STQA: A DATASET AND ROAD-SUBGRAPH-AUGMENTED LANGUAGE MODEL FOR SPATIOTEMPORAL TRAFFIC QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial-temporal question-answering is crucial in transportation scenarios. However, existing traffic models typically focus on fixed numerical tasks, while general-purpose language models struggle to integrate temporal dynamics, directed road network dependencies, and heterogeneous contextual evidence. We investigate whether diverse traffic-analysis tasks can be unified through a natural-language interface while retaining structured spatiotemporal evidence. To this end, we introduce STQA, a dataset of approximately 120,000 question–answer pairs spanning ten subtasks across forecasting, imputation, classification, and explanation. STQA aligns traffic observations with road-network context to support both numerical analysis and contextual explanations of traffic patterns and event responses. We further propose RoadSubLM, a road-subgraph-augmented language model that uses a directed graph encoder to represent local road-network context as compact graph soft tokens. Auxiliary supervision reconstructs adjacency and road distances from graph-token hidden states, encouraging road-network structure to remain recoverable after graph–language integration. On STQA, RoadSubLM outperforms the evaluated language-model and time-series-language baselines across all ten subtasks, with particularly large gains in imputation, spatial-propagation classification, and contextual explanation. The dataset and code are available at `https://anonymous.4open.science/r/STQA-66F3`.

1 INTRODUCTION

Spatiotemporal traffic data are fundamental to transportation applications, including traffic forecasting (Ma et al., 2026; Rühling Cachay et al., 2023), incident detection (Han et al., 2024b), and congestion analysis (Zhong et al., 2024; Jin et al., 2023). Advances in spatiotemporal modeling have improved the ability to capture temporal dynamics and dependencies among road segments. However, most existing approaches remain centered on numerical observations and predefined prediction or classification tasks, limiting their applicability to dynamic and interactive analysis.

Road-network traffic analysis supports a range of applications. There is a growing need for users to interact with spatiotemporal traffic data through natural language. For example, a traffic operator may ask why an accident at a particular road segment has caused greater disruption than similar incidents in the past. Answering such a question requires interpreting traffic dynamics together with road-network relationships and contextual conditions. This demand lies at the intersection of spatiotemporal analysis and natural language processing, gaining relevance with advances in large language models (LLMs) (Wen et al., 2023; Zhang et al., 2023; Li et al., 2024). However, despite increasing interest in cross-modal modeling (Lu et al., 2026b; Han et al., 2024a; Zhang et al., 2025a), systematic exploration of natural-language questions grounded in dynamic spatiotemporal traffic data over road networks remains limited. The lack of a standardized dataset for training and evaluating this capability further constrains progress.

To address this gap, we introduce STQA, a multi-task spatiotemporal question-answering dataset for road-network traffic analysis, as illustrated in Figure 1. STQA contains approximately 120,000 question–answer pairs organized into four task categories and ten subtasks: forecasting, imputation,

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

classification, and explanation. Each question specifies a target road node and an observation window. Its contextual evidence comprises timestamp information; static road attributes and nearby points of interest; weather conditions and weather events; traffic events; and neighboring traffic conditions. Evidence availability depends on the task, with information that directly reveals the target label withheld. Beyond numerical prediction and event identification, STQA includes explanations of temporal trends, differences across reference periods, and deviations from typical event responses. This dataset supports systematic investigation of how models interpret spatiotemporal evidence in response to natural-language queries.

Information from neighboring nodes is essential for spatiotemporal question answering; however, enabling LLMs to understand and exploit complex road-network topology is limited by a representation bottleneck. Although existing spatiotemporal models capture local dependencies and graph–language methods use soft-token interfaces, they do not directly resolve this bottleneck for road-network QA: long, heterogeneous node observations must fit within the context window without losing task-relevant dynamics or directed topology. To address this problem, we propose RoadSubLM, which selects a local subgraph around the target segment and encodes its neighborhood context with a directed graph encoder. A graph-to-language projector then maps the resulting node representations to compact graph soft tokens, which are provided to the LLM together with the question and textual context. To encourage these tokens to retain road connectivity and distance information, we apply auxiliary reconstruction objectives to the LLM hidden states at graph-token positions.

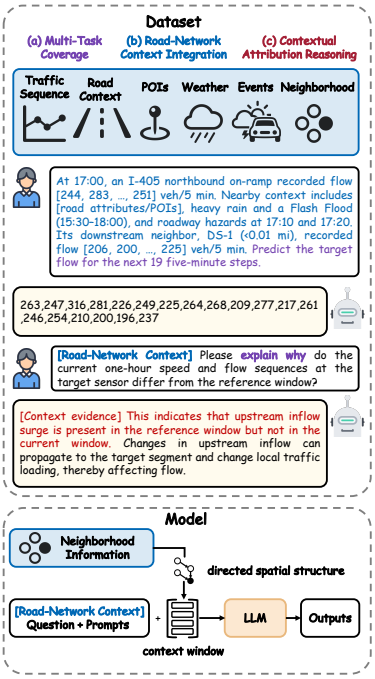


Figure 1: STQA: A Multi-Task QA Dataset Integrating Traffic Series and Road Networks, with a Compact-Token Model Bridging Spatiotemporal Graphs and LLMs

Our contributions are summarized as follows. First, we introduce STQA, a multi-task traffic question-answering dataset that couples traffic observations with road-network context and supports task-specific reasoning across forecasting, imputation, classification, and contextual explanation. Second, we propose RoadSubLM, a road-subgraph-augmented language model that converts neighborhood context into compact graph soft tokens, enabling efficient use of spatial evidence without lengthy textual serialization. Third, we introduce adjacency and road-distance reconstruction objectives to preserve and align road-network structure during graph–language integration. Across diverse traffic tasks, RoadSubLM consistently outperforms the evaluated baselines and achieves the best results on all 17 reported metrics.

2 Related Work

2.1 Spatiotemporal Traffic Modeling

Spatiotemporal traffic modeling has been widely studied for forecasting, imputation, and incident detection. Earlier approaches use statistical and recurrent models to capture temporal patterns, while graph-based methods incorporate dependencies among road segments. For example, DCRNN models traffic dynamics through diffusion on directed graphs (Li et al., 2017), whereas Graph WaveNet learns adaptive spatial dependencies alongside temporal patterns (Wu et al., 2019). More recently, pretrained language models have been adapted to traffic prediction through spatiotemporal representations (Liu et al., 2024b; Wu et al., 2022). Traffic datasets have also incorporated incident records and road attributes to support event-related analysis (Gou et al., 2024; 2026). These advances provide important foundations for understanding traffic dynamics and spatial relationships. However, such approaches are typically developed for predefined analytical tasks and output formats, limiting their flexibility in interactive settings where users request different analyses through natural language.

2.2 Large Language Models and Multimodal QA

Large language models have demonstrated strong capabilities in language understanding and generation, supporting tasks such as summarization and question answering. These capabilities have been extended to multimodal settings by connecting modality-specific encoders with language models. Vision–language frameworks combine visual representations with textual instructions to support image understanding and multimodal interaction (Li et al., 2023; Zhang et al., 2025b; Lu et al., 2026a). Related research introduces question answering for traffic-scene understanding and driving decisions (Qasemi et al., 2023; Sima et al., 2024). Beyond visual inputs, graph-oriented approaches integrate relational information into LLMs through graph–text projection and instruction tuning (Tang et al., 2024), while graph retrieval and soft prompting support question answering over larger textual graphs (He et al., 2024). Applying these methods to road-network observations requires representations that jointly capture multivariate traffic dynamics, road semantics, and directed spatial dependencies. However, integrating these heterogeneous sources into LLMs without excessively increasing input length remains challenging.

2.3 Towards Spatiotemporal Traffic QA

Although existing studies have explored combining textual information with time-series analysis, many use text primarily as auxiliary context to improve numerical forecasting (Liu et al., 2024a; Williams et al., 2024). These approaches enrich predictive inputs but do not directly support natural-language questions requiring different forms of analysis. Subsequent research investigates language-based querying and reasoning over temporal observations (Lee, 2022; Xie et al., 2024). In particular, Time-MQA unifies numerical tasks and open-ended reasoning through contextualized QA (Kong et al., 2025), while ITFormer introduces an industrial time-series QA dataset and a temporal–textual integration framework (Wang et al., 2025b). Broader benchmarks also examine LLM capabilities in spatiotemporal and urban analysis (Li et al., 2025). Nevertheless, systematic support for multi-task QA over road-network traffic observations remains limited, particularly when answers depend on local dynamics, neighboring conditions, and heterogeneous context. The limited availability of dedicated datasets connecting these observations with diverse questions and reference answers further constrains development and evaluation. To address this gap, we introduce STQA, a multi-task road-network QA dataset, together with RoadSubLM for integrating heterogeneous spatiotemporal evidence through local road subgraphs.

3 Methodology

3.1 Problem Formulation

Spatiotemporal traffic question answering encompasses four task categories: forecasting, imputation, classification, and explanation. These tasks require reasoning over traffic dynamics and road-network context, while involving diverse input and output formats. To unify them, we formulate traffic analysis as a single question-answering framework in which natural-language queries specify the task and all answers are expressed as text.

Formally, let $G = (\mathcal{V}, \mathcal{E})$ denote a directed road graph, where nodes represent sensors with static road descriptions and edges encode upstream–downstream connections. For a question Q targeting sensor v_q, let $G_q = (\mathcal{V}_q, \mathcal{E}_q)$ denote its local road subgraph, including the target and neighboring sensors. Each sensor v is associated with a record $S_v = (\mathbf{X}_v, x_v)$ comprising four types of information: flow, speed, occupancy, and node-specific static information. Here, $\mathbf{X}_v \in \mathbb{R}^{T_v \times 3}$ contains the three dynamic traffic variables over T_v observed time steps, while x_v describes the static information attached to the sensor. This representation applies to both the target and neighboring sensors. Let S_q denote the collection of their records for question Q. Let C_q denote the accompanying road-network context expressed as text. Given these inputs, the objective is to generate an answer A:

$$A = f_\theta(Q, S_q, C_q, G_q), \tag{1}$$

where f_θ models the interaction between spatiotemporal evidence and natural language. Depending on the task, A is a numerical sequence, a categorical label, or a natural-language explanation, all represented through the same textual output interface.

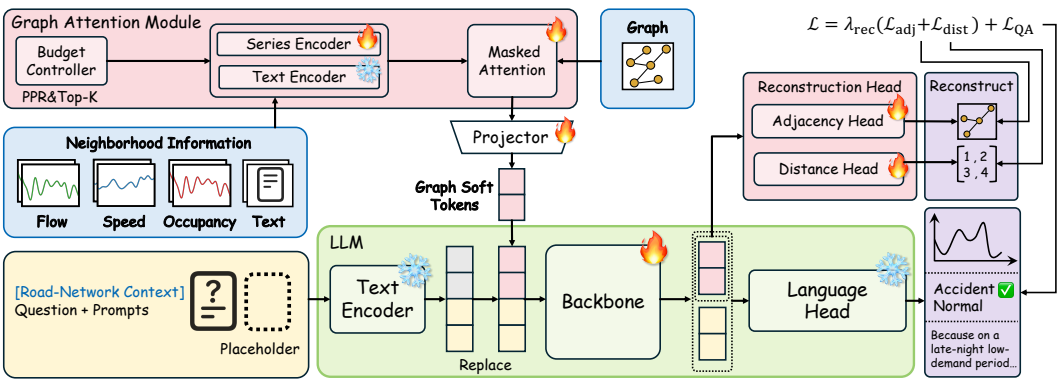


Figure 2: Overview of RoadSubLM. A graph attention module encodes neighborhood context through directed road connections. The resulting graph soft tokens replace placeholders in the input, which includes the road-network context, question and prompts. Auxiliary reconstruction heads encourage the LLM to retain directed adjacency and road-network distance information. Training proceeds through text-only adaptation, supervised graph–language alignment, and task-aware reinforcement learning.

We implement f_θ as RoadSubLM, an LLM-based question-answering model with a language backbone g_θ. The target sensor's dynamic observations and static information remain in the textual question, denoted by $\widetilde{Q} = \text{Prompt}(Q, S_{v_q})$. To incorporate additional evidence from neighboring road segments, RoadSubLM augments the backbone with a graph attention module. Let $\mathcal{N}_q$ denote the selected neighboring sensors, excluding the target, and S_q^{nbr} their associated records, including both dynamic observations and static information. The resulting mapping is

$$f_\theta(Q, S_q, C_q, G_q) = g_\theta\big(\widetilde{Q}, C_q, \Phi_{\text{graph}}(S_q^{\text{nbr}}, G_q)\big), \tag{2}$$

where Φ_{graph} comprises neighbor encoding, graph attention, and projection into the LLM embedding space. As illustrated in Figure 2, it combines neighboring sensors' static descriptions and dynamic traffic sequences into graph soft tokens $\mathbf{Z} = \Phi_{\text{graph}}(S_q^{\text{nbr}}, G_q)$. The target record enters through the textual input $\mathbf{T} = E_{\text{text}}(\widetilde{Q}, C_q)$ and is not encoded by the graph module. Graph tokens replace designated placeholders in $\mathbf{T}$, allowing the LLM to jointly process the target observations, road-network context, and neighboring evidence through its attention layers before autoregressively generating A. As shown in Figure 1, Q asks for future traffic flow, S_q provides target and neighboring sensor records, C_q contains different kinds of contextual information, and G_q specifies their road-network connections, together determining the predicted flow sequence A.

3.2 Model Architecture

Local spatiotemporal graph construction. To include relevant neighbors without expanding the input indefinitely, we retrieve candidates within bounded upstream and downstream hops from the target node. When the candidate set exceeds the node budget, road-distance-weighted personalized PageRank ranks nodes using forward and reverse propagation. We retain the target and add highly ranked candidates subject to budget, including directed paths to the target when available and affordable. The resulting subgraph is $G_q = (\mathcal{V}_q, \mathcal{E}_q)$. Appendix A specifies the propagation and selection rules. Each selected node supplies its static description and traffic observations from the configured historical window.

Heterogeneous node encoding. To capture both road semantics and traffic dynamics, we encode static descriptions and numerical observations separately before combining them:

$$\mathbf{h}_v^{(0)} = P_s(E_{\text{text}}(x_v)) + E_{\text{temp}}(S_v), \tag{3}$$

where x_v is node v's static description and $S_v \in \mathbb{R}^{T_v \times 3}$ contains its flow, speed, and occupancy observations. The frozen text embedding module E_{text} provides a cached static representation, which the trainable projection P_s maps to the graph space. The trainable temporal encoder E_{temp} projects each three-variable observation to a feature vector and adds sinusoidal temporal positions.

A one-dimensional convolution followed by the GELU activation function captures local changes, followed by masked temporal averaging and a projection to the graph space. Appendix B provides the precise operations.

Directed graph encoding. To preserve the direction of road-network interactions, the directed graph encoder E_{graph} restricts attention to incoming neighbors and each node itself. Two stacked layers combine query–key attention with residual updates and feed-forward transformations. We write $\mathbf{H} = E_{\text{graph}}(\{\mathbf{h}_v^{(0)}\}_{v\in\mathcal{V}_q}, \mathcal{E}_q)$ for the resulting node representations. Padded nodes are masked throughout propagation. Direction affects which messages can pass; reverse connections are not added automatically.

Road-subgraph-augmented language modeling. To expose spatial evidence to the LLM through a bounded number of tokens, the trainable graph-to-language projector P_g maps each node representation to the LLM embedding space:

$$\mathbf{Z} = P_g(\mathbf{H}), \tag{4}$$

where $\mathbf{Z}$ contains one soft token per selected node. The projector comprises layer normalization and two linear layers separated by GELU. These tokens replace dedicated input placeholders, preserving the selected-node order. The LLM receives them alongside Q and C_q to generate the answer.

Structural reconstruction heads. To encourage the LLM to retain connectivity, direction, and distance information, with the support of the theoretical proofs (Zhang et al., 2026), we attach two auxiliary prediction networks to its final hidden states at graph-token positions. We refer to these networks as reconstruction heads: one predicts directed adjacency, and the other predicts road-network distance. This supervision explicitly tests whether structural information remains recoverable after graph representations pass through the language backbone.

Let $\mathbf{u}_i$ denote the final LLM hidden state associated with selected node v_i. For each ordered pair of distinct nodes, the heads share the pair representation

$$\begin{aligned} &\mathbf{p}_{ij} = [\mathbf{u}_i; \mathbf{u}_j; \mathbf{u}_i - \mathbf{u}_j], \\ &\widehat{\mathbf{c}}_{ij} = \mathrm{softmax}(g_{\text{adj}}(\mathbf{p}_{ij})), \qquad \widehat{\ell}_{ij} = \mathrm{softplus}(g_{\text{dist}}(\mathbf{p}_{ij})), \end{aligned} \tag{5}$$

where $[\cdot;\cdot]$ denotes concatenation and g_{adj} and g_{dist} are separate two-layer multilayer perceptrons. The two entries of $\widehat{\mathbf{c}}_{ij}$ represent absent and present directed edges. The scalar $\widehat{\ell}_{ij}$ estimates $\log(1 + d_{ij})$, where d_{ij} is the road-network distance. Ordered pair features allow predictions for (i, j) and (j, i) to differ. These heads supervise internal representations; the language-model output head generates the QA response.

3.3 Training Process

Training proceeds through text-only adaptation, supervised graph–language alignment, and task-aware reinforcement learning. The pretrained LLM weights remain frozen throughout; LoRA parameters provide language-model adaptation.

Text-only supervised fine-tuning. To establish the shared QA interface before introducing graph tokens, we train LoRA on textual questions, available evidence, and reference answers. The answer-generation loss is

$$\mathcal{L}_{\text{ans}} = -\frac{1}{|A^*|}\sum_{t=1}^{|A^*|} \log p_\theta(a_t^* \mid Q, C_q, S_q, a_{<t}^*), \tag{6}$$

where $A^* = (a_1^*, \ldots, a_{|A^*|}^*)$ is the reference answer, a_t^* is its t-th token, and $a_{<t}^*$ denotes preceding answer tokens. The question Q, textual context C_q, and available observations S_q form the conditioning input; θ denotes the trainable model parameters. In this stage, observations are serialized as text. Prompt and padding tokens are excluded from supervision.

Supervised graph–language alignment. To make graph tokens useful for answering while retaining road structure, we jointly optimize answer generation and the reconstruction heads defined in Equation 5. The answer loss conditions on the graph-token-augmented input:

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \mathcal{L}_{\text{adj}} + \mathcal{L}_{\text{dist}}, \\ \mathcal{L}_{\text{align}} &= \lambda_a \mathcal{L}_{\text{ans}} + \lambda_r \mathcal{L}_{\text{rec}}, \end{aligned} \tag{7}$$

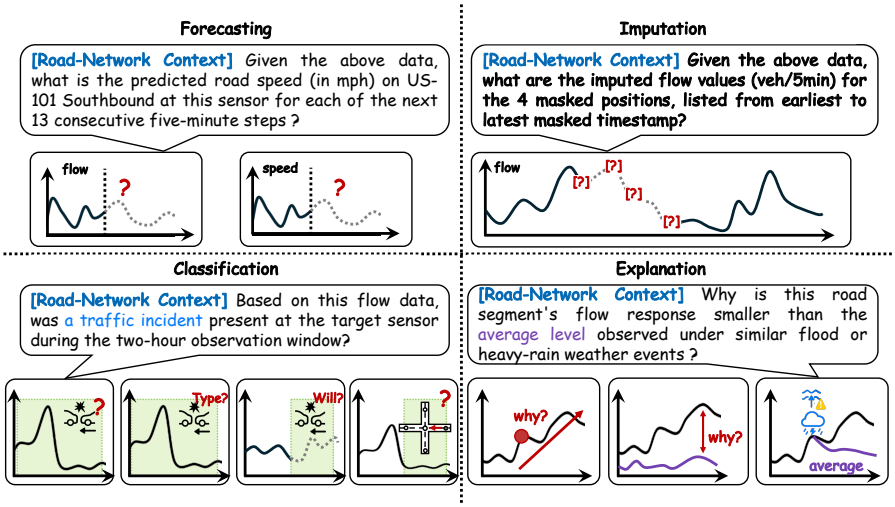


Figure 3: Overview of STQA's four task categories and ten subtasks: two forecasting tasks, one imputation task, four classification tasks, and three explanation tasks.

where $\mathcal{L}_{\text{adj}}$ is class-balanced two-class cross-entropy between $\widehat{\mathbf{c}}_{ij}$ and the directed edge label $c_{ij} = \mathbb{I}[(v_i, v_j) \in \mathcal{E}_q]$. The distance loss $\mathcal{L}_{\text{dist}}$ is the mean squared error between predicted and reference log-distances over valid pairs. Self-pairs and padded nodes are excluded. The coefficients λ_a and λ_r weight answer generation and reconstruction, respectively.

This stage updates P_s, E_{temp}, E_{graph}, P_g, both reconstruction heads, and the LLM LoRA parameters. The text embedding module remains frozen.

Task-aware reinforcement learning. To improve answer quality beyond supervised fine-tuning, we use Group Relative Policy Optimization (GRPO) with task-specific feedback. For each question, the model samples a group of responses and uses their relative rewards to update the policy without training a separate value model. For a response O with extracted answer A and reference answer A^*, the composite reward is

$$r(O) = 0.5r_{\text{format}}(O) + r_{\text{answer}}(A, A^*) + 0.1r_{\text{struct}}. \tag{8}$$

The format reward enforces the required response schema, the answer reward provides task-specific feedback for numerical, classification, and explanation outputs, and the structural reward evaluates connectivity, direction, and distance predictions. During GRPO, the graph modules and reconstruction heads remain frozen, while the LLM LoRA parameters are updated. Full reward definitions, component ranges, and scoring rules are provided in Appendix H.

4 Experiments

4.1 Dataset Description

STQA combines series data from the Los Angeles road network (Gou et al., 2024; 2026) with spatially and temporally aligned context. Each question specifies a target sensor and observation window. Data sources and construction are detailed in Appendix C.

As illustrated in Figure 3, STQA comprises four categories and ten subtasks. We use the task identifiers introduced below consistently in the figure and result tables; detailed window settings, labeling rules, and evaluation metrics are provided in Appendix C.2.

Forecasting and imputation. R1 predicts the target sensor's future flow sequence, while R2 predicts its future speed sequence, using historical observations and available context. R3 reconstructs a contiguous missing segment within an observed flow sequence.

Classification. C1 detects whether a traffic incident is present at the target sensor during the observation window, with labels *Incident present* and *No incident*. C2 classifies the observed traffic state into *Normal*, *Accidents*, or *Hazards*, distinguishing normal conditions, accident-type events,

Table 1: Mean results across repeated runs on the STQA test set over ten subtasks. MAE and RMSE are in the original data units (lower is better); accuracy, F1, and ROUGE-L are on a 0–100 scale (higher is better). Best results are shown in bold, and second-best results are underlined.

Model	Forecasting				Imputation		Classification								Explanation		
	R1		R2		R3		AF1		C1		C2		S1		T1	T2	T3
	MAE	RMSE	MAE	RMSE	MAE	RMSE	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	ROUGE-L	ROUGE-L	ROUGE-L
GPT-5.4 Nano	27.84	32.67	2.72	3.24	17.78	21.07	48.67	39.01	50.80	49.50	33.11	32.20	51.78	47.73	14.37	15.29	12.04
GPT-5.6 Luna	23.68	28.14	2.62	3.19	11.43	13.80	51.01	49.58	51.66	43.78	31.72	30.00	54.82	50.85	15.68	15.16	13.91
Claude Sonnet 5	25.32	30.01	2.66	3.24	12.56	15.10	50.20	48.72	50.51	40.41	31.29	27.03	59.46	59.23	13.67	13.90	12.16
Gemini 3.5 Flash Lite	26.46	31.12	2.58	3.15	13.35	16.04	51.61	47.94	51.50	42.32	31.45	23.04	57.15	54.96	17.39	19.77	16.56
OLMo-3-7B Instruct	23.09	27.41	2.29	2.80	13.42	16.07	51.51	46.10	52.40	48.02	33.02	25.40	52.76	41.78	49.05	58.74	34.86
Nemotron Nano-8B	23.18	27.55	2.28	2.78	13.29	15.96	51.82	46.85	51.82	46.66	32.09	24.86	51.11	38.90	57.85	58.57	39.30
DeepSeek-R1-Distill-Qwen-7B	24.02	28.39	2.29	2.79	13.30	15.98	51.91	50.54	52.80	51.95	32.22	22.54	63.73	60.32	46.06	57.89	34.64
ChatTime	24.84	29.28	2.29	2.81	10.51	12.57	52.27	51.23	54.53	53.15	35.11	32.02	70.67	69.81	54.28	48.11	32.19
RoadSubLM	**21.72**	**25.96**	**2.27**	**2.77**	**8.05**	**9.67**	**57.02**	**56.63**	**60.09**	**56.33**	**39.69**	**33.67**	**80.29**	**79.80**	**70.55**	**67.09**	**52.69**

and roadway hazards. AF1 predicts whether a traffic incident will occur within the specified future window, with *Yes* and *No* labels. S1 predicts whether congestion will propagate upstream from a downstream sensor to the target, operationally defined by whether an initially uncongested target becomes congested within the prediction window when downstream congestion is present. Event annotations that directly reveal detection or risk labels are excluded from the corresponding inputs.

Explanation. T1 explains observed flow or speed trends using the available contextual evidence. T2 compares a target window with a historical reference window and explains their traffic differences. T3 explains why the target sensor's response deviates from the average response level under same events. Together, these tasks assess the ability to interpret temporal changes, differences across periods, and context-dependent event responses. Reference answers explain observed traffic patterns through statistical comparisons and relevant road-network context.

4.2 Experimental Setup

To keep model training and testing independent and fair, we partition the dataset into training, validation, and test sets in an 80:5:15 ratio based on timestamps and spatial labels in the dataset. The splits use non-overlapping sensor groups and time ranges, with related samples across tasks assigned to the same split to prevent spatiotemporal and cross-task leakage.

We compare RoadSubLM with open-source and closed-source baselines. The open-source SFT baselines include OLMo-3-7B Instruct (Olmo et al., 2025), Nemotron Nano-8B (Bercovich et al., 2025), and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025), which are contemporary language models with parameter counts comparable to our backbone. We also include ChatTime (Wang et al., 2025a), an open-source time-series large language model, as a domain-specific SFT baseline. The closed-source baselines comprise GPT-5.4 Nano (OpenAI, a), GPT-5.6 Luna (OpenAI, b), Claude Sonnet 5 (Anthropic, 2026), and Gemini 3.5 Flash Lite (Google DeepMind, 2026). Appendix E provides the closed-source evaluation prompt, while Table 1 reports mean scores across repeated evaluation runs; complete mean–variance statistics are provided in Appendix D. Closed-source models use a temperature of 0.75 for stochastic generation.

RoadSubLM uses Qwen3-8B (Team, 2025) as its language backbone. It constructs a three-hop neighborhood and retains up to 20 nodes. Two direction-aware graph-attention layers encode the neighborhood context, including one-hour traffic histories, into 256-dimensional representations, which are projected into up to 20 tokens in the LLM's 4,096-dimensional embedding space. Training follows the three stages in Section 3.3: text-only supervised fine-tuning, supervised graph–language alignment, and task-aware reinforcement learning. The two supervised stages use AdamW with a learning rate of 5×10^{-5}, while GRPO samples four responses per prompt at temperature 0.75. Remaining architectural, optimization, and loss settings are provided in Appendix G.

4.3 Main Results

Table 1 shows that RoadSubLM achieves the best result on all 17 metrics. Relative to the strongest baseline for each metric, it reduces R3 MAE by 23.4%, improves S1 accuracy by 13.6%, and increases ROUGE-L by 14.2–34.1% across T1–T3. These gains demonstrate that the RoadSubLM can exploit structured road-network context while retaining strong numerical prediction performance.

Table 2: Ablation and cumulative training-stage results on the full STQA test set. The complete RoadSubLM configuration is reported in Table 1. MAE and RMSE are in the original units (lower is better); accuracy, F1, and ROUGE-L are on a 0–100 scale (higher is better). Best results are shown in bold, and second-best results are underlined.

Variant	Forecasting				Imputation		Classification								Explanation		
	R1		R2		R3		AF1		C1		C2		S1		T1	T2	T3
	MAE	RMSE	MAE	RMSE	MAE	RMSE	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	ROUGE-L	ROUGE-L	ROUGE-L
w/o graph module	22.90	27.26	2.75	2.98	10.35	12.40	55.72	**57.76**	54.55	54.65	34.82	**34.81**	69.46	69.27	60.67	58.43	45.60
w/o reconstruction heads	22.40	26.74	2.50	2.86	9.26	11.02	56.45	56.06	58.17	55.48	38.90	34.34	80.05	79.56	67.02	63.74	50.06
Text-only SFT (backbone)	24.12	28.48	2.28	**2.74**	13.63	14.69	54.71	55.58	51.17	49.80	33.35	32.78	55.87	55.87	53.98	59.59	47.46
Graph–language alignment	22.70	26.95	**2.24**	2.75	8.24	9.90	52.80	55.60	56.66	52.34	38.53	30.20	**81.60**	**81.45**	68.39	62.78	50.21
RoadSubLM (task-aware RL)	**21.72**	**25.96**	2.27	2.77	**8.05**	**9.67**	**57.02**	56.63	**60.09**	**56.33**	**39.69**	33.67	80.29	79.80	**70.55**	**67.09**	**52.69**

The baselines exhibit complementary but isolated strengths. The closed-source models show some numerical ability on forecasting and imputation, but their classification results remain mostly near chance level and their explanations align poorly with the reference reasoning. This likely reflects their emphasis on general-purpose language generation and the absence of specialized mechanisms for traffic-domain and spatiotemporal structure. Among the open-source baselines, ChatTime is competitive on imputation, indicating that a specialized time-series model can effectively capture temporal patterns, whereas the fine-tuned language models perform relatively well on forecasting and explanation, reflecting their ability to learn task-specific textual mappings. However, neither strong temporal modeling nor language modeling alone adequately captures directed road dependencies and heterogeneous road-network context. In contrast, RoadSubLM explicitly encodes neighborhood context as compact graph tokens, providing more balanced performance across all four task categories through a single generative interface. Section 4.4 examines the components responsible for these gains.

4.4 Ablation Study

Table 2 evaluates both architectural components and training stages. Relative to the complete RoadSubLM model, removing the graph module reduces S1 accuracy by 13.5% and lowers T1–T3 ROUGE-L by 12.9–14.0%. This confirms that neighborhood context strongly affects reasoning about spatial propagation along the road graph and explaining traffic patterns. Removing the reconstruction heads causes a generally smaller but consistent degradation, increasing numerical errors by 3.0–15.0% and reducing all three explanation scores by approximately 5.0%. This suggests that reconstruction facilitates graph–language alignment by preserving connectivity and distance information. The few F1 fluctuations indicate that these benefits are not uniform across every classwise tradeoff.

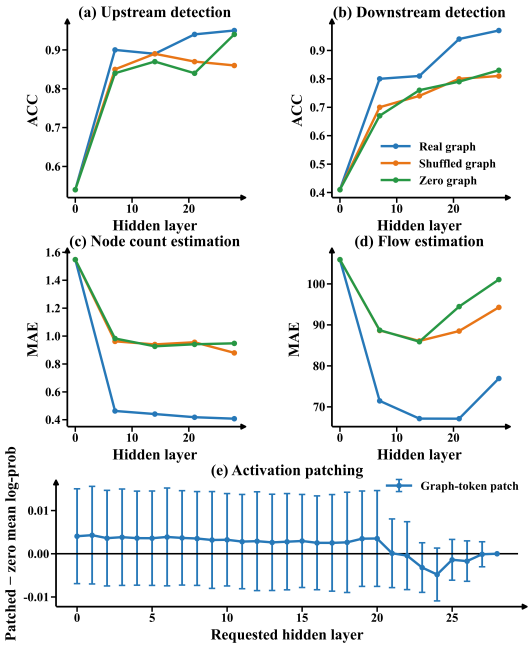


Figure 4: Linear probing and activation-patching.

The cumulative training-stage comparison follows the three-stage process in Section 3.3. Text-only SFT adapts the backbone to the shared QA interface and yields reasonable performance, but remains weak on some numerical predictions. Supervised graph–language alignment reduces R3 MAE and RMSE by 39.5% and 32.6%, respectively, and improves S1 accuracy by 46.1%, showing the value of explicitly aligning neighboring traffic and directed road relations with the language model. Task-aware reinforcement learning with GRPO then improves 13 of 17 metrics, further lowering R1 and R3 errors and improving three of four classification accuracies, with small tradeoffs on R2 and S1. Thus, supervised training establishes the response interface and graph alignment, whereas task-specific rewards sharpen numerical sensitivity and classification decisions.

4.5 Probing Graph Information Flow

To examine whether RoadSubLM incorporates graph evidence into its representations and uses it during answer generation, we conduct linear probing and activation patching on the test set. Model parameters remain frozen.

Linear probing. We train lightweight linear probes on hidden states immediately before answer generation to recover graph attributes. Zero-valued graph embeddings remove graph content while preserving token positions, and shuffled graphs break sample–graph correspondence while retaining graph values. Neighbor descriptions are excluded from the text to isolate the graph input. With the original graph, upstream- and downstream-node presence reaches approximately 95–96% accuracy, improving by 10–43% relative to the controls; node-count and mean-flow errors decrease by 53–57% and 19–24%, respectively (Figure 4(a)–(d)). These results show that the answer-generation representations encode sample-specific graph information beyond the presence of graph tokens alone.

Activation patching. Starting from a zero-graph run, we restore the original-graph hidden states at graph-token positions, one layer at a time, and compare reference-answer likelihoods with the unpatched run. Restoring these states increases mean log-probability per answer token by approximately 0.0025–0.0040, with the strongest effects in early and middle layers (Figure 4(e)). Together with the probing results, this intervention supports that graph evidence is both represented within the LLM and used to influence answer generation.

4.6 Efficiency Analysis

Table 3 reports inference efficiency with batch size 32, temperature 0.7, top-p 0.95, and a maximum of 128 generated tokens.

Table 3: Inference-efficiency comparison under the same generation configuration.

Model	Parameters (B)	Peak memory (GiB)	Load time (s)	Batch latency (s)	Sample latency (s)	Throughput (samples/s)
RoadSubLM	8.278	48.61	9.45	34.75	1.086	0.921
OLMo-3-7B Instruct	7.378	66.08	10.00	60.54	1.892	0.529
Nemotron Nano-8B	8.114	41.65	9.26	25.99	0.812	1.231
DeepSeek-R1-Distill-Qwen-7B	7.696	45.85	8.96	21.95	0.686	1.458
ChatTime	6.837	75.02	7.98	76.74	2.398	0.417

RoadSubLM has 8.278B parameters, only 2.0% more than the similarly sized baseline, indicating that the graph module introduces modest parameter overhead. Its memory use and generation speed remain within the same practical range as the comparably sized models: it is more efficient than OLMo and ChatTime. Overall, RoadSubLM retains competitive inference efficiency while providing substantially stronger performance across the STQA tasks.

5 Conclusion

In this work, we introduced STQA, a standardized traffic question-answering dataset characterized by multi-task coverage, road-network context, and contextual attribution. STQA spans forecasting, imputation, classification, and explanation, grounding questions in temporally aligned traffic data, directed neighbors, road semantics, weather, points of interest, and traffic events. Its explanation tasks further address traffic trends, historical differences, and deviations from typical event responses. Based on STQA, we proposed RoadSubLM, which encodes local road-subgraph context as compact graph tokens and uses structural reconstruction to preserve road relations.

RoadSubLM outperforms all evaluated baselines across the different metrics while maintaining balanced performance across task categories. Ablation studies and additional probing experiments further demonstrate that the model effectively represents and uses graph information, supporting the proposed design. Together, STQA and RoadSubLM provide a foundation for language-based interaction with structured spatiotemporal systems. Future work will extend the dataset's temporal coverage, strengthen the evaluation and validation of contextual explanations, and construct datasets for additional regions to assess geographic generalization.

USE OF AI TOOLS

Generative artificial intelligence tools were used in this study to assist with language polishing of the manuscript. During dataset construction, we used a large language model to convert raw traffic records and predefined task information into natural-language question–answer samples, allowing structured spatiotemporal information to be presented in a question-answering format. Specifically, DeepSeek was used to reformulate structured text into natural-language expressions. The traffic observations and task definitions underlying these question–answer pairs were derived from the data and annotation procedures described in this work, and all generated outputs were manually reviewed. The authors take full responsibility for the accuracy and integrity of the manuscript and dataset. We also accessed large language models through their APIs to evaluate them as baselines on STQA. Their generated answers were compared with the dataset's reference answers using the task-specific metrics described in the experiments.

REFERENCES

Anthropic. Introducing Claude Sonnet 5. Official model announcement, June 2026. URL https://www.anthropic.com/news/claude-sonnet-5. Published June 30, 2026. Accessed: 2026-09-26.

Marcello Barbella and Genoveffa Tortora. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317*, 2022.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, Eric Chung, and Chris Alexiuk. Llama-nemotron: Efficient reasoning models, 2025. URL https://arxiv.org/abs/2505.00949.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Greg Fall, David Kitzmiller, Sandra Pavlovic, Ziya Zhang, Nathan Patrick, Michael St. Laurent, Carl Trypaluk, Wanru Wu, and Dennis Miller. The office of water prediction's analysis of record for calibration, version 1.1: Dataset description and precipitation evaluation. *JAWRA Journal of the American Water Resources Association*, 59(6):1246–1272, 2023. doi: https://doi.org/10.1111/1752-1688.13143. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1752-1688.13143.

Google DeepMind. Gemini 3.5 Flash-Lite. Model card, Google DeepMind, July 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/. Published July 21, 2026. Accessed: 2026-09-26.

Xiaochuan Gou, Ziyue Li, Tian Lan, Junpeng Lin, Zhishuai Li, Bingyu Zhao, Chen Zhang, Di Wang, and Xiangliang Zhang. Xtraffic: A dataset where traffic meets incidents with explainability and more. *arXiv preprint arXiv:2407.11477*, 2024.

Xiaochuan Gou, Ziyue Li, Tian Lan, Junpeng Lin, Bingyu Zhao, Chen Zhang, Di Wang, Xiangliang Zhang, et al. Traffident: A dataset for understanding the interplay between traffic dynamics and incidents. *Advances in Neural Information Processing Systems*, 38, 2026.

Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 7(4):12–18, 2008.

Xiao Han, Zhenduo Zhang, Yiling Wu, Xinfeng Zhang, and Zhe Wu. Event traffic forecasting with sparse multimodal data. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 8855–8864, 2024a.

Xiaolin Han, Tobias Grubenmann, Chenhao Ma, Xiaodong Li, Wenya Sun, Sze Chun Wong, Xuequn Shang, and Reynold Cheng. Fdm: Effective and efficient incident detection on sparse trajectory data. *Information Systems*, 125:102418, 2024b.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907, 2024.

Guangyin Jin, Lingbo Liu, Fuxian Li, and Jincai Huang. Spatio-temporal graph neural point process for traffic congestion event prediction. In *Proceedings of the AAAI conference on artificial intelligence*, number 12, pp. 14268–14276, 2023.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. Time-mqa: Time series multi-task question answering with context enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29736–29753, 2025.

Yun Joon Jason Lee. A study of the effects of transmedia storytelling on active participation and language learning. *Journal of English Teaching through Movies and Media*, 23(2):29–40, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

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

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

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

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, 2024a.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024b.

Junkai Lu, Peng Chen, Xingjian Wu, Yang Shu, Chenjuan Guo, Christian S Jensen, and Bin Yang. Patra: Pattern-aware alignment and balanced reasoning for time series question answering. *arXiv preprint arXiv:2602.23161*, 2026a.

Yaxuan Lu, Guangyu Huo, Xiaohui Cui, Boyue Wang, Yong Zhang, and Zhiyong Cui. Emergency events traffic flow forecasting using text-prompt-guided multimodal large language models. *IEEE Transactions on Intelligent Transportation Systems*, 2026b.

Jiaming Ma, Binwu Wang, Guanjun Wang, Kuo Yang, Zhengyang Zhou, Pengkun Wang, Xu Wang, and Yang Wang. Less but more: Linear adaptive graph learning empowering spatiotemporal forecasting. *Advances in Neural Information Processing Systems*, 38:58382–58418, 2026.

NOAA U.S. Department of Commerce National Climatic Data Center, NESDIS. Storm events database, 2023. URL https://www.ncei.noaa.gov/stormevents/.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

OpenAI. GPT-5.4 nano. OpenAI API model documentation, a. URL https://developers.openai.com/api/docs/models/gpt-5.4-nano. Accessed: 2026-09-26.

OpenAI. GPT-5.6 Luna. OpenAI API model documentation, b. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed: 2026-09-26.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

Ehsan Qasemi, Jonathan M Francis, and Alessandro Oltramari. Traffic-domain video question answering with automatic captioning. *arXiv preprint arXiv:2307.09636*, 2023.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Advances in neural information processing systems*, 36:45259–45287, 2023.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 491–500, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12694–12702, 2025a.

Yilin Wang, Peixuan Lei, Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, Lei Jia, Yuanxiang Li, and Zhongyu Wei. Itformer: Bridging time series and natural language for multi-modal qa with large-scale multitask dataset. *arXiv preprint arXiv:2506.20093*, 2025b.

Yanjie Wen, Zhihong Li, Xiaoyu Wang, and Wangtu Xu. Traffic demand prediction based on spatial-temporal guided multi graph sandwich-transformer. *Information Sciences*, 643:119269, 2023.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. Context is key: A benchmark for forecasting with essential textual information. *arXiv preprint arXiv:2410.18959*, 2024.

Xinle Wu, Dalin Zhang, Miao Zhang, Chenjuan Guo, Shuai Zhao, Yi Zhang, Huai Wang, and Bin Yang. Autopinn: when automl meets physics-informed neural networks. *arXiv preprint arXiv:2212.04058*, 2022.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *CoRR*, abs/1906.00121, 2019. URL http://arxiv.org/abs/1906.00121.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*, 2024.

Jian Zhang, Yanyan Chen, Tao Wang, Chuan-Zhi Thomas Xie, and Yijun Tian. Mixture of spatial–temporal graph transformer networks for urban congestion prediction using multimodal transportation data. *Expert Systems with Applications*, 268:126108, 2025a.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, volume 2025, pp. 53285–53310, 2025b.

Yatao Zhang, Tianhong Zhao, Song Gao, and Martin Raubal. Incorporating multimodal context information into traffic speed forecasting through graph deep learning. *International Journal of Geographical Information Science*, 37(9):1909–1935, 2023.

Zhongjian Zhang, Xiao Wang, Mengmei Zhang, Jiarui Tan, and Chuan Shi. Toward graph-tokenizing large language models with reconstructive graph instruction tuning. In *Proceedings of the ACM Web Conference 2026*, pp. 430–441, 2026.

Xiaojing Zhong, Bin Pang, Feiqi Deng, and Xueyan Zhao. Hybrid stochastic control strategy by two-layer networks for dissipating urban traffic congestion. *Sci. China Inf. Sci*, 67(4):140204, 2024.

A Budgeted Neighborhood Selection

To limit graph-token length while retaining relevant spatial evidence, we first collect candidates within h_{up} incoming hops and h_{down} outgoing hops of the target v_q. If the candidate graph contains at most $N_{\max}$ nodes, it is retained without ranking. Otherwise, the following procedure applies.

A.1 Distance-Weighted Bidirectional Propagation

For a candidate directed edge (v_i, v_j), define $w_{ij} = \exp(-d_{ij}/\sigma_d)$, where d_{ij} is its available road-network distance in meters and $\sigma_d = 1000$ meters. The implementation uses a one-meter fallback when an edge has no usable distance; this is a ranking fallback, not a substitute reconstruction target. Normalize outgoing weights to obtain $P_{ij} = w_{ij}/\sum_k w_{ik}$. Rows without outgoing edges are zero. Construct $\mathbf{P}^{\rightarrow}$ on the original edges and $\mathbf{P}^{\leftarrow}$ on reversed edges, normalizing each separately.

For each direction $b \in \{\rightarrow, \leftarrow\}$, initialize $\mathbf{p}_b^{(0)} = \mathbf{e}_q$ and perform eight updates:

$$\mathbf{p}_b^{(k+1)} = \alpha \mathbf{e}_q + (1-\alpha)(\mathbf{P}^b)^\top \mathbf{p}_b^{(k)}, \qquad \alpha = 0.15, \tag{9}$$

where $\mathbf{e}_q$ assigns unit mass to the target node. These are finite-step personalized propagation scores; the implementation does not iterate to a convergence tolerance. Mass at dangling nodes is not redistributed, so scores need not sum to one.

The combined score for candidate v is

$$s(v) = \omega_{\mathrm{up}} p_{\leftarrow}^{(8)}(v) + \omega_{\mathrm{down}} p_{\rightarrow}^{(8)}(v), \tag{10}$$

where $\omega_{\mathrm{up}} = h_{\mathrm{up}}/(h_{\mathrm{up}} + h_{\mathrm{down}})$ and $\omega_{\mathrm{down}} = 1 - \omega_{\mathrm{up}}$. If both hop limits are zero, both weights are set to $1/2$. The experimental three-hop limits give equal directional weights. Reversing edges is used only for ranking; the selected graph retains the original road-edge directions.

A.2 Selection Under the Node Budget

Initialize the selected set as $\mathcal{U} = \{v_q\}$. Visit remaining candidates in decreasing order of $s(v)$, breaking ties by node identifier. For each candidate v, find a shortest-hop directed path from v to v_q in the candidate graph using breadth-first search. If such a path exists, let $\mathcal{P}(v)$ contain its nodes; otherwise, set $\mathcal{P}(v) = \{v\}$. Accept the candidate and its path only when $|\mathcal{U} \cup \mathcal{P}(v)| \le N_{\max}$. Otherwise, skip that candidate and continue. Stop when the budget is filled or all candidates have been considered. Finally, retain the original directed edges whose endpoints both belong to $\mathcal{U}$, together with their available distance targets.

This rule always retains the target and respects the node budget, but may return fewer than $N_{\max}$ nodes. The path check runs from candidate to target only. It therefore does not guarantee that every downstream node retains a directed path from the target after selection.

B Temporal Encoder Details

To encode a variable-length traffic sequence as one node vector, let $\mathbf{s}_{v,t} \in \mathbb{R}^3$ be the observation at step t and $\mathbf{m}_{v,t} \in \{0,1\}^3$ its effective availability mask. The mask incorporates both observed entries and requested channels. The temporal encoder computes

$$\begin{aligned}
\mathbf{e}_{v,t} &= W_{\mathrm{in}}(\mathbf{m}_{v,t} \odot \mathbf{s}_{v,t}) + \mathbf{b}_{\mathrm{in}} + \mathrm{PE}(t), \\
\mathbf{y}_{v,1:T_v} &= \mathrm{GELU}(\mathrm{Conv1D}_{k=3}(\mathbf{e}_{v,1:T_v})), \\
\overline{\mathbf{y}}_v &= \frac{\sum_t b_{v,t}\mathbf{y}_{v,t}}{\max(1, \sum_t b_{v,t})}, \\
E_{\mathrm{temp}}(S_v) &= b_v(W_{\mathrm{out}}\overline{\mathbf{y}}_v + \mathbf{b}_{\mathrm{out}}),
\end{aligned} \tag{11}$$

where $\odot$ is elementwise multiplication, PE is sinusoidal positional encoding, $b_{v,t}$ indicates at least one available channel, and b_v indicates at least one valid step. The convolution has kernel size three and padding one. Its feature dimension is 128; the output dimension is 256. Missing entries are zeroed before projection, and invalid steps are excluded from pooling. Positional features are still added before convolution, including at masked steps. The final b_v mask prevents the output projection bias from producing a dynamic representation for a node without observations.

C Dataset and Evaluation Details

C.1 Dataset Construction

STQA focuses on road-network traffic analysis in Los Angeles. Fixed traffic sensors define its spatial units, and traffic observations are recorded at five-minute intervals. Sensor locations and timestamps provide common references for aligning traffic sequences with static road attributes, directed road connectivity, surrounding land use, meteorological variables, weather events, and traffic incidents.

Traffic observations and incident records are obtained from XTraffic and TraffiDent (Gou et al., 2024; 2026). Gridded meteorological variables come from NOAA's Analysis of Record for Calibration (AORC) product (Fall et al., 2023), weather-event records come from NOAA Storm Events (National Climatic Data Center, 2023), and surrounding land-use and point of interest (POI) information comes from OpenStreetMap (OSM) (Haklay & Weber, 2008; OpenStreetMap contributors, 2017).

C.1.1 Static Features

Static semantic information covers sensor metadata and characteristics of the surrounding environment. Road metadata describe the road segment and associated infrastructure at each sensor. The principal fields include freeway number, travel direction, road type, design speed, road width, and special road use.

Information about the environment surrounding each sensor comes from OSM. Using OSMnx, we extract land-use features within 1 km of each sensor. These features comprise the area proportions of six land-use categories: residential, commercial, retail, industrial, education, and recreation ground. We also obtain counts for 14 categories of nearby points of interest, including shops, schools, stadiums, and hotels. To account for differences in the spatial influence of nodes, POIs are collected within 1 km for sensors on mainline or freeway and within 500 m for other sensors.

Land-use features and POIs are sensor-level static modalities and are treated as time invariant throughout the dataset period. They provide natural-language descriptions of static road segments and support the subsequent construction of factors related to commercial attraction, school travel, shopping, tourism, and freight activity.

C.1.2 Dynamic Traffic Data and Traffic Incidents

The dynamic traffic variables and traffic incidents are both drawn from the XTraffic dataset. The data cover the entire calendar year 2023 and comprise 105,120 five-minute time steps.

We first designate Los Angeles County, a representative county in terms of administrative division, as the target study area. Based on the dynamic traffic data, we then exclude sensor nodes whose overall missing rate exceeded 10%, and impute the remaining missing values via linear interpolation, yielding a final set of 2,211 retained nodes. Each time step thus comprises records from these 2,211 sensors across three dynamic channels: flow, occupancy, and speed.

To ensure that each event shares the same road environment with the relevant sensors and accurately captures the propagation semantics, we impose strict constraints on distance and driving direction, and match them using the nearest absolute milepost coordinates. Applying these matching rules yielded 100,805 incidents validly matched to 1,627 sensors, spanning incident types such as Hazard, NoInj, UnknInj, 1141, Fire, AHazard, CarFire, among others.

Whether an incident is active at a given time step is determined by its start timestamp. For incidents with anomalously long durations, the original start and end times are nonetheless retained for window labeling, while their natural-language descriptions uniformly state only that the duration exceeds 120 minutes. This treatment avoids presenting an exact, yet potentially delay-contaminated, duration as though it were a reliable fact.

C.1.3 METEOROLOGICAL DATA AND WEATHER EVENTS

Meteorological data come from the AORC.[1] They contain eight hourly variables for the full year of 2023, including surface precipitation, air temperature, and wind speed. For spatial alignment, each sensor is assigned the nearest grid cell according to its latitude and longitude. The meteorological data have an hourly resolution, whereas the traffic data have a five-minute resolution. Each hourly meteorological value is therefore mapped to the 12 traffic time steps within the same hour.

Weather events come from National Oceanic and Atmospheric Administration (NOAA) Storm Events. As with traffic incidents, we calculate the spherical distance between each weather-event location and each sensor. Because weather events can affect relatively large areas, an event is considered active at a sensor when its time interval covers the sample time and its nearest reported location is no more than 5 km from the target sensor. After filtering by time and region, we obtained a total of 40 weather event records, including Flash Flood, Debris Flow, Tornado, Flood, and Heavy Rain.

C.2 TASK DEFINITIONS AND METRICS

Forecasting. Flow forecasting (R1) and speed forecasting (R2) predict future numerical sequences. Historical and prediction lengths vary independently from 12 to 36 time steps, corresponding to 60–180 minutes.

Imputation. Flow imputation (R3) reconstructs a contiguous missing segment within an observed sequence. The complete sequence contains 24–72 time steps, and the missing segment contains 4–12 time steps. Errors are calculated only at masked positions.

Classification. Event detection (C1) determines whether a traffic incident is present during the observation window. Its two options are *Incident present* and *No incident*, encoded as A and B, respectively. Traffic-state classification (C2) distinguishes the source labels Normal, Accidents, and Hazards. Here, accident-type events and hazards are separate label categories; Hazards does not imply a uniform severity level. Both tasks use two hours of historical observations.

Future event-risk prediction (AF1) uses one hour of historical observations to assess the risk of a traffic incident within the next 30 minutes. Its two options are *Yes* and *No*, encoded as A and B, respectively. Congestion propagation prediction (S1) uses the same history and prediction lengths. It considers an initially uncongested target with downstream congestion and predicts whether the target subsequently becomes congested. This label operationalizes upstream congestion propagation without establishing a causal mechanism.

Explanation. Trend interpretation (T1) explains changes in observed traffic conditions. Historical comparison (T2) explains differences from corresponding historical periods. Event-response deviation (T3) explains departures from typical responses under comparable event conditions. Reference answers are constructed from identified factors and refined through constrained language rewriting. They represent evidence-based contextual explanations, rather than verified causal effects.

Numerical evaluation. Predicted numerical sequences must match the reference length to receive a regression score. For each valid QA sample, we calculate MAE and RMSE in the original units stored in the dataset. The reported metrics are arithmetic means of sample-level scores. Thus, each QA contributes equally, regardless of its sequence length.

Traffic flow is measured in vehicles per five-minute interval (veh/5 min), and speed is measured in miles per hour (mph).

Classification evaluation. We report accuracy and F1 using the parsed task labels. Invalid or missing classification answers are counted as incorrect.

Explanation evaluation. We report ROUGE-L (Barbella & Tortora, 2022) separately for T1, T2, and T3. When an aggregate explanation score is used for ablations, it averages sample-level scores across the three explanation subtasks.

[1]The NOAA AORC dataset was accessed on May 21, 2026, from `https://registry.opendata.aws/noaa-nws-aorc/`.

Table 4: Across-run means and variances for forecasting, imputation, and explanation. Lower MAE and RMSE are better; higher ROUGE-L is better.

Model	R1				R2				R3				T1		T2		T3	
	MAE		RMSE		MAE		RMSE		MAE		RMSE		ROUGE-L		ROUGE-L		ROUGE-L	
	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
GPT-5.4 Nano	27.84	2.2500	32.67	2.4100	2.72	0.2200	3.24	0.2600	17.78	0.5300	21.07	0.4800	14.37	0.1100	15.29	0.0600	12.04	0.1500
GPT-5.6 Luna	23.68	0.9500	28.14	0.9100	2.62	0.1300	3.19	0.1600	11.43	0.2000	13.80	0.2100	15.68	0.2600	15.16	0.2200	13.91	0.2100
Claude Sonnet 5	25.32	0.4900	30.01	0.5400	2.66	0.2700	3.24	0.2800	12.56	0.3900	15.10	0.4000	13.67	0.3500	13.90	0.8700	12.16	0.6800
Gemini 3.5 Flash Lite	26.46	0.8100	31.12	0.9000	2.58	0.4600	3.15	0.7400	13.35	0.3500	16.04	0.2700	17.39	0.0800	19.77	0.0500	16.56	0.0400
OLMo-3-7B Instruct	23.09	1.9494	27.41	2.1182	2.29	0.0108	2.80	0.0130	13.42	0.0009	16.07	0.0009	49.05	1.9874	58.74	15.3831	34.86	1.8155
Nemotron Nano-8B	23.18	0.4022	27.55	0.4370	2.28	0.0043	2.78	0.0052	13.29	0.1294	15.96	0.1250	57.85	3.4316	58.57	0.0472	39.30	1.6789
DeepSeek-R1-Distill-Qwen-7B	24.02	0.6700	28.39	0.7280	2.29	0.0136	2.79	0.0164	13.30	0.5900	15.98	0.5697	46.06	0.3765	57.89	1.2598	34.64	0.9110
ChatTime	24.84	1.5609	29.28	1.6148	2.29	0.0428	2.81	0.0525	10.51	0.3179	12.57	0.3239	54.28	0.9527	48.11	0.3685	32.19	0.8920
RoadSubLM	21.72	0.8551	25.96	0.8280	2.27	0.0477	2.77	0.0585	8.05	0.0251	9.67	0.0243	70.55	0.1663	67.09	0.6138	52.69	1.7716

Table 5: Across-run means and variances for classification. Accuracy and F1 are reported on a 0–100 scale; higher means are better.

Model	AF1				C1				C2				S1			
	Acc.		F1		Acc.		F1		Acc.		F1		Acc.		F1	
	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
GPT-5.4 Nano	48.67	1.1600	39.01	1.6100	50.80	3.8700	49.50	3.0500	33.11	2.3300	32.20	2.3700	51.78	2.7600	47.73	3.7900
GPT-5.6 Luna	51.01	1.4800	49.58	1.2200	51.66	0.8500	43.78	1.6800	31.72	1.7100	30.00	1.5300	54.82	1.9600	50.85	2.3800
Claude Sonnet 5	50.20	1.1200	48.72	0.6300	50.51	0.8500	40.41	1.1800	31.29	0.3200	27.03	0.6200	59.46	1.7100	59.23	0.9600
Gemini 3.5 Flash Lite	51.61	1.1600	47.94	1.0900	51.50	0.9700	42.32	1.2300	31.45	0.9700	23.04	2.1000	57.15	2.9600	54.96	2.4300
OLMo-3-7B Instruct	51.51	1.2484	46.10	1.1011	52.40	3.8232	48.02	5.0777	33.02	1.2484	25.40	1.8443	52.76	1.2484	41.78	1.2704
Nemotron Nano-8B	51.82	0.2081	46.85	0.1835	51.82	6.3114	46.66	8.3824	32.09	8.1146	24.86	11.9880	51.11	8.6001	38.90	8.7516
DeepSeek-R1-Distill-Qwen-7B	51.91	2.9823	50.54	2.6304	52.80	1.3178	51.95	1.7502	32.22	0.2081	22.54	0.3074	63.73	0.9016	60.32	0.9175
ChatTime	52.27	1.3350	51.23	1.1611	54.53	3.3207	53.15	4.4001	35.11	1.9400	32.02	2.1804	70.67	2.0315	69.81	1.8266
RoadSubLM	57.02	1.1928	56.63	0.8973	60.09	4.3450	56.33	5.7891	39.69	1.9339	33.67	1.8333	80.29	2.0683	79.80	2.6370

C.3 Evaluation Coverage

All models are evaluated on the same test set. Numerical metrics include only successfully parsed responses with the required sequence length, and scoring coverage is the number of such responses divided by the total number of evaluation questions. Invalid or missing classification answers remain in the denominator and are counted as incorrect. For the failed cases, multiple API requests are made, and the proportion of failed cases is ultimately less than 0.001.

D Across-Run Means and Variances

To assess result stability, we repeat evaluation runs under the same model and decoding configurations. Tables 4 and 5 report the mean and across-run variance for every metric in adjacent columns. The means are identical to those in Table 1. Variances are expressed in the squared units of the corresponding metrics; lower variance indicates greater stability.

E Prompt for API-Based Evaluation

We use the following prompt template to evaluate the API-based closed-source models. The placeholders are replaced with the static road-segment description, the task question, and its task-specific output requirements.

System Prompt

```
You are a traffic question-answering assistant. Use the supplied
road segment description to answer the question, and strictly follow
the output guidance.
```

User Prompt

```
## Road Segment Description
```

```
{road_segment_description}

## Question
{question}

## Output Guidance
{output_guidance}
```

F EXAMPLES FROM THE STQA DATASET

The following examples illustrate STQA's ten subtasks, with one example for each subtask. Each example retains the source static description, question, and reference answer; contextual evidence is not repeated outside these examples.

F.1 FLOW FORECASTING (R1)

Static description.

This is an off-ramp on southbound I-5 in Los Angeles County, a mountainous urban segment with a concrete surface and a posted speed limit of 70 mph. The surrounding area within 1,000 meters is 40% residential, with two supermarkets and four parking facilities nearby.

Question.

At 2023-11-20 14:00, the following readings were recorded on I-5 Southbound:
The target sensor observed a flow sequence of 118, 120, 121, 122, 122, 123, 123, 123, 125, 125, 125, 124 veh/5 min over the 12 five-minute steps ending at 14:55.
Its occupancy-ratio sequence (0-1) was 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07.
The speed sequence at the target was 61.90, 61.80, 61.80, 61.80, 61.70, 61.80, 61.70, 61.70, 61.70, 61.60, 61.50, 61.50 mph.

During the input window, the hourly weather record at 2023-11-20 14:00 reported precipitation of None, a temperature of 54.1°F (12.3°C), and moderate wind at 13.6 mph.
No active weather events overlapped the input window, and no traffic incidents were reported during this period.

From the upstream sensor US-1, located 2.74 mi away, the flow readings were 488, 490, 499, 501, 501, 497, 496, 501, 511, 511, 504, 496 veh/5 min.
Its speed readings were 59.40, 59.50, 59.50, 59.30, 59.30, 59.30, 59.20, 59.20, 59.20, 59.10, 58.90, 58.90 mph.
The occupancy-ratio sequence (0-1) for US-1 was 0.11, 0.11, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12.
No downstream sensors were present.

Given these observations, predict the road flow on I-5 Southbound for the next 60 minutes (12 consecutive five-minute steps). What will the flow be at each of the 12 steps?

Answer.

<Answer> 124,124,125,124,125,124,124,124,125,125,125,124 </Answer>

F.2 SPEED FORECASTING (R2)

Static description.

The I-405 Northbound on-ramp in Carson is a flat, urban concrete roadway with a 70 mph speed limit and 11.2 ft lanes, separated by a concrete barrier. Within a 1,000-meter radius, the land use is industrial 32%, commercial 21%, and residential 14%, with two fuel stations and one parking facility nearby.

Question.

At 2023-04-01 23:30, the following readings were recorded on I-405 Northbound:
Over the 60-minute observation window ending at 00:25, the target sensor recorded flow values of 18, 12, 12, 13, 12, 6, 12, 3, 10, 6, 9, 4 veh/5 min in consecutive five-minute steps.
The same sensor recorded occupancy ratios of 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 (0-1) across those steps.
Speed readings at the target sensor were 64.90, 65.00, 65.10, 65.00, 64.90, 65.00, 65.00, 65.20, 65.20, 65.00, 65.00, 65.00 mph in the same order.

Weather conditions during this window included an hourly record at 2023-04-01 23:00 with precipitation None, temperature 65.6°F (18.7°C), and wind 12.2 mph, and another at 2023-04-02 00:00 with precipitation None, temperature 64.9°F (18.3°C), and wind 11.5 mph. No weather events or traffic incidents overlapped the window.

Downstream sensor DS-1, located 1.43 mi ahead, recorded flow values of 119, 140, 158, 107, 118, 129, 74, 109, 109, 96, 77, 73 veh/5 min; speed values of 67.40, 66.30, 64.00, 62.20, 64.60, 65.00, 68.00, 67.30, 68.60, 69.20, 67.20, 67.40 mph; and occupancy ratios of 0.03, 0.04, 0.05, 0.03, 0.03, 0.04, 0.02, 0.03, 0.03, 0.03, 0.02, 0.02 (0-1) over the same 12 five-minute steps. No upstream sensors were present.

Given these observations, what is the predicted road speed on I-405 Northbound for each of the next 12 five-minute steps (i.e., the next 60 minutes)?

Answer.

<Answer> 65.10,64.90,65.20,65.20,65.30,65.00,65.00,64.70,64.80,65.20,65.20,65.10 </Answer>

F.3 Flow Imputation (R3)

Static description.

This is an urban off-ramp on I-210 Eastbound in Los Angeles, featuring a concrete barrier with a glare screen and a 70 mph posted speed limit. The surrounding area within 1,000 meters is predominantly undeveloped or green space, with a shopping mall, a school, and a parking facility located within 500 meters.

Question.

At 2023-02-25 02:20, the following readings were recorded on I-210 Eastbound:

The 24-step flow sequence (veh/5 min), with 4 masked values, is: 73, 59, 67, 62, 68, 57, [?], [?], [?], [?], 65, 49, 59, 52, 55, 57, 52, 50, 54, 53, 47, 65, 56, 53.

The occupancy-ratio sequence (0-1) is: 0.02, 0.01, 0.02, 0.02, 0.02, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.02, 0.01, 0.01, 0.02, 0.01, 0.01, 0.01, 0.01, 0.01, 0.02, 0.02, 0.01.

The speed sequence (mph) is: 63.80, 62.50, 63.20, 62.90, 63.40, 62.70, 61.90, 61.80, 62.50, 61.80, 63.50, 62.10, 62.80, 62.40, 62.20, 62.00, 61.50, 61.50, 61.70, 61.60, 61.80, 62.70, 61.30, 61.80.

The masked positions (1-based) are 7, 8, 9, and 10, corresponding to timestamps 2023-02-25 02:50, 2023-02-25 02:55, 2023-02-25 03:00, and 2023-02-25 03:05.

Weather conditions during the observation window included: at 2023-02-25 02:00, moderate rain at 0.18 in/h, temperature 47.9°F (8.9°C), and moderate wind at 18.1 mph; at 2023-02-25 03:00, moderate rain at 0.18 in/h, temperature 49.2°F (9.6°C), and strong wind at 25.1 mph; at 2023-02-25 04:00, moderate rain at 0.18 in/h, temperature 49.2°F (9.6°C), and strong wind at 28.1 mph.

A Flash Flood weather event was reported for LOS ANGELES from 2023-02-25 04:03 to 2023-02-25 05:00, with severity Minor, and the nearest matched event point is 4.81 km away. No traffic incidents overlap the input window.

What are the imputed flow values (veh/5 min) for the four masked positions, listed from earliest to latest timestamp?

Answer.

<Answer> 52,48,59,50 </Answer>

F.4 Event Detection (C1)

Static description.

This is a flat, urban segment of I-10 Westbound in Pomona, a freeway-to-freeway connector with a 70 mph speed limit, concrete pavement, and metal beam barriers. Within a 1,000-meter radius, the land use is predominantly undeveloped/green space, with a high density of 25 parking facilities and a few educational institutions nearby.

Question.

The observation window spans two hours and contains 24 consecutive five-minute steps, listed in chronological order. At 2023-01-14 15:00, the following readings were recorded on the target sensor: flow (veh/5 min) 42, 43, 40, 35, 42, 34, 29, 23, 46, 27, 25, 29, 35, 36, 37, 33, 34, 48, 40, 50, 36, 35, 41, 31; occupancy ratio (0-1) 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.02, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.02, 0.01, 0.02, 0.01, 0.01, 0.01, 0.01; speed (mph) 66.40, 65.90, 66.40, 67.10, 66.70, 67.50, 67.50, 66.50, 66.60, 67.00, 66.50, 67.10, 65.80, 66.80, 66.10, 65.70, 64.90, 65.10, 66.70, 65.60, 66.60, 66.30, 66.50, 66.40. Was a traffic incident present at this sensor during the two-hour observation window?
A. Incident present
B. No incident

Answer.

<Answer> A </Answer>

F.5 Traffic-State Classification (C2)

Static description.

This segment of SR-60 Eastbound is a flat, urban freeway-to-freeway connector with a 70 mph speed limit, running on a bridge deck with an 11-foot lane width and bridge barrier railing. The surrounding area within 1,000 meters is predominantly undeveloped or green space, with a high density of schools (10), moderate densities of parking facilities (4) and fuel stations (5), and one wholesale store and one hotel.

Question.

The observation window spans two hours and contains 24 consecutive five-minute steps, listed in chronological order. At 2023-10-30 03:15, the following readings were recorded on the target sensor: flow (veh/5 min) 58, 80, 68, 55, 74, 71, 58, 54, 60, 62, 58, 79, 56, 68, 79, 65, 69, 65, 65, 71, 71, 62, 82, 60; occupancy ratio (0-1) 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.02, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.04, 0.03; speed (mph) 64.00, 64.30, 63.60, 63.90, 63.60, 63.60, 63.70, 64.00, 63.60, 63.50, 63.80, 63.70, 63.90, 62.50, 63.80, 63.40, 63.70, 63.70, 63.70, 64.70, 63.60, 63.90, 63.90, 63.90. Based on these readings, what is the most likely traffic-disruption type during the two-hour observation window?
A. Normal
B. Accidents
C. Hazards

Answer.

<Answer> A </Answer>

F.6 Future Event-Risk Prediction (AF1)

Static description.

This is a flat, urban freeway-to-freeway connector on I-105 Westbound in Lynwood, Los Angeles County, with a posted speed limit of 70 mph and a bridge deck pavement. The surrounding area within 1,000 meters is residential 36% and industrial 6%, with few schools, high-density parking facilities, and few fuel stations nearby.

Question.

At 2023-08-21 01:35, the following readings were recorded on the target sensor location:
Observed flow over the past 12 five-minute steps: 10, 16, 9, 6, 11, 6, 6, 14, 9, 12, 8, 8 veh/5 min (average 10 veh/5 min).
Observed occupancy ratio over the past 12 five-minute steps: 0.00, 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, 0.01, 0.00, 0.01, 0.00, 0.00 (average 0.00).
Observed speeds over the past 12 five-minute steps: 67.50, 66.90, 68.50, 67.10, 67.80, 66.00, 67.20, 66.00, 67.10, 65.60, 67.50, 65.90 mph (average 66.92 mph).
Weather conditions include light rain at 0.08 in/h, temperature 73.1°F, and wind at 6.3 mph.
No neighboring sensor data is available.
Based only on these observations, will a new traffic accident begin at the target sensor location within the next 30 minutes?
Options: A) Yes, an accident will occur; B) No, an accident will not occur.

Answer.

<Answer> A </Answer>

F.7 Congestion Propagation Classification (S1)

Static description.

An urban on-ramp on I-5 Northbound in Los Angeles, this concrete road segment has a posted speed limit of 70 mph and is separated by a concrete barrier with a glare screen. The surrounding area within 1,000 meters is predominantly undeveloped or green space, with four parking facilities located within 500 meters.

Question.

At 2023-09-08 03:30, the following readings were recorded on the target sensor: flow values of 84, 63, 56, 70, 57, 68, 72, 57, 62, 74, 55, 74 veh/5 min; occupancy ratios of 0.03, 0.03, 0.02, 0.03, 0.02, 0.03, 0.02, 0.02, 0.02, 0.03, 0.02, 0.02; and speeds of 64.20, 62.50, 63.70, 63.60, 64.20, 61.60, 67.10, 63.90, 64.90, 64.60, 66.50, 67.70 mph. The downstream sensor DS-1 recorded flow values of 2, 4, 1, 0, 2, 2, 4, 4, 7, 2, 6, 0 veh/5 min; speeds of 63.70, 62.20, 60.00, 60.00, 57.00, 56.10, 42.80, 42.70, 37.90, 53.20, 36.30, 36.30 mph; and occupancy ratios of 0.00, 0.02, 0.01, 0.00, 0.02, 0.02, 0.03, 0.03, 0.04, 0.00, 0.06, 0.00. Based on this data, will the observed downstream congestion propagate upstream to the target sensor within the next hour?
A) Yes
B) No

Answer.

<Answer> B </Answer>

F.8 Explanation Tasks (T1–T3)

F.8.1 Trend Interpretation (T1)

Static description.

This is a flat, urban on-ramp for northbound SR-57 in Diamond Bar, featuring a 70 mph posted speed limit and concrete pavement with a concrete barrier and glare screen. The surrounding area within 1,000 meters is predominantly undeveloped or green space, with a few parking facilities located nearby.

Question.

At 2023-02-21 18:55, the following readings were recorded on SR-57 Northbound: The target sensor speed sequence over the past hour (18:00 to 18:55) was 36.10, 37.10, 37.80, 38.00, 38.90, 38.80, 40.50, 41.60, 42.70, 43.60, 45.40, 47.50 mph, and the flow sequence was 580.00, 583.00, 589.00, 594.00, 604.00, 604.00, 626.00, 620.00, 627.00, 623.00, 628.00, 635.00 veh/5 min. Weather conditions show no precipitation, temperature 60.0°F, and calm wind at 5.2 mph with no active weather events. No traffic incidents were reported at the target sensor. The downstream sensor DS-1 (located

less than 0.01 miles away) recorded flow values of 129.00, 127.00, 133.00, 133.00, 131.00, 128.00, 131.00, 117.00, 121.00, 100.00, 66.00, 60.00 veh/5 min; speed values of 34.00, 33.40, 33.60, 34.20, 35.20, 35.50, 36.40, 37.80, 38.50, 43.80, 53.00, 57.10 mph; and occupancy ratio values of 0.20, 0.18, 0.19, 0.19, 0.18, 0.18, 0.17, 0.15, 0.15, 0.10, 0.05, 0.05 . Why does the target sensor exhibit a trend of increasing flow and increasing speed?

Answer.

The target sensor exhibits a trend of increasing flow and increasing speed because the downstream sensor DS-1, located less than 0.01 miles away, shows a mean speed increase from 33.67 to 51.30 mph (delta +17.63), indicating neighbor recovery or dissipation propagation that can propagate spatially and affect target-segment operation, thereby affecting both speed and flow, together with the current time of 2023-02-21 18:55 satisfying a weekday evening-peak period (Tuesday 18:00), indicating that the evening commute peak is active and can concentrate return-trip demand, thereby affecting speed.

F.8.2 HISTORICAL COMPARISON (T2)

Static description.

The I-405 Southbound off-ramp in Hawthorne is a flat, urban bridge deck with a concrete barrier and an 11.2-foot lane width, where the median lane is designated as an HOV lane. The surrounding area within 1,000 meters is predominantly undeveloped/green space, with a hotel, four parking facilities, and two fuel stations nearby.

Question.

At 2023-07-24 03:55, the following readings were recorded on I-405 Southbound:

Current Window (2023-07-24 03:00 to 03:55):
– Weather: No precipitation, temperature 69.0°F (20.6°C), wind calm/light breeze at 7.6 mph, no active weather events.
– Target sensor speed sequence (mph): 51.20, 60.00, 53.00, 61.50, 53.70, 53.50, 48.80, 58.80, 51.50, 58.50, 52.70, 58.00.
– Target sensor flow sequence (veh/5 min): 7.00, 1.00, 2.00, 2.00, 3.00, 3.00, 5.00, 9.00, 5.00, 6.00, 14.00, 8.00.
– No reported traffic incidents at the target sensor.
– Upstream sensor US-1 (0.74 mi away): flow (veh/5 min): 43.00, 78.00, 81.00, 84.00, 80.00, 95.00, 107.00, 137.00, 151.00, 153.00, 136.00, 135.00; speed (mph): 68.90, 69.80, 68.20, 68.00, 68.00, 66.60, 68.60, 67.40, 67.10, 66.60, 67.80, 68.20; occupancy ratio (0-1): 0.01, 0.01, 0.02, 0.02, 0.02, 0.02, 0.02, 0.03, 0.03, 0.03, 0.03, 0.03.
– No downstream sensors.

Reference Window (2023-06-26 03:00 to 03:55):
– Weather: No precipitation, temperature 62.9°F (17.2°C), wind calm/light breeze at 11.9 mph, no active weather events.
– Target sensor speed sequence (mph): 60.70, 62.80, 62.00, 62.20, 61.80, 62.00, 62.10, 62.90, 59.20, 59.00, 59.10, 60.00.
– Target sensor flow sequence (veh/5 min): 14.00, 12.00, 13.00, 12.00, 12.00, 12.00, 11.00, 13.00, 21.00, 16.00, 17.00, 16.00.
– No reported traffic incidents at the target sensor.
– Upstream sensor US-1 (0.74 mi away): flow (veh/5 min): 46.00, 60.00, 63.00, 71.00, 91.00, 101.00, 104.00, 126.00, 145.00, 128.00, 153.00, 163.00; speed (mph): 67.20, 68.90, 69.20, 68.40, 69.60, 66.00, 68.40, 69.40, 70.70, 71.60, 70.20, 68.50; occupancy ratio (0-1): 0.01, 0.01, 0.01, 0.02, 0.02, 0.02, 0.02, 0.02, 0.03, 0.02, 0.03, 0.03.
– No downstream sensors.

Why do the current one-hour speed and flow sequences differ from the reference time window?

Answer.

Current-window evidence: mean upstream flow across 1 upstream sensor changes from 67.33 to 141.33 veh/5 min (delta +74.00). Reference-window evidence: mean upstream flow across 1 up-

stream sensor changes from 56.33 to 148.00 veh/5 min (delta +91.67). This indicates that upstream inflow surge is present in the reference window but not in the current window. Changes in upstream inflow can propagate to the target segment and change local traffic loading, thereby affecting flow.

F.8.3 EVENT-RESPONSE DEVIATION (T3)

Static description.

The I-210 Eastbound on-ramp in Duarte is a flat, urban bridge deck with a 70 mph speed limit and concrete barrier, featuring a median HOV lane. Within a 1,000-meter radius, land use is 32% residential and 17% retail, with two shopping malls, one hotel, seven parking facilities, and one fuel station nearby.

Question.

At 2023-03-22 08:55, the following readings were recorded on I-210 Eastbound: moderate rainfall observed at the sensor location is present, with precipitation at 0.13 in/h, temperature 46.5°F, wind 4.3 mph, and dense fog visibility. No traffic incidents are reported at the target sensor. The target sensor speed sequence over the past 12 five-minute steps (most-recent last) is: 16.10, 13.90, 14.80, 24.00, 40.20, 47.20, 50.80, 54.70, 56.60, 59.00, 56.90, 55.50 mph. One downstream sensor (DS-1, distance <0.01 mi) provides flow: 138.00, 108.00, 126.00, 104.00, 92.00, 93.00, 93.00, 76.00, 98.00, 109.00, 99.00, 101.00 veh/5 min; speed: 34.60, 38.30, 38.70, 46.50, 54.10, 56.20, 58.70, 59.40, 57.60, 53.00, 53.80, 54.90 mph; occupancy ratio: 0.14, 0.10, 0.12, 0.07, 0.06, 0.06, 0.06, 0.05, 0.07, 0.08, 0.07, 0.07. Why is this road segment's speed response clearly larger/stronger than the average level under similar moderate rainfall observed at the sensor location events?

Answer.

Under the same moderate rain condition, neighboring sensor sequences show no upstream sensors are available; 1 downstream sensors; downstream mean speed changes from 37.20 to 53.90 mph (delta +16.70); the target has 0 upstream and 1 downstream sensors. This indicates that neighbor recovery or dissipation propagation is active in the current window. Neighbor recovery can propagate spatially and affect target-segment operation, thereby affecting speed.

G IMPLEMENTATION DETAILS

G.1 GRAPH REPRESENTATION

The upstream and downstream hop limits are both three, and $N_{\max} = 20$. The retrieved dynamic window spans the preceding hour. The temporal feature dimension is 128, the graph dimension is 256, and the LLM embedding dimension is 4,096. Two directed attention layers each use one set of query, key, and value projections. The temporal and projection operations are defined in Appendix B and Section 3.2.

Training configuration. All training stages use AdamW with weight decay 0.01. Text-only SFT and supervised graph–language alignment use a learning rate of 5×10^{-5}; GRPO uses a learning rate of 10^{-7}. Text-only SFT uses a cosine scheduler with a warmup ratio of 0.03; the remaining stages use constant learning rates. LoRA uses rank 32, scaling factor 64, and dropout 0.05. GRPO samples four responses per prompt with temperature 0.75.

Loss and reward coefficients. Supervised graph–language alignment uses $\lambda_a = 2$ and $\lambda_r = 0.005$. The GRPO reward weights are 0.5 for format, 1 for answers, and 0.1 for structure. For numerical rewards, the manually specified reference scales (m_τ, b_τ) are $(17, 20)$ for R1, $(1.6294, 1.9647)$ for R2, and $(10.4295, 10.4295)$ for R3.

H GRPO Reward Definitions

H.1 Composite and Format Rewards

The three unweighted components in Equation 8 range over $[0, 0.75]$, $[0, 2]$, and $[0, 0.75]$ for format, answer, and structure, respectively. After applying the coefficients in Equation 8, the total reward ranges from 0 to 2.45.

The format reward evaluates the required response schema. We assign 0.10 for complete tags, 0.05 for nonempty sections, 0.05 for a reasoning length of 20–160 words, 0.10 for the ordered two-section format, and 0.10 for no nested XML tags. Satisfying all checks adds a further 0.35, giving a maximum format reward of 0.75. Checks apply to the assembled response, including inserted delimiters.

H.2 Task-Specific Answer Reward

For numerical tasks, MAE and RMSE are computed within each QA:

$$\begin{aligned} e_\tau &= 0.7\frac{\mathrm{MAE}(A, A^*)}{m_\tau} + 0.3\frac{\mathrm{RMSE}(A, A^*)}{b_\tau}, \\ r_{\text{answer}} &= 2\,\mathrm{clip}(1.5 - e_\tau, 0, 1). \end{aligned} \tag{12}$$

Here, m_τ and b_τ are manually specified reference error scales, rather than estimated dataset statistics. Their values are given in Appendix G. Unparseable answers or sequences with incorrect lengths receive zero. For classification, valid correct labels receive 2, valid incorrect labels receive 0.1, and invalid or missing answers receive zero. Explanation rewards equal twice the ROUGE-L score. An answer must include its closing delimiter to receive a nonzero answer reward.

H.3 Structural Reward

The structural reward evaluates whether the graph-token representations retain connectivity, direction, and road-distance information. These three components each contribute up to 0.25, giving a maximum structural reward of 0.75. The distance score is $\exp(-\,\mathrm{MAE}_d\,/100)$, where MAE_d is the distance error in meters; its contribution is multiplied by 0.25.