Disaster-Cast: Disaster-Aware POI Visit Forecasting with Multimodal LLM Fusion

Nurjahan
Department of Computer Science
Louisiana State University
Baton Rouge, Louisiana, USA
nurja1@lsu.edu

***Abstract*—Forecasting visits at specific points of interest (POIs) during and after natural disasters is important for emergency response, infrastructure planning, and understanding how communities recover. However, POI-level visit series are very short, noisy, and highly imbalanced, which makes both classical forecasting methods and recent prompt-based LLM approaches difficult to use reliably. In this work, we propose a hybrid spatiotemporal forecasting framework (Disaster-Cast) that combines three elements: (i) lightweight temporal encoders to capture short-term visit patterns, (ii) H3-based spatial embeddings to include local geographic context, and (iii) semantic POI information encoded through structured hard prompts to a LoRA-fine-tuned LLM. We evaluate our approach using SafeGraph mobility data from several Florida cities that were differently affected by Hurricane Ian. Our results show that simple 1D-CNN temporal encoders consistently perform better than recurrent models and often match or exceed classical baselines such as Auto-ARIMA and Prophet. In contrast, combining temporal, spatial, and LLM-based semantic features gives mixed outcomes: H3 embeddings help only in cities with more consistent spatial recovery, while LLM semantics and retrieval augmentation often do not improve over CNNs and sometimes even reduce accuracy. Rather than introducing a new state-of-the-art model, this work provides the first careful study of when temporal, spatial, and semantic signals actually help or hurt POI-level disaster mobility forecasting with LLM-based methods.**

***Index Terms*—Mobility Forecasting, Large Language Models, Disasters**

I. Introduction

In recent years, natural disasters have become more frequent and severe as the climate continues to change, disrupting daily life and reshaping how people move through cities. These disruptions make human mobility data increasingly important for understanding how cities function, plan infrastructure, and respond to shocks [1]. During extreme events, mobility becomes even more critical for emergency response, supply-chain continuity, and resilience planning, as population displacement strongly influences impact severity [2]. Under normal conditions, human movement is highly regular, with predictability estimates reaching up to 93% [3]. Disasters, however, break these regular patterns, producing abrupt behavioral changes and uneven spatial recovery [4]. Businesses reopen at different speeds, residents adjust daily routines, and neighborhoods experience varying levels of disruption. As a result, fine-grained visit forecasting at the Point-of-Interest (POI) level is essential for operational decision-making across sectors such as retail, transportation, and emergency management [5].

Human mobility prediction broadly refers to estimating future movements, either of individuals or aggregates, using spatiotemporal data [6]. Prior work has primarily examined tasks such as next-location prediction [6]–[9], user linking [10], time prediction [10], forecasting future traffic conditions at spatial locations [11], and modeling flows between regions [4]. In contrast, POI-level visit forecasting during a disaster presents a fundamentally different challenge: mobility becomes short-range, unstable, and highly sensitive to localized disruptions. Although several studies investigate POI-level visit prediction in normal, stable conditions, they do not address the volatility or spatial heterogeneity seen in disasters. Existing Large Language Model (LLM)-based approaches include natural-language translation models [12], fine-tuning foundation models [13], prompt-based forecasting [14], and prompt-mining strategies [5]. These methods, however, operate on random sample-level splits in which the same POIs may appear in both training and testing and focus exclusively on next-day horizons. As a result, they cannot generalize to unseen POIs or handle abrupt post-disaster behavioral shifts. Moreover, prior prompt-based techniques embed both POI semantics and raw visit sequences directly into natural-language descriptions [5], [12]–[14].

LLMs offer new opportunities for mobility forecasting because their semantic reasoning can complement numerical patterns. Hard prompts describing POI attributes or disaster context allow LLMs to generate high-level embeddings that capture functional and contextual similarities across locations. In our Disaster-Cast, numeric visit sequences are deliberately excluded from LLM prompts: temporal dynamics are handled by a dedicated trainable encoder, while the LLM processes only semantic metadata and disaster-related information. This separation improves interpretability, and enables genuine generalization capabilities to previously unseen POIs that prior LLM-based mobility forecasting frameworks do not provide. In this work, we propose a hybrid spatiotemporal forecasting framework that integrates three key components: a trainable temporal encoder, a trainable H3 spatial embedding, and a LoRA-augmented LLM that consumes structured hard prompts. The temporal encoder (Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), One-Dimensional Convo-

lutional Neural Network (1D-CNN), and GRU–CNN hybrid variants) captures short-term behavioral dynamics of visit counts. The H3 embedding represents each POI's geographic context using a single-resolution spatial index. The LLM component transforms a natural-language prompt, containing POI metadata and next-day prediction instructions, into a semantic embedding. Projection layers map temporal and spatial vectors into the LLM hidden dimension, and additive fusion combines all three signals into a unified representation, which is passed to a regression head. The model is trained end-to-end using LoRA adapters and a joint objective combining token-level cross-entropy and mean squared error (MSE), allowing the system to benefit from both semantic reasoning and numerical accuracy. Overall, this framework provides a clean, modular, and training-efficient approach for POI visit forecasting, and enables systematic comparison between temporal-only, temporal+H3, and fully fused architectures. Our contributions are:

- To the best of our knowledge, this is the first study to examine POI-level next-day forecasting in disaster settings over post-event horizons.
- We introduce a unified forecasting framework that combines trainable temporal encoders, spatial representations, and LLM-based semantic embeddings within a shared embedding space. The modular design enables each modality to be evaluated independently or in combination.
- We provide a systematic analysis of how temporal, spatial, and semantic signals contribute to next-day forecasting across post-disruption periods. Empirical results show that temporal encoders provide the strongest standalone performance, while spatial and semantic features offer selective benefits depending on underlying distributional characteristics.
- We offer an empirical framework that explains when multimodal fusion helps, when it does not, and why. The study highlights how sparsity, imbalance, and variability in POI visit patterns affect model behavior, offering insights that generalize beyond a single dataset or event.

The remainder of the paper is organized as follows. Section II reviews our related work. We describe our framework in Section III and report its experimental results in Section IV. We discuss our implications and limitations in Section VI, and Section VII concludes this paper with future research directions.

II. Related Work

Research relevant to this work spans three areas: (i) classical and deep-learning approaches for numerical and mobility time-series forecasting, (ii) LLM-based spatiotemporal forecasting and mobility analysis, and (iii) LLMs for general time-series forecasting. We discuss each in turn and highlight the limitations that motivate our multimodal LLM architecture for hurricane-aware POI visit forecasting.

A. Numerical and Deep Learning Approaches

Classical statistical models such as AutoRegressive Integrated Moving Average (ARIMA), Seasonal AutoRegressive Integrated Moving Average (SARIMA), and AutoRegressive Integrated Moving Average with Exogenous Variables [15]–[17] have long been used in traffic and mobility forecasting due to their interpretability and ability to capture seasonality. These models have been applied to subway passenger flows [18], [19] and congestion prediction [20], but they assume stable temporal patterns and therefore struggle under sudden behavioral changes caused by disasters or policy shifts. Ensemble methods such as Random Forests and Gradient Boosting [21] relax linear assumptions but still depend on feature engineering and do not naturally incorporate spatial relationships or contextual signals.

Deep learning approaches advanced mobility forecasting by learning richer temporal and spatial dependencies. Foundational architectures such as Recurrent Neural Networks (RNNs) [22], LSTMs [23], and GRUs [24] capture long-term temporal structure, while Seq2Seq models [25] and ST-RNN [26] integrate spatial distances and temporal intervals. Attention-based models, notably DeepMove [27], combine recurrent modeling with periodicity-aware attention for improved next-location prediction. More recent long-sequence forecasting architectures, including Informer [28], Autoformer [29], and FEDformer [30], introduce decomposition and autocorrelation modules for efficient long-range forecasting. Spatial-temporal GNNs (ST-GNNs) [31] extend deep learning models with explicit graph-based spatial reasoning. Surveys such as [32] summarize these developments and emphasize persistent challenges around robustness, multimodality, and generalization. None of the models work with sparse, short-term, noisy disaster induced data.

