VaT-BERT: Variational Tokenization and Correlation-Guided BERT for Spatio-Temporal Forecasting

Haoyu Jin, Rongyan Chen, and Xinyu Gu*
Beijing University of Posts and Telecommunications
Beijing, China
Email: {jinhaoyu, 030206cry, guxinyu}@bupt.edu.cn

***Abstract*—Traffic flow forecasting requires modeling temporal dynamics and inter-sensor interactions. Recent work reuses pretrained backbones for spatio-temporal prediction, but serializing sensors as tokens and applying decoder-style causal attention can couple inherently order-free spatial interactions to an arbitrary token order, inducing order bias. We propose VaT-BERT, an adaptation framework that repurposes a pretrained bidirectional encoder, DistilBERT, for traffic forecasting via a node-as-token representation and non-causal self-attention to capture global inter-node dependencies with reduced sensitivity to node order. To improve token quality and stability, we perform variational tokenization by pretraining a Temporal variational autoencoder (VAE) on per-node historical segments and feeding continuous latent tokens to the encoder. We further introduce correlation-prior-guided contrastive regularization to inject long-range correlation structure and stabilize node representations under global attention. Experiments on PEMS03/04/07/08 achieve the best MAE on all four datasets and the best RMSE on three datasets, while remaining competitive in MAPE; ablation and robustness results on PEMS04 further support the contribution of each component.**

***Index Terms*—traffic flow forecasting, spatio-temporal forecasting, pretrained transformers, bidirectional encoder, variational autoencoder, contrastive learning**

I. Introduction

Accurate traffic flow forecasting is a core capability of intelligent transportation systems (ITS), supporting congestion management, route planning, and mobility scheduling. Over the past decade, traffic forecasting models have evolved from purely temporal modeling to explicit spatio-temporal coupling. Spatio-temporal graph neural networks represent sensors as graph nodes and learn spatial dependencies together with temporal dynamics via graph convolution or attention. Representative examples include ASTGCN, which introduces spatio and temporal attention to capture dynamic dependencies [1], and Graph WaveNet, which alleviates the limitation of fixed graph structures through adaptive dependency modeling [2]. Transformer-based approaches (e.g., GMAN [3], STTN [4]) model extended dependencies, with subsequent works focusing on decoupled spatial-temporal relations [5], [6]. Overall, these advances highlight the importance of flexible, global inter-node interaction modeling, and motivate exploring more general-purpose sequence backbones for traffic forecasting.

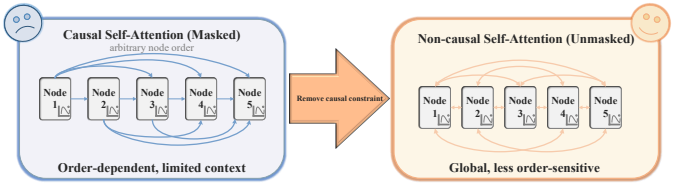

Fig. 1. Motivation on node interaction patterns. Causal (masked) self-attention couples interactions to an arbitrary token order, whereas non-causal self-attention enables global interactions among all node tokens. We instantiate the non-causal backbone with a pretrained bidirectional encoder (DistilBERT). This figure is conceptual and not the model architecture.

Motivated by strong pretrained Transformer representations, recent work reuses such backbones for time-series and traffic forecasting [7]–[9]. We focus on adapting a general-purpose pretrained encoder for traffic forecasting, rather than traffic-domain pretraining [10], [11].

As illustrated in Fig. 1, adapting pretrained Transformers to traffic networks faces three practical issues: (i) spatio tokens are order-free, whereas decoder-only backbones impose a fixed 1-D ordering with causal masks, inducing order-coupled interaction biases; (ii) noisy or anomalous measurements (e.g., sensor drift) can make tokens from simple linear projections unstable [12]; (iii) global self-attention, without structural inductive biases, may overfit spurious correlations or yield overly diffuse dependencies, harming generalization [13].

To bridge pretrained encoders with traffic-network modeling, we propose VaT-BERT, which uses *Node-as-Token* and forms a closed loop of (1) a pretrained *bidirectional* backbone (DistilBERT) for order-robust global inter-node interaction, (2) *Variational Tokenization* via a Temporal

This work was supported in part by the Beijing Natural Science Foundation (L242088), and in part by the National Natural Science Foundation of China (NSFC) under Grant 62201089.

