MUSE: MULTI-GRANULAR URBAN SEMANTIC ENVIRONMENT MODELING FOR ROAD-NETWORK-CONSTRAINED TRAJECTORY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human mobility trajectory generation synthesizes spatiotemporal traces that reflect real travel patterns, supporting urban planning, transportation analysis, and mobility research. Existing road-network-constrained generators primarily learn transitions between road segments from historical trajectories, leaving the urban environmental knowledge associated with these roads insufficiently explored. Points of interest (POIs) provide rich descriptions of urban functions, but turning heterogeneous, multi-scale POI information into road-aligned representations for generation remains challenging. To this end, we introduce **MUSE**, a Multi-granular Urban Semantic Environment modeling framework for integrating road-network topology with road-aligned environmental knowledge. It organizes open POI data into three complementary granularities: aggregate functional distributions (Macro), nearby POI instances (Local), and representative functional anchors (Landmark). A Road Query selectively aggregates instances and fuses these sources while preserving POI categories and source identities. Road and Environment autoregressive branches then learn from the same next-road targets, with gated environment-to-road interaction incorporating environmental history into road generation. Experiments on Beijing, Chengdu, and Xi'an show that **MUSE** improves the spatial and temporal distributional fidelity of generated trajectories, demonstrating the value of extracting road-aligned, multi-granular environmental knowledge and learning it jointly with road-network structure.

1 INTRODUCTION

Human mobility trajectories support urban research, public health analysis, and transportation planning. Privacy concerns, data access restrictions, and collection costs limit the availability and sharing of real trajectories. Trajectory generation learns mobility patterns from existing observations to synthesize supplementary data with similar spatiotemporal characteristics for research and simulation (Feng et al., 2020; Zhu et al., 2023; Li et al., 2026).

Existing studies include generation in coordinate space and generation constrained by road networks. We focus on the latter: given an origin road segment, a destination road segment, and a departure time, the model generates a sequence of connected road segments with timestamps. Related methods learn path distributions from road attributes, topology, regional context, and historical transitions (Jiang et al., 2023; Cao et al., 2025). However, these road-centered signals characterize the fine-grained functional environment surrounding individual roads only implicitly.

Travel is typically driven by work, study, or shopping, and urban functional spaces provide cues to these intentions: land use at origins and destinations is associated with trip purposes (Lee & Holme, 2015). The functional spaces encountered en route and their order also inform route choice, motivating the use of environmental context throughout the trip. The locations and categories of points of interest (POIs) describe these spaces; prior mobility studies use their functional statistics and nearby instances (Feng et al., 2020; Wei et al., 2025). Associating this information with roads and using it in trajectory generation poses three challenges (Figure 1).

Aligning environments with the trajectory task. Road-indexed trajectories do not naturally align with independently collected, unevenly distributed POIs (Figure 1a). A road segment may be near

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

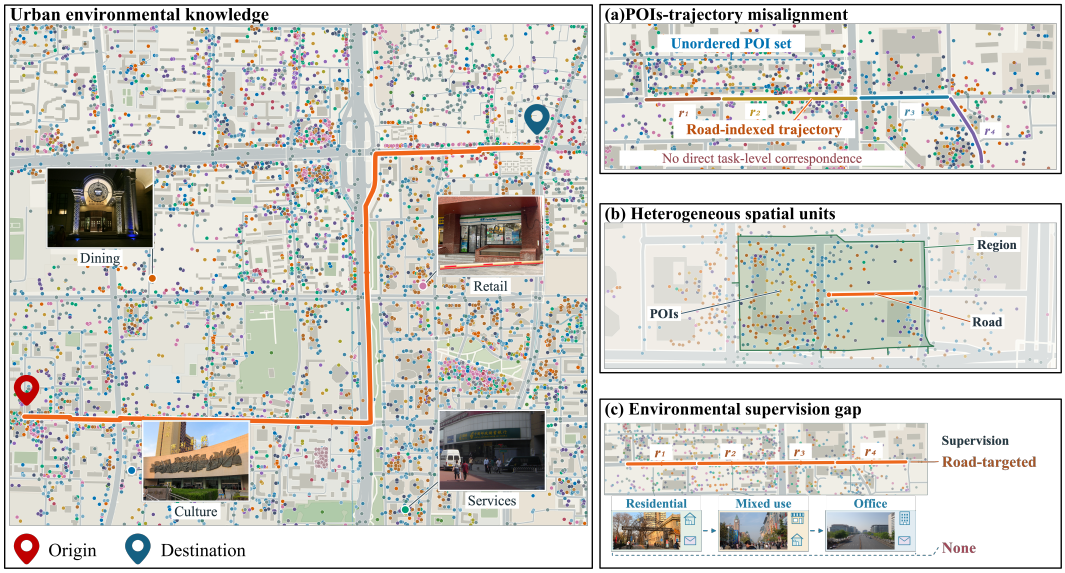

Figure 1: Urban environmental knowledge and its challenges in trajectory generation. Left: a trajectory through an urban environment rich in facilities. (a) Misalignment between unordered POIs and road-indexed trajectories. (b) Heterogeneity in the spatial and semantic granularities of POIs, roads, and regions. (c) Road-centered trajectory generation tasks lack direct supervision of environmental information.

multiple POIs, and the same POI may be near multiple road segments. The challenge is to organize these unordered proximity-based candidates into road-aligned environmental information for the trajectory task.

Balancing aggregate context and instance-level detail. POIs, roads, and regions differ in spatial and semantic granularity (Figure 1b). POIs also greatly outnumber roads and regions. Aggregate summaries can obscure local functions, while dense instance sets introduce redundancy and weakly related information. Joint modeling must preserve relevant functional details across these granularities.

Predicting mobility from the environmental view. Road and environmental sequences describe network traversal and changes in the functions encountered along the way, providing complementary views for mobility prediction (Figure 1c). Supervision centered on road prediction leaves the predictive role of environmental sequences implicit. The challenge is to give the environmental view an explicit predictive role and exploit its complementarity with the road view.

To address these challenges, we propose MUSE, which jointly learns road-aligned, multi-granular environmental representations and road sequences. Our contributions are:

- **Alignment of heterogeneous environmental knowledge.** We develop a unified representation mechanism that aligns unordered POIs with road segments and integrates complementary environmental granularities through road-conditioned aggregation, enabling heterogeneous urban knowledge to support trajectory generation.

- **A joint road–environment generation framework.** Our dual-view autoregressive framework learns complementary road transitions and environmental history, improving the spatiotemporal distributional fidelity of generated trajectories.

- **Empirical validation of generation quality.** Experiments in Beijing, Chengdu, and Xi'an show that MUSE outperforms the compared baselines on all four spatiotemporal distribution metrics, supporting the value of environmental modeling for trajectory generation.

2 Related Work

Human mobility trajectory generation. In coordinate space, DiffTraj (Zhu et al., 2023) learns GPS trajectory distributions through diffusion conditioned on trip attributes, and ControlTraj (Zhu et al., 2024) uses road topology to guide coordinate generation. TrajFlow (Li et al., 2026) combines flow matching with trajectory harmonization and reconstruction to support generation across urban, metropolitan, and nationwide scales. Producing discrete road-segment sequences from coordinate outputs requires map matching.

In road-network space, TS-TrajGen (Jiang et al., 2023) combines two-stage adversarial learning with A* search to generate connected paths. Diff-RNTraj (Wei et al., 2024) uses diffusion over a hybrid representation of discrete roads and continuous moving rates. HOSER (Cao et al., 2025) integrates road-, region-, and trajectory-level semantics with destination navigation, and MobilityGPT (Haydari et al., 2026) combines gravity-based sampling with reinforcement learning from trajectory feedback for road-trajectory generation.

Environmental knowledge in mobility modeling. Environmental knowledge supports mobility generation, trajectory representation learning, and next-location prediction. MoveSim (Feng et al., 2020) uses POI category distributions to describe functional similarity between locations, alongside geographic distance and historical transitions.

TransferTraj (Wei et al., 2025) integrates spatial, temporal, POI, and road-network modalities in a unified masking and recovery framework for learning across regions and tasks. CORE (Cao et al., 2026) is a trajectory representation learning framework that combines LLM-based multi-granular environmental semantics with context-aware route-choice modeling to learn trajectory embeddings for downstream tasks, including trajectory generation. AgentMove (Feng et al., 2025) combines individual mobility memory, urban structural knowledge, and collective patterns for next-location prediction. MUSE uses Road Query to align aggregate functional statistics and POI instances with individual roads, and jointly models road and environmental sequences autoregressively to incorporate multi-granular environmental knowledge into trajectory generation.

3 Methodology

MUSE models road-segment trajectories from two complementary views, Road and Environment. The model encodes road structure and the surrounding environment into road-indexed tokens and integrates historical information from both views through dual-branch autoregressive learning to generate road sequences and their corresponding time increments (Figure 2).

3.1 Preliminaries

Road network. A road network is represented as a directed graph $G = (\mathcal{R}, \mathcal{E})$, where $\mathcal{R}$ contains N road segments and $\mathcal{E} \subseteq \mathcal{R} \times \mathcal{R}$ represents legal directed transitions. A road r has a geographic polyline and a static attribute vector $\mathbf{f}_r$, such as road type, length, speed limit, and lane count. Its legal successors are $\mathrm{Succ}(r) = \{r' : (r, r') \in \mathcal{E}\}$, and all road attributes are denoted by $\mathbf{F}_{\mathcal{R}}$.

Trajectory. Map matching converts time-ordered geographic observations into a road-segment trajectory $\tau = ((r_1, t_1), \ldots, (r_L, t_L))$, where $(r_i, r_{i+1}) \in \mathcal{E}$. The time increment between consecutive records is $\Delta_{i+1} = t_{i+1} - t_i \in \mathbb{Z}_{\geq 0}$, measured in seconds.

Point of interest. Let $\mathcal{P}$ be the citywide POI set, with $p = (\mathbf{u}_p, c_p, b_p)$ representing a POI's coordinates, functional category, and subcategory. Road geometry and POI coordinates determine the environmental association set $\mathcal{P}_r \subseteq \mathcal{P}$ for road r; the same POI may be associated with several nearby roads.

Trajectory generation. Given G, $\mathbf{F}_{\mathcal{R}}$, $\mathcal{P}$, and a training set of real trajectories, we learn $p_\theta(\tau \mid \mathcal{I}, G, \mathbf{F}_{\mathcal{R}}, \mathcal{P})$, where $\mathcal{I} = (O, D, t_1)$ specifies the origin, destination, and departure time. The goal is to synthesize road-level trajectories whose spatial and temporal distributions resemble real mobility, predicting at each step from the candidate set $\mathcal{C}(r_i) = \mathrm{Succ}(r_i) \cup \{\mathrm{EOS}\}$, which comprises legal successors and a termination token.

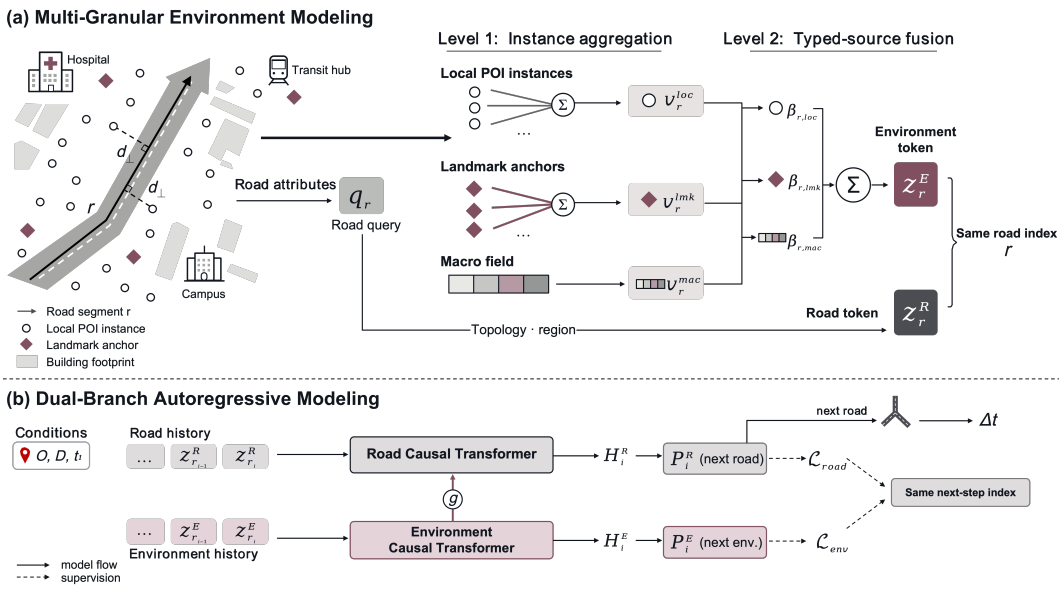

Figure 2: Overview of MUSE: (a) road-aligned environment modeling from multi-granular POI sources; (b) dual-branch autoregressive modeling with shared next-step supervision and gated environmental enhancement.

3.2 Road Tokenizer

The Road tokenizer integrates road-segment attributes, directed topology, and regional context. Field-wise encoding and fusion of discrete and continuous road attributes produce $\mathbf{q}_r \in \mathbb{R}^d$, which also serves as the road query for environmental aggregation. We learn a Node2Vec-style topology prior $\mathbf{v}_r^{\mathrm{n2v}}$ through second-order biased walks on the directed graph and a skip-gram objective (Grover & Leskovec, 2016); the pretrained vectors are frozen during trajectory learning and mapped to $\mathbf{n}_r$ through a trainable projection. Let $\rho(r)$ be the region containing road r in the urban region set $\mathcal{Z}$; its functional and urban-form attributes are encoded as $\mathbf{z}^{\mathrm{reg}}_{\rho(r)}$. Road attributes are first fused with the topology prior, then combined with the regional representation to form the Road token:

$$\bar{\mathbf{z}}_r^R = \mathrm{MLP}_{\mathrm{topo}}([\mathbf{q}_r; \mathbf{n}_r]), \qquad \mathbf{z}_r^R = \mathrm{LN}\Big(\mathrm{MLP}_{\mathrm{road}}([\bar{\mathbf{z}}_r^R; \mathbf{z}^{\mathrm{reg}}_{\rho(r)}])\Big). \tag{1}$$

The Road vocabulary $\mathbf{Z}^R \in \mathbb{R}^{(N+1)\times d}$ contains the road tokens and a learnable EOS vector. Field and encoding details are given in Appendix A.

3.3 Environment Tokenizer

To address the differences in spatial granularity shown in Figure 1b, the Environment tokenizer organizes POIs around roads into multi-granular information and uses two levels of road queries for instance aggregation and source fusion.

3.3.1 Road-Aligned Multi-Granular POI Environment

We construct three sources, Macro, Local, and Landmark, offline from the same cleaned static POI snapshot.

Macro: aggregate functional composition. Let $\mathcal{P}_r^{\mathrm{mac}}$ be the set of POIs within a road buffer of area A_r, and let $d_\perp(r,p)$ be the shortest distance from a POI to the road polyline. A distance-decayed density summarizes the local functional intensity of category k:

$$h_{r,k} = \frac{1}{A_r} \sum_{p \in \mathcal{P}_r^{\mathrm{mac}}} e^{-d_\perp(r,p)/\ell_{\mathrm{mac}}} \mathbb{I}(c_p = k). \tag{2}$$

Here ℓ_{mac} is the distance-decay scale. The normalized densities of all categories, total density, and a POI availability indicator form $\mathbf{x}_r^{\mathrm{mac}}$, which an MLP encodes into the source representation $\mathbf{v}_r^{\mathrm{mac}}$.

Local: nearby instances. Within the vicinity of a road, we combine road association, distance, and subcategory diversity to select up to K_{loc} nearby POIs into $\mathcal{P}_r^{\mathrm{loc}}$. Each instance retains its category, subcategory, and road-relative geometry (distance to the road and proximity of its projection to the road endpoints) to describe local functional and positional details.

Landmark: representative functional anchors. Based on the functional representativeness of candidate facilities, we use an LLM for semantic filtering, combining road context and candidate attributes to select representative facilities with clear functional identities, such as hospitals and transportation hubs. Up to K_{lmk} anchors form $\mathcal{P}_r^{\mathrm{lmk}}$, retaining their categories, subcategories, the same road-relative geometry, and LLM-assigned role and confidence labels. The selection rules and geometric definitions for the three sources are given in Appendix B.

3.3.2 Road-Query Alignment and Typed-Source Fusion

Both aggregation levels use scaled dot-product attention Attn, projecting the road query $\mathbf{q}_r$ through W_q^s and W_q^{src}, respectively.

Road-conditioned instance aggregation. Local and Landmark use independent parameters to jointly encode their respective semantic fields and road-relative geometry as $\mathbf{u}_{rp}^s \in \mathbb{R}^d$. For a nonempty source $s \in \{\mathrm{loc}, \mathrm{lmk}\}$, the valid instance encodings form the columns of $\mathbf{U}_r^s$ and are aggregated according to their relevance to the road query:

$$\mathbf{v}_r^s = W_o^s \,\mathrm{Attn}\big(W_q^s \mathbf{q}_r,\, W_k^s \mathbf{U}_r^s,\, W_v^s \mathbf{U}_r^s\big) + \mathbf{b}_o^s. \tag{3}$$

Typed-source fusion. We add a type embedding to each source, $\widetilde{\mathbf{v}}_r^s = \mathbf{v}_r^s + \mathbf{e}_s^{\mathrm{src}}$, and form the valid set $\mathcal{S}_r$ from Macro and the nonempty instance sources, whose representations form the columns of $\widetilde{\mathbf{V}}_r$. The source-level query learns the contribution of information at different granularities to the current road:

$$\mathbf{z}_r^E = \mathrm{MLP}_{\mathrm{src}}\Big(\mathrm{Attn}\Big(W_q^{\mathrm{src}} \mathbf{q}_r,\, W_k^{\mathrm{src}} \widetilde{\mathbf{V}}_r,\, W_v^{\mathrm{src}} \widetilde{\mathbf{V}}_r\Big)\Big). \tag{4}$$

Macro's availability indicator records missing POI information. Instance encoding, the expanded attention formulation, and the handling of empty sources are given in Appendix B. The Environment vocabulary $\mathbf{Z}^E \in \mathbb{R}^{(N+1)\times d}$ contains an independently learned EOS token and shares road indices with $\mathbf{Z}^R$.

3.4 Asymmetrically Coupled Dual-Branch Autoregressive Modeling

Road and Environment tokens arranged along a trajectory form synchronized sequences that describe the traversed road segments and their functional context, respectively (Figure 1c).

Synchronized sequence encoding. The shared temporal encoding $\boldsymbol{\tau}_i$ combines the preceding time increment, weekday, time-of-day slot, and circular features of seconds of day; the two branch inputs are $\mathbf{x}_i^R = \mathbf{z}_{r_i}^R + \boldsymbol{\tau}_i$ and $\mathbf{x}_i^E = \mathbf{z}_{r_i}^E + W_E^{\mathrm{time}} \boldsymbol{\tau}_i$. For a trajectory prefix of length T, a destination token is prepended at position zero, and the two causal branches encode their respective histories:

$$\begin{aligned} \mathbf{X}^k &= (\mathbf{z}_D^k, \mathbf{x}_1^k, \ldots, \mathbf{x}_T^k), \\ \mathbf{H}^k &= \big(\mathrm{CausalTransformer}_k(\mathbf{X}^k)_i\big)_{i=1}^T, \quad k \in \{R, E\}. \end{aligned} \tag{5}$$

Each branch uses a causal Transformer with relative spatiotemporal positional encoding (Vaswani et al., 2017), incorporating road midpoints and time of day into attention through rotary positional encoding (Su et al., 2024). Details of temporal encoding and the internal Transformer structure are given in Appendix C.

Environment-enhanced Road states. Environmental history is projected into the Road state space and enhances the Road state through an elementwise residual gate:

$$\begin{aligned} \mathbf{c}_i^E &= W_E \mathbf{H}_i^E, \qquad \mathbf{g}_i = \mathrm{sigmoid}(W_g[\mathbf{H}_i^R; \mathbf{c}_i^E] + \mathbf{b}_g), \\ \widetilde{\mathbf{H}}_i^R &= \mathrm{LN}(\mathbf{H}_i^R + \mathbf{g}_i \odot \mathbf{c}_i^E). \end{aligned} \tag{6}$$

A common mobility target. Each branch scores candidates in its own token space:

$$s^R_{i,c} = (W^R_o \widetilde{\mathbf{H}}^R_i)^\top \mathbf{z}^R_c, \qquad s^E_{i,c} = (W^E_o \mathbf{H}^E_i)^\top \mathbf{z}^E_c, \qquad c \in \mathcal{C}(r_i). \tag{7}$$

Applying softmax to the candidate scores yields P^R_i and P^E_i, with both branches predicting the next road segment or EOS. Through its own prediction objective, the Environment branch learns the relationship between environmental history and subsequent road choices.

3.5 Temporal Modeling and Joint Learning

Candidate-conditioned time prediction. The time head combines the hidden state $\widetilde{\mathbf{H}}^R_i$ after road–environment fusion with the Road token $\mathbf{z}^R_c$ of a nonterminal candidate and predicts a standardized log time increment $\widehat{y}_{i+1,c}$ through candidate modulation. Its formulation is given in Appendix D.1.

Joint objective. For a trajectory of length L, the next-road target is $y_i = r_{i+1}$ for $i < L$, and the final target is $y_L = \text{EOS}$. Teacher-forced training minimizes

$$\mathcal{L} = \lambda_R \mathcal{L}_{\text{road}} + \lambda_E \mathcal{L}_{\text{env}} + \lambda_\Delta \mathcal{L}_{\text{time}} + \lambda_{\text{trip}}(e) \mathcal{L}_{\text{trip}}. \tag{8}$$

$\mathcal{L}_{\text{road}}$ and $\mathcal{L}_{\text{env}}$ are mean next-road cross-entropies. $\mathcal{L}_{\text{time}}$ measures the SmoothL1 error between predicted and true standardized log increments over real transitions; $\mathcal{L}_{\text{trip}}$ compares the logarithms of predicted and true total trip durations to constrain accumulated timing error, with its weight $\lambda_{\text{trip}}(e)$ warmed up over training epochs. This objective jointly optimizes the trainable parameters of the tokenizers and sequence modules; time recovery and loss details are given in Appendix D.1.

3.6 Trajectory Generation

Given (O, D, t_1), we compute the two static vocabularies and initialize synchronized prefixes. At each step, we sample a candidate from P^R_i; if it is EOS, generation stops. Otherwise, we append road $\widehat{r}_{i+1}$, recover the time increment from the corresponding prediction $\widehat{y}_{i+1,\widehat{r}_{i+1}}$, and accumulate it to obtain the new timestamp. Both branches receive the new road and timestamp, and generation continues until EOS or a termination safeguard is triggered. The complete procedure is given in Appendix D.

4 Experiments

We evaluate trajectory generation in Beijing, Chengdu, and Xi'an, and use ablation studies to examine the roles of incorporating environmental information, multi-granular representations, and environmental sequence learning.

4.1 Experimental Setup

Data and task. We use trajectories and accompanying road and regional data for Beijing (BJ), Chengdu (CD), and Xi'an (XA) distributed through VecCity (Zhang et al., 2025)[1]. The trajectories in all three cities record vehicle travel and are represented as sequences of road segments with timestamps after map matching; the accompanying data include road geometry, connectivity, road-segment attributes, and regional information.

The POI data come from the Amap POI dataset published on the Peking University Open Research Data Platform (State Information Center, 2017)[2] and contain POI names, categories, latitude and longitude, addresses, and other information.

Each city is trained and evaluated separately using an 80/10/10 split in fixed file order. Generation is conditioned on the origin road, destination road, and departure time. Table 1 summarizes the datasets; POIs are counted within 1 km of the dataset road network, before road-wise selection.

[1] https://github.com/Beihang-BIGSCity/VecCity.

[2] https://doi.org/10.18170/DVN/WSXCNM.

Table 1: Dataset statistics. POIs are counted in the road-network study area (a 1 km buffer around the dataset roads), before road-wise selection and aggregation.

Statistic	Beijing (BJ)	Chengdu (CD)	Xi'an (XA)
Trajectories	1,018,312	559,729	384,618
Directed road segments	40,306	6,195	5,269
Directed connections	101,024	15,962	13,032
Urban regions	11,208	1,306	1,056
POIs in the study area	564,291	180,471	171,040

Baselines. We compare with Dijkstra and nine published generators: MoveSim (Feng et al., 2020), VOLUNTEER (Long et al., 2023), TS-TrajGen (Jiang et al., 2023), HOSER (Cao et al., 2025), MobilityGPT (Haydari et al., 2026), DiffTraj (Zhu et al., 2023), ControlTraj (Zhu et al., 2024), Diff-RNTraj (Wei et al., 2024), and TrajFlow (Li et al., 2026). These methods cover deterministic routing, GANs, VAEs, diffusion, flow matching, autoregressive generation, and search-based generation; experimental protocols and coordinate-space adaptation are described in Appendix H. Unreported metrics are denoted by a dash.

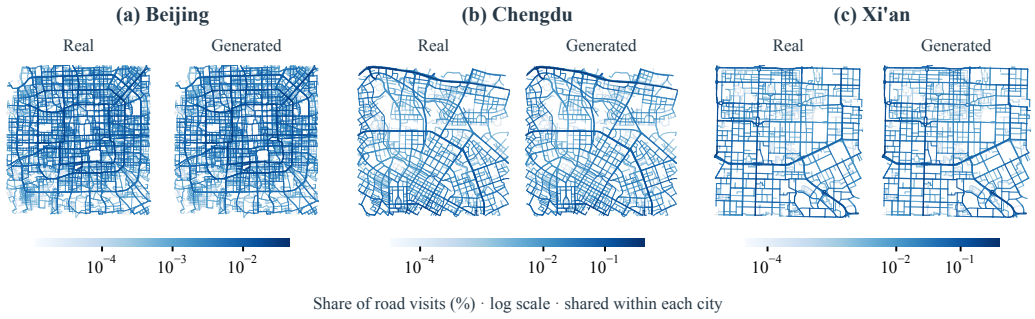

Figure 3: Real and generated road visits in three cities. Each pair shows held-out real trajectories (left) and three generated draws per trip condition (right). Colors indicate normalized visits with a shared logarithmic scale and full-network extent within each city; coincident opposite directions are combined.

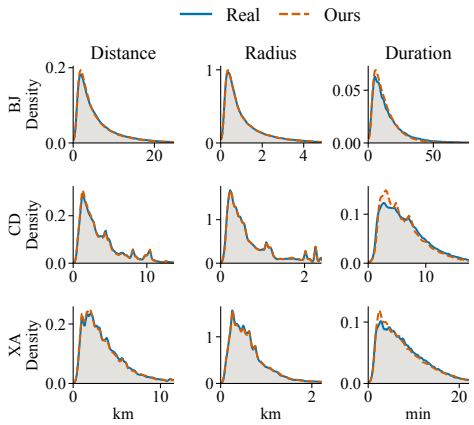

Figure 4: Central-range trip densities for BJ, CD, and XA, normalized over all trips.

Metrics. We evaluate generation quality in terms of travel scale, temporal characteristics, and regional visitation patterns. Distance measures the cumulative path length of a trip, reflecting travel distance; Radius measures the mean distance from trajectory locations to their geographic center, characterizing the spatial extent of travel. Together, they describe the spatial scale of mobility. Duration measures the elapsed time of a trip, reflecting its temporal characteristics; LocFreq measures the relative visitation frequencies of different regions, characterizing the spatial distribution of population mobility. We use Jensen–Shannon distance (JSD) to compare these four types of distributions between real and generated trajectory sets, with lower values indicating higher distributional fidelity. Metric definitions are given in Appendix E, and model and training settings in Appendix F.

4.2 Overall Performance

Across Beijing, Chengdu, and Xi'an, MUSE achieves the lowest JSD in all 12 city–metric combinations in Table 2, improving both trip-level spatiotemporal statistics and city-level visitation patterns. The advantage is especially large for Distance and Duration: relative to HOSER, the strongest baseline on these metrics, MUSE reduces JSD by 56.8%–63.8% and 64.6%–79.4%, respectively, across the three cities.

The quantitative gains are also visible in the spatial structure of the generated trajectories. Generated road-visitation patterns preserve the concentration on high-frequency roads and the sparse use of low-frequency areas in each city (Figure 3). Within the displayed ranges, the Distance, Radius, and Duration curves align with the major peaks and decay trends of the real distributions while retaining city-specific shapes (Figure 4). This distributional agreement accompanies variation in individual routes: in the three matched-OD cases in Appendix G, generated paths follow major corridors observed in real trips and include local route variations. Together, these observations connect the aggregate gains to realistic road-network use and trip characteristics, with route variation under shared origin–destination conditions.

Table 2: Overall trajectory generation performance on four macro-level distributions (JSD, lower is better). Dist. and Dur. abbreviate Distance and Duration. Best results are in **bold**; – denotes an unreported metric.

Method	Beijing (BJ)				Chengdu (CD)				Xi'an (XA)			
	Dist.	Radius	Dur.	LocFreq	Dist.	Radius	Dur.	LocFreq	Dist.	Radius	Dur.	LocFreq
Dijkstra	0.0623	0.0613	–	0.1974	0.1590	0.1244	–	0.1358	0.1282	0.0782	–	0.1458
MoveSim	0.2713	0.4490	–	0.2913	0.2001	0.5289	–	0.3342	0.2476	0.4343	–	0.3692
TS-TrajGen	0.0834	0.0441	–	0.0870	0.1010	0.0674	–	0.0519	0.0766	0.0315	–	0.0901
VOLUNTEER	0.3345	0.7280	0.7271	0.4443	0.2862	0.5932	0.3179	0.4270	0.2638	0.5963	0.3483	0.4819
DiffTraj	0.7753	0.0229	–	0.3628	0.6935	0.0994	–	0.3187	0.7375	0.0593	–	0.2998
ControlTraj	0.7501	0.0193	–	0.3209	0.6654	0.1021	–	0.3377	0.6305	0.0572	–	0.2794
Diff-RNTraj	0.4047	0.2144	–	0.3553	0.2809	0.1788	–	0.3283	0.2277	0.2097	–	0.3564
HOSER	0.0583	0.0215	0.1324	0.0587	0.0929	0.0647	0.1854	0.0449	0.0725	0.0316	0.1361	0.0519
MobilityGPT	0.2062	0.2436	–	0.2300	0.2893	0.3718	–	0.2554	0.2934	0.4039	–	0.2224
TrajFlow	0.5191	0.0519	–	0.3198	0.3466	0.1145	–	0.1635	0.3389	0.0720	–	0.1760
MUSE	**0.0211**	**0.0129**	**0.0469**	**0.0434**	**0.0353**	**0.0175**	**0.0439**	**0.0271**	**0.0313**	**0.0221**	**0.0280**	**0.0270**

4.3 Ablation Study

We compare four variants in Table 3. **w/o POI** removes the POI-based Environment tokenizer and the Environment branch, retaining the Road generator. **w/o Env Branch** concatenates the multi-granular representation with road features in one Road branch, removing independent environmental prediction and cross-branch interaction. **w/o Multi-granular** and **w/o Landmark** retain both branches and their interaction, using Macro alone and Macro with Local, respectively.

Table 3: Ablation study of POI information, the Environment branch, and multi-granular modeling across three cities (JSD, lower is better). Best results are in **bold**.

Method	Beijing (BJ)				Chengdu (CD)				Xi'an (XA)			
	Dist.	Radius	Dur.	LocFreq	Dist.	Radius	Dur.	LocFreq	Dist.	Radius	Dur.	LocFreq
w/o POI	0.0274	0.0164	0.0692	0.0553	0.0410	0.0231	0.0758	0.0301	0.0392	0.0249	0.0865	0.0304
w/o Env Branch	0.0249	0.0148	0.0584	0.0498	0.0384	0.0206	0.0617	0.0290	0.0358	0.0238	0.0589	0.0291
w/o Multi-granular	0.0232	0.0139	0.0527	0.0468	0.0368	0.0188	0.0516	0.0281	0.0339	0.0229	0.0417	0.0280
w/o Landmark	0.0219	0.0133	0.0486	**0.0420**	0.0359	0.0180	0.0462	**0.0267**	0.0323	0.0224	0.0321	**0.0266**
MUSE	**0.0211**	**0.0129**	**0.0469**	0.0434	**0.0353**	**0.0175**	**0.0439**	0.0271	**0.0313**	**0.0221**	**0.0280**	0.0270

Road-aligned, multi-granular context complements the transition patterns learned by the Road generator: both the full model and w/o Env Branch improve on w/o POI in all 12 city–metric comparisons. With the dual-branch architecture fixed, adding Local to Macro improves all four metrics across the three cities. Adding Landmark further reduces Distance, Radius, and Duration JSD, while LocFreq JSD increases slightly by 1.5%–3.3%. The full model improves all four metrics over the Macro-only variant, supporting the joint use of aggregate context and instance-level detail.

With multi-granular input fixed, the full model improves every metric over w/o Env Branch, reducing Duration JSD by 19.7%–52.5% across cities. These gains support the independent Environment branch, its explicit prediction objective, and Environment-to-Road interaction as a joint mechanism for using environmental history to guide subsequent road choices.

4.4 Analysis of Environmental Information Use

Figure 5 examines road-dependent environmental information allocation in Xi'an. At the entity level, the campus-access road primarily attends to the school rather than the similarly distant school gate or road exit, whereas the residential road favors the residential community over the closer pharmacy and road-related cues. These cases show that the Road Query can prioritize facilities consistent with road function among spatially proximate candidates.

At the sequence level, static road-source weights differ across trajectory positions. The combined Local and Landmark weight at origins and destinations is, on average, 5.23 percentage points higher than en route, while intermediate roads place relatively more weight on Macro functional composition. Across-road differences in source weights further show that roads form distinct combinations of Macro, Local, and Landmark information. These road-specific source mixtures complement the ablation evidence for multi-granular representations.

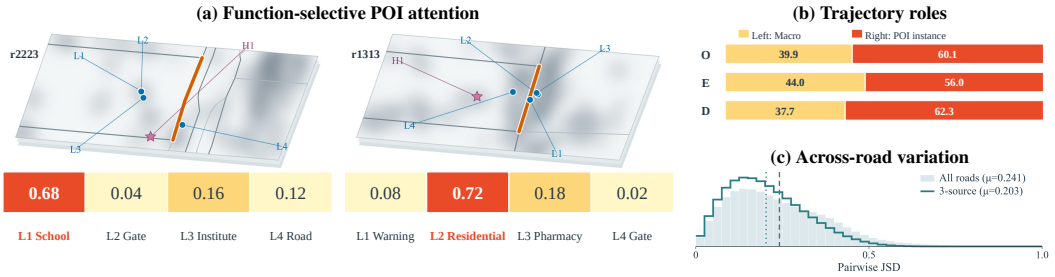

Figure 5: How does MUSE allocate environmental information? (a) Road context and Local entity attention; gray fields, blue circles, and magenta stars denote Macro, Local, and Landmark information. (b) Mean Macro and combined Local–Landmark weights across trajectory roles. (c) Distribution of pairwise JS distances between roads' source-weight vectors. "3-source" restricts comparisons to roads with all three sources available.

5 Conclusion

We introduced MUSE, a multi-granular urban semantic environment modeling framework that integrates road-network topology with road-aligned environmental knowledge. The framework associates heterogeneous environmental information with individual roads and jointly models road transitions and urban environmental context. MUSE improves the fidelity of generated trajectories in travel distance, spatial radius, trip duration, and location visitation distributions. Ablation and attention analyses further support the roles of multi-granular environmental representations and joint sequence learning. These results show that extracting road-associated urban knowledge can complement transition-pattern learning and support more faithful generation of urban mobility.

Limitations and future work. The current approach remains limited by imperfect temporal and spatial alignment among trajectory, road-network, and POI data, and does not yet fully capture the complex mechanisms through which urban environments influence travel decisions. Future work will combine temporally consistent urban data with richer behavioral signals to further investigate the complex dependencies and potential causal relationships among urban environments, travel intent, and road transitions, and to develop more interpretable and adaptive environment-aware trajectory generation methods.

AI USE STATEMENT

We used AI tools to assist with language polishing, translation, paragraph organization, and initial drafting. As part of the method, `gpt-5.6-sol` was used offline to select Landmark POIs and assign their role and confidence labels, as described in Appendix F.2. The authors are responsible for the final manuscript and reported results.

REFERENCES

Ji Cao, Tongya Zheng, Qinghong Guo, Yu Wang, Junshu Dai, Shunyu Liu, Jie Yang, Jie Song, and Mingli Song. Holistic semantic representation for navigational trajectory generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):40–48, 2025. doi: 10.1609/aaai.v39i1.31978.

Ji Cao, Yu Wang, Tongya Zheng, Jie Song, Qinghong Guo, Zujie Ren, Canghong Jin, Gang Chen, and Mingli Song. Capturing context-aware route choice semantics for trajectory representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 38(10):6408–6423, 2026. doi: 10.1109/TKDE.2026.3708608.

Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. Learning to simulate human mobility. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3426–3433, 2020. doi: 10.1145/3394486.3412862.

Jie Feng, Yuwei Du, Jie Zhao, and Yong Li. AgentMove: A large language model based agentic framework for zero-shot next location prediction. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1322–1338, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.61.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, 2016. doi: 10.1145/2939672.2939754.

Ammar Haydari, Dongjie Chen, Zhengfeng Lai, Michael Zhang, and Chen-Nee Chuah. MobilityGPT: Enhanced human mobility modeling with a GPT model. *IEEE Transactions on Intelligent Transportation Systems*, 27(1):1681–1694, 2026. doi: 10.1109/TITS.2025.3626357.

Wenjun Jiang, Wayne Xin Zhao, Jingyuan Wang, and Jiawei Jiang. Continuous trajectory generation based on two-stage GAN. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4374–4382, 2023. doi: 10.1609/aaai.v37i4.25557.

Minjin Lee and Petter Holme. Relating land use and human intra-city mobility. *PLOS ONE*, 10(10):e0140152, October 2015. doi: 10.1371/journal.pone.0140152. URL `https://doi.org/10.1371/journal.pone.0140152`.

Peiran Li, Jiawei Wang, Haoran Zhang, Xiaodan Shi, Noboru Koshizuka, Chihiro Shimizu, and Renhe Jiang. TrajFlow: Nation-wide pseudo GPS trajectory generation with flow matching models. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=BDOldEjwCE`.

Qingyue Long, Huandong Wang, Tong Li, Lisi Huang, Kun Wang, Qiong Wu, Guangyu Li, Yanping Liang, Li Yu, and Yong Li. Practical synthetic human trajectories generation based on variational point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4561–4571, 2023. doi: 10.1145/3580305.3599888.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

State Information Center. Map POI (point of interest) data. Peking University Open Research Data Platform, V2, 2017. URL `https://doi.org/10.18170/DVN/WSXCNM`.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Tonglong Wei, Youfang Lin, Shengnan Guo, Yan Lin, Yiheng Huang, Chenyang Xiang, Yuqing Bai, and Huaiyu Wan. Diff-RNTraj: A structure-aware diffusion model for road network-constrained trajectory generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):7940–7953, 2024. doi: 10.1109/TKDE.2024.3460051.

Tonglong Wei, Yan Lin, Zeyu Zhou, Haomin Wen, Jilin Hu, Shengnan Guo, Youfang Lin, Gao Cong, and Huaiyu Wan. TransferTraj: A vehicle trajectory learning model for region and task transferability. In *Advances in Neural Information Processing Systems*, volume 38, pp. 109941–109968. Curran Associates, Inc., 2025.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10524–10533, 2020.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wentao Zhang, Jingyuan Wang, Yifan Yang, and Leong Hou U. VecCity: A taxonomy-guided library for map entity representation learning. *Proceedings of the VLDB Endowment*, 18(8): 2575–2588, 2025. doi: 10.14778/3742728.3742749. URL https://doi.org/10.14778/3742728.3742749.

Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Jianqiao Yu. DiffTraj: Generating GPS trajectory with diffusion probabilistic model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 65168–65188, 2023. doi: 10.52202/075280-2843.

Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. ControlTraj: Controllable trajectory generation with topology-constrained diffusion model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4676–4687, 2024. doi: 10.1145/3637528.3671866.

Appendices A–D supplement the methodology with details of input encoding, POI construction, two-level attention, temporal modeling, training, and generation. Appendix E defines the macro-level evaluation metrics; Appendix F reports model, training, and LLM settings; Appendix G presents OD-conditioned micro-level results and route examples; and Appendix H describes the baseline protocols and coordinate-space adaptation.

A Road and Regional Inputs

Spatial data. The trajectory, road, and region resources follow the VecCity data representation (Zhang et al., 2025). Roads are directed segments with geographic polylines; region attributes describe urban function and built form. A fixed road index is used throughout trajectory records, structural vectors, POI banks, and the two token vocabularies. The map $\rho(r)$ associates each road with a region, whose attributes are encoded by the Road tokenizer.

Field-wise encoding. Table 4 lists the static fields. Each categorical field uses an independent embedding; each scalar field uses an independent affine projection. The field width is 16. For a concatenated input of width d_{in}, the road-attribute, regional, topology-fusion, and road–region fusion MLPs use

$$\text{Linear}(d_{\text{in}}, 2d) \longrightarrow \text{GELU} \longrightarrow \text{Linear}(2d, d) \longrightarrow \text{LN}. \tag{9}$$

Let $a_{r,j}^{\text{disc}}$, $a_{r,k}^{\text{cont}}$ and $b_{z,j}^{\text{disc}}$, $b_{z,k}^{\text{cont}}$ denote the discrete and normalized continuous fields of roads and regions, respectively. The attribute query and regional representation are

$$\begin{aligned} \mathbf{q}_r &= \text{MLP}_{\text{attr}}\left(\left[\Big\|_{j=1}^{J_d} \text{Emb}_j^{\text{attr}}(a_{r,j}^{\text{disc}}); \Big\|_{k=1}^{J_c} \text{Lin}_k^{\text{attr}}(a_{r,k}^{\text{cont}})\right]\right), \\ \mathbf{z}_z^{\text{reg}} &= \text{MLP}_{\text{reg}}\left(\left[\Big\|_{j=1}^{K_d} \text{Emb}_j^{\text{reg}}(b_{z,j}^{\text{disc}}); \Big\|_{k=1}^{K_c} \text{Lin}_k^{\text{reg}}(b_{z,k}^{\text{cont}})\right]\right). \end{aligned} \tag{10}$$

The attribute-module output $\mathbf{q}_r$ is used both to construct Road tokens and in the two levels of POI attention. Following Eq. (1), the road representation sequentially integrates topology and regional information and passes through a final LayerNorm to produce the Road token.

Table 4: Static fields used by the Road tokenizer. Each field has an independent embedding or projection before fusion.

Field group	Attributes
Road, categorical	Highway type, tunnel, bridge, roundabout, one-way indicator
Road, continuous	Length, lane count, speed limit, in-degree, out-degree, tortuosity
Region, categorical	Functional category, central-business-district indicator, urban-form type
Region, continuous	Functional mixture, density, building count, floor count, floor-area ratio, building coverage, perimeter, area

Scaling. Nonnegative road fields use a field-specific log transform $\log(1+x)/\log(1+s)$, where s is the corresponding scale. Regional functional mixture is retained on its original scale; the remaining nonnegative regional fields use log scaling. These transformations precede field encoding. Static road, region, and POI inputs are available for the city independently of trajectory supervision; time-increment normalization statistics are computed only from training trajectories.

Topology prior. Biased second-order walks follow outgoing edges of the directed road graph. A skip-gram objective learns road co-occurrence within walk contexts (Grover & Leskovec, 2016; Mikolov et al., 2013). The configuration uses walk length 40, 20 walks per road, return parameter 1, exploration parameter 2, context window 10, five uniform negative samples, and 20 pretraining epochs. The resulting vectors are row-wise ℓ_2 normalized and frozen during trajectory training. A trainable projection produces $\mathbf{n}_r = \text{LN}(W_n \mathbf{v}_r^{\text{n2v}} + \mathbf{b}_n)$, and the fusion layers also participate in trajectory training. Pretraining of the topology prior uses only the road graph.

B CONSTRUCTING ROAD-ALIGNED POI ENVIRONMENTS

Common POI bank. All three sources are constructed from the same deduplicated open POI snapshot (State Information Center, 2017). POI coordinates are represented in WGS-84, and spatial distances and buffer areas are evaluated in metric units. The city bank is cropped to the area within 1 km of the road network. Each retained POI has a category and subcategory under the common city taxonomy. Names and identifiers are used for offline selection and deduplication; the source-specific neural inputs are described below. The three sources select POIs according to their respective rules and share road indices.

B.1 POI CATEGORIES AND SUBCATEGORIES

For POI p, c_p denotes its top-level functional category and b_p its second-level subcategory. These fields are the first and second entries of the original semicolon-separated type string. For example, "education, science and culture; schools; higher-education institutions" yields the category "education, science and culture" and the subcategory "schools"; the third level remains part of the original type information and is not embedded separately. Each city builds category and subcategory vocabularies from the strings present in its cleaned POI bank, with Local and Landmark sharing the corresponding index mappings.

Table 5 lists the complete verified inventory for the Xi'an POI bank: 22 categories and 167 subcategories. The accompanying bilingual inventory maps each original Chinese label to its English description.

Table 5: Observed POI categories and subcategories in the Xi'an bank.

Category	Subcategories
Transport facilities	Transport-related services; Parking lots; Bus stops; Metro stations; Airport-related facilities; Ports and docks; Railway stations; Intercity coach stations.
Accommodation	Accommodation-related services; Hotels; Inns and guesthouses.
Sports and recreation	Leisure venues; