ST-VLM: A Spatial-to-Image Multimodal Spatial-Temporal Prediction Framework with Vision-Language Model

Tong Zhao[1,2], Junping Du[1,2*], Zhe Xue[1,2], Meiyu Liang[1,2], Aijing Li[1,2], Xiaolong Meng[1,2], Dandan Liu[1,2]

[1]School of Computer Science (National Pilot School of Software Engineering),Beijing University of Posts and Telecommunications

[2]Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia

{zhaotong, xuezhe, meiyu1210, aijing_li, mengxiaolong, luckydan6798}@bupt.edu.cn,{junpingdu}@126.com

Abstract

Spatial-temporal prediction plays a crucial role in various domains, including intelligent transportation and environmental monitoring. Although large language model has shown advantages in long-range dependency modeling and excellent generalization ability for prediction tasks, it has limited understanding of spatial-temporal features. Especially for spatial features, most existing methods still simplify the spatial-temporal prediction task into multiple independent temporal prediction tasks, failing to effectively encode the dynamic evolution of spatial relations. To address these problems, we propose ST-VLM (Spatial-Temporal Forecasting with Vision-Language Model), a novel framework that leverages visual representations to encode the dynamic spatial dependencies within spatial-temporal data and integrates multi-modal information to enhance prediction. This framework transforms spatial-temporal features into three modalities: vision, text, and time series, enhances cross-modal fusion through an attention-aware fusion mechanism in the first-layer of Vision-Language Model (VLM), optimizes multi-modal feature interaction via adaptive fine-tuning strategies. After fusion, the multi-modal embeddings are subsequently used for the final spatial-temporal prediction task. Extensive experiments demonstrate that ST-VLM achieves state-of-the-art performance across various datasets. In particular, the framework exhibits promising results in few-shot scenarios, verifying its strong generalization ability.

Code — https://github.com/ZtLyun/ST-VLM

Introduction

Spatial-temporal prediction serves as a critical foundation for analyzing and understanding dynamic systems such as intelligent transportation and urban planning (Song, Li, and Li 2017). Compared with temporal prediction, spatial-temporal prediction exhibits pronounced spatial heterogeneity, correlations, and dependency patterns. Traditionally, approaches such as Recurrent Neural Networks (RNNs) and Graph Convolutional Networks (GCNs) (Zonoozi et al. 2018) (Song et al. 2020) have been widely used, leveraging temporal sequences and fixed spatial topologies to make predictions. However, these methods often suffer from two major limitations: an inability to generalize to unseen scenarios or data distributions and a reliance on pre-defined graph structures which may not fully represent evolving spatial relationships.

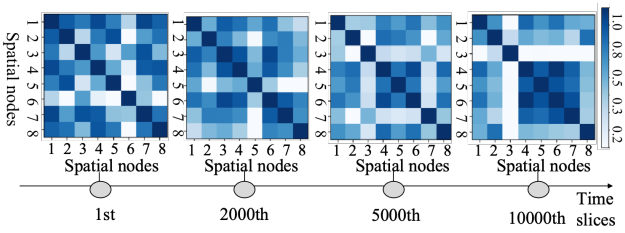

Figure 1: Illustration of complex changes in spatial relations over time

The advent of large language models (LLMs) has opened up new possibilities, as these models have shown remarkable generalization ability and flexibility in various domains, including sequential reasoning and complex pattern recognition (Liu et al. 2024a). Recent studies have demonstrated that LLMs exhibit strong performance in spatial-temporal prediction tasks, even under few-shot settings (Huang et al. 2025). Specifically, LLM-based frameworks can bypass the inflexible assumption of static spatial graphs, adjusting more flexibly to new spatial or temporal contexts (Zhong et al. 2025). Integrating LLMs into spatial-temporal prediction is thus a promising research direction that may overcome the bottlenecks of conventional architectures. Despite this progress, significant challenges remain when applying LLMs to spatial-temporal prediction, especially in effectively representing and modeling the unique spatial characteristics of such data and adapting to the inherently dynamic nature of real-world spatial relationships.

First, LLMs are not inherently suited for modeling non-sequential spatial structures, which limits their capability to accurately understand the essential spatial dependencies required for spatial-temporal prediction. Although recent studies demonstrate that integrating large language models (LLMs) into spatial-temporal prediction avoids the reliance of traditional methods (e.g., GCN, RNN) on fixed graph structures or local sliding windows (Wang et al. 2024), LLMs are natively designed for linear text

*Corresponding author.(junpingdu@126.com)

data (Huang et al. 2023a). Most existing frameworks attempt to encode spatial information or topology as sequential inputs for LLMs (Liu et al. 2024a), which remains suboptimal for learning structured or graph-based spatial relations. Therefore, how to enable large models to effectively comprehend and utilize spatial features in spatial-temporal data is still a fundamental challenge.

Second, for LLM-based framework, fixed topological representations cannot adapt to complex dynamically evolving correlations in real-world spatial-temporal data, and may introduce spurious or noisy dependencies that degrade prediction accuracy. In practice, the spatial relationships between nodes, such as upstream and downstream flows in traffic networks, change over time in response to shifting spatial dynamics (Yi et al. 2024). As shown in Figure 1, time-lagged correlations between eight traffic nodes in the PEMS04 dataset vary considerably across different time slices, illustrating the inherently dynamic nature of spatial dependencies. Existing enhancements—such as learnable adjacency matrices (Liu et al. 2023a) or spatial masking—still treat each node as an independent sequence, preventing effective modeling of temporal correlations within the spatial topology.