*Corresponding author.

VAE that maps per-node history into a smoother latent space and feeds deterministic latent tokens for stable continuous inputs, and (3) *correlation-prior-guided* contrastive regularization that injects long-range structural cues to stabilize representation geometry under global attention. The VaT-BERT project repository is publicly available at https://github.com/ther7777/VaT-BERT.

II. Methodology

A. Problem setup

We consider a traffic network with N sensors (nodes). At each time step t, we form a node-wise feature matrix

$$\mathbf{X}_t \in \mathbb{R}^{N\times C} \tag{1}$$

where each row corresponds to one node and $C = 3$ includes (i) traffic flow, (ii) normalized time-of-day $s_t \in [0, 1)$, and (iii) day-of-week index $w_t \in \{0, \dots, 6\}$. Let $v_t^{(i)}$ denote the flow value of node i at time t.

Given a history length H and prediction horizon P, the input window is

$$\mathbf{X}_{t-H+1:t} = \big[\mathbf{X}_{t-H+1}, \dots, \mathbf{X}_t\big] \in \mathbb{R}^{H\times N\times C} \tag{2}$$

and the forecasting target is

$$\mathbf{Y}_{t+1:t+P} = \big[\mathbf{v}_{t+1}, \dots, \mathbf{v}_{t+P}\big] \in \mathbb{R}^{P\times N} \tag{3}$$

The model learns $\widehat{\mathbf{Y}}_{t+1:t+P} = f_\Theta(\mathbf{X}_{t-H+1:t})$.

VaT-BERT performs value tokenization using only the flow channel. Specifically, for each node i, we extract its length-H flow segment $\mathbf{v}_{t-H+1:t}^{(i)} = [v_{t-H+1}^{(i)}, \dots, v_t^{(i)}] \in \mathbb{R}^H$. The time-of-day and day-of-week channels are not fed into the value encoder. Instead, we take the time context at the end of the input window (time t) and use (s_t, w_t) to index a periodic embedding table that is shared by all node tokens.

B. VaT-BERT

VaT-BERT adopts Node-as-Token by treating each node as one token. At each forecasting time t, we compress the length-H flow history of every node into a single token, producing a length-N token sequence per sample. This design aligns traffic forecasting with pretrained bidirectional encoders, whose strength lies in modeling global interactions over token sequences.

Each node token is constructed by a two-stream encoder: a value stream that tokenizes the node historical flow segment with a pretrained Temporal VAE encoder, and a temporal stream that injects periodic context via time-of-day and day-of-week embeddings. The fused node-token sequence is then processed by a bidirectional pretrained encoder to model global inter-node interactions. A lightweight prediction head outputs P-step forecasts. As illustrated in Fig. 2, the overall pipeline consists of the two-stream node tokenization, a bidirectional global-interaction backbone, and a lightweight prediction head.

a) Value stream (Temporal VAE tokenizer): For sample b and node i, the Temporal variational autoencoder (VAE) encoder maps the length-H flow segment $\mathbf{v}_{t-H+1:t}^{(b,i)}$ to a diagonal Gaussian posterior

$$\big(\boldsymbol{\mu}_t^{(b,i)}, \log \boldsymbol{\sigma}_t^{2(b,i)}\big) = E_\phi\Big(\mathbf{v}_{t-H+1:t}^{(b,i)}\Big) \tag{4}$$

where $\boldsymbol{\mu}_t^{(b,i)} \in \mathbb{R}^d$ and $\log \boldsymbol{\sigma}_t^{2(b,i)} \in \mathbb{R}^d$, and d is the token (latent) dimension.

In downstream forecasting, we use the posterior mean as a deterministic value token

$$\mathbf{e}_{\mathrm{val},t}^{(b,i)} = \boldsymbol{\mu}_t^{(b,i)} \in \mathbb{R}^d \tag{5}$$

which avoids sampling noise and yields stable, reproducible token representations.

b) Temporal stream (joint periodic embeddings): We learn a joint embedding table

$$\mathbf{T} \in \mathbb{R}^{D_{\mathrm{dow}}\times D_{\mathrm{tod}}\times d} \tag{6}$$

where D_{tod} is the number of time-of-day bins and $D_{\mathrm{dow}} = 7$ is the number of day-of-week categories.

