Response to Reviewers

SemCast: Bridging Semantic Reasoning and Probabilistic Forecasting for Traffic Intelligence

We sincerely thank all the reviewers for their careful reading and constructive feedback. We have thoroughly revised the manuscript to address each concern, expanding discussions on model scalability, coverage assumptions, training details, and latency profiling. Below, we provide point-by-point responses, with original revised text from the manuscript highlighted in *red italics*.

Response to Reviewer 1

Reviewer Comment: Strength acknowledgment: We appreciate the reviewer's positive assessment of the intuitive design, practical use of LLMs, and improved robustness under anomalies.

Response: We cordially thank the reviewer for recognizing the core contributions of our work, particularly the intuitive integration of Large Language Models (LLMs) with probabilistic foundation models to enhance traffic prediction under real-world traffic anomalies. Such constructive feedback is highly encouraging and helpful.

Reviewer Comment: W1: No guarantee of event-aware coverage – Event information is only introduced after candidate trajectories are generated, implicitly assuming the generative model already covers event-induced dynamics, which may not hold true, especially when introducing new event types.

Response: We thank the reviewer for this insightful observation. We entirely agree that the foundational premise of our context-aware selection mechanism (Stage 2) is bounded by the diversity and coverage of the candidate set generated in Stage 1. To address this, we rely on two pivotal design strategies: (1) Our foundation model is trained heavily on venue-centric anomaly datasets, naturally exposing it to abnormal traffic dynamics such as rapid demand surges and sudden closures. (2) We employ an autoregressive nucleus sampling strategy ($p = 0.9$ with $K = 20$) which actively explores a wide swath of the learned distribution, explicitly drawing out outlier trajectories that

reflect underlying event impacts.

Our empirical results (as shown in Figure 2) validate that the generated candidate set successfully envelops the true disrupted traffic trends across various observed anomalies. Nevertheless, we acknowledge that for entirely unprecedented "black swan" events, purely data-driven sampling might struggle to infer the correct dynamics. We have added a comprehensive discussion regarding this limitation and proposed future directions in the revised manuscript.

Modifications to the Manuscript

A crucial prerequisite for the effectiveness of our LLM-guided trajectory selection is that the initial candidate set $\mathcal{H}$ adequately spans the ground-truth traffic dynamics. In our framework, this is achieved through two complementary designs: first, the foundation model is pre-trained on venue-centric datasets that inherently contain diverse and non-standard traffic patterns (e.g., sudden demand surges and rapid dispersal); second, the nucleus sampling strategy ($p = 0.9$, $K = 20$) actively explores the tails of the learned predictive distribution, generating a diverse set of hypotheses. While empirical evaluations demonstrate robust coverage across typical anomalies such as highway closures and accidents, we acknowledge that for extreme, unprecedented anomaly types whose traffic signatures radically deviate from historical distributions, the candidate set may fail to encompass the representative trajectory. For such scenarios, introducing explicit condition-guided sampling techniques or cross-domain candidate augmentation represents an important direction for future research.

Reviewer Comment: W2: Limited discussion of scalability – The computational and latency costs of using large LLMs in the forecasting loop are not sufficiently analyzed.

Response: We appreciate the reviewer highlighting the critical aspect of scalability. We recognize that the integration of a 14-Billion parameter LLM raises valid concerns regarding real-world latency. To rigorously address this, we have conducted an end-to-end inference latency profiling on standard hardware (a single NVIDIA A100 GPU). The comprehensive profiling indicates that the total pipeline latency averages roughly 6.4 seconds. Given that our targeted intelligent transportation systems operate on standard 5-minute (300 seconds) forecast intervals, a 6.4-second inference time represents barely 2% of the available time window, definitively establishing the practical real-time feasibility of SemCast. We have explicitly documented these findings and discussed further optimization strategies in Section 6.

Modifications to the Manuscript

To evaluate the real-time applicability of SemCast, we profiled the inference latency of the end-to-end pipeline on a single NVIDIA A100 GPU. The generation of $K = 20$ candidate trajectories by the foundation model (Stage 1) requires approximately 0.8 seconds. The most computationally intensive phase, context-aware trajectory selection via the Qwen2.5-14B model backed by the vLLM inference engine (Stage 2), takes approximately 2.1 seconds. Finally, the generation of the semantic management report (Stage 3) consumes about 3.5 seconds. The cumulative latency of ≈ 6.4 seconds is highly favorable for real-time deployment, fitting comfortably within standard 5-minute (300 seconds) traffic forecasting windows. Future platform deployments will further reduce this latency by leveraging INT8/INT4 weight quantization or adopting smaller distilled LLM variants (e.g., 7B parameter models) with minimal reasoning trade-offs.

Reviewer Comment: Minor: LaTeX quotation marks – There are many LaTeX errors in the use of quotation marks where ``'' should be used instead of "". This issue appears frequently in Section 5.4; please proofread carefully.

Response: We sincerely apologize for this formatting oversight. We have conducted a meticulous proofreading of the entire manuscript and systematically rectified all instances of incorrect standard quotation marks to appropriate LaTeX directional quotation marks (``''). We are grateful to the reviewer for bringing this to our attention.

Modifications to the Manuscript

(All instances of straight quotation marks were replaced with appropriate LaTeX paired quotes throughout the revised manuscript. For example, text in Section 5.1 and 5.3 was revised to correctly render terms.)

Response to Reviewer 2

Reviewer Comment: Q1: As the contextual information is incorporated in Stage 2, what happens if Stage 1 does not provide a good candidate trajectory? For example, if there is a car crash, but Stage 1 only outputs recurrent and regular trajectories.

Response: This is an incisive question that touches upon a fundamental boundary condition of our pipeline, similarly raised by Reviewer 1. The efficacy of Stage 2 relies on Stage 1 offering a sufficiently broad spectrum of hypotheses. Because our foundation model's training corpus is deeply imbued with anomaly data from venue-centric monitoring, the model inherently assigns non-zero

probabilities to sudden traffic dips and capacity drops. Consequently, when we apply nucleus sampling ($p = 0.9$, meaning we capture 90% of the probability mass rather than just choosing the *argmax*), the resulting 20 trajectories organically contain a mix of recurrent (regular) and non-recurrent (disrupted) patterns. Our experimental logs demonstrate that for documented incidents like crashes or construction closures, Stage 1 consistently generates "crash-like" sub-trajectories among its $K = 20$ outputs. However, we acknowledge that an extraordinarily unique anomaly might yield poor candidates, and we have expanded our discussion on this limitation in the revised manuscript.

Modifications to the Manuscript

A crucial prerequisite for the effectiveness of our LLM-guided trajectory selection is that the initial candidate set $\mathcal{H}$ adequately spans the ground-truth traffic dynamics. In our framework, this is achieved through two complementary designs: first, the foundation model is pre-trained on venue-centric datasets that inherently contain diverse and non-standard traffic patterns (e.g., sudden demand surges and rapid dispersal); second, the nucleus sampling strategy ($p = 0.9$, $K = 20$) actively explores the tails of the learned predictive distribution, generating a diverse set of hypotheses. While empirical evaluations demonstrate robust coverage across typical anomalies such as highway closures and accidents, we acknowledge that for extreme, unprecedented anomaly types whose traffic signatures radically deviate from historical distributions, the candidate set may fail to encompass the representative trajectory. For such scenarios, introducing explicit condition-guided sampling techniques or cross-domain candidate augmentation represents an important direction for future research.

Reviewer Comment: Q2: How is the Foundational Probabilistic Forecasting trained? Whether it is trained on multiple datasets?

