FROM VALUES TO INTERACTIONS: ADAPTIVE GRAPH PRE-TRAINING FOR SPATIOTEMPORAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatiotemporal forecasting is inherently relational: future states depend not only on historical observations but also on how information is organized across interacting entities. Yet existing forecasting models typically learn such relations within task-specific predictors, while recent pre-training methods remain largely value-centric, emphasizing temporal reconstruction without explicitly encoding reusable cross-entity dependencies. We propose **AGPST**, an interaction-aware self-supervised framework that treats relational structure as reusable pre-training knowledge. AGPST combines data-adaptive graph aggregation with masked temporal modeling, enabling masked states to be reconstructed from both cross-entity evidence and long-range historical context. The resulting representations capture complementary relational and temporal information and can be transferred to heterogeneous downstream predictors through lightweight feature integration without modifying their internal architectures. Experiments on six traffic forecasting benchmarks show consistent improvements over representative task-specific and pre-training baselines across datasets, metrics, and prediction horizons. Ablation and cross-architecture studies further demonstrate the complementary benefits of adaptive interaction learning, graph propagation, and masked reconstruction, supporting the effectiveness and reusability of the learned representations.

1 INTRODUCTION

A future observation in a spatiotemporal system is rarely determined by its own history alone. In traffic networks, climate systems, energy grids, and public-safety applications, forecasting supports *decision-making* by anticipating how states jointly evolve across multiple interacting entities (Zhang et al., 2025; Cao et al., 2025). Traffic congestion, for example, is typically not caused by an isolated fluctuation at a single sensor, but emerges through the accumulation and propagation of traffic dynamics across interconnected road segments. Likewise, a local temperature anomaly becomes operationally significant only when it persists over time and develops coherently across neighboring regions (Marisca et al., 2024). These observations suggest that spatiotemporal forecasting is not merely a problem of extrapolating individual signals. More fundamentally, it requires models to organize evidence across spatial entities and temporal contexts so as to infer the collective dynamics that govern future system states.

This relational nature of forecasting gives rise to two closely coupled challenges. The first is *spatial interaction variability*. Physical proximity or a predefined topology provides a useful structural prior, but does not fully determine the effective dependency between entities. Geographically adjacent sensors may exhibit strong correlations during one period yet become weakly related during another, while distant locations may temporarily synchronize because of rerouted flows, demand changes, or external disruptions. We refer to this mismatch between predefined structure and effective dependency as **spatiotemporal inconsistency**. The second challenge is *temporal scale entanglement*. Short-term fluctuations, such as transient peaks, are embedded within longer-term rhythms spanning days, weeks, or seasons. These scales are not independent: long-term regimes regulate local variations, while accumulated local deviations may gradually reshape global patterns. Reliable forecasting therefore requires representations that simultaneously capture data-adaptive spatial interactions and long-range, multiscale temporal context.

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.

Existing methods only partially address these requirements. Classical statistical models (Fang et al., 2021; Song et al., 2020) rely heavily on stationarity or linearity assumptions and are consequently limited in complex non-stationary systems. Neural forecasting models combine graph neural networks with recurrent, convolutional, or attention-based temporal modules (Li et al., 2018; Yu et al., 2018; Wu et al., 2019; Guo et al., 2019), substantially improving their ability to capture spatial and temporal dependencies. Nevertheless, many of them rely on predefined or task-specific graph structures and learn relational patterns only through short-horizon supervised forecasting objectives. More recently, spatiotemporal pre-training methods (Gao et al., 2024) have exploited long historical sequences through masked reconstruction, demonstrating the value of reusable temporal representations. However, their objectives remain primarily *value-centric*: they emphasize recovering missing observations while spatial dependencies are often fixed, implicitly encoded, or separated from temporal representation learning. As a result, pre-training may effectively learn what historical signals look like without sufficiently capturing how cross-entity evidence should be organized to explain them.

This observation motivates a different view of spatiotemporal pre-training. Rather than regarding spatial structure as an auxiliary input to a temporal reconstruction model, we argue that *interaction-aware context itself should be learned as reusable forecasting knowledge*. The goal of pre-training is therefore not only to recover masked values, but also to construct representations in which long-range temporal patterns are interpreted through data-adaptive spatial interactions. Such representations can complement downstream predictors because they encode relational and temporal context that may be difficult to acquire from short supervised forecasting windows alone.

Following this perspective, we propose **AGPST**, an **A**daptive **G**raph-enhanced **P**re-training framework for **S**patiotemporal **T**ime-series forecasting. AGPST integrates adaptive graph representation learning with transformer-based masked autoencoding. The adaptive graph module relaxes the reliance on predefined spatial structures by learning interaction strengths directly from data, while graph-based aggregation organizes information across correlated entities. Meanwhile, patch-level masked reconstruction exposes the encoder to substantially longer temporal contexts and discourages trivial recovery from immediate observations, thereby promoting the joint encoding of spatial interactions and multiscale temporal dependencies. The resulting representations are subsequently fused with downstream forecasting models through a lightweight integration module, without modifying their internal architectures. In this way, AGPST separates reusable interaction-aware representation learning from task-specific prediction, allowing heterogeneous forecasting models to benefit from the same pre-trained spatiotemporal knowledge.

Our main contributions are summarized as follows:

- We reveal a mismatch between the relational nature of spatiotemporal systems and the value-centric objectives of existing pre-training methods, and reformulate pre-training as learning reusable *interaction-aware representations*.
- We propose AGPST, a unified framework combining data-adaptive graph learning, spatial aggregation, and patch-level masked temporal modeling to jointly capture cross-entity interactions and long-range dynamics.
- Experiments on six real-world benchmarks demonstrate consistent improvements over forecasting and pre-training baselines. AGPST also benefits recurrent, convolutional, linear, and Transformer predictors, validating its effectiveness and architectural generality.

