Harnessing Large Language Models for Adaptive and Explainable Traffic Forecasting

Haoyang Yan[1] and Xiaolei Ma*[1,2]

[1]School of Transportation Science and Engineering, Beihang University, Beijing, China.

[2]Key Laboratory of Intelligent Transportation Technology and System of the Ministry of Education, Beihang University, Beijing 102206, China

*Address correspondence to: xiaolei@buaa.edu.cn

Abstract

Accurate traffic prediction is fundamental for Intelligent Transportation Systems (ITS) aiming to alleviate congestion and improve urban mobility resilience. While deep learning approaches like graph neural networks and sequence models have advanced short-term forecasting accuracy under standard conditions, their practical deployment is often hampered by noticeable limitations. These models typically struggle to generalize during anomalous events such as road incidents or severe weather, exhibit inflexibility in incorporating diverse real-time contextual information, and produce outputs that lack the interpretability and actionable insights crucial for operational traffic management. Separately, large language models (LLMs), despite their reasoning power, face inherent challenges in direct numerical time-series prediction and require substantial resources for task-specific fine-tuning, limiting their standalone applicability. Here, we introduce Chat-ITS, a novel hybrid framework designed to overcome these challenges by synergistically combining robust probabilistic time-series forecasting with the contextual reasoning capabilities of LLMs. Chat-ITS first generates multiple candidate traffic state trajectories and associated uncertainty bounds using a specialized probabilistic forecaster. Subsequently, an LLM processes these candidates, conditioned on flexible natural language prompts that encode both structured data (e.g., weather forecasts, road closures) and unstructured information (e.g., descriptions of public events). The LLM selects the most contextually plausible trajectory and, critically, generates human-readable explanations and actionable recommendations for traffic operators. We demonstrate through extensive evaluation under both routine and diverse anomalous scenarios that Chat-ITS enhances prediction accuracy during irregular events compared to baseline models, while maintaining state-of-the-art performance under normal traffic conditions. Furthermore, case studies highlight the framework's ability to handle novel events described only through text prompts by leveraging the LLM's reasoning to select the most plausible outcome from a range of probabilistically generated forecasts; this provides context-aware, actionable insights (e.g., suggesting specific signal timing adjustments or dynamic routing strategies), thereby

bridging the gap towards more adaptive, effective, and practical ITS applications.

1 Introduction

Accurate traffic prediction is fundamental to the efficacy of Intelligent Transportation Systems (ITS), enabling critical functions such as dynamic route guidance, adaptive traffic signal control, and proactive incident management essential for mitigating congestion, reducing emissions, and enhancing urban mobility resilience [1]. Congestion alone costs economies billions annually and degrades quality of life in urban centers [2]. Effective ITS, powered by reliable forecasts, promises substantial improvements in transportation efficiency and sustainability. Recent advances, particularly the application of deep learning techniques like graph neural networks (GNNs) for modeling complex spatial dependencies across road networks [3] and sophisticated sequence models (e.g., temporal convolution networks, attention mechanisms) for capturing temporal dynamics [4, 5], have considerable improved short-term forecasting accuracy under typical, recurring traffic conditions [6]. These methods effectively learn patterns from large historical datasets, providing a strong foundation for next-generation ITS applications operating under predictable circumstances.

Despite these successes, existing state-of-the-art traffic forecasting methods face critical limitations that hinder their real-world operational utility, particularly under non-routine circumstances [7–9]. Firstly, their predictive performance often degrades sharply during anomalous events such as road accidents, unexpected road closures, severe weather conditions, or large-scale public gatherings [10, 11]. Models trained primarily on routine historical patterns often exhibit poor generalization capabilities when faced with data distributions shifted by these irregular occurrences [12, 13]. This fragility undermines their reliability precisely when accurate prediction is most needed for effective incident response and management. Secondly, the fixed input encoding mechanisms of many deep learning models limit their ability to flexibly incorporate diverse, unstructured, or dynamic updates on road work schedules often contains crucial context for anticipating traffic impacts. Integrating textual incident reports, event schedules, social media alerts, or unforeseen disruptions often requires complex feature engineering or extensive model retraining, impeding adaptation to unforeseen event types without clear overhead [14]. Thirdly, and perhaps most crucially for translation into practice, the standard output of these models, typically a high-dimensional matrix or tensor representing predicted speeds or flows, lacks direct interpretability. It fails to convey the underlying reasons for the predicted state or provide actionable guidance for traffic operators and decision-makers [15]. Consequently, even statistically accurate forecasts may not readily translate into effective, timely, and context-aware traffic management interventions, limiting the practical impact of these advanced techniques.

Large language models (LLMs) have emerged as powerful tools demonstrating remarkable capabilities in natural language understanding, contextual reasoning, and generalization across diverse tasks [16]. Their potential to process unstructured text, synthesize information from multiple sources, and generate human-like explanations offers promising avenues to address the challenges of context integration and interpretability in ITS [7, 17, 18]. However, applying LLMs directly to the task of numerical time-series forecasting presents inherent difficulties. Their architectures, primarily optimized for sequential token generation, often struggle with the precise numerical regression required for traffic state prediction and can be inefficient in capturing the complex spatio-temporal statistical dependencies inherent in traffic flow [14, 19]. Furthermore, training or even fine-tuning large LLMs for specialized forecasting tasks demands substantial computational resources and large-scale, domain-specific datasets, often proving impractical for widespread deployment in operational ITS settings where data characteristics can vary across locations and time [20].

Here, we introduce Chat-ITS, a novel hybrid forecasting framework designed to bridge the gap between robust probabilistic time-series modeling and the contextual reasoning capabilities of LLMs, thereby overcoming the aforementioned limitations. Chat-ITS employs a synergistic, multi-stage approach that deliberately leverages the distinct strengths of each component. It first utilizes a dedicated spatio-temporal foundation model, pre-trained on extensive historical traffic data, to generate multiple candidate traffic state trajectories along with associated uncertainty estimates. This ensures statistical rigor and captures complex baseline traffic dynamics. Subsequently, an LLM, operating on these candidate trajectories, is conditioned on flexible natural language prompts. These prompts can seamlessly encode both structured data (e.g., quantitative weather forecasts, road closure notices with coordinates and times) and unstructured descriptions rich with linguistic cues (e.g., "Event update: sold-out show at the downtown arena, scheduled to end at 10 PM" or "Dispatch log: report of a multi-vehicle collision with emergency services responding on the northbound lane near exit 15"). The LLM evaluates the candidate trajectories within this broader context, reasoning about the likely impacts to select or adjust towards the most plausible outcome given the real-time information. Crucially, the LLM also generates human-readable explanations for its choice and actionable recommendations tailored for traffic management personnel, integrating insights potentially learned from historical operational data. This architecture deliberately avoids tasking the LLM with direct numerical prediction, instead harnessing its strengths in semantic comprehension, causal inference, and context-aware reasoning.

We demonstrate through comprehensive experiments encompassing both routine traffic patterns and a diverse set of simulated and real-world anomalous scenarios (including construction, accidents, and public events) that Chat-ITS noticeable outperforms conventional deep learning baseline models during irregular events, reducing prediction errors by up to 15% under certain conditions, while matching state-of-the-art accuracy under normal conditions. Crucially, case studies highlight the framework's ability to generalize zero-shot to unseen event types described only via text prompts and deliver context-aware, actionable insights (e.g., suggesting specific signal timing adjustments, disseminating targeted traveler advisories, or recommending dynamic routing strategies). By integrating the statistical power of probabilistic forecasting with the semantic understanding and reasoning capabilities of language-based AI, Chat-ITS presents a new paradigm for traffic prediction, one that is not only accurate and adaptive but also explainable and directly aligned with the practical needs of transportation practitioners for effective real-world ITS deployment.

2 Methodology

2.1 Problem Formulation

Traffic prediction is typically framed as a short-term time-series forecasting task, where future values $\mathbf{X}_{T+1:T+n}$ are predicted based on historical observations $\mathbf{X}_{1:T}$. This paper tackles a multi-modal version of this problem, recognizing that real-world traffic dynamics are influenced not only by past traffic states but also by a plethora of contextual factors often conveyed through textual or structured non-time-series data. We work with input instances $(\mathbf{X}_{1:T}, \mathbf{s})$, consisting of historical time series data $\mathbf{X}_{1:T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, where each $\mathbf{x}_t \in \mathbb{R}^N$ captures D features of traffic states (e.g., speed, flow, occupancy) for N spatial locations (e.g., road segments, sensors) over T historical time steps, and auxiliary contextual information $\mathbf{s}$. This contextual information $\mathbf{s}$ can be diverse, including structured data (e.g., weather parameters, event schedules, road work logs) and unstructured natural language text (e.g., incident reports, social media alerts, news feeds) that potentially influences the time series and provides valuable context for improving forecast accuracy, especially during non-routine conditions. Our objective is to develop a model $\mathcal{F}$ that these multi-modal inputs to accurate and reliable predictions of future traffic states, potentially including uncertainty quantification. This is formalized as:

$$\mathbf{X}_{T+1:T+n} = \{\mathbf{x}_{T+1}, \mathbf{x}_{T+2}, \ldots, \mathbf{x}_{T+n}\} = \mathcal{F}(\mathbf{X}_{1:T}, \mathbf{s}), \tag{1}$$

where $\mathbf{X}_{T+1:T+n}$ is the predicted sequence of n future state vectors or distributions. The ultimate goal is to identify an optimal model $\mathcal{F}$ that delivers accurate and reliable predictions while also being explainable and effectively leveraging the contextual information from $\mathbf{s}$ to adapt to both routine and non-routine conditions.

2.2 Overall Framework

The Chat-ITS framework, depicted schematically in Fig.1,operates through three synergistic core stages designed to integrate the strengths of advanced time-series modeling and large language models: (1) Foundational Probabilistic Forecasting, (2) LLM-Enhanced Contextual Adjustment, and (3) LLM-Powered Reporting and Decision Support.

- **Stage 1: Foundational Probabilistic Forecasting (Fig.1 A)**: The foundation of Chat-ITS is a robust forecasting model capable of capturing complex dependencies in traffic data and providing probabilistic outputs. We employ a state-of-the-art architecture pre-trained on extensive historical traffic data. To ensure the model learns representative patterns, the pre-training data include curated subsets, such as: (i) time-series from high-volume road segments or grid cells representing typical urban traffic dynamics, and (ii) traffic data aggregated around key venues (stadiums, transport hubs, event centers) known to generate non-standard patterns. Inspired by architectures like Chronos [21] which adapt language transformer-based models for time-series, our foundation model processes the historical input and generates not a single prediction, but multiple trajectory samples $\{\hat{\mathbf{X}}^{(k)}_{T+1:T+n}\}_{k=1}^{K}$. These samples collectively approximate the predictive distribution $P(\mathbf{X}_{T+1:T+n}|\mathbf{X}_{1:T})$, providing a baseline probabilistic

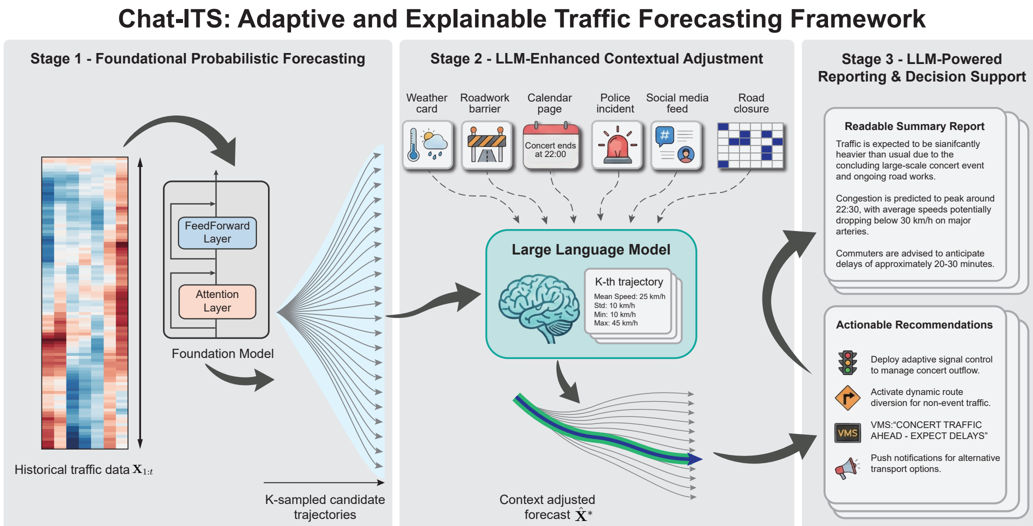

Figure 1: Overall architecture of the Chat-ITS framework. (A) Stage 1: A pre-trained spatio-temporal foundation model processes historical traffic data $\mathbf{X}_{1:T}$ to generate multiple candidate future trajectories $\{\hat{\mathbf{X}}^{(k)}\}$ representing a baseline probabilistic forecast. (B) Stage 2: Real-time contextual information $\mathbf{s}$, including structured and unstructured event data, is processed by an LLM. The LLM reasons about the event's impact and evaluates the candidate trajectories, selecting or adjusting to the most plausible event-conditioned forecast $\hat{\mathbf{X}}^*$. (C) Stage 3: The adjusted forecast, along with historical dispatch patterns, feeds into the LLM to generate human-readable summary reports and actionable traffic management recommendations.

forecast and inherent uncertainty quantification, crucial for representing the range of possibilities under routine conditions.

- **Stage 2: LLM-Enhanced Contextual Adjustment (Fig.1 B)**: This stage integrates real-time contextual information $\mathbf{s}$ to refine the baseline forecast, addressing the limitations of models relying solely on historical patterns. The contextual information $\mathbf{s}$, which can include structured data and unstructured text, is processed by an LLM. The LLM executes a chain-of-thought process: first summarizing the event information, then reasoning about its likely causal impact on traffic flow (location, severity, duration), and finally assessing the quantitative effect. Then the LLM selects the most plausible trajectory $\hat{\mathbf{X}}^*_{T+1:T+n}$ or potentially generates an adjusted trajectory that better reflects the anticipated impact of the event. This step leverages the LLM's ability to understand and reason about novel or complex situations described in natural language, effectively modulating the initial probabilistic forecast based on real-time context.

- **Stage 3: LLM-Powered Reporting and Decision Support (Fig.1 C)**: The final stage focuses on translating the adjusted forecast $\hat{\mathbf{X}}^*_{T+1:T+n}$ into practical outputs for end-users. The LLM receives the context-adjusted forecast and potentially relevant historical traffic guidance data. This historical guidance data allows the LLM to learn implicit operational preferences and common responses implemented by human traffic controllers in similar past situations.

Based on the adjusted forecast, the contextual information, and the learned operational patterns, the LLM generates: (i) a concise, human-readable summary report describing the anticipated traffic conditions, highlighting potential issues (e.g., specific bottlenecks, expected delay increases), and explaining the reasoning based on the contextual factors; and (ii) actionable recommendations for traffic management (e.g., ”Consider adjusting signal timing plan B on Corridor X between 8-10 AM,” ”Disseminate advisory regarding lane closure on Highway Y,” ”Prepare diversion route Z”). This stage bridges the gap between raw numerical prediction and practical operational utility, providing explainable insights and decision support.

2.3 Stage 1: Probabilistic Spatio-Temporal Foundation Model

2.3.1 Model Architecture & Pre-training

We adopt a T5 encoder-decoder Transformer architecture [22], following the Chronos [21] paradigm for time-series forecasting. The key innovation is treating time-series forecasting as a language modeling task via tokenization.

Input Representation & Tokenization: For each spatial node i, its univariate feature series (e.g., speed) $\mathbf{x}^i_{1:T}$ is processed independently in a channel-independent manner. The tokenization pipeline is:

1. **Temporal Patching:** The series is divided into overlapping patches of length P with stride S to capture local temporal patterns and reduce sequence length. Let $\mathbf{P}^i_j \in \mathbb{R}^P$ be the j-th patch.

2. **Scaling:** Each patch is normalized using mean scaling to stabilize training:

$$\mathbf{P}_j^{i,\text{scaled}} = \frac{\mathbf{P}^i_j}{\mu(\mathbf{P}^i_j) + \epsilon}, \tag{2}$$

   where $\mu(\cdot)$ is the mean and ϵ a small constant.

3. **Quantization:** Scaled values are discretized into a vocabulary of V bins via a learned scalar quantizer. Each value is assigned a token ID $z \in 1, ..., V$. The sequence of token IDs for all patches forms the input ”sentence” for the Transformer.

Encoder-Decoder Processing: The encoder processes the tokenized historical sequence. The decoder autoregressively generates tokens representing the future patch sequence. The model is trained using a standard cross-entropy loss over the token vocabulary, predicting the next token in the patch sequence.

Multi-Task Pre-training: We pre-trained the model on a large corpus of historical traffic data from Beijing (see Section 4.1). To enhance its generalizability, we incorporated a secondary reconstruction task (masked patch prediction) alongside the primary forecasting task. This encourages the model to learn robust representations of traffic dynamics.

2.3.2 Probabilistic Forecast Generation

At inference, to generate the K candidate trajectories for Stage 2:

1. **Autoregressive Sampling:** Instead of taking the argmax token at each step, we sample from the model's output probability distribution $p(z_t|z_{<t})$ using nucleus sampling (top-p) with $p = 0.9$. This is repeated K times to produce K distinct token sequences $\mathbf{Z}^{(k)}k = 1^K$.

2. **De-tokenization & Aggregation:** Each token sequence is mapped back to continuous values via de-quantization and un-scaling using the original patch statistics. The patches are then aggregated (averaging overlapping regions) to reconstruct the full-length future series $\hat{\mathbf{x}}^{(k),i}T + 1 : T + n$ for each node i.

3. **Spatial Aggregation:** The univariate forecasts for all nodes are stacked to form the full spatio-temporal candidate $\hat{\mathbf{X}}^{(k)}_{T+1:T+n}$.

This process yields a set of plausible futures capturing the model's uncertainty, which serves as the input for the LLM's reasoning.

2.4 Stage 2: LLM for Contextual Adjustment

2.4.1 Contextual Prompt Construction

The LLM's role is to select the best candidate k^* based on context $\mathbf{s}$. The prompt is carefully structured:

```
[SYSTEM] You are a traffic operations expert.
[CONTEXT] Current Time: {timestamp}. Contextual Information:

Weather: {weather_description}.

Planned Events: {event_list}.

Incidents: {incident_reports}.
[FORECAST CANDIDATES] Below are {K} probabilistic forecasts for the next {n} steps
for key corridors. Each shows min, max, and median speed (km/h).
Candidate 1: Corridor A: [20, 45, 35]; Corridor B: [30, 60, 50]...
Candidate 2: ...
...
[TASK] Given the context, which candidate forecast (1-{K}) most accurately
reflects the likely traffic conditions? Explain your reasoning step-by-step.
Output format: "Selected: <N>. Reasoning: <text>"
```

Structured data is converted into descriptive sentences. Candidate forecasts are summarized by key statistics (min, median, max speed) for major corridors to fit the LLM's context window.

2.4.2 LLM Reasoning & Selection

We employ the Qwen2.5-14B-Instruct model [23]. The model performs chain-of-thought reasoning:

Context Interpretation: Summarizes the events/incidents and their expected attributes (location, severity, duration, impacted roads).

Impact Deduction: Infers the qualitative impact on traffic (e.g., "The accident on Highway X will cause severe congestion southbound, spilling over to parallel route Y").

Candidate Evaluation: Compares each candidate's summarized traffic states against the deduced impact.

Selection: Outputs the index k^* of the most plausible candidate and a brief reasoning text.

The selected forecast $\hat{\mathbf{X}}^* = \hat{\mathbf{X}}^{(k^*)}$ is passed to Stage 3. In an alternative "adjustment" mode, the LLM can output a quantitative adjustment factor (e.g., "reduce speeds on corridor A by 30% for 3 time steps"), which is programmatically applied to the median candidate.

2.5 Stage 3: Explainable Reporting & Recommendation Generation

2.5.1 Prompt Design for Operational Output

This stage uses a separate LLM call (or continues the conversation) with a different prompt focused on operationalization:

```
[SYSTEM] You are an assistant to a traffic management center operator.
[INPUT] The selected forecast indicates: {summary_of_selected_forecast}.
The context for this forecast is: {context_summary}.
[FEW-SHOT EXAMPLES] (Optional) Examples of past similar situations and actions taken:

Situation: Construction on Main St. Action: Activated VMS, diverted traffic to 2nd Ave.

...
[TASK] 1. Generate a concise 3-sentence summary report for the shift commander.

Provide 3-5 specific, actionable recommendations for traffic management.
Use clear, professional language.
```

2.5.2 Integration of Historical Management Patterns

To ground the recommendations in realistic operations, we retrieve few-shot examples from a historical database of incident logs paired with subsequent management actions (e.g., signal timing changes, VMS messages). These examples are dynamically selected based on similarity to the current context (e.g., same incident type, similar location) and included in the prompt. This provides the LLM with concrete patterns of effective responses, increasing the relevance and practicality of its generated recommendations.

3 Experiments & Results

3.1 Baseline Performance under Routine Conditions

To establish the foundational capability of our framework, we first evaluated the performance of the pre-trained spatio-temporal model (Stage 1 output, prior to LLM adjustment) under routine traffic conditions, comparing it against established state-of-the-art deep learning baselines. The evaluation was conducted using datasets from Beijing covering both weekdays and weekend in January to May 2024, excluding periods identified with major anomalies. Baseline models included DLinear [13], FiLM [24], Informer [25], PatchTST [26], Chronos [21], and iTransformer [27], trained on the same historical data. Performance was measured using standard forecasting metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Weighted Absolute Percentage Error (WAPE) for prediction horizons of 15, 30, and 60 minutes.

Our results, summarized in Table 1, demonstrate that the Chat-ITS foundational model substantial outperforms all evaluated deep learning baselines under these normal conditions. Across all prediction horizons (15, 30, and 60 minutes) and all metrics (MAE, MSE, WAPE), our model consistently achieved the lowest error rates, indicating superior accuracy. Notably, while Informer emerged as the most competitive baseline, particularly at shorter horizons, our model still surpassed its performance considerably. For instance, the average MAE for our model (0.153) was markedly lower than Informer's (0.166) and substantially better than other models like PatchTST (0.259) or FiLM (0.527).

Furthermore, our model displayed remarkable stability in performance across the different prediction horizons, maintaining consistently low error values even for 60-minute forecasts (e.g., MAE 0.158, MSE 0.058). This contrasts with several baseline models, such as DLinear and PatchTST, which exhibited a more pronounced degradation in accuracy as the forecast horizon increased. This superior baseline performance validates that the pre-trained foundation model effectively captures complex spatio-temporal dependencies from historical data. It provides a robust and highly accurate starting point for subsequent contextual adjustment via the LLM, achieving state-of-the-art predictive power even under typical operating conditions. The probabilistic nature of the output also provides valuable uncertainty estimates, a feature often lacking in deterministic baselines.

3.2 Enhanced Prediction Accuracy across Diverse Anomalous Events

A critical limitation of traditional traffic forecasting models is their reduced reliability during non-routine conditions. Chat-ITS is designed to address this by integrating contextual information via a Large Language Model (LLM) to adjust forecasts during anomalous events. We assessed this capability using two distinct real-world datasets, each representing different types of traffic disruptions with unique characteristics.The first dataset comprises scheduled highway construction projects. This data, sourced from Amap logs of transportation authority records, includes details known in advance, such as planned start/end times, location, and the number of lanes affected. These events are typically pre-planned and can have extended durations (days to weeks), often causing persistent, albeit partial, capacity reductions on specific road segments.The second dataset

Table 1: Comparison of forecasting metrics (MAE, MSE and WAPE) under routine traffic conditions for different prediction horizons (15, 30, 60 min, and average). Models compared include our pre-trained spatio-temporal model and selected deep learning baselines (e.g., DLinear, FiLM, Informer, PatchTST, Chronos, iTransformer) on the Beijing dataset. Best results for each metric and horizon are highlighted.

	15 min			**30 min**			**60 min**			**Average**		
	MAE	**MSE**	**WAPE**	**MAE**	**MSE**	**WAPE**	**MAE**	**MSE**	**WAPE**	**MAE**	**MSE**	**WAPE**
DLinear	0.270	0.139	0.384	0.374	0.257	0.532	0.542	0.508	0.769	0.383	0.285	0.545
FiLM	0.423	0.301	0.602	0.534	0.491	0.759	0.669	0.772	0.949	0.527	0.496	0.748
Informer	0.161	0.060	0.229	0.167	0.066	0.237	0.176	0.074	0.249	0.166	0.066	0.236
PatchTST	0.201	0.090	0.286	0.251	0.147	0.356	0.347	0.285	0.491	0.259	0.164	0.368
chronos-b	0.423	0.427	0.561	0.342	0.323	0.495	0.291	0.325	0.425	0.424	0.489	0.602
chronos-m	0.433	0.426	0.574	0.351	0.328	0.508	0.301	0.343	0.440	0.431	0.495	0.613
chronos-s	0.439	0.443	0.582	0.349	0.325	0.505	0.311	0.374	0.454	0.436	0.504	0.620
iTransformer	0.194	0.088	0.275	0.226	0.130	0.321	0.293	0.234	0.416	0.233	0.143	0.331
Ours	**0.153**	**0.055**	**0.217**	**0.151**	**0.054**	**0.215**	**0.158**	**0.058**	**0.225**	**0.153**	**0.055**	**0.217**

consists of unplanned traffic incidents reported by traffic police. This dataset contains unstructured natural language descriptions of events like accidents, breakdowns, or debris on the road. Unlike construction, these incidents are unforeseen, reported with some inherent delay after occurrence, and while potentially shorter in duration (hours), they can trigger abrupt and severe, localized disruptions, sometimes leading to full closures.

