23.15**	50.25±1.09	46.35±0.92	↓ **7.76**	54.05±0.16	45.41±0.93	↓ **15.97**	51.96±1.02	43.35±0.36	↓ **16.56**	53.51±4.08	46.59±0.76	↓ **12.92**	53.11±4.55	**39.35**±1.11	↓ **25.94**
Intensity-22	MAE	62.66±0.59	50.59±0.04	↓ **19.26**	62.65±1.19	51.09±4.57	↓ **18.53**	61.34±0.83	62.36±1.83	↑ **1.66**	67.02±1.78	62.31±3.14	↓ **7.04**	68.04±0.65	60.90±1.19	↓ **10.49**	64.29±1.37	**50.53**±3.94	↓ **21.39**
	RMSE	113.21±1.13	**89.10**±7.36	↓ **21.30**	113.12±1.07	91.20±7.36	↓ **19.38**	111.50±0.40	102.84±2.01	↓ **7.78**	117.62±2.98	104.07±6.46	↓ **11.53**	116.81±0.13	102.06±2.49	↓ **12.63**	114.56±2.74	90.24±6.35	↓ **21.23**
	MAPE(%)	45.87±0.60	26.68±1.69	↓ **41.84**	30.77±0.48	26.62±3.11	↓ **13.50**	33.47±0.86	35.00±3.14	↑ **4.56**	38.23±0.99	35.81±1.34	↓ **6.33**	36.08±2.07	32.27±1.58	↓ **10.54**	32.31±1.36	**26.11**±2.67	↓ **19.18**
	MRE(%)	22.68±0.21	18.31±1.14	↓ **19.27**	22.68±0.43	18.49±1.66	↓ **18.48**	22.20±0.30	22.57±1.62	↑ **1.67**	24.26±0.64	22.55±1.14	↓ **7.04**	24.63±0.24	22.04±0.43	↓ **10.51**	23.27±0.49	**18.29**±1.43	↓ **21.41**

Table 8: Performance comparison of different models w/ and w/o ExoST (1-day scenario).

	Model	AGCRN			GWNet			GGNN			GRUGCN			STGCN			DCRNN		
	w/ ExoST	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)	✗	✓	Δ(%)
AQI-19	MAE	13.63±0.45	**9.33**±0.13	↓ **31.55**	10.92±2.01	11.57±0.32	↑ **5.95**	11.57±0.37	10.78±0.48	↓ **6.83**	11.63±0.08	11.65±0.39	↑ **0.17**	13.60±1.30	11.55±0.21	↓ **15.07**	13.80±0.46	10.25±0.19	↓ **25.72**
	RMSE	20.43±0.89	**14.24**±0.18	↓ **30.30**	17.17±5.31	17.79±0.37	↑ **3.61**	18.25±0.20	16.87±0.89	↓ **7.56**	18.44±0.04	18.43±0.38	↓ **0.05**	20.56±1.21	18.30±0.29	↓ **10.99**	21.07±0.43	16.28±0.06	↓ **22.73**
	MAPE(%)	60.71±0.41	**38.57**±0.32	↓ **36.80**	51.54±2.00	51.92±2.89	↑ **0.74**	56.99±1.81	45.09±5.46	↓ **20.88**	56.54±1.34	44.86±2.08	↓ **20.78**	68.55±12.26	48.33±3.28	↓ **29.50**	68.45±7.25	44.68±0.81	↓ **34.72**
	MRE(%)	52.74±1.81	**35.85**±0.71	↓ **32.03**	46.47±0.65	44.73±1.23	↓ **3.74**	45.17±0.99	41.67±1.67	↓ **7.75**	44.96±0.32	45.06±1.52	↑ **0.22**	52.59±5.03	44.68±0.81	↓ **15.04**	53.37±1.80	39.63±0.74	↓ **25.75**
*Speed-19**	MAE	3.92±0.01	3.79±0.01	↓ **3.32**	3.81±0.13	**3.70**±0.06	↓ **2.89**	21.81±3.90	4.03±0.26	↓ **81.52**	38.45±0.02	4.39±0.68	↓ **88.60**	5.53±0.40	4.34±0.36	↓ **21.52**	4.52±0.05	3.81±0.13	↓ **15.71**
	RMSE	31.78±0.02	31.73±0.02	↓ **0.16**	**30.92**±0.04	31.42±0.04	↑ **1.61**	43.31±3.43	31.08±0.12	↓ **28.30**	139.01±0.01	31.76±0.45	↓ **77.15**	38.60±0.65	32.19±0.60	↓ **16.61**	32.47±0.12	31.11±0.09	↓ **4.19**
	MAPE(%)	9.42±0.00	9.30±0.03	↓ **1.27**	**8.59**±0.21	9.27±0.21	↑ **7.33**	12.86±0.31	9.42±0.18	↓ **26.75**	99.84±0.01	10.12±0.50	↓ **89.87**	12.19±0.57	10.55±0.81	↓ **13.45**	10.95±0.13	9.25±0.20	↓ **15.52**
	MRE(%)	10.20±0.02	9.88±0.03	↓ **3.14**	9.91±0.16	**9.64**±0.16	↓ **2.72**	56.77±10.14	10.48±0.69	↓ **81.52**	100.07±0.05	11.42±1.78	↓ **87.59**	14.38±1.03	11.30±0.92	↓ **21.42**	11.76±0.12	9.91±0.34	↓ **15.73**
AQI-22	MAE	9.55±0.24	7.33±0.21	↓ **23.25**	8.58±0.41	7.98±0.19	↓ **6.99**	8.60±0.01	8.15±0.11	↓ **5.23**	8.13±0.11	7.55±0.09	↓ **7.13**	9.17±0.88	7.99±0.14	↓ **12.87**	9.11±1.32	**6.81**±0.30	↓ **25.25**
	RMSE	14.58±0.48	11.14±0.21	↓ **23.59**	13.44±0.62	12.26±0.10	↓ **8.78**	13.20±0.12	12.43±0.25	↓ **5.83**	12.84±0.08	12.17±0.29	↓ **5.22**	13.86±0.77	12.65±0.12	↓ **8.73**	13.97±1.56	**10.72**±0.58	↓ **23.26**
	MAPE(%)	**31.80**±0.70	44.90±0.88	↑ **41.19**	49.43±1.83	51.38±1.95	↑ **3.94**	55.74±1.04	53.39±2.41	↓ **4.21**	55.89±1.69	45.92±1.50	↓ **17.76**	61.74±11.81	50.05±2.16	↓ **18.93**	59.84±11.34	43.94±3.49	↓ **26.57**
	MRE(%)	51.10±1.28	30.23±1.11	↓ **23.23**	45.94±2.19	42.69±1.02	↓ **7.07**	46.01±0.06	43.63±0.60	↓ **5.17**	43.50±0.61	40.41±0.50	↓ **7.10**	49.06±4.70	42.76±0.75	↓ **12.84**	48.77±7.09	**36.44**±1.62	↓ **25.30**
Intensity-22	MAE	33.25±0.83	31.91±1.35	↓ **4.03**	34.00±3.74	35.94±4.97	↑ **5.71**	**29.74**±1.21	42.17±0.48	↑ **41.86**	38.91±1.68	42.04±3.37	↑ **8.04**	43.06±0.92	45.77±0.15	↑ **6.29**	35.91±2.16	32.78±2.80	↓ **8.72**
	RMSE	56.56±0.99	56.49±2.54	↓ **0.12**	57.40±4.89	63.08±8.58	↑ **9.90**	**48.88**±1.95	68.39±1.71	↑ **39.91**	70.08±4.91	68.89±5.45	↓ **1.70**	69.51±0.85	77.07±2.32	↑ **10.88**	60.87±4.59	58.63±4.43	↓ **3.68**
	MAPE(%)	18.53±0.75	**17.63**±0.97	↓ **4.86**	17.78±1.97	20.28±3.36	↑ **14.06**	19.26±0.95	22.95±0.95	↑ **19.16**	25.12±0.89	24.94±2.91	↓ **0.72**	24.23±1.85	25.15±2.09	↑ **3.79**	19.88±2.19	17.71±1.93	↓ **10.91**
	MRE(%)	12.05±0.30	11.57±0.49	↓ **3.98**	12.32±1.36	13.03±1.80	↑ **5.76**	**10.78**±0.44	15.28±0.18	↑ **41.74**	14.11±0.61	15.24±1.22	↑ **7.96**	15.61±0.33	16.59±0.06	↑ **6.28**	13.01±0.78	11.89±1.01	↓ **8.61**

Table 9: Performance under abnormal conditions with missing exogenous signals (1-day scenario).