2 Related Work

Spatiotemporal Series Forecasting Spatiotemporal forecasting predicts future system states by jointly modeling temporal dynamics and dependencies among spatial entities (Gao et al., 2024). Early recurrent models, including LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Chung et al., 2014), capture sequential regularities but omit explicit spatial interactions. Graph-based methods subsequently introduce message passing to encode structured dependencies (Li et al., 2018; Guo et al., 2019). Later studies improve expressiveness through learnable graphs, spectral or differential modeling (Yu et al., 2018; Fang et al., 2021; Han et al., 2021), and Transformer architectures (Xu et al., 2020; Liu et al., 2023b). However, most methods learn spatial relations inside task-specific predictors under short-window supervision, binding relational knowledge to a particular architecture

and objective. AGPST instead learns reusable interaction-aware representations from extended historical sequences before downstream forecasting.

Pre-training for Spatiotemporal Representation Masked pre-training is a general approach to transferable representation learning. BERT-style models recover masked tokens from context (Devlin et al., 2019; Lan et al., 2020), while BEiT and MAE reconstruct image patches to capture high-level structure (Bao et al., 2022; He et al., 2022). This paradigm has been extended to spatiotemporal data through STEP, GPT-ST, and STD-MAE (Shao et al., 2022; Li et al., 2023; Gao et al., 2024). Nevertheless, existing objectives remain predominantly *value-centric*: they recover missing observations while spatial relations are predefined, implicit, or separated from long-range temporal learning. Consequently, they emphasize what values should be reconstructed rather than how correlated entities jointly explain them. AGPST combines data-adaptive graph learning with patch-level masked modeling to encode cross-entity interactions and long-range temporal context in representations transferable across heterogeneous forecasting architectures.

3 PRELIMINARIES

Problem Setup. Let $\mathbf{X} \in \mathbb{R}^{N\times T\times F}$ denote a spatiotemporal sequence collected from N spatial entities over T time steps, where each observation has F features. Given a historical window $\mathbf{X}^H \in \mathbb{R}^{N\times T_h\times F}$, the forecasting task predicts the future sequence $\mathbf{Y} \in \mathbb{R}^{N\times T_f\times C}$ over the next T_f time steps.

Masked Pre-training. AGPST first learns reusable representations from a longer temporal segment $\mathbf{X}^L \in \mathbb{R}^{N\times k\times F}$. A binary mask $\mathbf{M} \in \{0,1\}^{N\times k\times F}$ is applied to obtain the visible input $\widetilde{\mathbf{X}}^L = \mathbf{M} \odot \mathbf{X}^L$, where zeros indicate masked entries. Let E_θ and D_ψ denote the pre-training encoder and decoder, respectively. The reconstruction objective is

$$\widehat{\mathbf{X}}^L = D_\psi\Big(E_\theta(\widetilde{\mathbf{X}}^L)\Big), \qquad \mathcal{L}_{\text{pre}} = \frac{1}{|\Omega|}\sum_{(i,j,c)\in\Omega} \ell_{\text{rec}}\left(\widehat{X}^L_{ijc}, X^L_{ijc}\right), \tag{1}$$

where $\Omega = \{(i,j,c) \mid M_{ijc} = 0\}$ denotes the masked positions.

Forecasting Transfer. The pre-trained encoder extracts a long-context representation $\mathbf{Z} = E_\theta(\mathbf{X}^L)$, which is combined with the recent history by a downstream predictor G_ϕ:

$$\widehat{\mathbf{Y}} = G_\phi\big(\mathbf{X}^H, \mathbf{Z}\big), \qquad \mathcal{L}_{\text{fore}} = \frac{1}{NT_fC}\sum_{i=1}^{N}\sum_{j=1}^{T_f}\sum_{c=1}^{C} \ell_{\text{fore}}\left(\widehat{Y}_{ijc}, Y_{ijc}\right). \tag{2}$$

This formulation separates long-context representation learning from task-specific forecasting.

4 METHODOLOGY

As illustrated in Figure 1, AGPST is built upon a central principle: long-context pre-training should capture not only temporal regularities, but also the cross-entity dependencies through which spatiotemporal evidence is organized. AGPST therefore separates reusable representation learning from task-specific forecasting through three coordinated components. First, an *Adaptive Graph Neural Network* learns a data-supported interaction operator and aggregates information across spatial entities. Second, an *Interaction-Aware Masked Pre-training Module* integrates these relational features with long-range temporal context by reconstructing masked temporal patches. Finally, a *Forecasting Integration Module* transfers the pre-trained representations to heterogeneous downstream predictors without modifying their internal architectures. This design allows long-term relational-temporal knowledge to complement the recent observations used by task-specific forecasting models.

4.1 ADAPTIVE GRAPH NEURAL NETWORK

Spatial entities in real-world systems interact through both physical connectivity and statistical dependence. Although a predefined graph can encode useful domain knowledge, its edges may not fully reflect the dependencies expressed in historical observations, particularly under demand

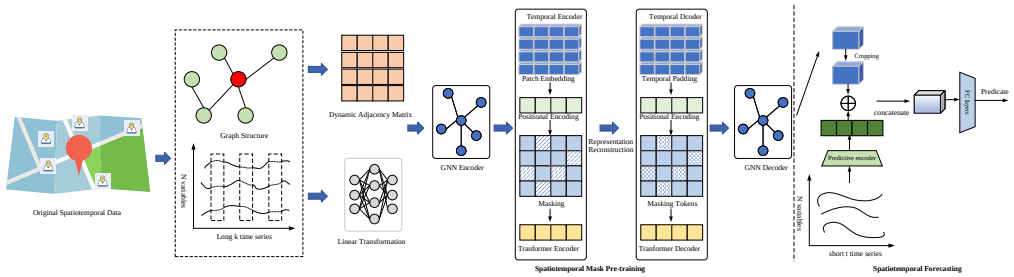


Figure 1: Overview of AGPST. The adaptive graph encoder learns data-supported interactions across spatial entities, while masked temporal pre-training captures long-range dependencies from extended historical sequences. The resulting interaction-aware representations are transferred to downstream forecasting models through lightweight feature fusion.

variation, seasonal effects, or system reconfiguration (Shen et al., 2023; Liu et al., 2023a; Cachay et al., 2023). We therefore introduce an adaptive graph neural network, termed AGNN, to learn a relational operator directly from data and use it to organize cross-entity information during pre-training.