To address these two challenges, we propose **ST-VLM**, a novel spatial-temporal multi-modal large language model (MLLM) framework that converts spatial-temporal data into multi-modal features and applies Vision Language Models (VLMs) for spatial-temporal prediction. To make spatial features more accessible to large models and capture dynamic spatial-temporal dependencies, we introduce a spatial-to-image module that models multi-dimensional correlations among spatial nodes, including linear, nonlinear, and time-delay relationships. These correlations are transformed into a multi-channel image and evolve dynamically over time, enabling the vision-language model (VLM) to effectively perceive and reason over complex spatial temporal patterns. Additionally, we design corresponding prompts and encode temporal features for each node, leveraging VLMs' cross-modal capabilities to collaboratively model temporal, visual, and textual information. To enhance the VLM's understanding of spatial-temporal multi-modal information, we introduce an attention-based adaptive fine-tuning strategy for VLM, which contains an adaptive bias to adaptively fuse and align spatial-temporal dependencies across different modalities. In summary, our key contributions are as follows:

- Our method is the first to deploy a multi-modal large language model for spatial-temporal prediction, achieving state-of-the-art results on multiple benchmark datasets.
- We propose a multi-dimensional spatial-to-image correlation mapping approach that explicitly and effectively represents the complex and dynamic evolution interactions between spatial nodes over time.
- We develop a novel self-attention-aware multi-modal fusion mechanism combined with adaptive Low-Rank Adaptation (LoRA) fine-tuning to improve the cross-modal alignment and fusion of spatial-temporal features in VLMs.

Related Work

Spatial-Temporal Prediction

Recent studies in spatial-temporal prediction (Jiang et al. 2021) (Makridakis and Hibon 1997) (Asif et al. 2013) primarily focused on temporal modeling, often overlooking spatial dependencies. Emerging approaches like Graph Convolutional Networks (GCN) (Ye et al. 2021) (Huang et al. 2023b)and Spatial-Temporal Graph Convolutional Networks (STGCN) (Yu, Yin, and Zhu 2017) are able to model both spatial and temporal features by applying convolution operations on graph structures. These methods face challenges in terms of high computational complexity and storage limitations. Attention-based models, such as ASTGCN (Guo et al. 2019), Transformer-based architectures (e.g., STAEformer (Liu et al. 2023a), iTransformer (Liu et al. 2023b)), enhance model flexibility and accuracy by adaptively learning spatial-temporal dependencies. These models come with high computational costs and are prone to overfitting when data is limited. More specifically, adaptive models like AGCRN (Bai et al. 2020) and PDFormer (Jiang et al. 2023) aim to dynamically adjust the model structure to better capture spatial-temporal patterns, while denoising models like STD-MAE (Gao et al. 2023) and SSTBAN (Guo et al. 2023) address issues related to noisy and missing data.

LLMs for Spatial-Temporal Analyse

Leveraging Large Language Models (LLMs) for spatial-temporal forecasting has emerged as a promising direction. Models like Time-LLM (Jin et al. 2023) and STLLM (Liu et al. 2024a) offer flexibility and power in capturing temporal and spatial patterns, while approaches like Time-VLM (Zhong et al. 2025) and STGLLM (Liu et al. 2024b) expand the scope by handling multi-modal and graph-based inputs. CasMLN (Wang et al. 2024) utilizes the fine-grained representation capabilities of LLMs, decoupling spatial and temporal features, making predictions through fusion strategies. Also, LLMGeovec (He, Nie, and Ma 2025) serves as a general-purpose spatial-temporal enhancer, utilizing LLMs to boost the representation learning of spatial-temporal data. For the generalization ability of LLMs, STD-PLM (Huang et al. 2025)introduces valuable functionality for data completion and forecasting, addressing real-world data issues. However, these models neglect the limited capacity of LLMs to comprehend topological structures, with the contribution of additional design modules or textual tokens becoming more prominent in facilitating the model's performance.

Methodology

Problem Formulation

We define a spatial-temporal sequence with N nodes and T time slices as $X = \{X_1, X_2, \cdots, X_T\} \in \mathbb{R}^{T \times N}$. Here $X_t = \{x_{t,1}, x_{t,2}, \cdots, x_{t,n}\}_{n=1}^{N}, x_{t,n} \in \mathbb{R}^k$, where $x_{t,n}$ is a k-dimensional vector representing the sequence of node n within time slice t, N is the overall number of spatial nodes, and time slice t contains k time steps. Given a spatial-temporal sequence in the past k time steps from N spatial nodes, our goal is to predict the future spatial-temporal data flow for the next k time steps across the same N nodes.