Let $u_t^{(b)} = \big\lfloor s_t^{(b)} \cdot D_{\mathrm{tod}} \big\rfloor$ and $w_t^{(b)} \in \{0, \dots, D_{\mathrm{dow}} - 1\}$. The temporal embedding is

$$\mathbf{e}_{\mathrm{tmp},t}^{(b)} = \mathbf{T}\big[w_t^{(b)}, u_t^{(b)}\big] \in \mathbb{R}^d \tag{7}$$

Since $s_t^{(b)}$ and $w_t^{(b)}$ are shared across nodes at time t, this embedding is duplicated across node tokens within the same sample.

c) Two-stream fusion and bidirectional encoding: For sample b and node i, we fuse two embeddings and project to the DistilBERT hidden size d_{plm}

$$\mathbf{z}_t^{(b,i)} = \mathbf{W}_f\,\mathrm{Concat}\Big(\mathbf{e}_{\mathrm{val},t}^{(b,i)}, \mathbf{e}_{\mathrm{tmp},t}^{(b)}\Big) + \mathbf{b}_f \in \mathbb{R}^{d_{\mathrm{plm}}} \tag{8}$$

Stacking over nodes forms a length-N token sequence that is fed into the DistilBERT encoder PLM$(\cdot)$ via inputs_embeds.

$$\mathbf{H}_t = \mathrm{PLM}(\mathbf{Z}_t) \in \mathbb{R}^{B\times N\times d_{\mathrm{plm}}} \tag{9}$$

Pretrained language models use positional embeddings to provide token identity and break permutation symmetry in self-attention. In our setting, the length-N token sequence corresponds to node indices rather than word order. To provide a consistent identity anchor for each node token, we replace the positional embedding table with a length-N node index embedding table. Unless stated otherwise, the node index embeddings are randomly initialized with the pretrained model initializer and kept frozen in Stage II.

We further adapt DistilBERT with low-rank adaptation (LoRA) modules inserted into the query and value projections of self-attention blocks (q_lin and v_lin), while keeping the original pretrained weights frozen. This preserves pretrained bidirectional interaction modeling and enables lightweight task adaptation.

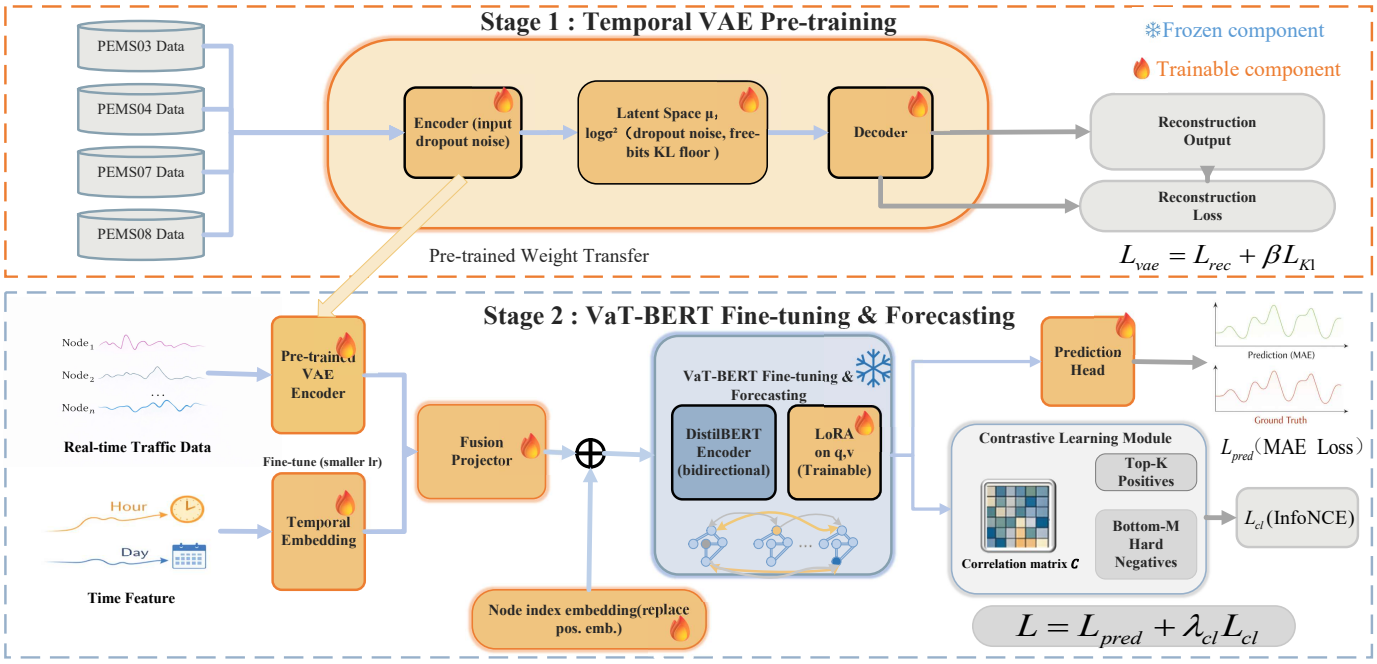

Fig. 2. Overview of VaT-BERT. (Left) Pretraining the Temporal VAE. (Right) Constructing two-stream node tokens and fine-tuning with correlation-guided contrastive regularization.

d) Prediction head: For sample b and node i, a node-wise head g_θ maps $\mathbf{H}_t[b,i]$ to the P-step forecast $\widehat{\mathbf{v}}^{(b,i)}_{t+1:t+P} \in \mathbb{R}^P$.

C. Training objectives

We train VaT-BERT in two stages. Stage 1 pretrains the Temporal VAE on length-H flow segments. Stage 2 fine-tunes the forecasting model end-to-end, including the VAE encoder with a smaller learning rate, while training the fusion projector, temporal embedding table, LoRA modules, and the prediction head.

a) Temporal VAE pretraining: We pretrain the Temporal VAE encoder on length-H flow segments. For a node segment $\mathbf{v} \in \mathbb{R}^H$, the encoder outputs a diagonal Gaussian posterior $q_\phi(\mathbf{z} \mid \mathbf{v}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$. We optimize a reconstruction term together with a variational regularizer to encourage a smooth and robust latent space, using a free-bits-style KL floor to avoid an overly small KL term. During VAE pretraining only, we additionally apply dropout noise to the input segment and to the sampled latent vector to improve robustness; these noises are disabled in downstream forecasting.

The VAE objective is

$$\mathcal{L}_{\text{vae}} = \mathcal{L}_{\text{rec}} + \beta \cdot \mathcal{L}_{\text{kl}} \tag{10}$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction term and $\mathcal{L}_{\text{kl}}$ is a KL regularizer with floor λ_{kl}. Optimization details are given in Sec. III-A.

b) Supervised forecasting loss: We define the forecasting loss on the inverse-transformed physical scale. Let $\mathcal{S}$ be the StandardScaler fitted on the training split of the flow channel and $\mathcal{S}^{-1}$ its inverse. We use mean absolute error (MAE) as the supervised objective, where $\widehat{v}^{(b)}_{t+p,i}$ and $v^{(b)}_{t+p,i}$ denote the predicted and ground-truth flows, respectively.

$$\mathcal{L}_{\text{pred}} = \frac{1}{BPN} \sum_{b=1}^{B} \sum_{p=1}^{P} \sum_{i=1}^{N} \left| \mathcal{S}^{-1}\left(\widehat{v}^{(b)}_{t+p,i}\right) - v^{(b)}_{t+p,i} \right| \tag{11}$$

c) Correlation-prior-guided contrastive regularization: We add a correlation-prior-guided contrastive learning (CL) regularizer $\mathcal{L}_{\text{cl}}$ weighted by λ_{cl}. We precompute a Pearson correlation matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ using the training split only. For node i, we define positives as the top-K correlated nodes excluding itself and negatives as the bottom-M correlated nodes excluding itself

$$\mathcal{P}(i) = \mathrm{TopK}\big(\mathbf{C}_{i,:} \backslash \{i\}\big) \qquad \mathcal{N}_{\text{neg}}(i) = \mathrm{BottomM}\big(\mathbf{C}_{i,:} \backslash \{i\}\big) \tag{12}$$

Given contextualized embeddings $\mathbf{H}_t \in \mathbb{R}^{B \times N \times d_{\text{plm}}}$, define $\mathbf{u}^{(b,i)}_t = \mathbf{H}_t[b,i]$ as the representation of node i in sample b. We use dot-product similarity $s^{(b)}_{i,j} = \langle \mathbf{u}^{(b,i)}_t, \mathbf{u}^{(b,j)}_t \rangle$. For each anchor i, we aggregate multi-positives by the mean embedding

$$\bar{\mathbf{u}}^{(b,i)}_t = \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \mathbf{u}^{(b,j)}_t \in \mathbb{R}^{d_{\text{plm}}} \tag{13}$$

and define $s^{(b)}_{i,\bar{i}} = \langle \mathbf{u}^{(b,i)}_t, \bar{\mathbf{u}}^{(b,i)}_t \rangle$. This objective increases the relative similarity between each anchor and its correlated positives while decreasing similarity to hard negatives, enforcing a correlation-consistent geometry in

TABLE I
TRAFFIC FORECASTING RESULTS ON FOUR BENCHMARKS (MAE / RMSE / MAPE(%)), AVERAGED OVER ALL HORIZONS.

Model	PEMS03			PEMS04			PEMS07			PEMS08		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Conventional baselines												
LSTM	20.62	33.54	28.94	26.81	40.74	22.33	29.71	45.32	14.14	22.19	33.59	18.74
ASTGCN	17.85	29.88	17.65	22.42	34.75	15.87	25.98	39.65	11.84	18.86	28.55	12.50
AGCRN	15.98	28.25	15.23	19.83	32.26	12.97	22.37	36.55	9.12	15.95	25.22	10.09
SSTBAN	15.90	26.11	17.27	18.88	31.10	12.56	20.17	33.45	8.93	14.39	24.19	10.06
PDFormer	14.74	25.59	15.35	18.31	29.97	12.10	19.83	32.87	8.53	13.58	23.51	9.05
iTransformer	19.48	31.20	17.84	22.56	35.21	16.29	24.70	37.91	11.40	20.05	31.90	11.99
PLM/LLM-based baselines												
OFA	20.96	33.43	19.11	27.37	42.99	17.97	30.53	47.51	12.98	21.89	34.63	13.30
STGLLM	15.26	**24.11**	15.73	20.00	32.11	13.69	21.98	35.02	9.72	15.53	24.74	10.15
ST-LLM	17.25	27.25	22.96	19.00	30.35	13.55	21.48	34.07	10.20	14.67	23.50	10.63
STD-PLM	14.59	25.36	14.92	18.16	30.21	**11.89**	19.25	32.84	**8.06**	13.31	23.19	8.84
VaT-BERT (ours)	**14.41**	25.79	**14.54**	**17.96**	**29.56**	12.05	**19.13**	**32.48**	8.08	**13.29**	**23.01**	**8.80**

the representation space. We minimize an InfoNCE (contrastive) loss with hard negatives $\mathcal{N}_{\text{neg}}(i)$

$$Z_i^{(b)} = \exp\Big(s_{i,\tilde{i}}^{(b)}/\tau\Big) + \sum_{k\in\mathcal{N}_{\text{neg}}(i)} \exp\Big(s_{i,k}^{(b)}/\tau\Big) \tag{14}$$

$$\mathcal{L}_{\text{cl}} = -\frac{1}{BN}\sum_{b=1}^{B}\sum_{i=1}^{N} \log \frac{\exp\Big(s_{i,\tilde{i}}^{(b)}/\tau\Big)}{Z_i^{(b)}} \tag{15}$$

where τ is the temperature. The downstream objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{cl}}\,\mathcal{L}_{\text{cl}} \tag{16}$$

III. EXPERIMENTS

A. Experimental setup

This section describes the evaluation protocol, baselines, and implementation details under a consistent and reproducible setting.

a) Datasets, metrics, and baselines: We evaluate on four traffic benchmarks (PEMS03/04/07/08) with 358/307/883/170 nodes, using train/validation/test split 6:2:2 and $H = 12 \to P = 12$ forecasting. We report MAE, RMSE, and MAPE averaged over all horizons. Baselines include classical models (LSTM), STGNNs (ASTGCN, AGCRN, SSTBAN, PDFormer, iTransformer), and PLM-based methods (OFA, ST-LLM, STGLLM, STD-PLM). We report three standard forecasting metrics: mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Unless otherwise stated, we average the errors over all prediction horizons.