Adaptive Interaction Modeling. Let $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{N \times d}$ denote two trainable node embedding matrices, where d is the embedding dimension. The data-adaptive interaction matrix is constructed as

$$\widehat{\mathbf{A}} = \mathrm{SoftMax}\left(\mathrm{ReLU}\left(\mathbf{E}_1 \mathbf{E}_2^\top\right)\right), \tag{3}$$

where $\mathrm{SoftMax}(\cdot)$ is applied row-wise. Consequently, each row of $\widehat{\mathbf{A}}$ defines a normalized distribution over potential information sources, and $\widehat{A}_{ij}$ measures the learned contribution of node j to node i. Unlike a hand-crafted graph that remains independent of the learning objective, $\widehat{\mathbf{A}}$ is jointly optimized with the pre-training model. It therefore provides a data-supported relational prior that complements, rather than merely reproduces, predefined physical connectivity.

The asymmetric parameterization using $\mathbf{E}_1 \mathbf{E}_2^\top$ also allows the influence from node j to node i to differ from that in the reverse direction. This property is suitable for spatiotemporal systems in which information propagation can be directional, such as traffic diffusion across road segments or load transmission across connected regions.

Interaction-Aware Aggregation. Given node representations $\mathbf{H}_p^{(\ell)} \in \mathbb{R}^{N \times D}$ for temporal patch p at graph layer ℓ, AGNN performs a GIN-inspired update (Xu et al., 2019):

$$\mathbf{H}_p^{(\ell+1)} = h_{\mathbf{\Theta}}^{(\ell)}\left(\left[\widehat{\mathbf{A}} + (1+\epsilon)\mathbf{I}\right]\mathbf{H}_p^{(\ell)}\right), \tag{4}$$

where $h_{\mathbf{\Theta}}^{(\ell)}(\cdot)$ is an MLP, $\mathbf{I}$ is the identity matrix, and ϵ controls the contribution of node-specific information. The term $\widehat{\mathbf{A}}\mathbf{H}_p^{(\ell)}$ aggregates evidence from data-supported neighbors, whereas $(1+\epsilon)\mathbf{I}\mathbf{H}_p^{(\ell)}$ preserves the identity and local state of each node. Their combination prevents relational aggregation from overwhelming node-specific dynamics.

After L_g graph layers, the spatial representation of patch p is obtained as

$$\mathbf{S}_p = \mathrm{AGNN}\left(\mathbf{H}_p^{(0)}; \widehat{\mathbf{A}}\right) = \mathbf{H}_p^{(L_g)}. \tag{5}$$

Rather than treating the graph as an auxiliary input, AGNN uses it as an explicit information-organization mechanism. The resulting representation records not only the state of each entity, but also the relational evidence collected from other entities before temporal modeling.

4.2 Interaction-Aware Masked Pre-training

Adaptive graph aggregation provides relational context, but short-window supervised forecasting alone may not be sufficient to internalize it into reusable representations. Forecasting objectives

are usually optimized using recent observations and relatively short prediction horizons, which can encourage models to exploit local continuity while underusing long-range dependencies (Meng et al., 2022; Baggag et al., 2021). AGPST therefore pre-trains the interaction encoder on extended historical segments through patch-level masked reconstruction.

Patch-Based Spatiotemporal Encoding. Given a long historical segment $\mathbf{X}^L \in \mathbb{R}^{N\times k\times F}$, we divide its temporal dimension into $P = k/L$ non-overlapping patches of length L. A convolutional projection maps the raw observations to patch-level embeddings:

$$\mathbf{U} = \mathrm{Conv2D}\left(\mathbf{X}^L\right) \in \mathbb{R}^{N\times P\times D}, \tag{6}$$

where $\mathbf{U}_p \in \mathbb{R}^{N\times D}$ denotes the representation of the p-th temporal patch.

Each patch is first processed by the adaptive graph encoder:

$$\mathbf{S}_p = \mathrm{AGNN}\left(\mathbf{U}_p; \widehat{\mathbf{A}}\right). \tag{7}$$

A learnable positional embedding $\mathbf{P}_p$ is then added to preserve temporal order:

$$\mathbf{Z}_p = \mathbf{S}_p + \mathbf{P}_p, \qquad p = 1, \ldots, P. \tag{8}$$

The resulting token $\mathbf{Z}_p$ contains both cross-entity relational context and patch-level temporal information. Compared with processing all k observations individually, patching compresses the temporal axis to P tokens, making extended historical contexts computationally manageable while preserving segment-level dynamics.

Masked Temporal Encoding. Let $\mathbf{M} \in \{0,1\}^{N\times P}$ be a binary patch mask with masking ratio r, where $M_{ip} = 1$ indicates that patch p of node i is visible. Rather than providing all tokens to the encoder, we retain only the visible subset:

$$\mathbf{Z}_{\mathrm{vis}} = \{\mathbf{Z}_{ip} \mid M_{ip} = 1\}. \tag{9}$$

The visible tokens are processed by a Transformer encoder:

$$\mathbf{H}_{\mathrm{vis}} = f_{\mathrm{enc}}\left(\mathbf{Z}_{\mathrm{vis}}\right), \tag{10}$$

which models temporal dependencies among the observed patches. Because masking removes complete temporal segments, the encoder cannot recover missing information by directly copying the corresponding local observations.

For reconstruction, the encoded visible tokens are restored to their original positions, while masked positions are filled with a shared learnable token $\mathbf{e}_{\mathrm{mask}}$. After adding decoder positional embeddings $\mathbf{P}^d$, the decoder input is

$$\mathbf{Q} = \mathrm{Restore}\left(\mathbf{H}_{\mathrm{vis}}, \mathbf{M}, \mathbf{e}_{\mathrm{mask}}\right) + \mathbf{P}^d. \tag{11}$$

A lightweight Transformer decoder then predicts the representations of all patches:

$$\widehat{\mathbf{Z}} = f_{\mathrm{dec}}\left(\mathbf{Q}\right). \tag{12}$$

Masked Reconstruction Objective. Let

$$\Omega = \{(i,p) \mid M_{ip} = 0\}$$

denote the set of masked patch positions. The pre-training objective is computed only over these positions:

$$\mathcal{L}_{\mathrm{pre}} = \frac{1}{|\Omega|} \sum_{(i,p)\in\Omega} \left\| \widehat{\mathbf{Z}}_{ip} - \mathbf{Z}_{ip} \right\|_2^2. \tag{13}$$

Restricting the loss to masked patches prevents visible tokens from dominating optimization and directs the model toward context-based inference. More importantly, the target token $\mathbf{Z}_{ip}$ has already incorporated graph-based relational evidence through Equation equation 7. The reconstruction task therefore does not learn temporal regularities in isolation; it learns to recover relationally enriched temporal states from incomplete observations.