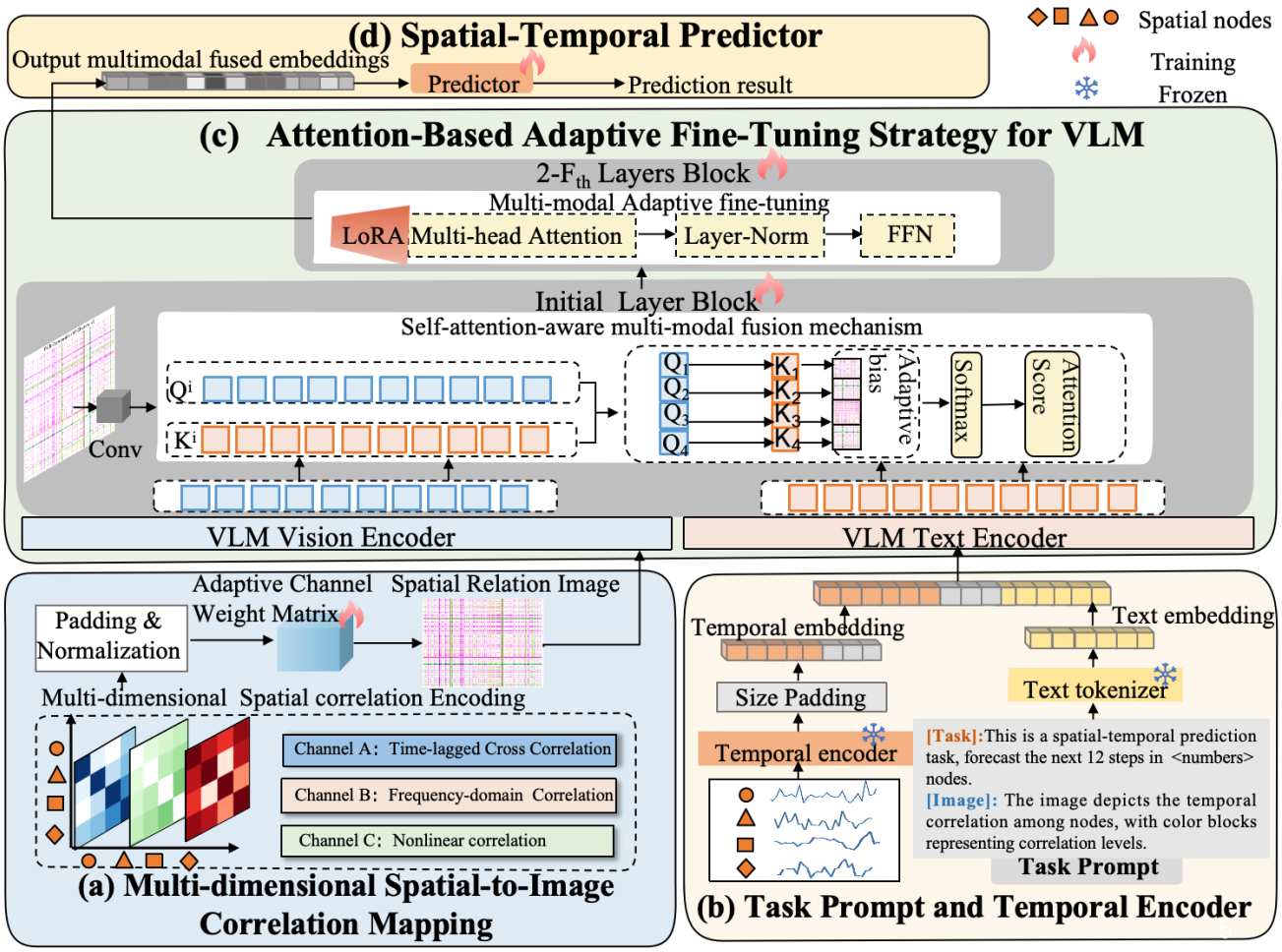

Figure 2: Overview of the ST-VLM framework

Framework Overview

As shown in Figure 2, to capture complex dependencies and better understand the evolving patterns in in spatial-temporal data, we propose **ST-VLM**, which leverages multiple modalities to enhance the spatial-temporal representations and adaptively fine-tunes a pre-trained vision-language model (VLM) for the prediction task.

This framework includes a *Multi-dimensional Spatial-to-Image correlation Mapping* module that dynamically transforms the changing relationships among nodes into multi-channel images, capturing higher-order and multi-dimensional spatial patterns. Task-specific prompts are crafted to describe these images, which will be used in subsequent modules to assist the VLM in understanding the spatial correlations. To better understand time dependencies and provide task-specific guidance, we design a *Task Prompt and Temporal Encoder* module, encoding the temporal characteristic of individual nodes into sequential tokens, allowing textual prompts and temporal data to be jointly processed. To fuse the complementarity of vision, text, and temporal modalities and capture fine-grained interaction of spatial-temporal dependencies, we incorporate an *Attention-Based Adaptive Fine-Tuning Strategy for VLM* that involves an self-attention-aware fusion mechanism within the initial self-attention layer of VLM and an adaptive fine-tuning with LoRA. These fused multi-modal embeddings are then processed by a *Spatial-Temporal Predictor* to generate precise prediction.

Multi-dimensional Spatial-to-Image Correlation Mapping

Static topological or Euclidean distance embeddings often fail to capture dynamic, nonlinear, and lagged relationships between spatial nodes (e.g., lead-lag effects). Therefore, we design the Spatial-to-Image module to represent spatial interactions from multi-dimension: nonlinear correlation, time-lagged cross-correlation, and frequency domain correlation. These perspectives represent the similarities between the temporal sequences at different spatial nodes, while nonlinear correlation primarily captures the asynchronous similarity of temporal sequences, time-lagged cross-correlation evaluates the similarity between two temporal sequences

with a time delay, and frequency domain correlation assesses the similarity by comparing the frequency components of the temporal sequences. As the sequence advances, this module adaptively converts the inter-node relationships into a 3-channel RGB image, the pixel values in each row represent multi-dimensional dependencies between the current spatial node and all of the others, preserving complex patterns interpretable by VLM.

Nonlinear Correlation. We use Dynamic Time Warping (DTW) to quantify the nonlinear similarity between node pairs at each time slice. Given a sequence of temporal data for node a over i time steps at time slice t, denoted as $x_t = (x_1, x_2, \ldots, x_i)$ and node b across j time steps $y_t = (y_1, y_2, \ldots, y_j)$, the DTW cost matrix $\mathbf{C}(x_i, y_j)$ is computed recursively as:

$$\mathbf{C}(x_i, y_j) = D(i,j) + \min\{C(i-1,j),\, C(i,j-1),\, C(i-1,j-1)\}, \quad (1)$$

where $D(i,j)$ is the Euclidean distance between elements x_i and y_j. The optimal warping path and its corresponding cumulative distance between the two temporal sequences are obtained as the final DTW distance, reflecting the minimal alignment cost.

Time-lagged Cross Correlation. It quantifies the temporal dependency by assessing how one sequence, with a lag, correlates with another sequence. For sequences x_t at node a and y_t at node b, the normalized cross-correlation at lag τ is defined as:

$$\mathbf{R}_{xy}(\tau) = \frac{\sum_{\text{i=1}}^{I-\tau}(x_i - \bar{x})(y_{i+\tau} - \bar{y})}{\sqrt{\sum_{i=1}^{I}(x_i - \bar{x})^2 \sum_{\text{t=1}}^{J}(y_i - \bar{y})^2}}, \quad (2)$$

where $\bar{x}$ and $\bar{y}$ are means, since the lengths of the two sequences are identical, $\tau \in [-I+1, I-1]$. Positive τ indicates y lags behind x, and vice versa. The resulting coefficients are assembled into matrices, where each spatial node processes its corresponding row in the matrix.

Frequency-domain Correlation. Fast Fourier Transform (FFT) is applied to the sequence of each node to obtain amplitude spectra. The Pearson correlation coefficient $\mathbf{P}_{(\text{FFT}(x_t),\text{FFT}(y_t))}$ between these spectra reflects frequency domain relationships among nodes. These correlations represent the cross-spectral coherence of amplitude distributions across spatial nodes in the frequency domain.

Multi-dimensional Channel Fusion. To overcome the information limitations of a single channel and enhance the complementary capabilities of multi-dimensional channels, we design a fusion layer that assembles the three matrices into a $N \times N$ correlation volume, forming a 3-channel image $\mathbf{M}_{\text{multi-dimensional}} = [\mathbf{C}_{N\times N}, \mathbf{R}_{N\times N}, \mathbf{P}_{N\times N}] \in \mathbb{R}^{N\times N\times 3}$, where C represents DTW matrix, R represents time-lagged matrix, and P represents frequency-domain matrix. Since the sizes of these images vary, they are resized via padding, aligned at the top-left corner. As DTW distances are unbounded, channels are normalized. To enhance feature learning, the resized and normalized multi-channel matrix is then convolved with a adaptive 3×3 weight matrix W_{af}, formulated as $\mathbf{M}'_{\text{multi-dimensional}} = \mathbf{M}_{\text{multi-dimensional}} \times W_{af}$, producing a weighted spatial pattern for subsequent processing.

Task Prompt and Temporal Encoder

To augment the model's comprehension of the data from both textual and temporal dimensions, we integrate these two modalities, thereby capitalizing on the distinct information they offer from complementary perspectives. To facilitate VLM understanding, we design task-specific prompt details about pixel colors derived from the spatial images and correlation strength, such as: *"This is a spatial-temporal prediction task, forecast the next 12 steps in 307 nodes; the image depicts the temporal correlation among nodes, with color blocks representing correlation levels."* The prompt is processed and serialized using the BERT tokenizer (Vaswani et al. 2017) $\mathbf{E}_p = [e_1, e_2, \ldots, e_l]$.

For each node, we encode sequences of t_{th} time slice with k time steps into vectors via an LSTM-based (Graves 2012) temporal encoder:

$$\mathbf{E}_t = \{e(s_n)\}_{n=1}^{N} \in \mathbb{R}^{N\times \text{length}}, \quad (3)$$

where each $e_{(s_n)}$ is a length-dimensional embedding representing a specific time slice at node n. To match the input size expected by the VLM, we pad these sequences to a fixed length and concatenate them with corresponding textual tokens. After embedding the temporal sequence, we concatenate the resulting vector with the prompt vector of the N nodes in each time slice to facilitate cross-modal fusion in subsequent modules. The fusion process can be represented as $\mathbf{M}_{concat} = [\mathbf{E}_t, \mathbf{E}_p]$.

Attention-Based Adaptive Fine-Tuning Strategy for VLM

To adaptively fuse the representation of visual, textual, and temporal modalities, we perform an attention-based adaptive fine-tuning strategy for the pre-trained VLM. This module achieves complex dynamic interaction and alignment of multiple modalities through an self-attention-aware multi-modal fusion mechanism and an adaptive fine-tuning. By leveraging an self-attention-aware multi-modal fusion mechanism, this module enables the model to selectively focus on the complex and dynamic interaction aspects of visual, textual, and temporal modalities, reconstructs the initial layer's attention mechanism with an adaptive attention score bias. The designed adaptive attention score bias compel the model to concentrate on pivotal spatial-temporal associations, modulate the gradient distribution across subsequent layers, and expedite the learning of alignment. Moreover, we apply an adaptive fine-tuning in multi-attention heads, enabling the model to capture intricate dependencies and understand complex relationships between visual, textual, and temporal features.