b) Baselines: To cover representative families of traffic forecasting methods, we include classical sequence models (LSTM) [14], spatio-temporal graph models (ASTGCN [1], AGCRN [15]), recent strong spatio-temporal Transformer/STGNN baselines (SSTBAN [16], PDFormer [6], iTransformer [7]), and PLM/LLM-inspired forecasting baselines (OFA [17], ST-LLM [8], STGLLM [18]), with STD-PLM [9] as a strong PLM-based reference. For fair comparison under a consistent protocol, we follow the same dataset splits and the $12 \to 12$ forecasting setting. Unless otherwise specified, the baseline results in Table I are directly cited from STD-PLM [9], where these methods are evaluated under the same split and protocol.

c) Implementation details: Results are averaged over five seeds. We use DistilBERT with LoRA (rank $r = 32$, $\alpha = 16$, dropout 0.1) applied to query and value projections. Node-index embeddings replace positional embeddings and are kept frozen. Token dimension $d = 256$; temporal stream uses joint periodic embeddings with $D_{\text{tod}} = 288$.

Stage I: The Temporal VAE uses a two-block 1D CNN encoder ($1 \to 16 \to 32$ channels) mapping flow history to $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$. Pretrained on pooled PEMS data with $\mathcal{L}_{\text{vae}} = \mathcal{L}_{\text{rec}} + \beta \cdot \mathbb{E}[\max(\lambda_{\text{kl}}, \mathcal{L}_{\text{kl}})]$ using DropToken/DropLatent rate 0.1, $\beta = 0.01$, $\lambda_{\text{kl}} = 0.5$.

Stage II: Fine-tune with $\rho_{\text{vae}} = 0.2$, up to 300 epochs with early stopping. Contrastive regularization uses $K = 1$, $M = 16$, $\tau = 0.2$, $\lambda_{\text{cl}} = 0.1$.

B. Main results

We benchmark VaT-BERT against diverse baselines under the standard $12 \to 12$ protocol. As summarized in Table I, VaT-BERT achieves the lowest MAE on all four datasets and yields the best RMSE on three datasets (PEMS04/07/08). For MAPE, VaT-BERT is best on PEMS03 and PEMS08, and is within 0.16 and 0.02 absolute MAPE of the best method on PEMS04 and PEMS07, respectively.

Against the strong PLM baseline STD-PLM, VaT-BERT reduces MAE/RMSE from 18.16/30.21 to 17.96/29.56 on PEMS04 and from 19.25/32.84 to 19.13/32.48 on PEMS07; on PEMS08, it also improves all three metrics (13.29/23.01/8.80 vs. 13.31/23.19/8.84). Compared with spatio-temporal baselines such as PDFormer and SSTBAN, VaT-BERT also shows consistent advantages, especially on PEMS04/07/08. In contrast, time-only sequence models (LSTM, iTransformer) and general sequence-to-sequence baselines

TABLE II

Ablation on backbone choice (bidirectional vs. causal) and pretraining (PEMS04). Abbrev.: w/o = without; freeze PLM = freezing pretrained encoder. Best is highlighted in the main block.

Setting	MAE	RMSE	MAPE(%)
VaT-BERT (DistilBERT, pretrained)	**17.96**	**29.56**	**12.05**
GPT2-causal (12L)	18.66	30.45	12.58
GPT2-causal (6L)	19.50	31.54	13.29
w/o pretrain (rand init)	18.40	29.77	12.59
freeze PLM (train others)	18.25	30.10	12.16
Keep original positional embeddings	18.14	29.78	12.20
Node perm.: VaT-BERT	17.95	29.57	12.05
Node perm.: GPT2-causal (6L)	19.71	31.67	13.67

TABLE III

Ablation on VAE-based tokenization and the two-stream design (PEMS04). Abbrev.: cap-match = using a deterministic CNN value encoder with a matched convolutional structure and output dimension to the VAE encoder's μ-path. Best is highlighted.

Setting	MAE	RMSE	MAPE(%)
VaT-BERT	**17.96**	**29.56**	**12.05**
CNN value encoder (cap-match)	18.01	29.63	12.16
w/o temporal stream	18.36	29.75	12.36
Temporal-only (w/o value stream)	24.22	40.43	16.25

(OFA) yield substantially larger errors, indicating that accurate traffic forecasting benefits from explicit inter-node interaction modeling rather than relying on temporal correlations alone.