4.3 FORECASTING INTEGRATION MODULE

The objective of AGPST is to complement existing forecasting models rather than replace them with another task-specific architecture. During downstream forecasting, the pre-trained encoder supplies long-context relational-temporal knowledge, while the original predictor continues to model the recent system state.

Long-Context Representation Extraction. Given an extended historical sequence $\mathbf{X}^L$ of length T_{long}, the pre-trained encoder produces

$$\mathbf{Z}^L = E_\theta\left(\mathbf{X}^L\right) \in \mathbb{R}^{N\times P\times D}. \tag{14}$$

We retain the representations associated with the most recent P_r patches, since they are temporally aligned with the input window of the downstream predictor. These representations are reshaped into

$$\mathbf{Z}' = \text{Reshape}\left(\mathbf{Z}^L_{:,\,P-P_r+1:P,\,:}\right) \in \mathbb{R}^{N\times P_r D}. \tag{15}$$

Integration with Downstream Predictors. We use Graph WaveNet (GWNet) (Wu et al., 2019) as the primary downstream predictor. Let

$$\mathbf{H}_\phi = G^{\text{enc}}_\phi\left(\mathbf{X}^H\right) \tag{16}$$

denote the hidden representation generated by the downstream encoder from the recent historical window $\mathbf{X}^H \in \mathbb{R}^{N\times T_h\times F}$. The short-context representation $\mathbf{H}_\phi$ captures the current forecasting state, whereas $\mathbf{Z}'$ provides relational-temporal knowledge learned from a substantially longer history.

We concatenate the two representations and apply a lightweight fusion network:

$$\widetilde{\mathbf{H}} = \text{MLP}_{\text{fuse}}\left([\mathbf{H}_\phi \| \mathbf{Z}']\right), \tag{17}$$

where $\|$ denotes feature concatenation. The original output module of the downstream predictor then produces the future sequence:

$$\widehat{\mathbf{Y}} = G^{\text{out}}_\phi\left(\widetilde{\mathbf{H}}\right) \in \mathbb{R}^{N\times T_f\times C}. \tag{18}$$

The downstream model is optimized using mean absolute error:

$$\mathcal{L}_{\text{fore}} = \frac{1}{NT_fC}\sum_{i=1}^{N}\sum_{t=1}^{T_f}\sum_{c=1}^{C}\left|\widehat{Y}_{itc} - Y_{itc}\right|. \tag{19}$$

The same representation-level interface can be applied to predictors with different inductive biases. In addition to GWNet, we instantiate AGPST with DCRNN, MTGNN, STID, and STAEformer in our experiments. This evaluation examines whether the gain originates from reusable pre-trained knowledge rather than a favorable coupling with one particular forecasting architecture.

Computational Complexity. Constructing the adaptive interaction matrix in Equation equation 3 requires $O(N^2d)$ operations. Applying graph aggregation to P temporal patches costs $O(PN^2D)$ per graph layer under dense interaction modeling. Let $P_{\text{vis}} = (1-r)P$ denote the expected number of visible patches per node. When masked tokens are removed before temporal encoding, Transformer self-attention requires $O(NP_{\text{vis}}^2D)$ operations; otherwise, the conservative upper bound is $O(NP^2D)$. Ignoring layer numbers and constant factors, the pre-training complexity is therefore

$$O\left(N^2d + PN^2D + NP_{\text{vis}}^2D\right). \tag{20}$$

Since $P = k/L$, patching substantially reduces the temporal sequence length relative to processing all k observations directly. Pre-training optimization is performed once and reused across downstream models. At forecasting time, AGPST introduces one long-context encoding pass and a lightweight fusion network, while leaving the internal computation of the downstream predictor unchanged.

Table 1: Forecasting performance on four PeMS traffic-flow datasets. Lower is better. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	PEMS03			PEMS04			PEMS07			PEMS08		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Transformer	17.50	30.24	16.80	23.83	37.19	15.57	26.80	42.95	12.11	18.52	28.68	13.66
DCRNN	18.18	30.31	18.91	19.63	31.24	13.52	25.30	38.58	11.66	15.21	24.11	10.04
STGCN	17.49	30.12	17.15	22.70	35.55	14.59	25.38	38.78	11.08	18.02	27.83	11.40
ASTGCN	17.69	29.66	19.40	22.93	35.22	16.56	28.05	42.57	13.92	18.61	28.16	13.08
GWNet	19.85	32.94	19.31	18.74	30.32	13.10	26.85	42.78	12.12	14.55	23.53	9.31
STSGCN	17.48	29.21	16.78	21.19	33.65	13.90	24.26	39.03	10.21	17.13	26.80	10.96
STFGNN	16.77	28.34	16.30	19.83	31.88	13.02	22.07	35.80	9.21	16.64	26.22	10.60
STGODE	16.50	27.84	16.69	20.84	32.82	13.77	22.99	37.54	10.14	16.81	25.97	10.62
PDFormer	14.94	25.39	15.82	18.32	29.97	12.10	19.83	32.87	8.53	13.58	23.51	9.05
STAEformer	15.35	27.55	15.18	18.22	30.18	11.98	19.14	32.60	8.01	13.46	23.25	8.88
RAST	15.20	25.81	16.12	18.39	29.93	14.43	19.52	32.73	8.23	14.16	23.33	9.27
STEP	14.22	24.55	14.42	18.20	29.71	12.48	19.32	32.19	8.12	14.00	23.41	9.50
STD-MAE	13.80	24.43	13.96	17.80	29.25	11.97	18.65	31.44	7.84	13.44	22.47	8.76
AGPST	**13.61**	**23.91**	**13.10**	**16.73**	**28.39**	**11.10**	**17.78**	**30.78**	**7.24**	**12.97**	**22.13**	**7.91**

Table 2: Forecasting performance on the traffic-speed datasets. Each entry reports MAE/RMSE/MAPE. Lower is better.

Dataset	Horizon	GWNet	STEP	PDFormer	STD-MAE	AGPST
PEMS-BAY	3	1.30/2.73/2.71	1.26/2.73/2.59	1.23/2.62/2.56	1.32/2.83/2.78	**1.15/2.58/2.50**
	6	1.63/3.73/3.73	1.55/3.58/3.43	1.53/3.53/3.42	1.64/3.79/3.71	**1.47/3.49/3.37**
	12	1.99/4.60/4.71	1.79/4.20/4.18	1.77/4.20/4.17	1.91/4.43/4.51	**1.71/4.11/4.05**
METR-LA	3	2.69/5.15/6.99	2.61/4.98/6.60	2.83/5.45/7.77	2.62/5.02/6.70	**2.59/4.90/6.55**
	6	3.08/6.20/8.47	2.96/5.97/7.96	3.20/6.46/9.19	2.99/6.07/8.04	**2.92/5.88/7.91**
	12	3.51/7.28/9.96	3.37/6.99/9.61	3.62/7.47/10.91	3.40/7.07/9.59	**3.31/6.88/9.53**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and Metrics. We evaluate AGPST on six public traffic forecasting benchmarks. **METR-LA** and **PEMS-BAY** contain traffic-speed measurements collected from road sensors (Jagadish et al., 2014; Cao et al., 2020), whereas **PEMS03**, **PEMS04**, **PEMS07**, and **PEMS08** contain traffic-flow observations from the California Performance Measurement System (Song et al., 2020). Dataset statistics are provided in Appendix B. We evaluate prediction accuracy using mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE), where lower values indicate better performance. Additional details are reported in **Appendix** A.

Baselines. We compare AGPST with representative methods covering recurrent and convolutional forecasting, spatiotemporal graph learning, Transformer-based modeling, and spatiotemporal pre-training. The baselines include DCRNN (Li et al., 2018), STGCN (Yu et al., 2018), ASTGCN (Guo et al., 2019), GWNet (Wu et al., 2019), STSGCN (Song et al., 2020), STFGNN (Li & Zhu, 2021), STGODE (Fang et al., 2021), PDFormer (Jiang et al., 2023), STAEformer (Liu et al., 2023b), STEP (Shao et al., 2022), STD-MAE (Gao et al., 2024), and RAST (Ruan et al., 2026).

Implementation Details. We split PEMS03, PEMS04, PEMS07, and PEMS08 into training, validation, and test sets using a 6:2:2 ratio, and adopt a 7:1:2 split for METR-LA and PEMS-BAY, following prior work (Li et al., 2018; Song et al., 2020; Fang et al., 2021). The forecasting horizon is fixed to 12. AGPST uses two graph layers, four Transformer encoder layers, one Transformer decoder layer, and four attention heads. The graph and temporal hidden dimensions are 96 and 32, respectively. Unless otherwise specified, the patch length and masking ratio are set to $L = 12$ and $r = 0.25$. Models are optimized using MAE. All experiments are implemented in PyTorch 1.9.0 with Python 3.9 and conducted on an NVIDIA RTX 4090 GPU with 24GB memory.

5.2 Overall Performance

Tables 1 and 2 compare AGPST with representative baselines on traffic-flow and traffic-speed forecasting benchmarks across different graph scales, temporal patterns, and prediction horizons. The results support three main observations: 1) ***AGPST consistently improves performance across datasets and metrics***, achieving the best overall results on all four traffic-flow benchmarks under MAE, RMSE, and MAPE. The agreement across these complementary metrics suggests that the gains are not driven by a particular error measure, but reflect more accurate and stable representations across datasets with different graph sizes and temporal dynamics; 2) ***the improvements remain stable across forecasting horizons***, with consistent advantages on both traffic-speed datasets from short- to long-term prediction. This trend is particularly meaningful at longer horizons, where recent observations become less informative and accumulated uncertainty increases, indicating that the long-context representations learned during pre-training provide additional evidence beyond local temporal continuity; and 3) ***AGPST outperforms both task-specific and pre-training baselines***, suggesting that its advantage does not arise solely from a stronger downstream predictor or longer historical context. Instead, jointly learning data-adaptive cross-entity interactions and long-range temporal structure enables relational and temporal evidence to be encoded into reusable representations that complement heterogeneous forecasting models.

5.3 Component Analysis

To assess the contribution of each component, we construct three variants: **w/o Adaptive Graph** replaces the learned interaction matrix with a fixed graph, **w/o GNN** replaces graph aggregation with an MLP, and **w/o Masking** removes patch masking during pre-training.

Table 3: Framework ablation. Each entry reports MAE/RMSE/MAPE.

Variant	PEMS03	PEMS04	PEMS07	PEMS08
AGPST	**13.61/23.91/13.10**	**16.73/28.39/11.10**	**17.78/30.78/7.24**	**12.97/22.13/7.91**
w/o Adaptive Graph	14.98/25.11/14.87	18.21/29.91/13.17	19.11/32.11/8.99	14.10/23.87/9.25
w/o GNN	15.11/26.72/15.31	19.38/30.81/15.93	18.90/33.85/10.48	14.12/24.72/10.38
w/o Masking	15.62/25.87/15.91	18.39/30.93/14.87	18.79/33.28/10.17	14.37/24.81/10.31

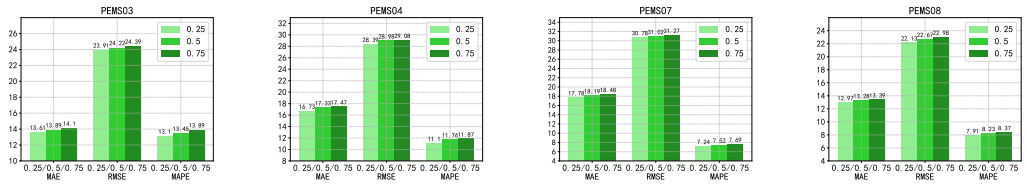
As shown in Table 3, removing any component consistently increases all three errors across the four datasets. Replacing the learned interaction matrix with a fixed graph degrades performance despite retaining graph propagation, showing that predefined topology cannot fully capture the dependencies expressed in historical observations. The decline of **w/o GNN** further indicates that learning interaction weights alone is insufficient. Graph aggregation is required to transform these relations into node-level contextual evidence: adaptive interaction learning determines *where* useful information originates, while graph propagation determines *how* it is integrated. Removing masking also reduces accuracy because complete inputs allow the encoder to rely on local continuity. Patch masking instead requires missing states to be inferred from long-range temporal context and correlated entities. Overall, the three components form a coherent pipeline: adaptive interaction learning identifies informative dependencies, graph aggregation organizes cross-entity evidence, and masked reconstruction internalizes it into reusable representations.

5.4 Efficiency Analysis

Table 4: Measured pre-training runtime (ms) under identical hardware settings. Lower is better.

Dataset	#Nodes	STEP	STD-MAE	AGPST
PEMS03	358	108	50	**31**
PEMS04	307	73	34	**27**
PEMS07	883	516	142	**99**
PEMS08	170	62	48	**35**

As shown in Table 4, AGPST achieves the lowest measured pre-training runtime across all four datasets under identical hardware settings. The advantage remains evident on PEMS07, the largest graph considered, suggesting that adaptive interaction encoding does not dominate the computational cost as graph scale increases. This efficiency mainly comes from patching long sequences into shorter token streams and restricting temporal encoding to visible patches, while the globally learned interaction matrix avoids repeated graph construc-



(a) PEMS03 (b) PEMS04 (c) PEMS07 (d) PEMS08

Figure 2: Effect of the masking ratio r on forecasting MAE/RMSE/MAPE. Lower is better.

tion at each time step. Overall, AGPST improves representation quality while maintaining practical efficiency; however, these results reflect empirical runtime under the reported implementation and hardware rather than a universal asymptotic advantage.

5.5 Hyperparameter Analysis

We examine the sensitivity of AGPST to the masking ratio r and the pre-training sequence length k on PEMS03, PEMS04, PEMS07, and PEMS08. The forecasting horizon is fixed to 12, while all remaining settings are kept unchanged to isolate the effect of each hyperparameter.

Masking Ratio. As shown in Figure 2, forecasting performance improves with moderate masking but degrades when the masking ratio becomes too large. With insufficient masking, reconstruction remains easy because the encoder can rely on nearby observations and short-range continuity, limiting the use of broader temporal context and cross-entity dependencies. Increasing r strengthens contextual inference and encourages recovery from information distributed across time and interacting entities. In contrast, excessive masking removes too much observable evidence, increasing reconstruction ambiguity and weakening representation reliability. The best performance at $r = 0.25$ therefore reflects a balance between reconstruction difficulty and contextual sufficiency. The smooth degradation around this setting further suggests that AGPST is not highly sensitive to moderate changes in the masking ratio.

Pre-training Length. Increasing the pre-training sequence length k initially improves forecasting because a broader historical window exposes the encoder to longer-range temporal regularities and a richer set of cross-entity co-evolution patterns. This additional context is particularly useful for learning reusable representations that cannot be fully inferred from the short input windows used by downstream predictors. As k continues to increase, however, the marginal benefit diminishes: distant observations become increasingly redundant or weakly relevant to the current forecasting state, while longer sequences also increase computational cost. The resulting trend therefore suggests that effective pre-training depends on providing sufficient historical coverage without introducing excessive redundancy. Detailed results are reported in **Appendix** B.

Additional Analyses. AGPST consistently improves recurrent, convolutional, linear, and Transformer-based predictors, and remains stronger than STD-MAE-enhanced variants, indicating that its gains are not solely due to longer historical context. Qualitative reconstruction and forecasting examples further illustrate the behavior of the learned relational-temporal representations. Full results are provided in **Appendix** C and **Appendix** D.

6 Conclusion

We proposed **AGPST**, an interaction-aware pre-training framework for spatiotemporal forecasting. By combining data-adaptive graph aggregation with masked temporal modeling, AGPST jointly captures cross-entity dependencies and long-range context. Experiments on six benchmarks demonstrate consistent gains across datasets, metrics, and horizons, while ablation and compatibility studies confirm the complementary components and cross-architecture reusability of the learned representations. These results highlight the importance of jointly learning temporal regularities and relational organization for effective spatiotemporal pre-training.

Reproducibility Statement

Implementation details, dataset splits, evaluation protocols, and training configurations are provided in Section 5 and the appendix. Additional dataset statistics, hyperparameter analyses, runtime evaluations, cross-architecture compatibility results, and qualitative reconstruction and forecasting analyses are reported in the appendix. These materials are intended to facilitate independent reproduction of the reported experiments.

AI Use Statement

Generative AI tools were used solely for language polishing and improving the clarity and readability of the manuscript. All problem formulations, methodological designs, experimental settings, analyses, and reported results were developed, conducted, and verified by the authors.

References

Abdelkader Baggag, Sofiane Abbar, Ankit Sharma, Tahar Zanouda, Abdulaziz Al-Homaid, Abhiraj Mohan, and Jaideep Srivastava. Learning spatiotemporal latent factors of traffic via regularized tensor factorization: Imputing missing values and forecasting. *IEEE Trans. Knowl. Data Eng.*, 33(6):2573–2587, 2021. doi: 10.1109/TKDE.2019.2954868. URL https://doi.org/10.1109/TKDE.2019.2954868.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8df90a1440ce782d1f5607b7a38f2531-Abstract-Conference.html.

Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Spectral temporal graph neural network for multivariate time-series forecasting. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow forecasting. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 11463–11471. AAAI Press, 2025. doi: 10.1609/AAAI.V39I11.33247. URL https://doi.org/10.1609/aaai.v39i11.33247.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014. URL http://arxiv.org/abs/1412.3555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. Spatial-temporal graph ODE networks for traffic flow forecasting. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge*

Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021, pp. 364–373. ACM, 2021.

Haotian Gao, Renhe Jiang, Zheng Dong, Jinliang Deng, Yuxin Ma, and Xuan Song. Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence, IJCAI, Jeju Island, South Korea, August 3-9,*, 2024.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 922–929. AAAI Press, 2019.

Liangzhe Han, Bowen Du, Leilei Sun, Yanjie Fu, Yisheng Lv, and Hui Xiong. Dynamic and multi-faceted spatio-temporal deep learning for traffic speed forecasting. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 547–555. ACM, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 15979–15988. IEEE, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Commun. ACM*, 57(7):86–94, 2014.

Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 4365–4373. AAAI Press, 2023.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Mengzhang Li and Zhanxing Zhu. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 4189–4196. AAAI Press, 2021.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Zhonghang Li, Lianghao Xia, Yong Xu, and Chao Huang. GPT-ST: generative pre-training of spatio-temporal graph neural networks. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Fan Liu, Weijia Zhang, and Hao Liu. Robust spatiotemporal traffic forecasting with reinforced dynamic adversarial training. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos,

Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (eds.), *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pp. 1417–1428. ACM, 2023a. doi: 10.1145/3580305.3599492. URL https://doi.org/10.1145/3580305.3599492.

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 *ACM International Conference on Information and Knowledge Management, CIKM, Birmingham, United Kingdom,*, pp. 4125–4129. ACM, 2023b.

Ivan Marisca, Cesare Alippi, and Filippo Maria Bianchi. Graph-based forecasting with missing data through spatiotemporal downsampling. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=uYIFQOtb58.

Chuizheng Meng, Hao Niu, Guillaume Habault, Roberto Legaspi, Shinya Wada, Chihiro Ono, and Yan Liu. Physics-informed long-sequence forecasting from multi-resolution spatiotemporal data. In Luc De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pp. 2189–2195. ijcai.org, 2022. doi: 10.24963/IJCAI.2022/304. URL https://doi.org/10.24963/ijcai.2022/304.

Weilin Ruan, Xilin Dang, Ziyu Zhou, Sisuo Lyu, and Yuxuan Liang. A retrieval augmented spatio-temporal framework for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46):39163–39172, 2026. doi: 10.1609/aaai.v40i46.41264.

Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *KDD ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1567–1577. ACM, 2022.

Tong Shen, Yang Li, and José M. F. Moura. Forecasting COVID-19 dynamics: Clustering, generalized spatiotemporal attention, and impacts of mobility and geographic proximity. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, pp. 2892–2904. IEEE, 2023. doi: 10.1109/ICDE55515.2023.00221. URL https://doi.org/10.1109/ICDE55515.2023.00221.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 914–921. AAAI Press, 2020.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 1907–1913. ijcai.org, 2019.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *CoRR*, abs/2001.02908, 2020.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 3634–3640. ijcai.org, 2018.

Yudong Zhang, Xu Wang, Xuan Yu, Zhaoyang Sun, Kai Wang, and Yang Wang. Drawing informative gradients from sources: A one-stage transfer learning framework for cross-city spatiotemporal

forecasting. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 1147–1155. AAAI Press, 2025. doi: 10.1609/AAAI.V39I1.32102. URL https://doi.org/10.1609/aaai.v39i1.32102.

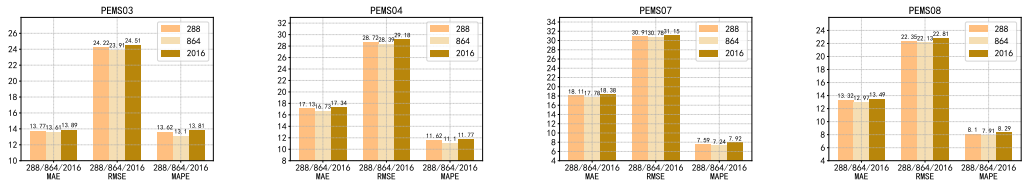
A DATASETS

Table 5: Statistics of datasets.

Datasets	#Sensors	#Time Steps	#Time Interval
PEMS03	358	5min	26208
PEMS04	307	5min	16992
PEMS07	883	5min	28224
PEMS08	170	5min	17856
METR-LA	207	5min	34272
PEMS-BAY	325	5min	52116

We evaluate AGPST on six widely used public traffic benchmarks covering both traffic speed and traffic flow forecasting. **METR-LA** records traffic speeds measured by loop detectors deployed on the Los Angeles County road network (Jagadish et al., 2014). **PEMS-BAY** is a large-scale traffic speed dataset collected from the California Performance Measurement System (PeMS) (Cao et al., 2020). In addition, **PEMS03**, **PEMS04**, **PEMS07**, and **PEMS08** are traffic flow datasets obtained from the same PeMS system and have been extensively used as standard benchmarks in spatiotemporal forecasting (Song et al., 2020). The detailed statistics of all datasets are summarized in Table 5.

B EFFECT OF PRE-TRAINING LENGTH (k).



(a) PEMS03 (b) PEMS04 (c) PEMS07 (d) PEMS08

Figure 3: Hyper-parameter study on K.

Figure 3 illustrates the effect of varying the pre-training sequence length. Increasing k from 432 to 864 consistently improves forecasting performance across all datasets, suggesting that short pre-training sequences fail to expose the model to sufficient long-range temporal context and evolving spatial interaction patterns. The best performance is achieved at $k = 864$, which provides a balanced temporal coverage for learning transferable spatiotemporal representations. Further increasing k yields diminishing returns and introduces mild performance fluctuations, likely due to temporal redundancy and increased optimization complexity. Importantly, performance does not degrade sharply as k varies, demonstrating that AGPST is robust to the choice of pre-training length. We therefore adopt $k = 864$ as the default setting.

C COMPATIBILITY WITH DOWNSTREAM PREDICTORS

A desirable pre-trained representation should provide useful information beyond the inductive bias of a particular downstream architecture. We therefore examine whether the benefits of AGPST persist across heterogeneous forecasting models. Specifically, we integrate the same pre-trained representation with five representative backbones: DCRNN for recurrent graph modeling, MTGNN

and GWNet for graph-based temporal convolution, STID for linear forecasting, and STAEformer for Transformer-based prediction. For each backbone, Table 6 compares the original model with its STD-MAE-enhanced and AGPST-enhanced variants under identical downstream settings.

Table 6: Compatibility with downstream predictors. Each entry reports MAE/RMSE/MAPE, where lower values indicate better performance.