Datasets		*AQI-19*				*Speed-19**				*Intensity-22*			
Masking Strategy		MAE	RMSE	MAPE(%)	MRE(%)	MAE	RMSE	MAPE(%)	MRE(%)	MAE	RMSE	MAPE(%)	MRE(%)
Zero	20%	9.98±0.24	15.19±0.06	40.86±0.02	38.59±0.01	3.81±0.01	31.80±0.04	9.28±0.01	9.91±0.01	33.74±3.24	62.25±7.82	17.81±1.15	12.23±1.18
	40%	10.41±0.26	16.22±0.52	41.00±0.34	40.28±1.03	3.79±0.01	31.77±0.02	**9.23**±0.01	**9.84**±0.01	33.70±0.40	58.59±2.51	19.63±0.77	12.22±0.14
	60%	10.33±0.22	16.26±0.56	41.71±0.01	39.98±0.01	3.80±0.01	31.75±0.05	9.31±0.15	9.90±0.12	36.22±0.39	65.67±3.71	19.90±0.68	13.13±0.14
	80%	10.99±0.15	17.10±0.29	47.36±4.28	42.50±0.58	3.81±0.03	31.78±0.06	9.41±0.13	9.93±0.08	35.96±1.23	64.86±4.13	20.03±1.02	13.04±0.44
Random	20%	9.96±0.12	15.40±0.28	40.04±0.02	38.53±0.01	3.80±0.01	31.80±0.06	9.25±0.01	9.88±0.02	32.91±2.61	60.30±5.91	18.01±0.93	11.93±0.94
	40%	10.40±0.31	16.12±0.33	41.52±0.41	40.24±0.01	3.82±0.03	31.79±0.04	9.41±0.01	9.95±0.08	**31.40**±1.10	**54.03**±1.13	17.89±1.27	**11.38**±0.40
	60%	10.66±0.28	16.91±0.53	43.23±0.55	41.23±0.01	3.80±0.04	31.76±0.05	9.31±0.15	9.88±0.09	34.34±2.71	61.96±9.15	18.84±1.25	12.44±0.98
	80%	11.13±0.14	17.67±0.19	43.96±0.01	43.05±0.55	**3.78**±0.01	**31.72**±0.03	**9.23**±0.01	**9.84**±0.02	34.68±2.80	64.10±8.08	19.05±0.96	12.57±1.02
No Masking		**9.33**±0.13	**14.24**±0.18	**38.57**±0.32	**36.09**±0.49	3.79±0.01	31.73±0.02	9.30±0.03	9.88±0.03	31.91±1.35	56.49±2.54	**17.63**0.97	11.57±0.49

Table 10: Performance under abnormal conditions with missing exogenous signals (2-day scenario).

Datasets		*AQI-19*				*Speed-19**				*Intensity-22*			
Masking Strategy		MAE	RMSE	MAPE(%)	MRE(%)	MAE	RMSE	MAPE(%)	MRE(%)	MAE	RMSE	MAPE(%)	MRE(%)
Zero	20%	10.48±0.21	15.73±0.23	44.37±1.28	40.54±0.82	3.84±0.01	31.86±0.05	9.35±0.04	9.99±0.04	**50.06**±6.65	**89.00**±13.90	**26.06**±2.03	**18.12**±2.40
	40%	11.05±0.07	16.78±0.13	44.64±1.19	42.76±0.28	3.82±0.03	31.85±0.05	9.33±0.06	9.95±0.08	53.31±0.02	94.52±0.57	29.38±0.70	19.29±0.01
	60%	11.22±0.45	17.21±0.75	46.81±1.28	43.42±1.72	3.83±0.02	31.84±0.04	9.38±0.08	9.97±0.06	54.33±7.93	96.30±14.73	28.86±3.89	19.66±2.87
	80%	12.28±0.00	19.05±0.00	51.12±0.00	47.52±0.00	3.82±0.02	31.83±0.05	9.41±0.10	9.94±0.06	56.17±5.74	99.93±10.67	29.99±3.51	20.33±2.07
Random	20%	10.44±0.09	15.93±0.04	43.41±1.98	40.41±0.34	3.83±0.00	31.87±0.07	**9.32**±0.05	9.96±0.02	51.75±7.53	92.27±15.45	26.69±2.90	18.73±2.73
	40%	10.96±0.12	16.74±0.27	45.10±1.23	42.43±0.47	3.83±0.02	31.85±0.04	9.42±0.08	9.97±0.06	50.93±9.79	89.54±18.50	26.92±5.07	18.44±3.54
	60%	11.06±0.08	16.95±0.24	45.94±1.28	42.81±0.32	3.82±0.02	31.84±0.04	9.37±0.07	9.96±0.05	51.76±4.47	92.00±10.11	27.44±2.26	18.73±1.62
	80%	12.28±0.00	19.05±0.00	51.12±0.00	47.52±0.00	3.82±0.01	31.82±0.06	**9.32**±0.04	9.94±0.03	53.20±6.37	94.41±14.30	28.13±2.32	19.26±2.30
No Masking		**10.04**±0.19	**15.12**±0.26	**42.62**±0.52	**38.84**±0.75	**3.80**±0.01	**31.78**±0.02	**9.32**±0.03	**9.89**±0.03	50.59±0.04	89.10±7.36	26.68±1.69	18.31±1.14

C.3 More Result on Mechanism Study

C.3.1 Ablation Study.

Table 11: Ablation study from the data perspective on *AQI-19* for 2-day forecasting, where P, F, and D represent Past, Future, and Data exogenous variables respectively.

P	F	D	MAE	RMSE	MAPE (%)	MRE (%)
✓	-	-	14.71±0.98	21.53±0.82	74.15±8.44	56.91±3.79
-	✓	-	10.71±0.40	16.05±0.55	45.33±2.40	41.44±1.54
-	-	✓	14.42±0.81	21.34±1.01	69.23±8.53	55.81±3.10
✓	-	✓	14.35±0.24	21.24±0.52	71.57±1.48	55.52±0.92
-	✓	✓	10.32±0.02	15.61±0.02	43.22±0.13	39.93±0.08
✓	✓	-	10.27±0.33	15.67±0.57	**42.59**±0.92	39.73±1.27
✓	✓	✓	**10.04**±0.19	**15.12**±0.26	42.62±0.52	**38.84**±0.75

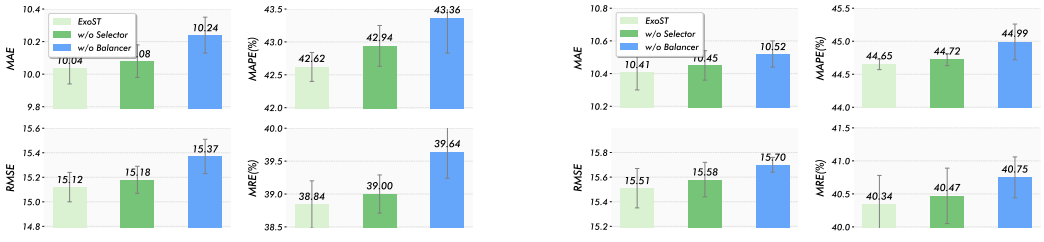
Table 12: Ablation study from the data perspective on *AQI-19* for 3-day forecasting, where P, F, and D represent Past, Future, and Data exogenous variables respectively.

P	F	U	MAE	RMSE	MAPE (%)	MRE (%)
✓	-	-	14.26±0.78	21.13±0.77	71.30±6.43	55.23±3.02
-	✓	-	10.94±0.28	16.25±0.38	46.23±1.60	42.38±1.08
-	-	✓	14.55±0.63	21.42±0.60	70.16±9.39	56.33±2.44
✓	-	✓	14.20±0.10	20.83±0.35	69.91±0.12	55.02±0.38
-	✓	✓	10.54±0.10	15.83±0.12	44.59±0.19	40.84±0.39
✓	✓	-	10.49±0.28	15.85±0.47	**43.71**±0.70	40.62±1.08
✓	✓	✓	**10.41**±0.12	**15.51**±0.16	44.65±0.08	**40.34**±0.47