In the initial cross-modal interaction layer of VLM, we modify the multi-head self-attention by incorporating a spatial bias based on the image channel features. With the input of multi-dimensional image representations, along with temporal and textual embeddings into the VLM, the model proceeds to the multi-modal interaction layer after passing through the vision and text encoders. Specifically, the Query

($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) matrices are computed as:

$$\mathbf{Q}_h = \text{LoRALinear}(\mathbf{M}_{\text{concat}})\ \mathbf{W}_h^Q, \quad (4)$$

$$\mathbf{K}_h = \text{LoRALinear}(\mathbf{M'}_{\text{multi-dimensional}})\ \mathbf{W}_h^K, \quad (5)$$

$$\mathbf{V}_h = \mathbf{M'}_{\text{multi-dimensional}}\ \mathbf{W}_h^V, \quad (6)$$

where $\mathbf{M}_{\text{concat}}$ contains concatenated temporal and textual embeddings, and $\mathbf{M'}_{\text{multi-dimensional}}$ represents the visual features. The LoRA modules (LoRALinear) enable efficient low-rank adaptation. Since the channel pixel values in the image can most intuitively and adaptively show the mutual correlation between spatial-temporal features, we design attention-aware fusion strategy, using the channel feature values of the image as an adaptive attention score bias to update the initialization attention score, which serves as a priori value to help dynamically accelerate the convergence of inter-modal alignment, and adaptively changing with the association relationship. The attention score is adjusted as:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top + \mathbf{A}_{\text{bias}}}{\sqrt{d_k}}\right)\mathbf{V}, \quad (7)$$

where $\mathbf{A}_{\text{bias}}$ is an attention score bias that adaptively varies with spatial-temporal dependencies derived from a 1D convolution on the image matrix:

$$\mathbf{A}_{\text{bias}} = f_{\text{PadConv}}(\mathbf{M'}_{\text{multi-dimensional}}), \quad (8)$$

The image features are zero-padded in sequence length and normalized. This bias guides the attention mechanism to focus on mutually correlated regions, accelerating convergence. Furthermore, to better align and extract task-relevant features from multi-modal inputs, we extend the adaptive fine-tuning process by focusing on the optimization of attention parameters, ensuring that the model's attention layers are finely tuned to the intricate relationships between visual, textual, and temporal features. Updating the query and key weights via low-rank matrices:

$$\mathbf{H}(\mathbf{Q}^l) = \mathbf{MHA}(\mathbf{X}^l\mathbf{W}_q^{l,\text{LoRA}}) = \mathbf{MHA}(\mathbf{X}^l\left(\mathbf{W}_q^l + \mathbf{B}_q^l\mathbf{A}_q^l\right)), \quad (9)$$

$$\mathbf{H}(\mathbf{K}^l) = \mathbf{MHA}(\mathbf{X}^l\mathbf{W}_k^{l,\text{LoRA}}) = \mathbf{MHA}(\mathbf{X}^l\left(\mathbf{W}_k^l + \mathbf{B}_k^l\mathbf{A}_k^l\right)), \quad (10)$$

where l indexes layers, MHA represents multi-head attention and X indexes the input feature matrix. Subsequently, the fine-tuned VLM outputs a multi-modal fused embedding $\mathbf{E}_{\text{multi-modal}}$.

Spatial-Temporal Predictor

After fine-tuning, each multi-modal fused embedding $\mathbf{E}_{\text{multi-modal}}$ contains the multi-modal information of all nodes over the current time slice (K time steps). These embeddings are then fed into a spatial-temporal predictor to predict future spatial-temporal data in the next time slice. The predictor head, functioning as a decoder with LSTM layers and fully connected layers, generates predictions for all nodes over the next time slice. The modal is optimized by minimizing the loss function below:

$$\hat{X}_{t,n} = \text{Linear}(\text{LSTM}_{\text{decoder}}(\mathbf{E}_{\text{multi-modal}})), \quad (11)$$

$$\mathbf{L}_{\text{Pre}} = \text{MAE}_{\text{avg}} = \frac{1}{N\times K}\sum_{n=1}^{N}\sum_{t=1}^{K}\left|\hat{X}_{t,n} - X_{t,n}\right|, \quad (12)$$

where $\hat{X}_{t,n}$ is the predicted value for the $n-th$ node at the t time slice, $X_{t,n}$ is the true value for the $n-th$ node at the t time slice, N represents the number of nodes.

The complete training process of ST-VLM is outlined in Algorithm 1.

Algorithm 1: Training Process of ST-VLM

Input: Traffic feature $X_{t,n}$ in the historical t_{th} time slice with N nodes, and all hyperparameters.
Output: Trained ST-LLM for spatial-temporal prediction.

```
for each epoch do
    for each batch X_{t,n} in training data do
        E_S^{n×n} ← Spatial correlation embedding
        E_T ← Temporal and text embedding
        for each layer i in the VLM network do
            VLM Vision encoder ← E_S^{n×n}
            VLM Text encoder ← E_T
            if i ≤ F (first layers) then
                Use Equation (7) to calculate H^{i+1}
            else
                H^{F+U} ← Apply LoRA Fine-Tuning use Equation(9,10)
        E_multi-modal ← Fusion of multi-modal outputs
        X̂_{(t+1),n} ← Pass into predictor head to get predictions
        Update all learnable parameters by minimizing L_Pre in Equation (12)
```

Experiments

Datasets and Experiment Setup

Datasets. To evaluate the proposed ST-VLM model, we conduct extensive experiments on four california traffic datasets (PEMS03, PEMS04, PEMS07, PEMS08) (Chen et al. 2001).
Baselines and Metrics. We evaluate our ST-VLM model on established spatial-temporal prediction benchmarks, including basic spatial-temporal prediction models and LLM-enhanced models. Basic models contain LSTM (Song et al. 2020), TCN (Lan et al. 2022), STGCN (Yu, Yin, and Zhu 2017), ASTGCN (Guo et al. 2019), GWNet (Wu et al. 2019), SSTBAN (Guo et al. 2023), STAEformer (Liu et al. 2023a), AGCRN (Bai et al. 2020), PDFormer (Jiang et al. 2023), STD-MAE (Gao et al. 2023), iTransformer (Liu et al. 2023b). LLM-enhanced models contain OFA (Zhou et al. 2023), STLLM (Liu et al. 2024a), STGLLM (Liu et al. 2024b), STD-PLM (Huang et al. 2025).The performance metrics are mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE).
Settings. We partitioned the PEMS dataset into training, validation, and test sets with a ratio of 6:2:2. All experiments were conducted using ViLT(Kim, Son, and Kim 2021) as the foundational architecture for multi-modal large language modeling (MLLM).A time slice contains 12 time steps. The model was trained with a batch size of 32 and optimized using the Adam optimizer. We implemented an adaptive learning rate scheduler with patience in 10 epochs and 0.5 reduc-

Category	Model	PEMS03			PEMS04			PEMS07			PEMS08		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Basic	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%
	TCN	19.31	33.24	19.86%	31.11	37.25	15.48%	32.68	42.23	14.22%	22.69	35.79	14.04%
	STGCN	17.25	30.86	18.11%	21.70	34.72	14.02%	24.31	37.81	11.08%	18.02	27.23	11.60%
	ASTGCN	17.01	29.71	17.21%	22.81	35.03	15.63%	25.46	38.98	11.97%	18.67	28.69	12.38%
	GWNet	19.02	33.08	19.21%	25.39	40.01	17.08%	26.91	42.11	12.09%	19.32	31.14	12.70%
	AGCRN	15.16	28.12	15.61%	20.03	32.11	13.02%	22.29	36.61	9.55%	16.11	25.34	10.25%
	SSTBAN	16.01	26.32	17.21%	18.90	31.10	12.56%	20.17	33.45	8.93%	14.33	24.16	10.25%
	PDFormer	14.82	25.72	15.40%	18.32	30.01	12.21%	19.81	32.93	8.55%	13.61	23.63	9.04%
	iTransformer	19.51	31.22	17.89%	22.69	35.25	16.33%	24.60	37.94	11.33%	20.03	31.93	11.96%
	STAEformer	15.12	27.35	15.11%	18.25	30.19	11.99%	19.18	32.71	8.21%	13.59	23.31	8.89%
	STD-MAE	13.81	24.41	13.92%	17.79	29.27	11.98%	18.67	31.48	7.87%	13.48	22.44	8.77%
LLM-based	OFA	20.98	33.46	19.12%	27.39	43.02	17.99%	30.51	47.50	12.91%	21.89	34.67	13.31%
	STGLLM	15.27	24.21	15.73%	20.00	32.21	13.70%	21.99	35.01	9.79%	15.58	24.72	10.16%
	STLLM	17.28	27.31	22.89%	18.97	30.31	13.53%	21.49	34.09	10.22%	14.68	23.52	10.73%
	STD-PLM	14.61	25.37	14.91%	18.17	30.19	11.90%	19.27	32.83	8.05%	13.33	23.21	8.88%
	ST-VLM(Ours)	**11.66**	**19.19**	**12.03%**	**15.61**	**25.71**	**11.11%**	**17.32**	**29.87**	**7.56%**	**12.23**	**19.79**	**8.19%**

Table 1: Overall Performance on PEMS datasets

tion factor. The training epoch is 100 with a 30 epochs early stopping mechanism. We implement the model with the PyTorch toolkit on a Linux server with a NVIDIA RTX A6000 GPU.

Overall Performance

Table 1 presents the performance comparison between our proposed ST-VLM and baseline models. The best results are highlighted in bold, while underline denotes the second-best results.

The analysis of the results presented in table 1 leads to the following conclusions: (1) Both conventional sequential models (e.g., LSTM, TCN, iTransformer) demonstrate the lowest prediction accuracy, highlighting the limitation of using only a single temporal dimension for spatial-temporal forecasting. (2) Although early LLM-based approaches have demonstrated superior overall performance in spatial-temporal prediction compared to conventional methods (e.g., STGCN, ASTGCN), their predictive accuracy remains constrained by inherent limitations in spatial feature. Compared with some of the latest spatial-temporal autoencoding method (e.g., STD-MAE), the existing LLM based methods still remain gaps. (3) ST-VLM significantly outperforms all baseline methods across all datasets and evaluation metrics. This demonstrates the clear advantage of leveraging visual representations to enhance VLM's ability in understanding, and fusing multi-modal information for spatial-temporal prediction tasks. The outstanding performance indicate that visual guidance enables more effective modeling of complex spatial-temporal patterns compared to traditional non-visual approaches.

model	Ratio	PEMS04			PEMS08		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE
ST-VLM	5%	24.43	37.21	19.75	20.57	30.93	16.33
	10%	22.17	33.36	16.26	17.62	27.48	12.64
	20%	18.18	29.23	12.76	14.08	23.03	10.13
STD-PLM	5%	27.81	42.37	20.58	22.63	34.39	16.89
	10%	25.09	38.66	17.68	19.81	31.39	13.31
	20%	21.19	33.98	14.02	16.49	27.14	10.88

Table 2: Few-shot performance

Few-shot Performance

To evaluate the generalization ability of ST-VLM under few-shot learning conditions, we conduct a few-shot experiment on PEMS04 and PEMS08 datasets, as presented in Table 2. We compare ST-VLM against the baseline model STD-PLM, which leveraging LLM and demonstrating strong generalization capabilities. This comparison was made across various training sample ratios (5%, 10%, 20%), evaluating the models using three widely-used metrics: MAE, RMSE, and MAPE. The results demonstrate the performance of ST-VLM when trained with limited data.