Backbone	Representation	PEMS04	PEMS08
DCRNN	Original	19.63/31.24/13.52	15.21/24.11/10.04
	+ STD-MAE	18.65/30.09/13.07	14.50/23.38/9.36
	+ AGPST	**18.01/29.13/12.46**	**13.82/22.31/8.69**
MTGNN	Original	19.17/31.70/13.37	15.18/24.14/10.20
	+ STD-MAE	18.72/31.03/12.72	14.84/23.58/9.58
	+ AGPST	**18.12/30.33/11.96**	**14.12/22.21/8.73**
STID	Original	18.35/29.86/12.50	14.21/23.35/9.32
	+ STD-MAE	17.93/29.43/12.11	13.53/22.60/8.97
	+ AGPST	**17.03/28.29/11.20**	**12.12/21.44/8.21**
STAEformer	Original	18.22/30.18/11.98	13.46/23.25/8.88
	+ STD-MAE	17.92/29.37/12.11	13.30/22.51/8.82
	+ AGPST	**17.12/28.45/11.32**	**12.21/21.78/8.01**
GWNet	Original	18.74/30.32/13.10	14.55/23.53/9.31
	+ STD-MAE	17.80/29.25/12.97	13.44/22.47/8.76
	+ AGPST	**16.73/28.39/11.10**	**12.97/22.13/7.91**

Table 6 reveals two consistent patterns. First, adding AGPST reduces MAE, RMSE, and MAPE for every backbone on both datasets. The gains extend beyond graph-based recurrent and convolutional models to the linear STID and Transformer-based STAEformer, indicating that the learned representation does not depend on a specific message-passing operator or temporal encoder. Instead, it provides complementary relational-temporal information that remains useful under substantially different downstream inductive biases.

Second, AGPST consistently improves upon the corresponding STD-MAE-enhanced variants. This comparison is particularly informative because both approaches provide downstream models with representations learned from extended historical sequences. The additional gains of AGPST therefore cannot be attributed solely to access to longer temporal context. Rather, they are consistent with the hypothesis that explicitly organizing cross-entity information during pre-training contributes relational evidence beyond temporal masked reconstruction alone. This distinction is important for the motivation of AGPST: its benefit arises not merely from enlarging the historical receptive field, but from learning how information across entities can be organized within that extended context.

Overall, these results demonstrate *cross-architecture compatibility*: the same pre-trained representation complements recurrent, convolutional, linear, and Transformer-based predictors without redesigning their internal forecasting mechanisms. We emphasize that this experiment evaluates reuse across architectures within the same dataset and should not be interpreted as evidence of cross-dataset or cross-domain transfer.

D Qualitative Analysis

We complement the quantitative evaluation with qualitative analyses of both masked reconstruction and downstream forecasting. These visualizations are intended to illustrate the behavior of the learned representations rather than serve as independent evidence of overall forecasting performance.

D.1 Masked Reconstruction

Figure 4 compares the original series with the output reconstructed by AGPST during masked pre-training. The reconstructed signal follows the major temporal trends and recurring variations of the original sequence while remaining aligned across different temporal regimes. Importantly, the reconstruction also tracks several relatively rapid changes rather than collapsing to a uniformly smooth trajectory. This behavior suggests that the model does not rely exclusively on short-range continuity when recovering masked observations.

The reconstruction mechanism helps explain this behavior. Because part of the temporal sequence is unavailable to the encoder, missing states must be inferred from the remaining historical evidence



Figure 4: Masked reconstruction on PEMS03 (Sensor@0).

together with information aggregated across related entities. Consequently, successful reconstruction requires the representation to preserve temporal structure while exploiting complementary cross-entity context. The close qualitative agreement in Figure 4 is therefore consistent with the intended role of interaction-aware masked pre-training: learning representations that retain useful relational and temporal information under partial observation. We stress that this visualization is illustrative; the quantitative effect of masking and graph aggregation is evaluated separately through the ablation experiments in the main paper.

D.2 Forecasting Behavior

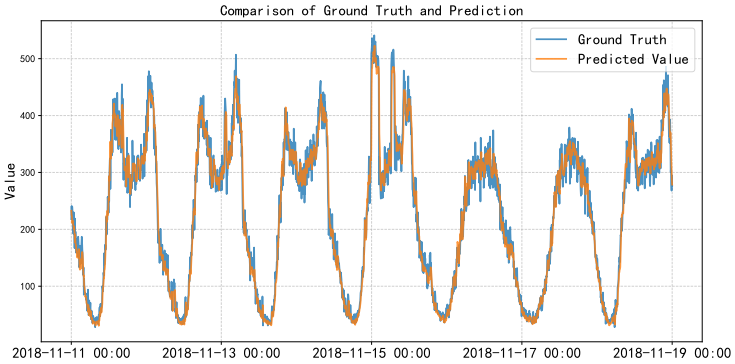


Figure 5: Forecasting results on PEMS03 (Sensor@0).

Figure 5 presents the downstream forecasting trajectory for the same sensor over the displayed evaluation interval. The predictions broadly follow the ground-truth dynamics, capturing recurring fluctuations while also responding to several relatively abrupt changes. Deviations become more visible around extreme peaks, where forecasting is intrinsically more difficult, but the predicted trajectory remains aligned with the overall temporal evolution rather than degenerating into an overly smoothed pattern.

This behavior is consistent with the quantitative results reported in the main paper. The downstream predictor primarily observes the recent forecasting window, whereas AGPST additionally supplies representations learned from a longer relational-temporal context. Such complementary information can be particularly useful when the immediate history alone is insufficient to characterize the subsequent evolution. The visualization should therefore be interpreted as an intuitive illustration of how the pre-trained representation complements local forecasting signals, rather than as standalone evidence of robustness or generalization.

D.3 DISCUSSION

Taken together, the two visualizations provide complementary views of AGPST. The reconstruction example illustrates that the pre-training representation preserves meaningful temporal structure when observations are partially hidden, while the forecasting example shows how this representation can complement recent observations during downstream prediction. These qualitative observations are consistent with the quantitative ablations and cross-architecture results: adaptive interaction learning organizes cross-entity evidence, masked reconstruction encourages inference beyond directly observed values, and the resulting relational-temporal representation can be reused by downstream predictors. Because the examples are drawn from a single sensor, we treat them as interpretive illustrations rather than evidence of general performance; the latter is established by the aggregate results across datasets, metrics, and forecasting horizons.