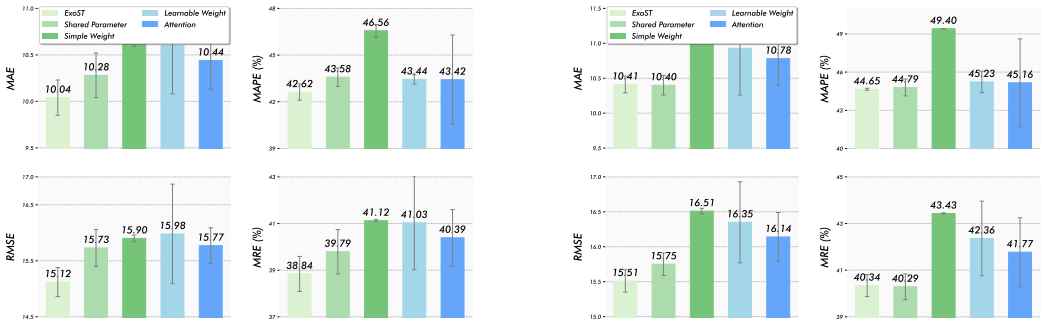
From the data perspective, we also evaluated the impact of exogenous variables on ExoST's performance for 2-day and 3-day forecasting on the *AQI-19* , as shown in Tables 11 and Tables 12. We have further observation: ❶ Consistent with the 1-day findings, future covariates outperform past covariates, while data variables alone produce limited accuracy but enhance performance when combined with others, with the best results achieved by integrating all variables, demonstrating ExoST's ability to synthesize diverse temporal information. ❸ In 2-day forecasting, combining past and future covariates significantly improves percentage-based metrics, indicating a balanced contribution from historical and forward-looking signals. ❷ For 3-day forecasting, combining future and data covariates almost matches the full model, highlighting the significant role of the immediate temporal context in short-term predictions. ❹ Across 1-day to 3-day horizons, the most critical trend is increasing the importance of the past covariates as the forecast horizon extends, complementing the dominant predictive power of future covariates.

C.3.2 Strategy Study.

We first formalize the four strategies in detail through formulas:



(a) Ablation (2-day) (b) Ablation (3-day)



(c) Strategy (2-day) (d) Strategy (3-day)

Figure 13: Combined analysis on *AQI-19* dataset. (a)-(b) Ablation studies; (c)-(d) Strategy studies.

- *Sharing Parameter*: A single spatio-temporal encoder is re-used for both exogenous streams:

$$\mathbf{Y}^p = \phi_{\text{st}}^{\text{shared}}\big(\mathbf{X}^{p'}\big), \qquad \mathbf{Y}^f = \phi_{\text{st}}^{\text{shared}}\big(\mathbf{X}^{f'}\big)$$

  The forecast is obtained by averaging the two outputs:

$$\hat{\mathbf{Y}} = \frac{1}{2}\mathbf{Y}^p + \frac{1}{2}\mathbf{Y}^f.$$

- *Simple Weight*: Each stream is processed by its own encoder 4.1. The final prediction employs fixed, equal weights:

$$\hat{\mathbf{Y}} = \frac{1}{2}\,\mathbf{Y}^p + \frac{1}{2}\,\mathbf{Y}^f.$$

- *Learnable Weight*: Let $\mathbf{w}_{init} \in \mathbb{R}^2$ be a learnable vector, normalized by softmax:

$$\mathbf{w} = \text{Softmax}(\mathbf{w_{init}}), \qquad w_0 + w_1 = 1.$$

  Given the two outputs of the backbone, the fusion is

$$\hat{\mathbf{Y}} = w_0\,\mathbf{Y}^p + w_1\,\mathbf{Y}^f + \big(\mathbf{Y}^p + \mathbf{Y}^f\big),$$

  where the last term is a residual connection.

- *Attention*: Let $\mathbf{Y}^p, \mathbf{Y}^f$ denote the intermediate hidden states produced by the two branches before the final read-out layer. For each directional attention we use shared projection matrices $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{H\times H}$. *Future* $\rightarrow$ *Past*

$$\mathbf{Q}^f = \mathbf{Y}^f\mathbf{W}_q, \quad \mathbf{K}^p = \mathbf{Y}^p\mathbf{W}_k, \quad \mathbf{V}^p = \mathbf{Y}^p\mathbf{W}_v,$$

$$\boldsymbol{\alpha}_{f\to p} = \text{softmax}\left(\frac{\sum_{h=1}^{H}\big(\mathbf{Q}^f_{:,t,h}\cdot\mathbf{K}^p_{:,t,h}\big)}{\sqrt{H}}\right),$$

$$\mathbf{A}_{f\to p} = \boldsymbol{\alpha}_{f\to p}\odot\mathbf{V}^p,$$

$$\mathbf{Y}^p_{\text{enh}} = \mathbf{Y}^p + \mathbf{A}_{f\to p}.$$

Past → Future

$$\mathbf{Q}^p = \mathbf{Y}^p\mathbf{W}_q, \quad \mathbf{K}^f = \mathbf{Y}^f\mathbf{W}_k, \quad \mathbf{V}^f = \mathbf{Y}^f\mathbf{W}_v,$$
$$\boldsymbol{\alpha}_{p\to f} = \mathrm{softmax}\left(\frac{\sum_{h=1}^{H}\left(\mathbf{Q}^p_{:,t,h}\cdot\mathbf{K}^f_{:,t,h}\right)}{\sqrt{H}}\right),$$
$$\mathbf{A}_{p\to f} = \boldsymbol{\alpha}_{p\to f}\odot\mathbf{V}^f,$$
$$\mathbf{Y}^f_{\text{enh}} = \mathbf{Y}^f + \mathbf{A}_{p\to f}.$$

Final Prediction

$$\hat{\mathbf{Y}} = \frac{1}{2}\mathbf{Y}^p_{\text{enh}} + \frac{1}{2}\mathbf{Y}^f_{\text{enh}},$$

As Figures 13 show, we can observe that: ❶ In the 2-day horizon, ExoST maintains its advantage, while *Simple Weight* and *Learnable Weight* exhibit amplified weaknesses: the former's static weighting fails to handle increased uncertainty in longer forecasts, and the latter, despite its adaptable weights, amplifies noise from future exogenous signals due to lacking day-to-day context modeling. ❷ Although all methods show increasing errors across horizons of 1-day to 3-day, ExoST consistently suppresses error accumulation better than alternatives.

D More Discussion

D.1 Impact Statement

This paper presents work whose goal is to advance the field of machine learning for physical world modeling. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

D.2 Current Limitation

In this paper, we propose a new framework for modeling exogenous variables in spatio-temporal forecasting: ExoST. This "select, then balance" design addresses the inconsistent effects of variables and the imbalanced effects of types. Extensive experiments demonstrate the superiority of our method in terms of effectiveness, universality, and robustness. Although we have made some attempts in this direction, there are still some limitations that need to be addressed:

- First, our method assumes that spatio-temporal exogenous variables are available in real-time and are synchronized with the target system. In reality, it may not be possible to acquire real-time exogenous information at all times. Even when available, differing equipment may lead to inconsistent and mismatched sampling rates. While we conducted preliminary experiments using random masking to simulate asynchronous and non-real-time data, and our model showed some robustness, a deeper investigation into this issue is left for future work.
- Second, our model currently assumes that spatio-temporal exogenous variables are numerical sensor signals. However, real-world scenarios often involve diverse, heterogeneous exogenous information, such as language and visual information (Chen et al., 2024; Luo et al., 2024; Chen et al., 2025a). Recent research has explored using Large Language Models (LLMs) for time series (Jin et al., 2024b;c; Pan et al., 2024; Sun et al., 2024; Jia et al., 2024; Chang et al., 2025; Liu et al., 2025b) and spatio-temporal forecasting (Liu et al., 2024; Li et al., 2024; Liu et al., 2025a; Xu & Liu, 2025), hoping that LLMs can gain numerical prediction capabilities through various alignment techniques. However, some studies have raised significant doubts about these approaches (Tan et al., 2024). Different from these views, we believe that language can be more effectively viewed as sparse, discrete exogenous auxiliary information. The key challenge lies in how to leverage this information to enhance various ST tasks.

D.3 Future Direction

Based on these analyses, we identify two key areas for future work:

- For spatio-temporal forecasting tasks, we believe the core challenge is to use the semantic representations of exogenous variables to provide an auxiliary spatio-temporal backbone network with interventional information, allowing it to respond to dynamic changes. This approach is distinct from using language models as the primary prediction backbone.
- For spatio-temporal reasoning tasks (Han et al., 2025; Li et al., 2025; Hettige et al., 2025; Ni et al., 2026), we believe the core essence is to activate the intrinsic capabilities of language models to analyze the underlying causes of ST changes and provide logical knowledge-based answers.