C. Ablation studies

1) Bidirectional backbone and pretraining: Table II shows that replacing the bidirectional encoder with a causal decoder degrades all metrics, and node permutation results indicate VaT-BERT is substantially less sensitive to node order (MAE 17.96→17.95) than GPT2-causal (6L) (MAE 19.50→19.71). This suggests that imposing a left-to-right causal mask under an arbitrary node serialization introduces an order-coupled directional constraint on cross-node information flow. Random initialization leads to a clear performance drop, indicating that adapting a pretrained backbone is more effective than learning from scratch. Freezing the pretrained encoder also reduces performance, showing that parameter-efficient backbone adaptation contributes beyond optimizing surrounding components. Replacing original positional embeddings with node index embeddings further improves performance.

2) Variational tokenization and two-stream inputs: Table III evaluates the contribution of each input stream and the effect of VAE-based tokenization. Removing the value stream (Temporal-only) results in a large performance drop, suggesting that per-node historical flow segments provide the primary predictive signal. Removing the temporal stream also hurts performance, indicating that explicit time-of-day and day-of-week context provides additional conditioning information. Replacing the pretrained VAE tokenizer with a deterministic CNN yields worse performance, suggesting that VAE-based tokenization provides more stable representations. This robustness tendency also appears under additive Gaussian perturbations. Compared with the cap-matched CNN encoder, the VAE tokenizer yields lower MAE/RMSE at both $\sigma = 0.1$ (18.23/29.86 vs. 18.34/29.99) and $\sigma = 0.2$ (18.68/30.46 vs. 18.77/30.64), suggesting that the variational latent space better tolerates noisy sensor inputs.

TABLE IV

Representative PEMS04 MAE under correlation-guided contrastive regularization. All nonzero-CL settings outperform the w/o CL baseline.

Setting	MAE
w/o CL	18.10
$K = 1$, $M = 8$, $\lambda_{\text{cl}} = 0.10$	17.95
$K = 4$, $M = 32$, $\lambda_{\text{cl}} = 0.10$	17.98
$K = 1$, $M = 16$, $\lambda_{\text{cl}} = 0.05$	18.01
$K = 1$, $M = 16$, $\lambda_{\text{cl}} = 0.10$	17.95
$K = 1$, $M = 16$, $\lambda_{\text{cl}} = 0.30$	17.95

3) Correlation-prior-guided contrastive regularization: As shown in Table IV, representative settings across different positive/negative sampling choices and contrastive weights all improve over the w/o CL baseline. The MAE varies only within a narrow range, supporting that the contrastive term is beneficial without delicate hyperparameter tuning. We use $(K, M) = (1, 16)$ and $\lambda_{\text{cl}} = 0.1$ as practical defaults.

IV. Related Work

Spatio-temporal forecasting commonly models sensors as nodes and learns spatial dependencies jointly with temporal dynamics. Typical baselines include STGNNs such as ASTGCN and Graph WaveNet, while more recent attention-based models such as GMAN and PDFormer strengthen long-range interaction modeling [1]–[3], [6]. Pretraining has been explored to provide transferable representations for traffic forecasting, including traffic-domain masked modeling [10] and general masked time-series pretraining [11]; in contrast, our goal is to adapt a general-purpose pretrained encoder without relying on traffic-specific pretraining. VAE-based methods have been used for traffic imputation and compact latent representation learning [19], [20], while contrastive learning has been adopted to improve spatio-temporal representations through more structure-aware objectives [21], [22]. Our method combines these directions by using a VAE as a tokenizer for stable node representations and by injecting correlation-aware supervision into a pretrained bidirectional backbone.

V. Conclusion

VaT-BERT adapts a pretrained bidirectional encoder for inter-node modeling with variational tokenization and correlation-prior-guided contrastive regularization. Evaluated on PEMS03/04/07/08, VaT-BERT achieves the best MAE on all four datasets and the best RMSE on three datasets, while ablations confirm the contribution

of each design choice. Ablations further show that non-causal bidirectional attention reduces node-order sensitivity, while the VAE tokenizer and temporal stream provide complementary gains. The correlation-guided contrastive term remains stable across tested sampling choices and weights, making the method practical to tune.

References

[1] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *AAAI Conference on Artificial Intelligence*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:58565947

[2] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for deep spatial-temporal graph modeling," in *International Joint Conference on Artificial Intelligence*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:173990443

[3] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A graph multi-attention network for traffic prediction," *ArXiv*, vol. abs/1911.08415, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208158373

[4] M. Xu, W. Dai, C. Liu, X. Gao, W. Lin, G.-J. Qi, and H. Xiong, "Spatial-temporal transformer networks for traffic flow forecasting," *ArXiv*, vol. abs/2001.02908, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:210116815

[5] Z. Shao, Z. Zhang, W. Wei, F. Wang, Y. Xu, X. Cao, and C. S. Jensen, "Decoupled dynamic spatial-temporal graph neural network for traffic forecasting," *ArXiv*, vol. abs/2206.09112, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249889254

[6] J. Jiang, C. Han, W. X. Zhao, and J. Wang, "PDFormer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction," *ArXiv*, vol. abs/2301.07945, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255999991

[7] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "iTransformer: Inverted transformers are effective for time series forecasting," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=JePfAI8fah

[8] C. Liu, K. H. Hettige, Q. Xu, C. Long, S. Xiang, G. Cong, Z. Li, and R. Zhao, "ST-LLM+: Graph enhanced spatio-temporal large language models for traffic prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 8, pp. 4846–4859, 2025.

[9] Y. Huang, X. Mao, S. Guo, Y. Chen, J. Shen, T. Li, Y. Lin, and H. Wan, "STD-PLM: understanding both spatial and temporal properties of spatial-temporal data with PLM," in *Proceedings of the AAAI Conference on Artificial Intelligence*, ser. AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i11.33286

[10] K. Jin, J. A. Wi, E. Lee, S. Kang, S. Kim, and Y. Kim, "TrafficBERT: Pre-trained model with large-scale data for long-range traffic flow forecasting," *Expert Syst. Appl.*, vol. 186, p. 115738, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238739037

[11] J. Dong, H. Wu, H. Zhang, L. Zhang, J. Wang, and M. Long, "SimMTM: a simple pre-training framework for masked time-series modeling," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/fc400d1a5f7394d73d6c5e8f17107ae1-Abstract-Conference.html

[12] J. Qian, S. Zhang, Y. Pian, X. Chen, and Y. Liu, "Spatiotemporal subspace variational autoencoder with repair mechanism for traffic data imputation," *Neurocomputing*, vol. 617, p. 128948, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274205637

[13] L. Pan, Q. Ren, Z. Li, and X. Lv, "Rethinking spatial-temporal contrastive learning for urban traffic flow forecasting: multi-level augmentation framework," *Complex & Intelligent Systems*, vol. 11, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:275129021

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/ce1aad92b939420fc17005e5461e6f48-Abstract.html

[16] S. Guo, Y. Lin, L. Gong, C. Wang, Z. Zhou, Z. Shen, Y. Huang, and H. Wan, "Self-supervised spatial-temporal bottleneck attentive network for efficient long-term traffic forecasting," in *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*. IEEE, 2023, pp. 1585–1596. [Online]. Available: https://doi.org/10.1109/ICDE55515.2023.00125

[17] T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin, "One fits all: Power general time series analysis by pretrained LM," 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*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/86c17de05579cde52025f9984e6e2ebb-Abstract-Conference.html

[18] L. Liu, S. Yu, R. Wang, Z. Ma, and Y. Shen, "How can large language models understand spatial-temporal data?" 2024. [Online]. Available: https://arxiv.org/abs/2401.14192

[19] G. Boquet, J. L. Vicario, A. Morell, and J. Serrano, "Missing data in traffic estimation: A variational autoencoder imputation method," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2882–2886, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:146065401

[20] G. Boquet, A. Morell, J. Serrano, and J. L. Vicario, "A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection," *Transportation Research Part C: Emerging Technologies*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216337658

[21] X. Liu, Y. Liang, C. Huang, Y. Zheng, B. Hooi, and R. Zimmermann, "When do contrastive learning signals help spatio-temporal graph forecasting?" *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:253265402

[22] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Wu, J. Zhang, and Y. Zheng, "Spatio-temporal self-supervised learning for traffic flow prediction," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. [Online]. Available: https://doi.org/10.1609/aaai.v37i4.25555