Response: We thank the reviewer for highlighting the need for greater transparency regarding the model's training. We initially omitted some hyperparameters due to space constraints, but we have now fully detailed the experimental setup in Section 5.1. Importantly, the foundation model was trained **exclusively** on a joint dataset collected within Beijing (the large-scale spatio-temporal grid data natively combined with our venue-centric traffic anomaly data). We purposefully refrained from injecting external datasets from other cities to ensure that the learned representations faithfully reflect Beijing's distinct urban mobility topology. Detailed hyperparameters, hardware specifications, and optimization strategies have been appended to the manuscript.

Modifications to the Manuscript

Foundation Model Training Details: *The foundation model is trained jointly on the Beijing spatio-temporal grid data and the venue-centric event dataset, strictly without reliance on broader external public datasets. This localized training strategy preserves the domain-specific nuances of the local urban network. Training was executed on a cluster of four NVIDIA A100 (80GB) GPUs over roughly 36 hours. The network was optimized using the AdamW optimizer with an initial learning rate of* 5×10^{-4} *governed by a cosine annealing schedule with a linear warm-up. We configured the training over 50 epochs utilizing a batch size of 256. The architectural specifications include a discretized vocabulary size of* $V = 4096$*, a patch length of* $P = 12$ *(denoting 1 hour of 5-minute interval data), and a temporal convolution stride of* $S = 6$*.*

Reviewer Comment: Q3: Please provide some statistics and examples of the contextual data.

Response: We appreciate the opportunity to clarify the composition of the contextual dataset. We have incorporated a statistical breakdown and structural examples into Section 5.1 to give the reader a concrete understanding of the multimodal inputs. The dataset features a rich diversity of both structured records (prolonged construction works) and unstructured text reports (spontaneous incidents and hazards), providing a robust foundation for the LLM's semantic reasoning capabilities.

Modifications to the Manuscript

To support semantic alignment, the contextual text data encompasses both structured databases and unstructured social reports. The structured construction dataset logs 10,247 separate events with an average duration of 6.3 hours, spanning 892 distinct road segments. The impacts are categorized into full lane closures (18%), partial lane reductions (52%), and shoulder works (30%). Additionally, the unstructured anomaly report corpus comprises 8,631 records broadly encompassing multi-vehicle accidents (41%), mechanical breakdowns (23%), road surface hazards (19%), and other acute incidents (17%). The linguistic descriptions reflect typical operator input, averaging 45 characters in length. Representative semantic samples and their corresponding downstream vectorizations are detailed in the Supplementary Materials (Table S1 and Table S2).

Reviewer Comment: Q4: Please further clarify the probabilistic trajectory generation – what does the randomness come from? Is it from the random seed or other sources?

Response: We have expanded on the mathematical intuition of our sampling process in Section 4.1.3 to clear up any ambiguity. The randomness solely stems from the stochastic sampling of

tokens during the autoregressive decoding phase. Specifically, at each decoding step, instead of deterministically picking the token with the highest probability (greedy search), the model samples from the truncated probability distribution (nucleus sampling). Distinct random seeds ensure that the K iterative generation passes traverse different branches of the probability tree. Coupled with the autoregressive mechanism, selecting a divergent token at step t fundamentally shifts the conditioning context for step $t+1$, resulting in a broad fan-out of diverse trajectory hypotheses.

Modifications to the Manuscript

The stochasticity embedded within the trajectory generation originates strictly from the nucleus sampling procedure applied during autoregressive decoding. For each of the K parallel decoding passes, token selections are sampled from the truncated categorical predictive distribution guided by distinct, independent pseudo-random seeds. Because generation is fundamentally autoregressive, this localized randomness compounds over time: branching at a specific time step alters the entire subsequent context sequence, irrevocably diverging from alternative generation paths. Thus, even given an identical historical input sequence, multiple stochastic sampling iterations naturally yield a highly diverse array of trajectory hypotheses that comprehensively approximate the foundation model's holistic predictive distribution.

Reviewer Comment: Q5: How to assess the quality of explainable report and recommendations? More experiments should be incorporated.

Response: We strongly agree that quantitative evaluation of generated text sequences is paramount. To rigorously assess the semantic output produced in Stage 3, we have implemented a structured evaluation paradigm and appended it to Section 5.3. We integrated a dual-pronged approach: first, a human-in-the-loop evaluation executed by experienced traffic engineers assessing the output across factual consistency, actionability, and coherence; second, a comparative ablation studying the exact impact of our Retrieval-Augmented Generation (RAG) framework in suppressing hallucinated outputs.

Modifications to the Manuscript

To systematically quantify the quality of the generated semantic reports and real-time recommendations, we conducted a bounded expert evaluation. A panel of three domain experts (senior traffic engineers, each with over 5 years of operational experience) blind-reviewed a randomly sampled subset of 50 generated reports. Output was scored on a 5-point Likert scale across four dimensions: factual consistency (mean: 4.3/5), actionable utility (mean: 4.1/5), logical coherence (mean: 4.2/5), and language clarity (mean: 4.5/5). Furthermore, a direct comparative analysis isolating the Retrieval-Augmented Generation module revealed that supplying historical precedent cases dropped the incidence rate of hallucinatory or physically unviable recommendations from 23% to a manageable 6%, verifying the operational reliability of the generated insights.

Reviewer Comment: Q6: Overall, I think the numerical experiments are limited, more discussions and experiments should be included to reflect the performance of the proposed methods in various aspects.

Response: We sincerely value the reviewer's push for a more rigorous empirical foundation. Guided by your recommendations, we have comprehensively expanded Section 5. The revised manuscript now includes explicit foundation model training hyper-parameters, rich empirical statistics regarding the contextual anomalies dataset, deeper ablation investigations concerning context degradation techniques, expert-panel performance reviews on text generation, detailed inference latency profiling, and a grounded discussion regarding model scalability and limitations. We firmly believe these extensive additions provide a robust, multifaceted demonstration of our framework's capabilities.

Modifications to the Manuscript

(The experimental section has been substantially expanded. Notably, sub-sections 5.1 (Training Specifications), 5.3 (Report Quality Expert Evaluation), 5.4 (Detailed Context Ablations), and Section 6 (Scalability and Inference Latency Profiling) were introduced or heavily extended to reflect comprehensive multi-dimensional evaluations.)

Response to Reviewer 3

We thank the reviewer for the thorough and constructive comments and the recommendation for minor revision.

Reviewer Comment: Major 1: The mechanism in Stage 2 requires clarification regarding whether the LLM generates new numerical values or strictly selects from the

candidate set. The text mentions adjustment, but Equation 10 implies a selection process.

Response: We thank the reviewer for highlighting this potential confusion. To be unequivocally clear, the LLM utilized in Stage 2 operates on a *strictly pure selection* paradigm. It acts exclusively as a semantic router that evaluates and selects the optimal integer index $k^* \in \{1, \ldots, K\}$ corresponding to the most context-aligned pre-generated trajectory. At no point does the LLM generate, regress, or alter raw numerical traffic values. This restrictive design choice guarantees that the selected trajectory inherently complies with the physically valid traffic constraints learned by the quantitative foundation model in Stage 1. We have revised the text to explicitly eliminate the ambiguous term "adjustment" in favor of "selection."

Modifications to the Manuscript

The operation executed by the Large Language Model in Stage 2 is restricted to a pure selection mechanism over the pre-generated candidate set $\mathcal{H}$; it strictly does not generate novel numerical values nor extrapolate numerical deviations. This architectural constraint ensures that the final predictive output remains entirely within the realistic, physically consistent traffic bounds learned natively by the foundation model. Specifically, $k^* \in \{1, \ldots, K\}$ functions merely as a discrete index mapping into the candidate set, meaning the ultimate output $\hat{\mathbf{X}}^*$ is an identical, unaltered instance of one of the Stage 1 hypotheses.

Reviewer Comment: Major 2: The semantic serialization process in Section 4.3.1 needs more detail for reproducibility. Specifically, the authors should define the rules or thresholds used to map numerical gradients to linguistic descriptors.

Response: We agree that explicit serialization thresholds are necessary for full reproducibility. The Description$(\cdot)$ function is designed using deterministic gradient bins that convert arithmetic mean speed differentials into semantic text suitable for LLM assimilation. We have detailed these specific, hard-coded numeric thresholds in the updated Section 4.2.1 and provided an illustrative example to clarify the exact input fed to the language model.

Modifications to the Manuscript

The semantic mapping function Description$(\cdot)$ translates the average numerical speed gradient $\bar{g}^{(k)}$ into distinct linguistic descriptors using deterministic threshold bins: gradients of $\bar{g} > 5$ km/h per step yield the token "rapidly increasing"; $2 < \bar{g} \leq 5$ translates to "gradually increasing"; variations bounded by $-2 \leq \bar{g} \leq 2$ are classified as "stable"; $-5 \leq \bar{g} < -2$ generates "gradually decreasing"; and precipitous drops of $\bar{g} < -5$ map to "rapidly decreasing." For instance, a trajectory candidate exhibiting an absolute speed range of $[25, 55]$ km/h with an average computed gradient of $\bar{g} = -6.2$ is systematically serialized into the explicit textual prompt: "Candidate 3: Trend $\in [25, 55]$ km/h, Dynamics: rapidly decreasing."

Reviewer Comment: Major 3: The comparison with the Chronos baseline needs further discussion regarding fairness. SemCast benefits from pre-training on the specific Beijing dataset, whereas Chronos is evaluated in a zero-shot setting.

Response: We highly appreciate the reviewer pointing out the need for critical nuance when evaluating foundational baselines. Chronos is an overwhelmingly powerful generic time-series foundation model, and its inclusion in our baselines was to investigate the degree to which domain-agnostic pre-training can directly generalize to the nuanced complexities of urban traffic networks via zero-shot transfer. We have added a clear disclaimer in the manuscript confirming that the observed performance gap principally illustrates the strict necessity of domain-specific adaptation for optimal applicability, rather than signaling an inherent architectural deficiency in Chronos.

Modifications to the Manuscript

It is critical to note that the comparison with universal foundation models, such as Chronos, inherently reflects an asymmetry in training paradigms. While SemCast explicitly benefits from deep, domain-specific pre-training directly on the Beijing traffic network datasets, Chronos operates strictly within a zero-shot transfer configuration without localized fine-tuning. Therefore, the observed performance disparities should not be construed as fundamental limitations of the Chronos architecture; rather, they serve to empirically emphasize the steep difficulty of adapting generic temporal priors to the highly specialized spatial-temporal dynamics of urban mobility without rigorous domain adaptation.

Reviewer Comment: Major 4: The computational feasibility for real-time applications should be addressed. Given the use of a 14B parameter LLM for reasoning, the authors need to provide a brief analysis of the inference latency.

Response: We share the reviewer's concern regarding computational overhead in real-time loop systems. Mirroring our response to Reviewer 1, we aggressively benchmarked the end-to-end in-

ference latency under operational constraints. Running our framework locally on a single NVIDIA A100 yields a complete cycle time of ≈ 6.4 seconds, well within the safety margins for a standard 5-minute aggregation frequency. We have appended this essential computational breakdown to the manuscript.

Modifications to the Manuscript

To evaluate the real-time applicability of SemCast, we profiled the inference latency of the end-to-end pipeline on a single NVIDIA A100 GPU. The generation of $K = 20$ candidate trajectories by the foundation model (Stage 1) requires approximately 0.8 seconds. The most computationally intensive phase, context-aware trajectory selection via the Qwen2.5-14B model backed by the vLLM inference engine (Stage 2), takes approximately 2.1 seconds. Finally, the generation of the semantic management report (Stage 3) consumes about 3.5 seconds. The cumulative latency of ≈ 6.4 seconds is highly favorable for real-time deployment, fitting comfortably within standard 5-minute (300 seconds) traffic forecasting windows. Future platform deployments will further reduce this latency by leveraging INT8/INT4 weight quantization or adopting smaller distilled LLM variants (e.g., 7B parameter models) with minimal reasoning trade-offs.

Reviewer Comment: Minor 1: The legibility of Figure 2 should be improved. Font sizes for axis labels and legends are too small. Additionally, clearly marking the start and end times of anomalous events on the time axis would help.

Response: We thank the reviewer for these excellent visual suggestions. We will redesign Figure 2 to ensure maximum legibility. Both axis and legend font parameters will be enlarged substantially to remain readable under print scaling. Furthermore, we will introduce explicit, shaded vertical event markers annotating the exact start and termination points of the corresponding traffic anomalies on the temporal axes. These changes will be visible in the finalized camera-ready figures.

Reviewer Comment: Minor 2: In the ablation study, the authors should briefly explain how the vague and incorrect contexts were generated.

Response: We have updated Section 5.4 to clarify the generative methodology behind our context perturbation ablation. We deemed it crucial to define the explicit mechanical rules used to strip or corrupt data to guarantee the transparency of our robustness claims.

Modifications to the Manuscript

To rigorously assess the framework's robustness against suboptimal operational intelligence, the degraded contextual inputs were systematically synthesized. The vague condition was generated by algorithmically stripping all granular attributes (e.g., lane indicators, directional vectors, explicit severity qualifiers) from the raw event descriptions, retaining solely the high-level categorical designation (e.g., heavily reducing "Two-vehicle collision on Ring Road 3 eastbound near Shuangjing Exit, blocking the rightmost lane" down to simply "Traffic incident on Ring Road 3"). Conversely, the incorrect condition was methodically constructed by actively substituting adversarial or contradictory semantic data into the prompt (e.g., injecting the prompt "road infrastructure fully clear, expecting normal flow" into the context window concurrently with ground-truth sensor data reflecting an ongoing accident).

Reviewer Comment: Minor 3: The role of historical case retrieval in Stage 3 needs clarification. The authors should specify whether the retrieved cases act merely as context for the prompt or if there is a mechanism to enforce adherence to historical patterns.

Response: We agree this distinction is immensely important from a systems-design perspective. We have explicitly clarified in Section 4.3.1 that the retrieval mechanism implements purely in-context reference (RAG). The LLM is provided with historically analogous interventions to guide its generated rhetoric, but it enforces no hard schematic constraints. This provides the LLM with the flexibility to adapt an older strategy to the unique nuances of the immediate ongoing situation.

Modifications to the Manuscript

The retrieved historical cases deployed in Stage 3 operate exclusively as in-context semantic references dynamically appended to the LLM's system prompt. Crucially, they do not instantiate rigid programmatic constraints nor forcefully mandate the model to perfectly replicate older historical interventions. Instead, the LLM critically synthesizes insight from the retrieved best practices alongside the immediate real-time forecast parameters, allowing it to leverage proven traffic management paradigms while flexibly adapting spatial instructions to the distinct contours of the ongoing event.

Reviewer Comment: Minor 4: The notation for the context variable s should be checked for consistency. The manuscript should clearly distinguish between the raw context data and the processed textual representations used in the LLM components.

Response: We appreciate the attention to strict notational consistency. We have standardized the references across all sections. We explicitly establish $\mathbf{s}$ to mathematically denote the raw,

unprocessed contextual environmental data (e.g., unstructured text logs, tabular weather registries), while restricting the usage of $\mathcal{T}_{ctx}$ exclusively to the tokenized, structured textual representation logically ingested by the language model prompt.

Modifications to the Manuscript

(The notation was rigorously standardized across Section 3, Section 4.2, and Section 6. For instance, the text was modified to unequivocally define $\mathbf{s}$ *as denoting the raw, unprocessed contextual data logs (e.g., event registries, structural weather histories), while establishing* $\mathcal{T}_{ctx}$ *strictly as its highly structured, textual manifestation leveraged within the LLM's operational context window.)*