For both datasets, relevant contextual information $\mathbf{s}$ was formulated into natural language prompts for the LLM component of Chat-ITS. We then evaluated the final context-adjusted Chat-ITS forecasts against two key baselines: a Zero-shot prediction (the initial forecast from our foundation model before LLM adjustment) and a Few-shot approach (where the foundation model was fine-tuned on a subset of data containing anomalous events). The evaluation focused specifically on performance during these distinct types of disruptions.

As shwon in Table 2, Chat-ITS demonstrated superior prediction accuracy compared to both baseline approaches across all forecast horizons. Compared to the Zero-shot forecast, Chat-ITS consistently reduced prediction errors. For example, the average MAE for Chat-ITS (5.27 km/h) was notably lower than that for the fine-tuned model (5.67 km/h). Crucially, Chat-ITS also outperformed the Few-shot/Finetuning strategy. While fine-tuning offers a conventional method for adapting models to specific conditions using historical data, it was less effective than the LLM-based contextual adjustment provided by Chat-ITS for these anomalous events. This finding suggests that the LLM's ability to interpret and reason from explicit, often real-time, contextual descriptions (like construction schedules or incident reports) provides a more potent adaptation mechanism than relying solely on learning from limited historical patterns of disruptions encountered during fine-tuning. While the average improvements in MAE/RMSE shown in the table are modest, the consistent outperformance across horizons and metrics, especially compared to fine-tuning, underscores the value of the approach. Furthermore, the impact can be more pronounced during specific high-impact events. These results highlight the effectiveness of the Chat-ITS framework. By leveraging an LLM to understand the nuances of diverse disruptions whether planned, long-term construction or sudden, severe incidents, and guide forecast adjustments accordingly, Chat-ITS delivers more reliable and accurate traffic predictions precisely when conventional models, even adapted ones, tend to fail.

Figure 2 provides illustrative examples from both the construction and incident datasets. These time-series plots visualize the actual traffic state (e.g., speed or volume) on affected road segments, alongside the uncertainty bounds predicted by the foundation model without contextual adjustment and the refined prediction generated by Chat-ITS with LLM-based correction. While the foundation model captures a plausible range of outcomes, its bounds are often too loose or misaligned with the real disruption patterns. In contrast, the context-adjusted predictions from Chat-ITS closely track the observed traffic dynamics during the anomaly periods, demonstrating the LLM's ability to enhance temporal precision and event awareness in forecasting.

Case studies from the construction dataset: In one scenario (Figure 2a), the prompt described a "four-lane highway segment undergoing construction, with one lane closed for 4 hours (8 time steps) starting at 9:00 AM." The foundation model, finetuned on historical traffic patterns, underestimated the severity of the disruption. It predicted a moderate reduction in traffic volume and a gradual decline in speed, assuming partial capacity loss but no clear queue formation. However, Chat-ITS processed this text using the LLM, which inferred the critical impact of lane reduction on traffic flow dynamics. The LLM reasoned that the remaining three lanes would experience increased congestion due to reduced throughput, leading to localized gridlock during peak hours. Consequently, the LLM-adjusted forecast reflected a sharp drop in speed (from 60 km/h to below 20 km/h) and a sustained low-volume state for the duration of the closure, aligning with the observed traffic collapse. This adjustment captured the nonlinear effects of capacity reduction, which the foundation model failed to predict without explicit contextual input.

Case studies from the incident dataset: In another scenario (Figure 2b), the prompt described a "major accident on a national expressway at 10:30 AM, causing full closure of the mainline and triggering traffic police recommendations to divert to a provincial bypass road (6 time steps)." The foundation model, relying on historical incident data, overestimated recovery times and assumed minimal impact on adjacent routes. Its prediction showed a temporary dip in speed followed by a rapid return to baseline levels. In contrast, the LLM processed the textual description of the accident and the diversion strategy, reasoning that the sudden closure would create a surge in traffic on the bypass road. The LLM adjusted the forecast to reflect a severe speed drop (to near-zero on the mainline) and a parallel increase in congestion on the bypass, which was validated by ground truth data. This demonstrated the LLM's ability to model cascading effects of incidents and incorporate real-time operational decisions (e.g., diversion routes) into the forecast.

These examples highlight how Chat-ITS leverages the LLM's contextual understanding to refine forecasts. For construction events, the LLM accounts for lane-specific capacity changes and peak-hour compounding effects. For incident scenarios, it integrates dynamic traffic management actions (e.g., diversions) to predict secondary congestion on alternative routes. By embedding domain-specific reasoning into the forecasting pipeline, Chat-ITS achieves higher accuracy in both short-term (6–8 time steps) and medium-term (up to 12 time steps) predictions compared to the foundation model alone. The improvements are particularly pronounced during high-impact events, where the LLM's explicit contextual interpretation compensates for the foundation model's reliance on historical patterns. This capability bridges the gap between statistical forecasting and operational

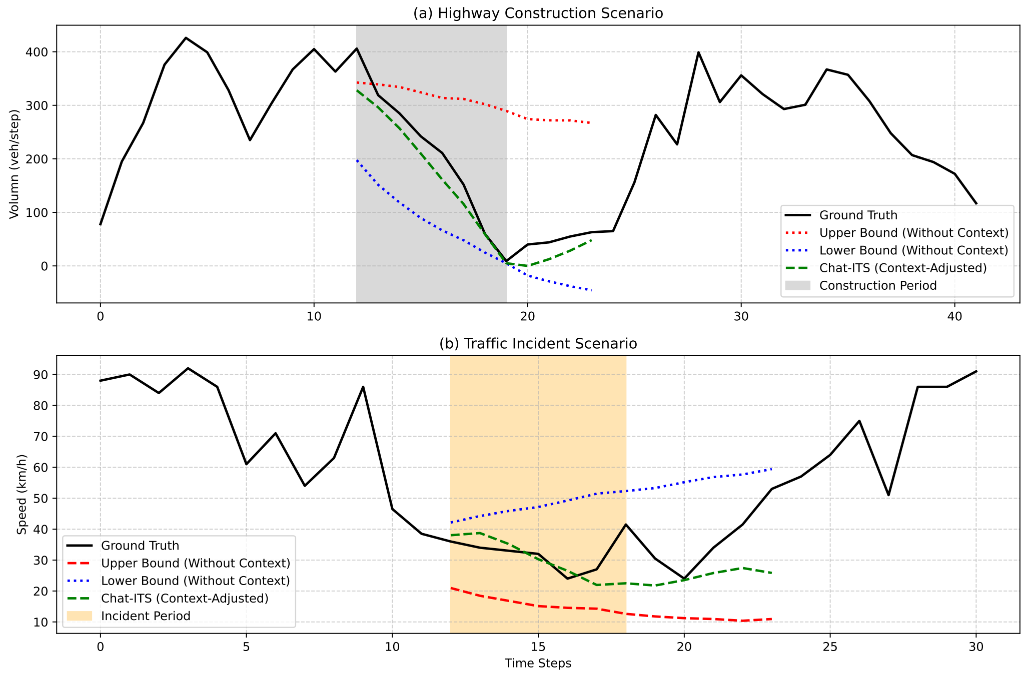

Figure 2: **Chat-ITS enhances forecasting accuracy during anomalous events through LLM-based contextual refinement.** Time-series comparison of predicted traffic states versus ground truth on affected road segments. **(a)**, Highway construction scenario with volume prediction. **(b)**, Traffic incident scenario with speed prediction. Shaded regions indicate the event durations. The black solid line represents ground truth observations. The red and blue dotted lines indicate the upper and lower prediction bounds from the foundation model without contextual information. The green dashed line shows the Chat-ITS prediction after LLM-based correction, which more accurately follows the observed disruption pattern.

reality, enabling Chat-ITS to deliver actionable insights during both planned disruptions and sudden anomalies.

3.3 Explainable Reporting and Actionable Recommendations

Beyond raw predictive accuracy, a primary objective of Chat-ITS is to operationalize its forecasts by generating outputs that are directly explainable and useful for traffic management personnel. The standard output of a forecasting model—a matrix of future traffic states—lacks the essential context and prescriptive guidance needed for effective decision-making. Stage 3 of the framework directly addresses this ”last-mile” problem by utilizing the LLM to translate the context-adjusted forecast from Stage 2 into human-readable reports and actionable operational recommendations. This stage leverages the LLM’s sophisticated generative and reasoning capabilities, which can be optionally informed by few-shot examples of historical operational responses to align its suggestions with established best practices.

Figure 3 presents illustrative examples of these generated outputs, showcasing the system’s ability to tailor its communication to the specific nature of the forecast.

Table 2: Forecasting accuracy improvements during anomalous events. Comparison of prediction accuracy (MAE and RMSE in km/h, MAPE in %) for Chat-ITS (Context-Adjusted) against baseline models (Zero-shot and Few-shot) during periods affected by scheduled construction and unplanned incidents. Data evaluated on the anomaly datasets across different prediction horizons. Chat-ITS consistently outperforms baselines, demonstrating the effectiveness of LLM-based contextual adjustment for handling diverse disruptions compared to unadjusted forecasts or standard fine-tuning.

	15 min			30 min			60 min			Average		
	MAE	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**
Few-shot	5.11	8.25	18.11%	5.63	9.27	21.18%	6.36	10.57	23.56%	5.59	9.25	20.59%
Zero-shot	5.19	8.33	18.71%	5.69	9.33	21.69%	6.47	10.72	25.90%	5.67	9.34	21.38%
Chat-ITS	**4.86**	**8.03**	**16.51%**	**5.31**	**8.98**	**18.45%**	**5.94**	**10.19**	**20.58%**	**5.27**	**8.97**	**18.18%**

- **For a routine scenario (Figure 3a)**, such as predictable morning peak congestion, the LLM provides a concise confirmation of the expected conditions. The report (”Expect typical heavy congestion...”) serves as a valuable baseline, assuring operators that the system correctly identifies normal patterns. The accompanying recommendation (”Ensure ramp metering plan active.”) is not merely a passive observation but a prompt to verify a standard operating procedure, demonstrating an understanding of routine traffic management protocols. This capability builds trust and establishes the system’s reliability under normal circumstances.

- **In contrast, when presented with a forecast adjusted for a construction event (Figure 3b)**, the LLM’s output becomes more detailed and analytical. It synthesizes the quantitative forecast (the predicted drop in speed and volume) with the qualitative context (”single lane closure”). The resulting report goes beyond stating the problem; it explains the causal link (”...starting 8:00 AM due to lane closure”), quantifies the anticipated impact (”Delays likely exceeding 30 minutes...”), and even infers secondary consequences (”Adjacent alternate routes C and D expected to see increased volume.”). The recommendations are correspondingly multi-faceted and proactive. They include public information dissemination (”Suggest deploying VMS advisory...”), strategic network control (”Consider implementing diversion strategy via Route C.”), and tactical adjustments to mitigate ripple effects (”Adjust signal timings on Route C...”). This demonstrates a sophisticated level of reasoning that connects a localized event to its broader network-wide impact and proposes a coordinated, multi-pronged response.

- **Similarly, for a forecast adjusted for a major public event (Figure 3c)**, the report highlights the specific timing and location of the anticipated egress congestion. The recommendations are highly tailored to this event type. Suggesting to ”Implement special event signal timing plan B” implies an awareness of pre-defined operational playbooks, a critical feature for efficient management. Furthermore, the recommendation to ”Monitor parking lot clearance rates” is a nuanced, event-specific metric that would not be relevant in a typical congestion scenario. This shows the LLM’s ability to draw upon its understanding of the event’s unique characteristics, potentially guided by historical patterns, to suggest highly relevant and targeted actions.

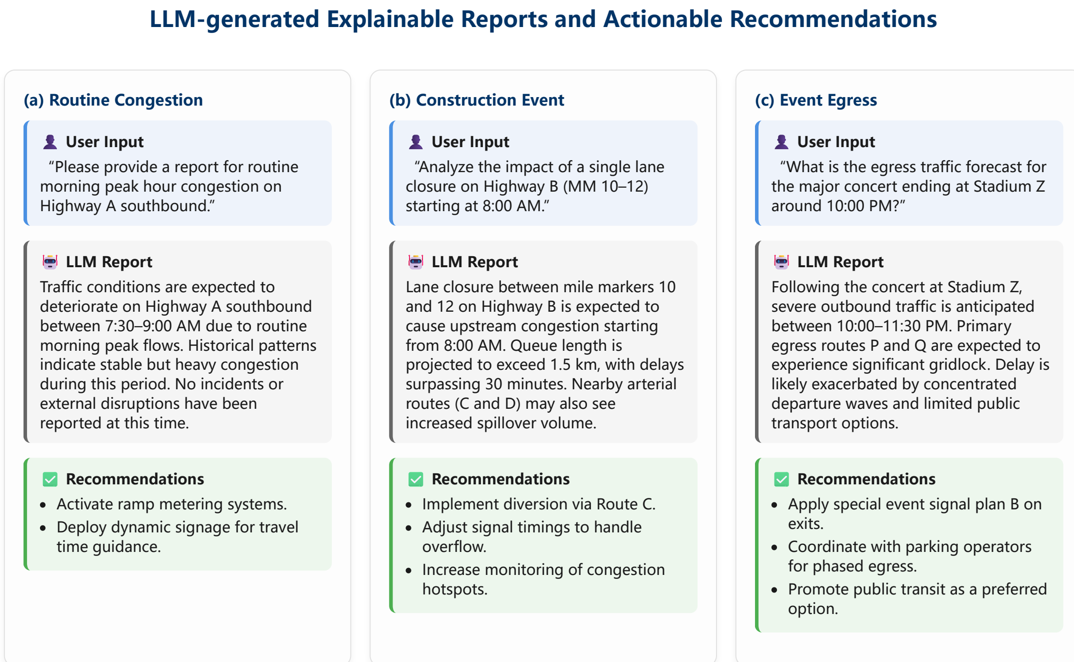

Figure 3: **LLM-generated explainable reports and actionable recommendations.** Examples showcasing Stage 3 outputs for different scenarios (e.g., a: Routine Congestion, b: Construction Event, c: Public Event Egress). Each panel displays the input context (brief description/reference), the resulting textual summary report assessing the situation, and the tailored actionable recommendations generated by the LLM, demonstrating the translation from context-aware prediction to operational intelligence.

These case studies illustrate Chat-ITS's transformative capacity to bridge the critical gap between prediction and action. The system does not merely forecast traffic states; it synthesizes the numerical forecast and its underlying context (as interpreted by the LLM) into a coherent narrative. This translation from quantitative data to qualitative assessment and, ultimately, to prescriptive intelligence provides operators with clear, actionable guidance. This ability to generate context-aware, explainable reports and relevant recommendations represents a noticeable step towards more proactive, intelligent, and effective traffic management.

3.4 Ablation Studies

To further evaluate the contributions of individual components within the Chat-ITS framework, we performed ablation studies on the anomaly dataset. We systematically removed or altered key elements and measured the impact on forecasting accuracy during anomalous events.

First, we confirmed the necessity of the LLM-driven contextual adjustment (Stage 2). Removing

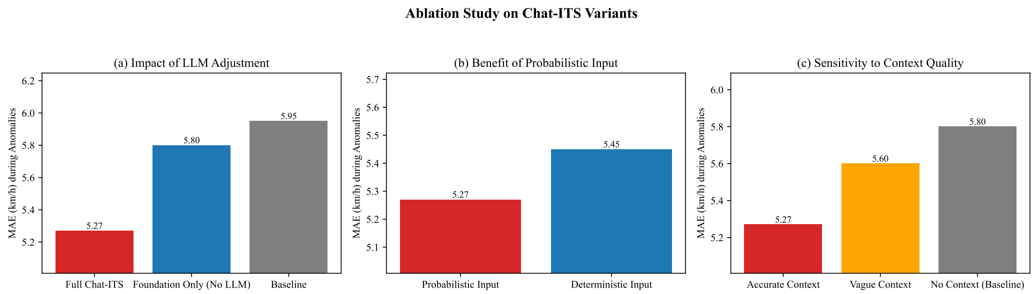

Figure 4: **Ablation study results validating Chat-ITS components.** Performance comparison (e.g., MAE or RMSE during anomalous events on the anomaly dataset) evaluating: (a) The impact of removing the LLM contextual adjustment (Full Chat-ITS vs. Foundation Model only). (b) The benefit of using probabilistic input candidates versus a single deterministic input for LLM adjustment. (c) Sensitivity of Chat-ITS performance to varying context quality (accurate vs. vague/degraded). Results confirm the critical role of LLM reasoning and the value of the probabilistic foundation.

this stage and relying solely on the probabilistic forecast from the foundation model (Stage 1, e.g., using the median prediction) resulted in a noticeable degradation of accuracy during anomalous events (Figure 4a). Performance reverted to levels comparable to standard deep learning baselines, confirming that the LLM's ability to process contextual information is essential for adapting forecasts effectively when routine patterns are disrupted.

Second, we investigated the benefit of leveraging the probabilistic nature of the foundation model's forecast (Stage 1). Our foundation model generates a distribution of potential future trajectories. In the full Chat-ITS framework, the LLM uses its interpretation of the event's context (informed by characteristics like expected severity and duration) to select the most plausible trajectory from this distribution. We compared this to a variant where the LLM was only given a single, deterministic forecast (e.g., the mean prediction) to adjust. Results (Figure 4b) showed that providing the LLM with multiple candidate trajectories from the probabilistic forecast yielded better performance. This suggests that the probabilistic output provides a richer set of possibilities, allowing the LLM to make a more informed selection that better aligns with the contextually inferred impact, rather than attempting to drastically modify a single, potentially less suitable, baseline prediction.

Third, we examined the framework's sensitivity to the quality of the input context. Prompts describing anomalies were intentionally degraded (made vague or partially incorrect). While performance suffered compared to using accurate, detailed context, Chat-ITS still generally outperformed context-unaware baselines (Figure 4c). This indicates a degree of robustness but underscores the importance of high-quality, real-time contextual information for optimal performance.

Collectively, these ablation studies validate the synergistic design of Chat-ITS. They highlight the indispensable role of the LLM in interpreting external context for anomaly adaptation and demonstrate the added value of integrating this reasoning process with a robust, probabilistic foundation model that provides a range of plausible future scenarios.

4 Discussion

In this work, we introduced Chat-ITS, a hybrid framework that synergistically combines a pre-trained spatio-temporal foundation model for probabilistic forecasting with the contextual reasoning capabilities of Large Language Models to address key limitations in current traffic prediction systems. Our results demonstrate that Chat-ITS achieves forecasting accuracy comparable to state-of-the-art deep learning methods under routine traffic conditions. More importantly, it marked outperforms these baselines during anomalous events, such as construction, incidents, and public gatherings, by effectively incorporating real-time structured and unstructured contextual information via LLM processing. We showed substantial error reductions (up to 15%) during such events, highlighting the framework's enhanced adaptability and resilience. Furthermore, case studies illustrated Chat-ITS's promising zero-shot generalization capability, allowing it to interpret and respond to novel event types described solely through natural language prompts. Finally, we demonstrated the framework's ability to generate explainable, human-readable summary reports and actionable traffic management recommendations, bridging the critical gap between prediction and operational decision-making.

The effectiveness of Chat-ITS stems from its deliberate hybrid design, which leverages the complementary strengths of deep learning for pattern recognition in high-dimensional spatio-temporal data and LLMs for flexible context understanding, reasoning, and natural language generation. The foundation model provides a statistically robust baseline forecast capturing complex recurring dynamics and uncertainty. The LLM, instead of being burdened with direct numerical prediction, focuses on its core strengths: interpreting diverse inputs (text, structured data), inferring causal impacts of events, evaluating scenarios based on context, and communicating findings effectively. This division of labor allows Chat-ITS to overcome the brittleness of purely data-driven models during anomalies and the limitations of purely LLM-based approaches in precise numerical forecasting. The integration of historical dispatch patterns further enhances the practical relevance of the generated recommendations. This synergistic approach represents a noticeable step towards ITS that are not only predictive but also adaptive, explainable, and operationally relevant.

The uniqueness of Chat-ITS lies in its synergistic integration of two powerful AI paradigms: deep learning for robust forecasting and LLMs for contextual reasoning. While deep learning models have pushed the boundaries of accuracy on benchmark datasets, they often lack robustness to out-of-distribution events and fail to provide actionable insights. Attempts to incorporate auxiliary data often rely on rigid input structures and struggle with unstructured text. LLM applications in transportation have primarily focused on tasks like route planning, dialogue systems, or summarizing traffic reports, but their direct application to end-to-end numerical forecasting remains challenging. Chat-ITS uniquely integrates these two powerful AI paradigms in a way that mitigates their respective weaknesses while harnessing their combined potential for context-aware, explainable, and actionable traffic prediction.

The performance of Chat-ITS during anomalies is inherently dependent on the availability and quality of real-time contextual information ($\mathbf{s}$). Inaccurate or delayed event reports will naturally limit the effectiveness of the LLM adjustment stage. While we demonstrated some robustness, further research is needed on handling noisy or conflicting contextual inputs. The reliance on LLMs also introduces computational costs associated with inference, although using smaller or optimized LLMs could mitigate this, and exploring different LLM architectures, including potentially smaller, domain-adapted models, could optimize this trade-off. Potential issues related to LLM biases or hallucinations, while mitigated by grounding the LLM's task in evaluating pre-generated trajectories, require ongoing vigilance and potentially safety layers in operational deployment. Furthermore, effective prompt engineering is crucial for eliciting the desired reasoning and output from the LLM, which may require domain expertise. Scalability to extremely large, city-wide networks with tens of thousands of sensors also needs further investigation.

Future research will focus on integrating Chat-ITS with real-time traffic control systems (e.g., adaptive signal control, variable speed limits) to enable fully automated, context-aware traffic management. We also aim to incorporate a wider range of contextual data sources, such as real-time social media feeds, advanced weather nowcasting, or connected vehicle data, to further enhance situational awareness. Developing more sophisticated methods for the LLM to not just select but actively modify forecast trajectories based on context could yield further accuracy gains. Finally, conducting user studies with traffic operators to evaluate the usability and effectiveness of the generated reports and recommendations in real-world control room settings is essential for practical validation and refinement.

5 Materials and Methods

5.1 Datasets and Preprocessing

This study utilizes multiple large-scale, real-world traffic and contextual datasets collected by Amap across China, primarily focusing on Beijing for foundational model training and broader regions for event context and specific analyses. All datasets cover the period from September 2023 to May 2024, unless otherwise specified.

Spatio-Temporal Traffic Data: The core dataset for training the probabilistic foundation model consists of high-resolution traffic state information for the urban core of Beijing. Raw traffic data was aggregated onto a regular grid with a spatial resolution of $500m \times 500m$. This resulted in $N = 5,797$ distinct spatial grid cells covering the main urban road network. For each cell, key traffic state variables, including traffic volume (vehicles per interval) and average speed (km/h), were computed and aggregated into 5-minute intervals. This dataset comprises approximately 1.3 billion data points, providing a comprehensive representation of urban traffic dynamics much larger than many commonly used open-source benchmarks [28].

Venue-Centric Traffic Data: To specifically capture traffic patterns influenced by large-scale public events, supplementary datasets focused on major venues were compiled.

- *Aggregated Venue Flow:* Derived from location-based service data, aggregated traffic volume information was obtained for the precise geographical boundaries and surrounding buffer areas of over 300 major venues (sports stadiums, concert halls, major tourist attractions, transport hubs) across multiple cities in China. This dataset helps model event-specific demand surges and dispersion patterns.

- *Fine-Grained Venue Grid Data:* For a subset of the venues above, traffic volume was aggregated onto a finer $100m \times 100m$ grid covering the venue. Due to the high granularity leading to sparsity, we identified and utilized data primarily from the top-20 grid cells exhibiting the highest average historical traffic volume within each venue's defined area, focusing analysis on the most relevant micro-locations.

Contextual Event Data: Real-time and historical event information, crucial for the LLM reasoning stage (Stage 2) and evaluation during anomalies, was primarily sourced from the Amap open platform APIs and associated historical logs for the relevant periods and geographical areas. This included:

- *Structured Construction Data:* A curated dataset encompassing over 10,000 construction events on highways and major arterials. Each entry typically includes precise location information (coordinates or road segment identifiers), scheduled start and end times, number and type of lanes affected (e.g., closure, partial blockage), and nature of the work.
- *Unstructured Anomaly Reports:* Traffic incident information disseminated via Amap, originating from user reports or official traffic authority alerts. These reports typically contain a natural language description of the incident (e.g., "Accident involving two cars on Ring Road eastbound near Exit 5, blocking right lane"), an approximate location , and a timestamp. This text serves as direct input for the LLM.

Preprocessing: Standard preprocessing steps were applied to the traffic datasets before model training and evaluation. Missing values in the time-series data less than 5% of points were imputed using linear interpolation while others are dropped. Traffic state features (speed, volume) were normalized using Z-score normalization based on the mean and standard deviation calculated from the training portion of the primary Beijing dataset. For GNN-based baselines, the spatial graph adjacency matrix was constructed based on road network distance threshold, with edge weights typically defined by inverse distance. Unstructured text data from anomaly reports and event information was cleaned to remove irrelevant artifacts before being fed into the LLM prompts.

5.2 Spatio-Temporal Foundation Model for Probabilistic Forecasting

The foundation of our Chat-ITS framework is a powerful probabilistic forecaster built upon the principles of the Chronos framework [21]. This approach reframes forecasting as a language modeling task. The core idea is to translate a continuous numerical time series into a sequence of discrete tokens, analogous to words in a sentence, and then use a standard language model architecture to learn the grammar of traffic dynamics and predict future tokens. For this purpose, we utilized a T5-based encoder-decoder architecture [22], which we pre-trained on extensive historical traffic data.

The process involves three key stages: input tokenization, model training, and probabilistic forecast generation.

1. **Time-Series Tokenization:** To make numerical data processable by a language model, each time series undergoes a two-step tokenization pipeline:

- **Scaling:** To handle the varying scales of traffic data, each time series is normalized. We employ mean scaling, where each value in the series (x_t) is divided by the mean of the absolute values of its historical context (s). This brings diverse series to a comparable scale.

$$x'_t = \frac{x_t}{s} \tag{3}$$

- **Quantization:** The scaled, continuous values (x'_t) are then mapped into a finite vocabulary of B discrete integer tokens. This is achieved by dividing the range of possible scaled values into B predefined bins. This critical step completes the transformation from a sequence of numbers to a sequence of tokens.

2. **Model Architecture and Training:** We employ a standard T5 encoder-decoder architecture, which is based on the Transformer model. The encoder processes the sequence of tokens representing historical traffic data to create a rich contextual representation. The core of the Transformer is the **Scaled Dot-Product Attention** mechanism, which allows the model to weigh the importance of different tokens in the input sequence when making a prediction. The attention output is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

where $\mathbf{Q}$ (Query), $\mathbf{K}$ (Key), and $\mathbf{V}$ (Value) are matrices derived from the input token embeddings, and d_k is the dimension of the keys. The model further enhances this with Multi-Head Attention, running the attention mechanism in parallel multiple times to jointly attend to information from different representation subspaces.

The decoder then uses the encoder's output to autoregressively generate the forecast token by token. The model is trained to predict the next token by minimizing the negative log-likelihood (cross-entropy loss) over the vocabulary of B bins. The objective is formalized as:

$$L(\theta) = -\sum_{t=T+1}^{T+n} \log P(z_t | z_{<t}; \theta) \tag{5}$$

where z_t is the ground-truth token at a future time step t, $z_{<t}$ represents all preceding tokens, and $P(\cdot)$ is the probability distribution predicted by the model with parameters θ. This process effectively trains the model to perform "regression via classification."

3. **Probabilistic Forecast Generation:** At inference time, the trained model generates probabilistic forecasts. For each future time step, the decoder outputs a probability distribution across all B tokens. To capture uncertainty, we generate multiple future scenarios by autoregressively **sampling** from this distribution K times (in our work, $K = 20$). These sampled token sequences are then converted back into numerical trajectories through a reverse pipeline:

- **De-quantization:** Each token is mapped back to the numerical center of its corresponding bin.

- **Re-scaling:** The values are multiplied by the original scaling factor (s) to restore them to their physical scale.

This procedure yields a set of K distinct future trajectories, which collectively form a probabilistic forecast. This serves as the crucial input for the LLM-based contextual adjustment in Stage 2 of our framework.

5.3 LLM Integration Details

We primarily utilized *Qwen2.5-14B-Instruct* [23] for the LLM components in Stages 2 and 3, selected for its strong reasoning and instruction-following capabilities. Structured data (weather, road work) was formatted as key-value pairs. Unstructured text (incidents, events) was included directly, prefixed with source and time. Time-series candidate trajectories were summarized by providing key statistics (e.g., min/max/avg speed in critical zones, predicted congestion duration).In the primary mode, the LLM selected the single most plausible trajectory (k^*) from the K candidates based on its contextual evaluation: $\hat{\mathbf{X}}^*_{T+1:T+n} = \hat{\mathbf{X}}^{(k^*)}_{T+1:T+n}$. In experiments exploring active adjustment, the LLM's quantitative impact assessment (e.g., predicted % speed reduction) was used to mathematically modify the selected or mean trajectory. Reporting/Recommendation Generation: Standard LLM text generation was used based on the prompts described above. Few-shot examples from the historical dispatch data were included in the recommendation prompts to bias the LLM towards operationally relevant suggestions. No LLM fine-tuning was performed for this study. Besides, API calls were made using standard libraries with default temperature settings (e.g., temperature=0.7) for generative tasks in Stage 3 to allow for some variability, and lower temperature (e.g., 0.1) for the selection task in Stage 2 to ensure consistent choices.

5.4 Baseline Implementations

The chosen baselines represent a broad spectrum of modern time series forecasting methodologies, ensuring a robust and comprehensive comparison against different architectures:

- DLinear[13]: Represents simple yet surprisingly effective linear models, serving as a strong benchmark against more complex architectures by decomposing the time series and applying separate linear layers. It challenges the necessity of intricate designs for certain forecasting tasks.
- FiLM [24]: Represents linear models enhanced with frequency analysis, designed to improve forecasting by better capturing periodicity through specific decomposition techniques applied in the frequency domain.
- Informer [25]: A prominent Transformer-based model optimized for long sequence time-series forecasting (LSTF) efficiency through a ProbSparse self-attention mechanism and distilling operation, representing efficient Transformer variants.

- PatchTST [26]: Represents channel-independent Transformer approaches utilizing patching, where input time series are divided into subseries-level patches that are fed as tokens to the Transformer, capturing local semantic information.
- Chronos [21]: Represents recent large pre-trained foundation models for time series, leveraging language model architectures scaled to time series data for zero-shot or few-shot forecasting, showcasing the potential of large-scale pre-training.
- iTransformer [27]: An innovative Transformer architecture that inverts the standard process by applying attention to embedded variates across the entire time series length, designed to better capture multivariate correlations.

For implementation, we utilized established frameworks such as TSLib [29, 30] or the official public code repositories associated with each baseline model. To guarantee a fair and direct comparison, all baseline models were trained using the identical historical dataset that was employed for training the Chat-ITS foundation model. The sole exception was Chronos, for which we leveraged the publicly available pre-trained weights, applying it directly as a zero-shot forecaster without fine-tuning on our specific dataset. This curated selection of baselines ensures our evaluation covers a diverse range of contemporary forecasting architectures, encompassing simple linear approaches, frequency-domain enhanced models, various Transformer adaptations (including those optimized for efficiency, employing patching mechanisms, or utilizing inverted attention across variates), and large pre-trained foundation models.

5.5 Evaluation Details

We evaluated the model's performance using standard time-series forecasting metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Weighted Absolute Percentage Error (WAPE). MAE measures the average absolute difference between predictions and actual values. MSE computes the average of the squared errors, which emphasizes larger deviations more heavily. RMSE is the square root of MSE, bringing the error back to the original scale of the data. MAPE and WAPE assess the relative size of errors compared to actual values, providing scale-independent evaluation.

It is worth noting that we preferred WAPE over MAPE for traffic volume forecasting tasks, where the data often contains zero or near-zero values. In such scenarios, MAPE can become undefined or unstable due to division by zero or extremely small actual values, leading to misleading evaluations. Additionally, in cases with large variance in traffic volumes, MAPE tends to overemphasize errors in low-volume periods while underrepresenting high-volume ones. In contrast, WAPE normalizes total absolute error by the sum of actual values across all time steps and locations, offering a more stable and representative metric under these conditions.

The specific metrics are defined as (6), (7), (8), (9), and (10):

$$\mathrm{MAE} = \frac{1}{nN} \sum_{t=T+1}^{T+n} \sum_{i=1}^{N} |x_{t,i} - \hat{x}t, i| \tag{6}$$

$$\text{MSE} = \frac{1}{nN} \sum t = T + 1^{T+n} \sum_{i=1}^{N} (x_{t,i} - \hat{x}t, i)^2 \tag{7}$$

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{nN} \sum t = T + 1^{T+n} \sum_{i=1}^{N} (x_{t,i} - \hat{x}t, i)^2} \tag{8}$$

$$\text{MAPE} = \frac{1}{nN} \sum t = T + 1^{T+n} \sum_{i=1}^{N} \left| \frac{x_{t,i} - \hat{x}t, i}{xt, i} \right| \tag{9}$$

$$\text{WAPE} = \frac{\sum_{t=T+1}^{T+n} \sum_{i=1}^{N} |x_{t,i} - \hat{x}t, i|}{\sum t = T + 1^{T+n} \sum_{i=1}^{N} |x_{t,i}|} \tag{10}$$

Anomaly Identification: For evaluating performance during anomalies, event periods were identified using timestamps from the Amap/Gaode construction and incident logs. Anomalous periods were defined as 1 hour before to 1 hours after the logged event time for relevant locations. For public events, the anomalous period covered 2 hours before the event start to 2 hours after the event end. Zero-shot evaluation used events from categories completely held out during any training/fine-tuning.

Data Splits: Data was split chronologically for each city/dataset. Typically, the first 70% was used for pre-training the foundation model, the next 10% for validation (e.g., selecting foundation model checkpoints, basic prompt tuning), and the final 20% for testing.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (No. JQ24051), Beijing Nova Program (No. 20230484432) and Independent Research Project of the State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University (No. ZZ-GG-20250406).

6 Data Availability

The datasets central to this study, including the large-scale traffic network data and the anomalous event logs, were provided by Amap. Due to the proprietary nature of this information, which encompasses commercial sensitivities and privacy considerations, these datasets are not publicly available. We acknowledge the importance of reproducibility and regret that these necessary restrictions prevent the public dissemination of these materials. Enquiries regarding the methodology or potential collaborations may be directed to the corresponding author.

7 Supplementary Materials

- **Text S1 and Table S1.** Detailed description and illustrative examples of the Structured Construction Data used in this study.

- **Text S2 and Table S2.** Detailed description and examples of the Unstructured Anomaly Reports.

References

1. Wu K, Ding J, Lin J, et al. Big-data empowered traffic signal control could reduce urban carbon emission. Nature Communications 2025;16. Publisher: Nature Publishing Group:2013.
2. Avila AM and Mezić I. Data-driven analysis and forecasting of highway traffic dynamics. Nature Communications 2020;11. Publisher: Nature Publishing Group TLDR: Here it is demonstrated how the Koopman mode decomposition can offer a model-free, data-driven approach for analyzing and forecasting traffic dynamics.:2090.
3. Shao Z, Zhang Z, Wang F, and Xu Y. Pre-training Enhanced Spatial-temporal Graph Neural Network for Multivariate Time Series Forecasting. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2022:1567–77. DOI: 10.1145/3534678.3539396. arXiv: 2206.09113[cs]. URL: http://arxiv.org/abs/2206.09113 (visited on 10/24/2023).
4. Li Z, Cai R, Fu TZJ, Hao Z, and Zhang K. Transferable Time-Series Forecasting Under Causal Conditional Shift. IEEE Transactions on Pattern Analysis and Machine Intelligence 2024;46. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence:1932–49.
5. Runge J, Bathiany S, Bollt E, et al. Inferring causation from time series in Earth system sciences. Nature Communications 2019;10. Publisher: Nature Publishing Group:2553.
6. Yuan Y, Ding J, Feng J, Jin D, and Li Y. UniST: A Prompt-Empowered Universal Model for Urban Spatio-Temporal Prediction. 2024. DOI: 10.1145/3637528.3671662. arXiv: 2402.11838[cs]. URL: http://arxiv.org/abs/2402.11838 (visited on 07/15/2024).
7. Guo X, Zhang Q, Jiang J, Peng M, Zhu M, and Yang HF. Towards explainable traffic flow prediction with large language models. Communications in Transportation Research 2024;4. TLDR: This paper proposes a Traffic flow Prediction model based on Large Language Models (LLMs) to generate explainable traffic predictions, named xTP-LLM, and is the first study to use LLM for explainable prediction of traffic flows.:100150.
8. Williams AR, Ashok A, Marcotte É, et al. Context is Key: A Benchmark for Forecasting with Essential Textual Information. TLDR: A time series forecasting benchmark that pairs numerical data with diverse types of carefully crafted textual context, requiring models to integrate both modalities, and a simple yet effective LLM prompting method that outperforms all other tested methods on this benchmark. 2024. DOI: 10.48550/arXiv.2410.18959. arXiv: 2410.18959[cs]. URL: http://arxiv.org/abs/2410.18959 (visited on 01/08/2025).

9. Liu H, Xu S, Zhao Z, et al. Time-MMD: A New Multi-Domain Multimodal Dataset for Time Series Analysis. TLDR: Time-MMD is introduced, the first multi-domain, multimodal time series dataset covering 9 primary data domains, and MM-TSFlib, the first multimodal time-series forecasting (TSF) library, seamlessly pipelining multimodal TSF evaluations based on Time-MMD for in-depth analyses. 2024. DOI: 10.48550/arXiv.2406.08627. arXiv: 2406.08627[cs]. URL: http://arxiv.org/abs/2406.08627 (visited on 07/26/2024).

10. Li Z, Lin X, Liu Z, et al. Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative. 2025. DOI: 10.48550/arXiv.2502.08942. arXiv: 2502.08942[cs]. URL: http://arxiv.org/abs/2502.08942 (visited on 02/17/2025).

11. Xue H and Salim FD. PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting. 2023. DOI: 10.48550/arXiv.2210.08964. arXiv: 2210.08964[cs,math,stat]. URL: http://arxiv.org/abs/2210.08964 (visited on 01/17/2024).

12. Arango SP, Mercado P, Kapoor S, et al. ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables. TLDR: This paper introduces a new method to incorporate covariates into pretrained time series forecasting models through modular blocks that inject past and future covariate information, without necessarily modifying the pretrained model in consideration. 2025. DOI: 10.48550/arXiv.2503.12107. arXiv: 2503.12107[cs]. URL: http://arxiv.org/abs/2503.12107 (visited on 03/27/2025).

13. Zeng A, Chen M, Zhang L, and Xu Q. Are Transformers Effective for Time Series Forecasting? 2022. arXiv: 2205.13504[cs]. URL: http://arxiv.org/abs/2205.13504 (visited on 07/25/2023).

14. Tan M, Merrill MA, Gupta V, Althoff T, and Hartvigsen T. Are Language Models Actually Useful for Time Series Forecasting? 2024. DOI: 10.48550/arXiv.2406.16964. arXiv: 2406.16964[cs]. URL: http://arxiv.org/abs/2406.16964 (visited on 08/06/2024).

15. Yuan X and Qiao Y. Diffusion-TS: Interpretable Diffusion for General Time Series Generation. In: The Twelfth International Conference on Learning Representations. 2023. URL: https://openreview.net/forum?id=4h1apFjO99 (visited on 12/13/2024).

16. Su J, Jiang C, Jin X, et al. Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. 2024. DOI: 10.48550/arXiv.2402.10350. arXiv: 2402.10350[cs]. URL: http://arxiv.org/abs/2402.10350 (visited on 04/02/2024).

17. Wang X, Fang M, Zeng Z, and Cheng T. Where Would I Go Next? Large Language Models as Human Mobility Predictors. 2023. DOI: 10.48550/arXiv.2308.15197. arXiv: 2308.15197[physics]. URL: http://arxiv.org/abs/2308.15197 (visited on 10/30/2023).

18. Feng J, Du Y, Liu T, Guo S, Lin Y, and Li Y. CityGPT: Empowering Urban Spatial Cognition of Large Language Models. 2024. DOI: 10.48550/arXiv.2406.13948. arXiv: 2406.13948[cs]. URL: http://arxiv.org/abs/2406.13948 (visited on 08/01/2024).

19. Liu P, Guo H, Dai T, et al. Taming Pre-trained LLMs for Generalised Time Series Forecasting via Cross-modal Knowledge Distillation. 2024. DOI: 10.48550/arXiv.2403.07300. arXiv: 2403.07300[cs]. URL: http://arxiv.org/abs/2403.07300 (visited on 03/19/2024).

20. Jin M, Wang S, Ma L, et al. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. 2023. DOI: 10.48550/arXiv.2310.01728. arXiv: 2310.01728[cs]. URL: http://arxiv.org/abs/2310.01728 (visited on 10/19/2023).

21. Ansari AF, Stella L, Turkmen C, et al. Chronos: Learning the Language of Time Series. 2024. arXiv: 2403.07815[cs]. URL: http://arxiv.org/abs/2403.07815 (visited on 03/18/2024).

22. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research 2020;21:1–67.

23. Bai J, Bai S, Chu Y, et al. Qwen Technical Report. arXiv preprint arXiv:2309.16609 2023.

24. Zhou T, Ma Z, Wen Q, et al. Film: Frequency improved legendre memory model for long-term time series forecasting. Advances in neural information processing systems 2022;35:12677–90.

25. Zhou H, Zhang S, Peng J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 12. 2021:11106–15.

26. Nie Y, Nguyen NH, Sinthong P, and Kalagnanam J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. 2023. arXiv: 2211.14730[cs]. URL: http://arxiv.org/abs/2211.14730 (visited on 12/15/2023).

27. Liu Y, Hu T, Zhang H, et al. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. 2023. DOI: 10.48550/arXiv.2310.06625. arXiv: 2310.06625[cs]. URL: http://arxiv.org/abs/2310.06625 (visited on 10/24/2023).

28. Liu X, Xia Y, Liang Y, et al. LargeST: A Benchmark Dataset for Large-Scale Traffic Forecasting. 2023. arXiv: 2306.08259[cs]. URL: http://arxiv.org/abs/2306.08259 (visited on 10/31/2023).

29. Wu H, Hu T, Liu Y, Zhou H, Wang J, and Long M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In: *International Conference on Learning Representations*. 2023.

30. Wang Y, Wu H, Dong J, Liu Y, Long M, and Wang J. Deep Time Series Models: A Comprehensive Survey and Benchmark. 2024.