From Table 2, we can observe that ST-VLM consistently outperforms STD-PLM across all three training ratios for both datasets. When trained with 5% of the training samples, ST-VLM achieves a performance close to STGCN, and even outperforms LSTM trained on the full dataset. When the training sample size reaches 20%, the performance of ST-VLM surpasses that of most traditional methods trained on the full data samples. This validates the strong generalization capability of our model. Even in data-sparse scenarios, ST-VLM is able to achieve superior performance, demonstrating its effectiveness and robustness in handling limited data.

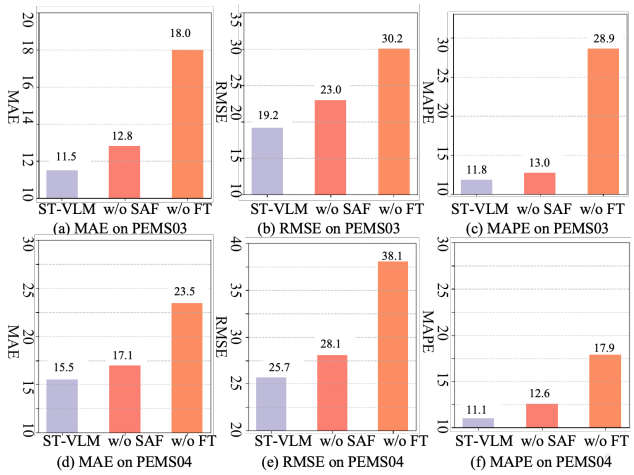

Figure 3: Ablation study on PEMS03 and PEMS04

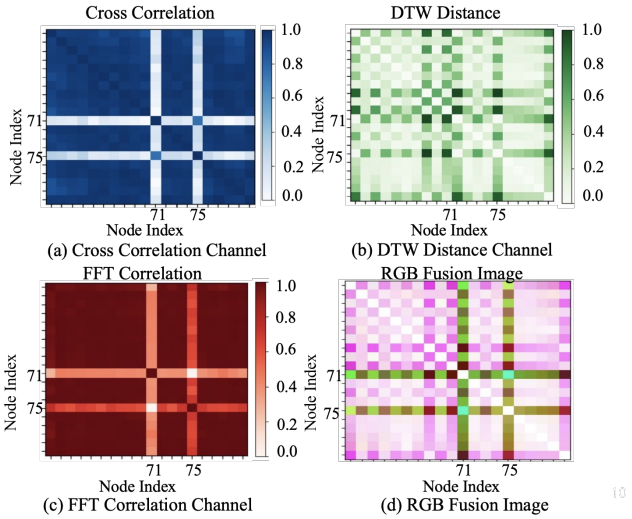

Figure 4: Visualizing spatial correlation across nodes

Ablation Study

To further validate the effectiveness of the proposed ST-VLM model, we conduct an ablation study. This study systematically evaluates the contribution of different model components by removing or modifying specific elements of the model. we designed a set of variants as 1) w/o SAF: remove the self attention fusion mechanism in initial layer of VLM, only use LORA to fin-tune VLM, 2) w/o FT: remove the fine-tuning strategy of VLM.

Figure 3 shows that the fine-tuning strategy is crucial for ST-VLM. The model's performance in w/o FT is significantly lower compared to the full ST-VLM model. This suggests that although we transform the spatial relations into images, fine-tuning is still essential. In fact, after applying adaptive LoRA fine-tuning, the model's performance already surpasses the second-best baseline. The w/o SAF shows performance comparable to ST-VLM, yet it is the self-attention-aware multi-modal fusion mechanism that enables the model to achieve its optimal performance.

Case Study

ST-VLM can effectively leverages the Spatial-to-Image module to utilize visual representations for capturing distinct spatial-temporal patterns across different spatial nodes. As shown in Figure 4, we visualize the correlations among 20 sensors in PEMS04 for the first time slice (the first 12 time steps). Based on the visualization results, it is apparent that the tensors corresponding to sensor 71 and sensor 75 exhibit distinctive behavior with weaker pixel values compared to the others.

As shown in Figure 5, the traffic flows corresponding to these sensors are extracted over the first 100 time steps. As expected, they represent two different traffic patterns. Figure 5(a) shows that sensors 71 and 75 consistently exhibit distinct trends, markedly deviating from the temporal patterns of nearby nodes. Figure 5(b) further compares the traffic patterns of sensors 71, 75, 69, and 67, where the distance from the center reflects traffic flow magnitude. Overall, sensors 71 and 75 show low, minimal-fluctuation traffic near the center, while sensors 69 and 67 display a distinct cyclical pattern with multiple peaks. This highlights the effectiveness of visual representations in reflecting different spatial patterns. Additionally, on sensors 71 and 75, ST-VLM achieves average RMSE reductions of 2.5 and 2.9 in the prediction task compared to STD-MAE and STD-PLM respectively, highlighting its superior ability to capture patterns in low-traffic and low-variance spatial nodes.

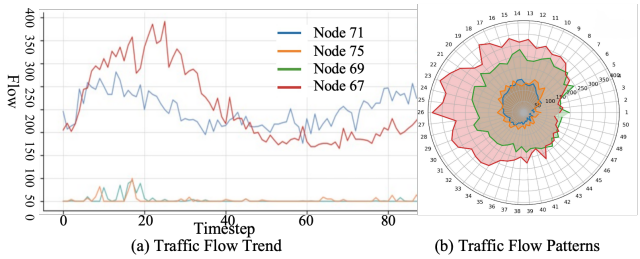

Figure 5: Different time patterns for different sensors

Conclusion

In this paper, we propose ST-VLM, a spatial-temporal prediction framework based on multi-modal large language models, integrating time series, images, and textual data. The framework enhances spatial understanding through spatial-to-image transformation and a fusion strategy for multi-modal complementary learning. Our method achieves state-of-the-art performance in prediction tasks, demonstrating effectiveness in few-shot scenarios and in capturing diverse spatial-temporal dependencies, offering a promising direction for spatial-temporal prediction with MLLMs.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (2023YFF0725103), National Natural Science Foundation of China (U22B2038, 62192784, 62422202).