B. LLM-Based Spatiotemporal and Mobility Forecasting

The application of LLMs to mobility forecasting has evolved in several stages. Early work reprogrammed mobility time series into natural-language descriptions so that pretrained LLMs could reason about mobility patterns. AuxMobLCast [13] demonstrated that next-day POI visit prediction is possible using only textualized inputs, though the model lacked explicit temporal and spatial representations. Prompt-based approaches extended this idea by framing forecasting as a text-generation task. PromptCast [33], [34] used fixed templates combining POI metadata and raw visit sequences, allowing LLMs to compete with neural baselines for next-day prediction. Prompt-Mining [5] further introduced entropy-guided prompt search and chain-of-thought refinement, improving prompt-only forecasting. However, these methods operate entirely on textualized numerical data and remain limited to stable, routine conditions.

LLMs have been explored for semantic mobility reasoning rather than forecasting. Prior work includes zero-shot next-location prediction using narrative trajectories [6], intention-aware prompting and memory mechanisms [10], synthetic trajectory generation [35], [36], semantic trajectory classification [37], and contextual mobility analysis [38]. Surveys [39], [40] highlight that these methods rarely incorporate structured temporal or spatial encodings and typically evaluate

under stable, non-disruptive settings. DisasterMobLLM [7] considers disaster-era next-location prediction but does not address aggregated POI visit forecasting or integrate trainable temporal and H3 spatial embeddings. No existing LLM-based approach unifies these components for disaster-aware POI-level forecasting.

C. LLMs for General Time-Series Forecasting

Beyond mobility, LLMs have also been adapted for general time-series forecasting. Methods such as Time-LLM [41], Chronos [42], TEMPO [43], and LLM4TS [44] show that pre-trained LLMs can model temporal dynamics through prompt-based reprogramming, discretization of numerical sequences, prefix tuning, and multimodal fusion. These models exhibit strong zero-shot and few-shot generalization in domains such as energy, weather, and finance. However, these methods assume dense, continuous, and relatively stable temporal patterns. They do not address sparse and noisy POI-level visit sequences during disasters. While informative for understanding how LLMs can be adapted to numerical sequences, these time-series LLMs do not directly address the spatially uneven, volatile recovery patterns that arise after hurricanes.

III. Methodology

A. Problem Formulation

Let $\{POI_1, \ldots, POI_P\}$ denote a collection of P points of interest. For each POI p, we observe a sequence of aggregated visit counts over the past n time steps:

$$\mathbf{v}^{(p)}_{k-n+1:k} = \left[v^{(p)}_{k-n+1}, v^{(p)}_{k-n+2}, \ldots, v^{(p)}_{k} \right], \tag{1}$$

where $v^{(p)}_t$ is the number of visits to POI p on day t.

Each POI is accompanied by a compact set of contextual features. We use (i) a semantic description derived from the POI's location name, business category, and city, and (ii) a spatial identifier given by a single-resolution H3 index. Additionally, deterministic temporal metadata such as the series start date and the hurricane landfall date, which provides coarse-grained situational context for the forecasting window. The goal is to predict the future visit count at horizon $h \geq 1$. Formally, the task is to learn a mapping

$$f : \left(\mathbf{v}^{(p)}_{k-n+1:k},\ s_p,\ g_p\right) \longrightarrow \hat{v}^{(p)}_{k+h}, \tag{2}$$

where s_p denotes the semantic features and g_p the H3-based spatial identifier.

This formulation is model-agnostic and accommodates multiple short-term forecasting horizons (e.g., $h = 14$ or $h = 21$).

B. Our Proposed Framework

Figure 1 illustrates the architecture of our proposed Disaster-Cast model. The architecture is built around a multimodal embedding design that integrates temporal, spatial, and semantic information into a unified representation suitable for LLLM reasoning. The model consists of three main components. First, a temporal encoder summarizes the historical visit trajectory of each POI into a compact embedding that

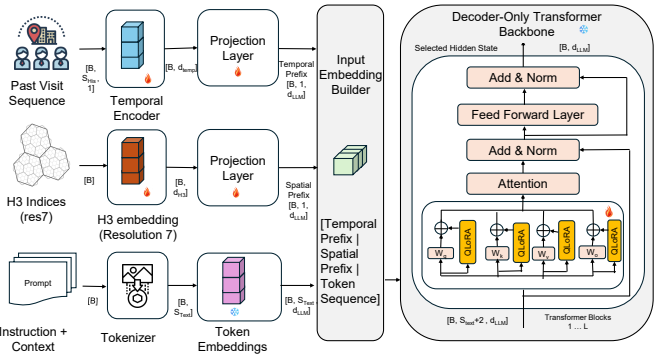


Fig. 1: Architecture of our Disaster-Cast Model

captures short- and long-term mobility patterns. Second, a spatial module encodes the POI's geospatial characteristics through a learnable H3 embedding, providing location-aware signals that complement temporal dynamics. Third, a prompt encoder converts structured natural-language instructions and POI metadata into token embeddings using the pretrained tokenizer of the LLM. These three embeddings are projected into a common latent space and concatenated by the Input Embedding Builder to form a single prefix-augmented sequence. This unified sequence is then fed into a pretrained decoder-only transformer, which is adapted to the forecasting task using parameter-efficient fine-tuning with low-rank adapters. The transformer jointly attends to all input components and produces a hidden representation from which the final visit prediction is generated. In the following subsections, we elaborate on each component of the framework in detail.

C. Temporal Feature Encoder

To model temporal dependencies in historical mobility patterns, we employ a trainable temporal encoder based on either a GRU, LSTM, or 1D convolutional architecture. Given the input sequence $v^p_{k-n+1:k} \in \mathbb{R}^{n\times 1}$, the encoder produces a fixed-length temporal representation

$$h^p_{\text{temp}} \in \mathbb{R}^{d_{\text{temp}}}. \tag{3}$$

This latent vector is subsequently projected into the embedding dimension of the language model through a linear transformation,

$$p_1 = W_{\text{temp}} h^p_{\text{temp}}, \tag{4}$$

resulting in a prefix token $p_1 \in \mathbb{R}^{d_{\text{LLM}}}$ that summarizes the dynamics of past visit behaviors.

D. Spatial Embedding via H3 Indexing

Each POI is associated with an H3 index h_p at a fixed resolution. We treat the H3 index as a categorical spatial identifier and learn a trainable embedding

$$h^p_{\text{H3}} \in \mathbb{R}^{d_{\text{H3}}}, \tag{5}$$

which allows the model to represent coarse spatial locality patterns shared by POIs within the same H3 cell. A projection

layer maps this spatial embedding into the language model's hidden space,

$$p_2 = W_{\text{H3}} h^p_{\text{H3}}, \tag{6}$$

yielding a spatial prefix token $p_2 \in \mathbb{R}^{d_{\text{LLM}}}$. This token enables the transformer to condition its predictions on location-dependent visitation patterns, even though the H3 embedding itself is learned purely from categorical indices.

E. Prompt Token Embeddings

To incorporate semantic attributes and task instructions, we design a structured natural-language prompt containing elements such as POI category, location name, and the forecasting objective. The prompt is processed using the tokenizer of the pretrained model, producing a sequence of token embeddings

$$E_{\text{text}} \in \mathbb{R}^{S_{\text{text}} \times d_{\text{LLM}}}. \tag{7}$$

These embeddings encode categorical information and provide the model with flexible interpretability through natural language.

F. Unified Embedding Construction

The final input to the transformer is constructed by concatenating the two prefix tokens with the prompt token embeddings:

$$X = [p_1; p_2; E_{\text{text}}], \tag{8}$$

resulting in an input sequence of length $S_{\text{text}} + 2$ in the shared latent space $\mathbb{R}^{d_{\text{LLM}}}$. This representation allows the transformer to jointly attend to temporal, spatial, and semantic cues within a single sequence of embeddings.

G. Parameter-Efficient Fine-Tuning

We adapt the pretrained decoder-only language model such as meta-llama/Meta-Llama-3-8B [45] and mistralai/Mistral-7B-v0.1 [46] using parameter-efficient fine-tuning with the QLoRA framework [47]. Rather than updating all model weights, QLoRA introduces low-rank matrices into the attention projection layers, modifying the weight matrix W as

$$W' = W + BA, \tag{9}$$

where A and B are learnable matrices of rank $r \ll d_{\text{LLM}}$. This approach dramatically reduces memory usage while enabling the model to specialize in forecasting mobility patterns. Given the input sequence X, the transformer generates a final hidden representation, from which a linear regression head produces the forecasted visit count $\hat{v}^p_{k+h}$.

H. Training Objective

Given the unified multimodal input sequence X, the model produces a hidden representation h^p associated with POI p. A linear prediction head maps this representation to a forecasted visit count,

$$\hat{v}^p_{k+h} = W_{\text{out}} h^p + b_{\text{out}}. \tag{10}$$

We train the model to minimize the error between the predicted and true future visit counts v^p_{k+h} using a mean squared error objective:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{p=1}^{N} (\hat{v}^p_{k+h} - v^p_{k+h})^2. \tag{11}$$

For architectures that include the pretrained language model, we additionally apply a generative objective that trains the decoder to autoregressively generate the numeric forecast in natural-language form. This is implemented using a causal language modeling loss:

$$\mathcal{L}_{\text{gen}} = -\mathbb{E}_{(X,y)} \big[\log p_\theta(y \mid X) \big], \tag{12}$$

where y is the tokenized representation of the target count.

The final loss combines the two components:

$$\mathcal{L} = \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{gen}}, \tag{13}$$

where λ_{reg} controls the contribution of the regression term. In ablation settings without the language model, only the regression objective is optimized.

IV. Experiments

Our experiments investigate the effectiveness of the proposed framework for disaster-aware mobility forecasting and address four research questions:

- **RQ1**: Do lightweight temporal encoders (e.g., 1D-CNNs) outperform classical and recurrent models after Hurricane Ian? We hypothesize that CNNs offer better accuracy and stability (**H1**) because post-disaster visit sequences exhibit sharp local fluctuations rather than long-range dependencies.
- **RQ2**: When does geospatial structure improve forecasting accuracy? We hypothesize that H3 embeddings help only where POIs show spatially coherent recovery, and may add noise in cities with heavy-tailed or irregular patterns (**H2**).
- **RQ3**: Can pretrained LLMs act as effective POI semantic encoders with structured prompts? We hypothesize modest gains (**H3**) when POI semantics correlate with recovery behavior, but limited benefit otherwise.
- **RQ4**: How does distributional heterogeneity across cities affect model performance? We hypothesize that low-variance cities favor simpler models, while volatile or heavy-tailed cities amplify RMSE and challenge recurrent models (**H4**). Multimodal gains are expected only when temporal, spatial, and semantic signals align with city-level structure.

A. Datasets

We evaluate our forecasting models using SafeGraph mobility data [48] from five major Florida cities affected by Hurricane Ian in 2022: Miami, Orlando, Jacksonville, Tampa, and Cape Coral. SafeGraph provides anonymized, weekly aggregated POI-level records, including daily visit counts, business metadata (e.g., NAICS classifications), and persistent geospatial identifiers. We extract a continuous three-week window from September 19 to October 9, covering (i) a stable

TABLE I: POI visit-count percentage distribution by city for Day 14 and Day 21.

	Day 14 Visit Ranges (%)										Day 21 Visit Ranges (%)									
City	**0**	**1–5**	**6–10**	**11–20**	**21–50**	**51–100**	**101–200**	**201–500**	**501–1000**	1000+	**0**	**1–5**	**6–10**	**11–20**	**21–50**	**51–100**	**101–200**	**201–500**	**501–1000**	1000+
Cape Coral	41.8	42.3	7.6	4.5	2.3	1.0	0.5	0.0	0.0	0.0	35.8	40.2	8.8	7.1	5.6	1.5	0.9	0.2	0.0	0.0
Jacksonville	40.1	38.5	9.1	6.4	4.1	1.2	0.4	0.1	0.0	0.0	39.9	38.3	8.9	6.7	4.2	1.3	0.4	0.1	0.0	0.0
Miami	38.7	40.5	9.5	6.3	3.6	0.8	0.3	0.2	0.0	0.0	39.7	41.0	8.8	5.9	3.3	0.7	0.3	0.2	0.0	0.0
Orlando	34.8	38.3	9.9	7.7	5.9	2.0	0.9	0.3	0.1	0.1	36.0	36.2	9.6	7.8	6.4	2.2	1.0	0.5	0.1	0.1
Tampa	39.8	40.6	8.7	5.9	3.5	1.0	0.3	0.2	0.1	0.0	40.3	39.2	9.0	6.1	3.7	1.1	0.3	0.2	0.0	0.0

TABLE II: Dataset statistics: number of POIs in train, validation, and test splits for each city.

City	Train	Val	Test	Total
Tampa	8,574	2,858	2,858	14,290
Cape Coral	1,237	413	413	2,063
Jacksonville	8,838	2,946	2,947	14,731
Miami	12,142	4,048	4,048	20,238
Orlando	9,516	3,172	3,172	15,860

pre-landfall baseline, (ii) the landfall week when Hurricane Ian struck Florida on September 28, and (iii) the early recovery phase.

After filtering, the dataset contains 67,182 POIs: Miami (20,238; 30.1%), Orlando (15,860; 23.6%), Jacksonville (14,731; 21.9%), Tampa (14,290; 21.3%), and Cape Coral (2,063; 3.1%). As shown in Table I, the visit-count distributions vary sharply across cities and reflect their different levels of hurricane exposure. Cape Coral, the point of direct landfall, exhibits the most compressed distribution: on Day 14, 84.1% of POIs fall within the 0–5 visit range, and virtually no POIs exceed 200 visits. Jacksonville and Miami, farther from the storm center, show similar low-visit dominance (78–80% across the 0 and 1–5 bins combined), with only modest mass in the mid-volume ranges. Tampa and Orlando, situated along the inland storm track, also display high sparsity (73–80% in the 0–5 range) but with noticeably heavier tails; in particular, Orlando contains the largest concentration of high-visit POIs (maximum 42,732), including small but non-negligible proportions above 500 and 1,000 visits. By Day 21, the distributions widen slightly but preserve the same overall structure: Cape Coral remains the most heavily disrupted, Jacksonville and Miami show moderate spread, and Orlando retains the broadest and most skewed visit profile driven by extreme high-traffic venues.

B. Dataset Preprocessing

POIs with missing or corrupted `visits_by_day` values are removed to ensure complete three-week sequences. For each POI, we consolidate metadata such as business name, NAICS category, subcategory, and census block group by selecting the unique non-null value associated with its Placekey to ensure stable semantic information. Each Placekey encodes location as `What@Where`, where the "Where" component embeds geospatial position through Uber's H3 system. We extract the H3 index at a single resolution and use it as the POI's spatial identifier, providing a uniform grid-based representation for downstream spatial modeling. The final processed dataset contains, for every POI: (i) a complete three-week visit-count sequence, (ii) consolidated semantic metadata, (iii) a single-resolution H3 identifier, and (iv) temporal markers such as sequence start date and hurricane landfall date. These standardized records form the input to all forecasting models. Finally, we split the samples into train, validation, and test sets using a 60/20/20 ratio (see Table II).

C. Implementations

To ensure that our findings are not tied to a single language-model architecture, we employ two state-of-the-art open-source LLMs, Llama and Mistral, as semantic backbones. Below, we summarize the LLM backbones, tuning strategy, prompt templates, and baseline configurations.

a) LLM Backbone and Tuning Strategy.: We adopt Llama and Mistral as our two pretrained LLMs. Llama is a decoder-only transformer with 8 billion parameters across 32 layers and a hidden size of 4,096, offering a strong balance between capacity and computational efficiency. Mistral is a 7-billion-parameter decoder-only transformer with 32 layers and a hidden size of 4,096, incorporating grouped-query attention (GQA) and sliding-window attention for efficient long-context inference. Both models support parameter-efficient fine-tuning (PEFT) and run reliably on HPC environments using QLoRA. For fairness, we apply the same supervised instruction-tuning setup to both LLMs, ensuring that any performance differences arise from the modeling framework rather than the backbone itself.

In our multimodal models, we use a single fixed instruction-style prompt for supervised tuning. Unlike the zero-shot baselines, the prompt excludes numerical visit counts: the temporal encoder receives the raw 13- or 20-day visit sequence directly, while the LLM is conditioned only on POI metadata (name, NAICS category, city) and disaster context (landfall date, target date), followed by a short instruction such as "Predict the number of visitors on the target date." The ground-truth count is appended as the target response during training, and the same template is used at inference time without the answer.

b) Zero-shot LLM Baselines.: To assess how pretrained LLMs behave without fine-tuning, we construct three zero-shot baselines that differ in the amount of contextual information included in the prompt (Table VI in Appendix Section). Template A is a minimalist temporal prompt containing only the raw daily visit sequence. Template B adds disaster context, including the POI's city, time-series start date, and hurricane landfall time. Template C provides the full context,

additionally incorporating POI semantics such as name and business category. These templates allow us to examine how LLM predictions evolve as temporal, disaster, and semantic cues are incrementally introduced. After generation, LLM outputs are parsed using a strict regex pattern (`PREDICTION: <number>`), and converted to the final predicted integer count.

c) Statistical and Neural Baselines.: We compare against two non-LLM baselines: per-POI statistical models (Auto-ARIMA, Prophet), which forecast from each POI's recent 13–20 day history using small nonseasonal fits and trend-only models with mean fallback, and global neural models (LSTM, GRU, 1D-CNN, GRU+CNN), trained for 3 epochs with Adam (10^{-3}), batch size 32, and MSE loss. All baselines operate solely on raw visit sequences and provide standardized comparisons for evaluating our multimodal framework.

D. Evaluation Metrics

Let y_i denote the true visit count for POI i and $\hat{y}_i$ the corresponding prediction. To evaluate forecasting accuracy across small, medium, and high-volume POIs, we employ three widely used regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error(RMSE), and Root Mean Squared Logarithmic Error (RMSLE). Meanwhile, The coefficient of determination (R^2) is used in this paper only as a diagnostic indicator of model behavior and is therefore not reported as a primary performance metric.

a) Mean Absolute Error (MAE).: MAE measures the average magnitude of prediction error and is robust to outliers:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|. \tag{14}$$

b) Root Mean Squared Error (RMSE).: RMSE penalizes larger deviations more strongly, making it sensitive to misestimation of high-traffic POIs:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}. \tag{15}$$

c) Root Mean Squared Logarithmic Error (RMSLE).: RMSLE measures error in log space and emphasizes accurate estimation of relative change:

$$\mathrm{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log(1+\hat{y}_i) - \log(1+y_i)\right)^2}. \tag{16}$$

d) Coefficient of Determination (R^2).: R^2 quantifies the proportion of variance in y explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{17}$$

where $\bar{y}$ is the mean visit count.

E. Reproducibility, Tools, and Hyperparameter Settings

To ensure reproducibility, we fix all random seeds to 42, apply deterministic computation where compatible with quantized training, and use identical preprocessing and evaluation procedures across all models. Experiments are conducted on NVIDIA A100 (80 GB) GPUs on the LONI HPC cluster, with development managed through remote execution tools and environment isolation via `pyenv`. All neural models, including temporal encoders and our multimodal LLM architectures—are trained for 3 epochs using AdamW with a learning rate of 1×10^{-4} and batch size 4; temporal encoders use hidden size 64, and H3 embeddings are 64-dimensional. The implementation relies on PyTorch, Hugging Face Transformers, and PEFT for QLoRA fine-tuning. Additional tools, configurations, and hyperparameters appear in Appendix Table VII.

V. Results

We present our results in three parts. First, we analyze the performance of our multimodal model with the statistical and neural baselines. Next, we evaluate zero-shot LLM behavior under different prompts. Finally, we study how temporal, spatial, and semantic components affect our multimodal models.

A. Performance Comparison of Baselines and Our Multimodal Disaster-Cast Model

Table III reports Day-14 and Day-21 forecasting performance across all statistical baselines, neural temporal encoders, and our multimodal Disaster-Cast framework. Among the standalone neural models, lightweight 1D-CNNs remain the most stable and broadly effective. CNNs achieve the best MAE in Jacksonville (Day-14), Miami, Orlando (Day-14), and Tampa (Day-14), and consistently avoid the error drift observed in GRU and LSTM. This behavior aligns with the structure of post-hurricane POI data, where most venues exhibit short-lived fluctuations within a 0–20 visit range. CNNs capture these local temporal patterns without accumulating long-horizon dependencies, whereas recurrent models tend to amplify noise or underpredict high-variance trajectories.

City-level differences are strongly tied to underlying volume distributions. Cape Coral, dominated by low-magnitude POIs, shows simple autoregressive patterns; here, Auto ARIMA performs best in RMSE for Day-14 and in both MAE and RMSE for Day-21. Orlando, on the opposite extreme, contains the heaviest-tailed visit volumes, with counts reaching 30–40K and numerous locations above 100 visits. Such extremes destabilize recurrent models, especially at longer horizons, resulting in substantial RMSE and RMSLE spikes. CNN remains competitive on Day-14, but Day-21 dynamics become smoother and more autoregressive, allowing Auto ARIMA to surpass neural models. Miami presents a milder heavy tail, where CNN reliably outperforms GRU and LSTM, while Tampa contains large venues (e.g., stadiums, airport terminals) that appear in the test set but not in validation, causing recurrent models to collapse toward conservative predictions and enabling Auto ARIMA and CNN to perform more consistently.

TABLE III: Performance of statistical, neural, and multimodal models for Day-14 and Day-21 POI visit forecasting. Best MAE, RMSE, and RMSLE per city–horizon shown in **blue bold**.

City	Model	Day 14			Day 21		
		MAE	RMSE	RMSLE	MAE	RMSE	RMSLE
Cape Coral	Auto ARIMA	3.02	**6.89**	0.81	**3.05**	**6.08**	0.72
	Prophet	3.06	8.85	0.81	4.36	11.96	0.93
	CNN	2.89	8.16	0.71	4.42	8.55	0.94
	GRU	**2.83**	8.45	**0.70**	4.12	14.26	**0.66**
	GRU+CNN	2.96	8.17	0.73	3.83	9.78	0.78
	LSTM	3.00	9.01	**0.70**	4.60	15.82	0.75
	Disaster-Cast	2.90	7.57	0.72	4.74	13.49	0.85
Jacksonville	Auto ARIMA	3.16	7.10	0.79	3.51	20.47	0.79
	Prophet	2.45	**6.14**	0.66	3.90	21.11	0.84
	CNN	**2.44**	6.38	0.62	3.14	19.68	0.71
	GRU	3.19	12.13	0.69	3.29	21.21	0.67
	GRU+CNN	3.00	7.44	0.77	**2.66**	**19.63**	**0.59**
	LSTM	2.98	11.90	0.67	3.29	21.85	0.67
	Disaster-Cast	2.63	9.57	**0.59**	3.14	20.46	0.65
Miami	Auto ARIMA	4.02	23.35	0.83	3.55	15.38	0.79
	Prophet	2.95	13.18	0.73	3.74	16.78	0.79
	CNN	**2.37**	**7.58**	0.63	**2.35**	**7.67**	**0.59**
	GRU	3.32	27.45	0.69	3.96	63.33	0.65
	GRU+CNN	2.55	13.82	**0.60**	2.77	16.39	0.61
	LSTM	3.16	26.69	0.62	3.87	63.62	0.61
	Disaster-Cast	2.92	22.83	0.62	3.52	60.25	**0.59**
Orlando	Auto ARIMA	4.66	23.44	0.75	**5.78**	**79.12**	0.75
	Prophet	4.76	26.91	0.69	8.99	166.09	0.76
	CNN	**3.72**	**13.47**	0.69	7.77	150.96	0.60
	GRU	5.73	71.26	0.65	24.57	831.23	0.63
	GRU+CNN	3.65	14.39	**0.64**	6.97	98.14	0.76
	LSTM	6.08	70.97	0.67	25.36	832.69	0.65
	Disaster-Cast	5.29	67.34	**0.64**	24.64	827.21	**0.59**
Tampa	Auto ARIMA	4.21	76.34	0.75	3.61	23.09	0.76
	Prophet	4.40	72.04	0.71	3.80	22.98	0.80
	CNN	**3.87**	64.94	0.70	**3.50**	**19.84**	0.69
	GRU	4.91	85.69	**0.65**	3.18	24.96	0.67
	GRU+CNN	3.89	**57.04**	0.70	**3.06**	22.11	**0.62**
	LSTM	4.96	86.20	**0.65**	3.35	25.63	**0.62**
	Disaster-Cast	4.67	82.70	0.71	3.46	23.75	0.79

Across all cities, the RMSLE metric reveals an additional insight: it reflects proportional error rather than domination by a few large POIs. Under this metric, our multimodal Disaster-Cast system is competitive with the strongest baselines and often performs best on the most heterogeneous cities. For instance, Disaster-Cast achieves the lowest or tied-lowest RMSLE in Jacksonville (Day-14), Orlando (both horizons), and Miami (Day-21), even when its MAE or RMSE are not the absolute best.

Overall, CNN remains the strongest standalone temporal baseline across most cities, Auto ARIMA excels when POI sequences exhibit smoother autoregressive structure or contain very large venues absent from training, and Disaster-Cast provides meaningful improvements in percentage-normalized errors (RMSLE) under the most challenging, heavy-tailed conditions.

B. Zero-Shot LLM Baselines

Table IV reports zero-shot forecasting results for Llama-3 8B and Mistral 7B across three prompt templates that introduce progressively richer context. The results show a clear pattern: the simplest prompt (Template A) works best, while adding hurricane or semantic text (Templates B and C) usually makes predictions worse. Across cities, Mistral performs better than Llama on the plain numerical prompt. This is expected because Mistral's sliding-window attention naturally focuses on the most recent numbers and ignores most of the extra text. Llama uses global attention and is more sensitive to descriptive words in the prompt. When we add hurricane-related context, Llama often gives too much weight to that information and "overreacts," producing large errors, especially in Orlando. Mistral is less affected because it does not attend strongly to earlier text. Llama sometimes performs slightly better in the most detailed semantic prompt (Template C), especially in cities where business category carries useful information. Its larger training corpus and global attention help it use that semantic detail.

TABLE IV: Zero-shot LLM performance using three prompt templates for Day-14 and Day-21 POI visit forecasting. Best values per city–horizon in **blue bold**.

City	Model	Temp	Day 14			Day 21		
			MAE	RMSE	RMSLE	MAE	RMSE	RMSLE
Cape Coral	Llama	A	7.32	22.04	1.23	8.36	33.68	1.17
		B	**4.52**	14.58	1.35	7.94	25.66	1.64
		C	5.00	15.50	1.36	10.17	29.79	1.71
	Mistral	A	4.91	**12.27**	**1.13**	**3.54**	**5.74**	**0.94**
		B	13.04	26.61	1.67	9.83	32.49	1.28
		C	5.01	17.17	1.27	7.11	22.06	1.60
Jacksonville	Llama	A	5.03	13.27	1.04	**5.51**	**25.84**	**1.05**
		B	5.72	20.04	1.49	6.08	27.26	1.48
		C	8.19	56.48	1.50	11.73	51.22	1.54
	Mistral	A	**4.37**	**12.64**	**0.57**	5.62	26.34	**1.05**
		B	10.36	29.79	1.50	7.99	35.75	1.23
		C	4.67	14.87	1.25	5.66	25.87	1.39
Miami	Llama	A	5.65	30.57	1.01	**5.46**	30.20	**0.96**
		B	5.86	34.29	1.42	6.36	65.83	1.41
		C	6.70	38.18	1.45	6.69	**28.39**	1.43
	Mistral	A	**5.12**	31.20	**0.96**	**5.46**	30.13	**0.96**
		B	10.60	34.81	1.51	7.62	31.37	1.18
		C	5.18	**24.91**	1.21	6.65	67.56	1.32
Orlando	Llama	A	6.80	28.31	1.00	16.00	301.57	1.04
		B	10.44	79.29	1.69	27.52	772.49	1.80
		C	10.20	44.71	1.67	30.64	780.01	1.80
	Mistral	A	**5.69**	**26.95**	**0.94**	**7.28**	**64.28**	**0.97**
		B	14.06	119.43	1.50	14.09	231.24	1.18
		C	9.27	78.87	1.42	21.54	421.68	1.66
Tampa	Llama	A	6.58	75.88	1.04	6.25	35.30	1.01
		B	7.30	88.04	1.43	6.25	31.01	1.47
		C	8.24	89.89	1.46	7.90	41.99	1.49
	Mistral	A	**5.49**	**75.72**	**0.97**	5.91	34.05	**1.00**
		B	12.80	113.77	1.54	8.62	37.19	1.25
		C	7.41	89.23	1.30	**5.79**	**29.78**	1.38

C. Ablations of Our Multimodal Framework

Table V compares CNN-only models with CNN+H3 spatial variants and full multimodal models incorporating LLM-derived semantic embeddings. The CNN-only models again provide the strongest and most stable performance across cities. Adding H3 spatial information yields mixed results: spatial context helps in cities where neighboring POIs recover coherently (e.g., Jacksonville and Tampa at Day-21), but harms performance in cities with heavy-tailed or heterogeneous activity patterns, such as Miami and Orlando. In these environments, nearby POIs behave differently, and spatial similarity does not translate into predictive value.

Introducing LLM semantic embeddings in the CNN+H3+LLM models consistently degrades performance, often sharply. Business semantics encoded by Llama or

TABLE V: Ablation study comparing CNN-only, CNN+H3, and CNN+H3+LLM variants for Day-14 and Day-21 POI visit forecasting. Best MAE, RMSE, and RMSLE per city–horizon shown in **blue bold**.

City	Model	Day 14			Day 21		
		MAE	RMSE	RMSLE	MAE	RMSE	RMSLE
Cape Coral	CNN	**2.89**	8.16	**0.71**	4.42	8.55	0.94
	CNN+H3	3.15	**6.51**	0.83	3.69	**6.93**	0.86
	CNN+H3+Llama	2.90	7.57	0.72	4.74	13.49	0.85
	CNN+H3+Mistral	2.92	6.55	0.76	**3.70**	8.81	**0.78**
Jacksonville	CNN	**2.44**	6.38	0.62	3.14	19.68	0.71
	CNN+H3	2.84	**6.36**	0.74	**2.86**	**19.54**	**0.65**
	CNN+H3+Llama	2.63	9.57	**0.59**	3.14	20.46	**0.65**
	CNN+H3+Mistral	4.18	13.78	1.01	3.24	21.14	0.66
Miami	CNN	**2.37**	**7.58**	0.63	2.35	**7.67**	**0.59**
	CNN+H3	2.39	7.83	0.65	**2.32**	10.41	0.60
	CNN+H3+Llama	2.92	22.83	**0.62**	3.52	60.25	**0.59**
	CNN+H3+Mistral	4.85	26.30	1.18	3.49	59.78	**0.59**
Orlando	CNN	**3.72**	**13.47**	0.69	**7.77**	**150.96**	0.60
	CNN+H3	3.83	13.54	0.74	11.19	284.05	0.70
	CNN+H3+Llama	5.29	67.34	**0.64**	24.64	827.21	**0.59**
	CNN+H3+Mistral	5.85	65.74	0.83	25.17	828.19	0.67
Tampa	CNN	3.87	**64.94**	0.70	3.50	19.84	**0.69**
	CNN+H3	**3.82**	65.20	**0.67**	**2.82**	**16.56**	**0.69**
	CNN+H3+Llama	4.67	82.70	0.71	3.46	23.75	0.79
	CNN+H3+Mistral	4.59	81.09	0.72	3.54	23.58	0.85

Mistral do not meaningfully correlate with short-horizon post-disaster visit behavior. Instead, they introduce high-variance signals that distract the model from the strong temporal cues captured by CNN. This effect is especially pronounced in cities with extreme spikes, where LLM semantics amplify noise rather than provide useful structure.

D. Overall Comparison

Taken together, the results show a clear hierarchy of model effectiveness. The strongest baselines typically the 1D-CNN or Auto ARIMA, offer the most reliable and accurate predictions across cities and horizons. Zero-shot LLMs, even under optimized prompting, perform far below these baselines (except Orlando-21 using Mistral) and remain unstable. Our multimodal variants confirm that incorporating semantic LLM features does not outperform purely temporal or temporal+spatial models; instead, semantic embeddings introduce noise that overwhelms short-range temporal dynamics.

E. Category Level Error Analysis

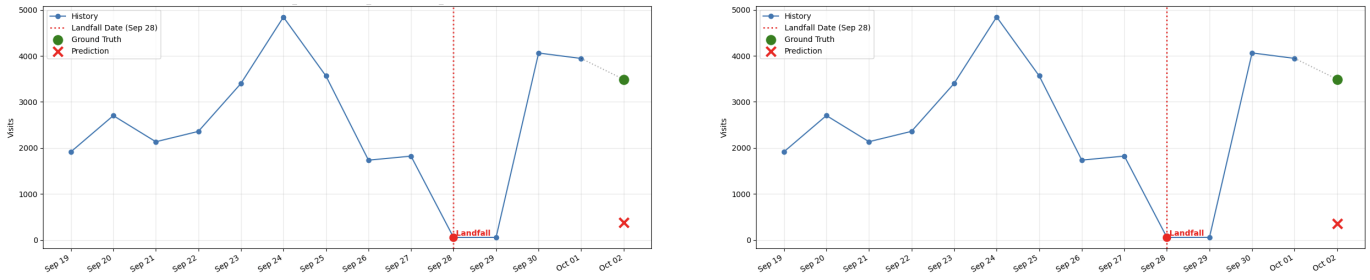
Category-level analysis shows that most POI groups exhibit low median errors, but a few categories especially amusement and theme parks, airports, stadiums, and large malls, have inflated mean errors driven by a small number of extreme outlier POIs. These categories contain venues with unusually high visit volumes and highly irregular, hurricane-driven recovery patterns, which are fundamentally different from the smooth, moderate trajectories found in typical POIs such as restaurants, groceries, and pharmacies. As a result, the mean error overstates model failure, while the median better reflects typical forecasting performance.

Figures 2 and Figure 3 present case studies illustrating the model's largest prediction errors across cities and horizons. In Orlando (Day-14 and Day-21), both Llama and Mistral severely underpredict the post-hurricane surges at Disney Springs and Walt Disney World Resort, where visits rebound sharply to several thousand or tens of thousands following landfall. Miami and Jacksonville show similar patterns: large event-driven or tourism-linked venues experience abrupt spikes that neither model anticipates, leading to large absolute errors. Tampa exhibits the same behavior at Raymond James Stadium and the international airport. Together, these examples demonstrate that the worst errors arise from a small number of high-volume, highly volatile POIs rather than widespread model deficiencies.

VI. Discussion

Our experiments were designed to answer RQ1–RQ4 by testing temporal, spatial, and semantic signals under different modeling choices. Overall, the results mostly confirm our hypotheses. First, supporting H1, lightweight temporal encoders, especially 1D-CNNs perform the best across almost all cities and horizons. After a disaster, POI visits change sharply from day to day, and CNNs capture these short-range fluctuations more effectively than RNNs or classical models. Second, in line with H2, spatial information helps only when POIs recover in a geographically consistent pattern. H3 embeddings improve results in cities like Jacksonville and Tampa but hurt performance in Miami and Orlando, where nearby POIs often behave very differently. This shows that spatial features are useful only when spatial correlation is strong. Third, although H3 allowed the possibility that LLM semantics might offer small benefits when aligned with recovery behavior, our results do not support this. In practice, LLM-derived semantic embeddings introduce high variance and rarely match short-horizon mobility dynamics. As a result, they tend to degrade or destabilize forecasting instead of helping. Finally, consistent with H4, model performance depends heavily on the underlying data distributions. Cities with narrow, low-variance behavior (e.g., Cape Coral) are easier to forecast and favor simpler models, while heavy-tailed, volatile cities (e.g., Orlando, Miami) challenge recurrent networks and inflate RMSE. These distributional differences explain why no single model works best everywhere. Overall, our study shows that temporal signals carry the most predictive value, spatial signals help only in specific conditions, and semantic signals from LLMs require much more careful integration to avoid harming performance.

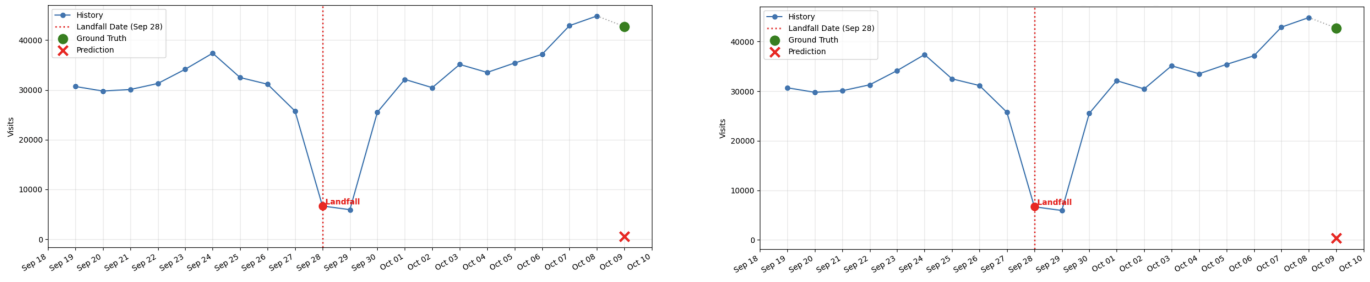
This study has several limitations. First, our analysis focuses on a single extreme event such as Hurricane Ian (2022) and a specific regional context in Florida. While Ian provides a rich and highly disruptive case study, relying on one event limits our ability to assess whether the observed behaviors of temporal, spatial, and semantic signals generalize across different disaster types and impact geometries. Second, the SafeGraph mobility data used in this study may contain sampling biases inherent to device-based collection, including uneven penetration rates across demographic groups and POI categories. These biases may introduce noise in ground-truth visit counts and affect model stability. Third, although parameter-efficient fine-tuning reduces computational cost, LLM-based



(a) Llama prediction

(b) Mistral prediction

Fig. 2: Comparison of Llama and Mistral predictions for the Day-14 horizon at *Disney Springs (Malls category): POI- zzw-223@8fy-8kn-j35* in Orlando. The ground-truth value is 3,490. LLaMA predicts 375 and Mistral predicts 352, producing absolute errors of 3,115 and 3,138 respectively.



(a) Llama prediction

(b) Mistral prediction

Fig. 3: Comparison of Llama and Mistral predictions for the Day-21 horizon at *Walt Disney World Resort: POI-zzw-222@8fy-fjg-b8v* in Orlando. Actual visits surge to 42,732 during the late-stage recovery period. Llama predicts 561 while Mistral predicts 369, resulting in absolute errors of 42,171 and 42,363 respectively.

methods remain substantially more resource-intensive than statistical models and lightweight neural baselines. Finally, our evaluation weights all POIs equally, though real-world decision-making often prioritizes critical infrastructure such as hospitals, grocery stores, and fuel stations.

VII. Conclusion

In this paper, we presented a comprehensive study of POI-level visit forecasting during disaster recovery and introduced a hybrid spatiotemporal–semantic framework that integrates temporal encoders, H3 spatial structure, and LLM-based semantic prompts. Across five Florida cities and two post-hurricane horizons, our experiments provide the first systematic analysis of when temporal, spatial, and semantic signals help or hinder LLM-driven mobility forecasting. The results show that lightweight temporal encoders such as 1D-CNNs remain strong and often outperform recurrent models; that spatial features are beneficial only in cities with coherent recovery patterns; and that naïve semantic prompting provides little value without adaptation. While our full fusion architecture did not consistently surpass simpler baselines, the analysis highlights clear modality-specific failure modes and conditions under which LLMs can meaningfully contribute.

Future work will broaden the empirical scope by incorporating additional hurricane events such as Hurricane Idalia (2023), enabling cross-event generalization and testing whether the conditions under which temporal, spatial, and semantic features help or hurt, remain consistent across different disaster geometries. Expanding the framework to incorporate richer contextual data (e.g., outage maps, flood depth, evacuation flows) and alternative mobility sources may reduce sampling bias and improve robustness. Further, exploring distilled or domain-adapted LLMs could reduce computational overhead, while risk-aware metrics would allow evaluation that reflects societal priorities.

References

[1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018, human mobility: Models and applications. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037015731830022X

[2] X. Lu, L. Bengtsson, and P. Holme, "Predictability of population displacement after the 2010 haiti earthquake," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 576–11 581, 2012. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1203882109

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science (New York, N.Y.)*, vol. 327, pp. 1018–21, 02 2010.

[4] J. Tang, J. Wang, J. Li, P. Zhao, W. Lyu, W. Zhai, L. Yuan, L. Wan, and C. Yang, "Predicting human mobility flows in response to extreme urban floods: A hybrid deep learning model considering spatial heterogeneity," *Computers, Environment and Urban Systems*, vol. 113, p. 102160, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0198971524000899

[5] H. Xue, T. Tang, A. Payani, and F. D. Salim, "Prompt mining for language models-based mobility flow forecasting," in *Proceedings of the 32nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24)*. New York, NY, USA: ACM, 2024. [Online]. Available: https://doi.org/10.1145/3678717.3691232

[6] X. Wang, M. Fang, Z. Zeng, and T. Cheng, "Where would i go next? large language models as human mobility predictors," 2024. [Online]. Available: https://arxiv.org/abs/2308.15197

[7] Y. Tang, H. Wang, X. Fan, and Y. Li, "Predicting human mobility in disasters via llm-enhanced cross-city learning," 2025. [Online]. Available: https://arxiv.org/abs/2507.19737

[8] J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin, "Deepmove: Predicting human mobility with attentional recurrent networks," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 1459–1468. [Online]. Available: https://doi.org/10.1145/3178876.3186058

[9] D. Yang, B. Fankhauser, P. Rosso, and P. Cudre-Mauroux, "Location prediction over sparse user mobility traces using rnns: flashback in hidden states!" in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, ser. IJCAI'20, 2021.

[10] L. Gong, Y. Lin, X. Zhang, Y. Lu, X. Han, Y. Liu, S. Guo, Y. Lin, and H. Wan, "Mobility-llm: learning visiting intentions and travel preferences from human mobility data with large language models," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2025.

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

[12] H. Xue, F. D. Salim, Y. Ren, and C. L. A. Clarke, "Translating human mobility forecasting through natural language generation," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1224–1233. [Online]. Available: https://doi.org/10.1145/3488560.3498387

[13] H. Xue, B. P. Voutharoja, and F. D. Salim, "Leveraging language foundation models for human mobility forecasting," in *Proceedings of the 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22)*. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1–9. [Online]. Available: https://doi.org/10.1145/3557915.3561026

[14] H. Xue and F. D. Salim, "Promptcast: A new prompt-based learning paradigm for time series forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 6851–6864, 2024.

[15] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, ser. Holden-Day Series in Time Series Analysis and Digital Processing. San Francisco, CA, USA: Holden-Day, 1970.

[16] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice Hall, 1994.

[17] A. Pankratz, *Forecasting with Dynamic Regression Models*. New York, NY, USA: John Wiley & Sons, 1991.

[18] E. Chen, Z. Ye, C. Wang, and M. Xu, "Subway passenger flow prediction for special events using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1109–1120, 2019.

[19] Y. Chen, X. Li, M. Liu, L. Wu, and Y. Hu, "Subway passenger flow prediction for special events using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 3, pp. 636–645, 2016.

[20] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal arima model with limited input data," *European Transport Research Review*, vol. 7, no. 3, p. 21, 2015.

[21] A. Tziorvas, G. S. Theodoropoulos, and Y. Theodoridis, "Shared micro-mobility demand forecasting using gradient boosting methods," in *Proceedings of the Workshops of the EDBT/ICDT 2025 Joint Conference (EDBT/ICDT-WS 2025)*, ser. CEUR Workshop Proceedings, M. Boehm and K. Daudjee, Eds., vol. 3946. Barcelona, Spain: CEUR-WS.org, Mar. 2025, 7th International Workshop on Big Mobility Data Analytics (BMDA). [Online]. Available: https://ceur-ws.org/Vol-3946/BMDA-4.pdf

[22] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[24] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014. [Online]. Available: https://arxiv.org/abs/1406.1078

[25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 3104–3112.

[26] Q. Liu, S. Wu, L. Wang, and T. Tan, "Predicting the next location: a recurrent model with spatial and temporal contexts," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, p. 194–200.

[27] J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin, "Deepmove: Predicting human mobility with attentional recurrent networks," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 1459–1468. [Online]. Available: https://doi.org/10.1145/3178876.3186058

[28] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," 2021. [Online]. Available: https://arxiv.org/abs/2012.07436

[29] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," 2022. [Online]. Available: https://arxiv.org/abs/2106.13008

[30] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," 2022. [Online]. Available: https://arxiv.org/abs/2201.12740

[31] K.-H. N. Bui, J. Cho, and H. Yi, "Spatial-temporal graph neural network for traffic forecasting: An overview and open research issues," *Applied Intelligence*, vol. 52, no. 3, p. 2763–2774, Feb. 2022. [Online]. Available: https://doi.org/10.1007/s10489-021-02587-w

[32] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, "A survey on deep learning for human mobility," *ACM Comput. Surv.*, vol. 55, no. 1, Nov. 2021. [Online]. Available: https://doi.org/10.1145/3485125

[33] H. Xue and F. D. Salim, "Promptcast: A new prompt-based learning paradigm for time series forecasting," *arXiv preprint arXiv:2210.08964*, 2023. [Online]. Available: https://arxiv.org/abs/2210.08964

[34] ——, "Promptcast: A new prompt-based learning paradigm for time series forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 6851–6864, 2024.

[35] A. Kobayashi, N. Takeda, Y. Yamazaki, and D. Kamisaka, "Modeling and generating human mobility trajectories using transformer with day encoding," in *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*, ser. HuMob-Challenge '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 7–10. [Online]. Available: https://doi.org/10.1145/3615894.3628506

[36] S. Li, T. Tran, H. Lin, J. Krumm, C. Shahabi, L. Zhao, K. Shafique, and L. Xiong, "Geo-llama: Leveraging llms for human mobility trajectory generation with constraints," in *2025 26th IEEE International Conference on Mobile Data Management (MDM)*, 2025, pp. 20–31.

[37] Y. Luo, Z. Cao, X. Jin, K. Liu, and L. Yin, "Deciphering human mobility: Inferring semantics of trajectories with large language models," *arXiv preprint arXiv:2405.19850*, 2024. [Online]. Available: https://arxiv.org/abs/2405.19850

[38] C. Shao, F. Xu, B. Fan, J. Ding, Y. Yuan, M. Wang, and Y. Li, "Beyond imitation: Generating human mobility from context-aware reasoning with large language models," *CoRR*, vol. abs/2402.09836, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.09836

[39] Z. Zhang, Y. Sun, Z. Wang, Y. Nie, X. Ma, R. Li, P. Sun, and X. Ban, "Large language models for mobility analysis in transportation systems: A survey on forecasting tasks," 2025. [Online]. Available: https://arxiv.org/abs/2405.02357

[40] Q. Long, S. Liu, N. Cao, Z. Ren, W. Ju, C. Fang, Z. Zhu, H. Zhu, and Y. Zhou, "A survey of large language models for traffic forecasting: Methods and applications," *Authorea Preprints*, 2025.

[41] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen, "Time-LLM: Time series forecasting by reprogramming large language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[42] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor *et al.*, "Chronos: Learning the language of time series," *arXiv preprint arXiv:2403.07815*, 2024.
[43] D. Cao, F. Jia, S. O. Arik, T. Pfister, Y. Zheng, W. Ye, and Y. Liu, "Tempo: Prompt-based generative pre-trained transformer for time series forecasting," *arXiv preprint arXiv:2310.04948*, 2023.
[44] C. Chang, W.-Y. Wang, W.-C. Peng, and T.-F. Chen, "Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 3, Apr. 2025. [Online]. Available: https://doi.org/10.1145/3719207
[45] M. AI, "Meta llama 3 8b," https://huggingface.co/meta-llama/Meta-Llama-3-8B, 2024, accessed: 2025-10-10.
[46] ——, "Mistral 7b v0.1," https://huggingface.co/mistralai/Mistral-7B-v0.1, 2023, accessed: 2025-10-10.
[47] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314
[48] SafeGraph, "Safegraph data," https://www.safegraph.com, 2020, accessed: 2025-09-10.

APPENDIX

A. Distribution of Visit Volumes

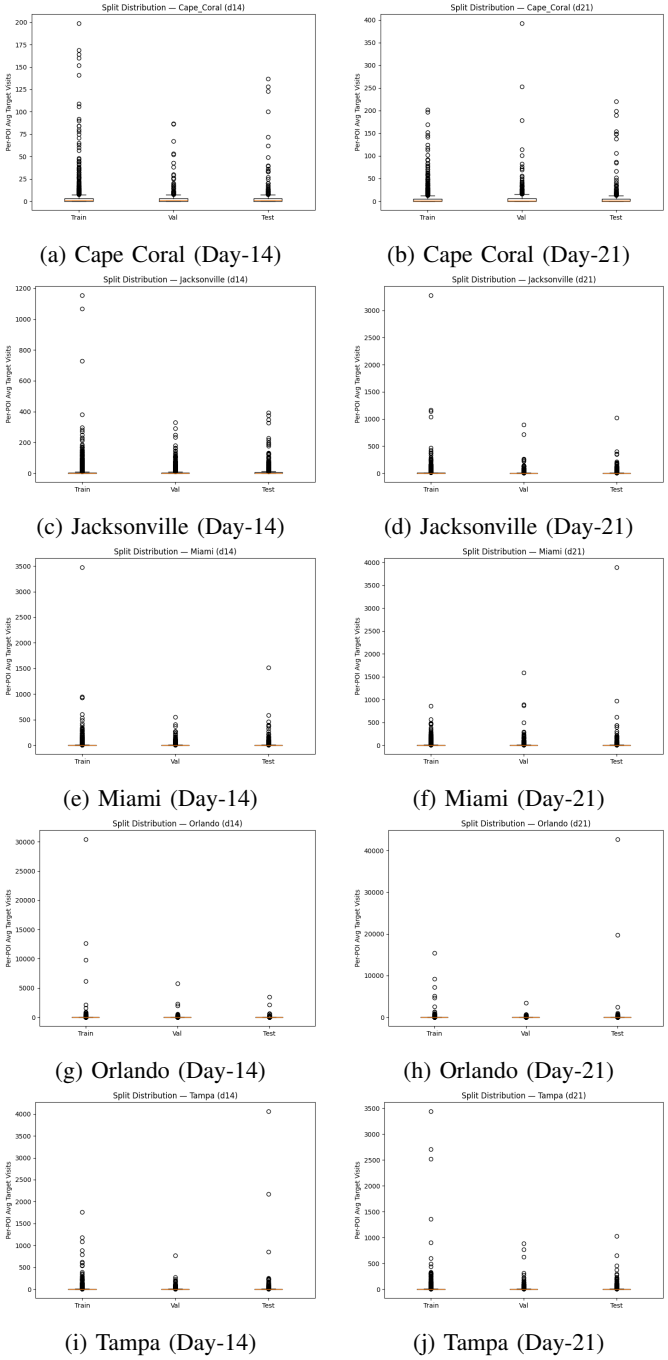
Figure 4 reports the distribution of POI visit volumes across the five Florida cities for Day-14 and Day-21 horizons. These boxplots illustrate strong skewness, heavy tails, and volatility patterns that affect forecasting difficulty.

B. Zero-Shot Prompt Templates

Table VI presents the full zero-shot prompt templates (A: Minimalist, B: Hurricane-aware, C: Full Context) used for LLM baselines. All prompt content is shown verbatim.

TABLE VI: Zero-shot LLM prompt templates: A (Minimalist), B (Hurricane-aware), and C (Full Context).

Templ.	Prompt (Full Snippet)
A	`You are an expert time-series forecaster. Given the historical daily visit counts, predict the next day's visit count. Return ONLY: PREDICTION: <number>. DailyVisits: {values}. PREDICTION:`
B	`You are an expert in disaster-aware mobility forecasting. Predict the visit count for the target date using the provided historical daily visits and the hurricane landfall time. Do NOT explain your reasoning. Return ONLY: PREDICTION: <number>. Location: {location_name}. City: {city}. SeriesStart: {series_start}. HurricaneLandfall: {landfall}. TargetDate: {target_date}. DailyVisits: {values}. PREDICTION:`
C	`You are an expert in human-mobility forecasting after natural disasters. Predict the visit count for the target date by combining temporal trends, hurricane disruption, and POI semantics. Return ONLY: PREDICTION: <number>. === CONTEXT === LocationName: {location_name}. BusinessCategory: {business_category}. City: {city}. SeriesStartDate: {series_start}. HurricaneLandfallDate: {landfall}. TargetDateToPredict: {target_date}. PastDailyVisits: {values}. === END === PREDICTION:`



(a) Cape Coral (Day-14) (b) Cape Coral (Day-21)
(c) Jacksonville (Day-14) (d) Jacksonville (Day-21)
(e) Miami (Day-14) (f) Miami (Day-21)
(g) Orlando (Day-14) (h) Orlando (Day-21)
(i) Tampa (Day-14) (j) Tampa (Day-21)

Fig. 4: Visit-volume distributions across five Florida cities for Day-14 and Day-21 horizons.

C. Tools and Hyperparameters

Table VII summarizes the core tools, platforms, and hyperparameters used throughout the study.

TABLE VII: Core tools, platforms, and hyperparameters used in the study.

Category	**Setting**
Programming Environment	Python 3.12; `pyenv`
Frameworks	PyTorch; HF Transformers; PEFT (QLoRA)
LLM Backbones	Llama-3 8B (4-bit), Mistral 7B v0.1
Temporal Encoders	GRU / LSTM / 1D-CNN (64 units)
Spatial Modeling	Trainable H3 embeddings (64-dim)
Fusion Strategy	Prefix fusion into LLM hidden space
Optimizer	AdamW
Batch Size	4
Learning Rate	1×10^{-4}
Training Epochs	3
Quantization	QLoRA (NF4 4-bit)
Zero-Shot Decoding	Greedy; max_new_tokens = 30
Data Processing	NumPy, pandas, scikit-learn
Visualization	Matplotlib, Seaborn
Hardware	NVIDIA A100 (80 GB) GPUs (LONI HPC)