References

Asif, M. T.; Dauwels, J.; Goh, C. Y.; Oran, A.; Fathi, E.; Xu, M.; Dhanya, M. M.; Mitrovic, N.; and Jaillet, P. 2013. Spatiotemporal patterns in large-scale traffic speed prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2): 794–804.

Bai, L.; Yao, L.; Li, C.; Wang, X.; and Wang, C. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33: 17804–17815.

Chen, C.; Petty, K.; Skabardonis, A.; Varaiya, P.; and Jia, Z. 2001. Freeway performance measurement system: mining loop detector data. *Transportation research record*, 1748(1): 96–102.

Gao, H.; Jiang, R.; Dong, Z.; Deng, J.; Ma, Y.; and Song, X. 2023. Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting. *arXiv preprint arXiv:2312.00516*.

Graves, A. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37–45.

Guo, S.; Lin, Y.; Feng, N.; Song, C.; and Wan, H. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 922–929.

Guo, S.; Lin, Y.; Gong, L.; Wang, C.; Zhou, Z.; Shen, Z.; Huang, Y.; and Wan, H. 2023. Self-supervised spatial-temporal bottleneck attentive network for efficient long-term traffic forecasting. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 1585–1596. IEEE.

He, J.; Nie, T.; and Ma, W. 2025. Geolocation representation from large language models are generic enhancers for spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17094–17104.

Huang, J.; Zhang, X.; Mei, Q.; and Ma, J. 2023a. Can llms effectively leverage graph structural information through prompts, and why? *arXiv preprint arXiv:2309.16595*.

Huang, Q.; Shen, L.; Zhang, R.; Ding, S.; Wang, B.; Zhou, Z.; and Wang, Y. 2023b. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. *Advances in Neural Information Processing Systems*, 36: 46885–46902.

Huang, Y.; Mao, X.; Guo, S.; Chen, Y.; Shen, J.; Li, T.; Lin, Y.; and Wan, H. 2025. Std-plm: Understanding both spatial and temporal properties of spatial-temporal data with plm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11817–11825.

Jiang, J.; Han, C.; Zhao, W. X.; and Wang, J. 2023. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 4365–4373.

Jiang, R.; Yin, D.; Wang, Z.; Wang, Y.; Deng, J.; Liu, H.; Cai, Z.; Deng, J.; Song, X.; and Shibasaki, R. 2021. Dl-traff: Survey and benchmark of deep learning models for urban traffic prediction. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 4515–4525.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, 5583–5594. PMLR.

Lan, S.; Ma, Y.; Huang, W.; Wang, W.; Yang, H.; and Li, P. 2022. Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *International conference on machine learning*, 11906–11917. PMLR.

Liu, C.; Yang, S.; Xu, Q.; Li, Z.; Long, C.; Li, Z.; and Zhao, R. 2024a. Spatial-temporal large language model for traffic prediction. In *2024 25th IEEE International Conference on Mobile Data Management (MDM)*, 31–40. IEEE.

Liu, H.; Dong, Z.; Jiang, R.; Deng, J.; Deng, J.; Chen, Q.; and Song, X. 2023a. Staeformer: Spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting. *arXiv preprint arXiv:2308.10425*.

Liu, L.; Yu, S.; Wang, R.; Ma, Z.; and Shen, Y. 2024b. How can large language models understand spatial-temporal data? *arXiv preprint arXiv:2401.14192*.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023b. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*.

Makridakis, S.; and Hibon, M. 1997. ARMA models and the Box–Jenkins methodology. *Journal of forecasting*, 16(3): 147–163.

Song, C.; Lin, Y.; Guo, S.; and Wan, H. 2020. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 914–921.

Song, Q.; Li, D.; and Li, X. 2017. Traffic prediction based route planning in urban road networks. In *2017 Chinese Automation Congress (CAC)*, 5854–5858. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, F.; Zhu, G.; Yuan, C.; and Huang, Y. 2024. Llm-enhanced cascaded multi-level learning on temporal heterogeneous graphs. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 512–521.

Wu, Z.; Pan, S.; Long, G.; Jiang, J.; and Zhang, C. 2019. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*.

Ye, J.; Sun, L.; Du, B.; Fu, Y.; and Xiong, H. 2021. Coupled layer-wise graph convolution for transportation demand prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4617–4625.

Yi, Z.; Zhou, Z.; Huang, Q.; Chen, Y.; Yu, L.; Wang, X.; and Wang, Y. 2024. Get rid of isolation: A continuous multi-task spatio-temporal learning framework. *Advances in Neural Information Processing Systems*, 37: 136701–136726.

Yu, B.; Yin, H.; and Zhu, Z. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Zhong, S.; Ruan, W.; Jin, M.; Li, H.; Wen, Q.; and Liang, Y. 2025. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *arXiv preprint arXiv:2502.04395*.

Zhou, T.; Niu, P.; Sun, L.; Jin, R.; et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36: 43322–43355.

Zonoozi, A.; Kim, J.-j.; Li, X.-L.; and Cong, G. 2018. Periodic-CRN: A convolutional recurrent model for crowd density prediction with recurring periodic patterns. In *Ijcai*, volume 18, 3732–3738.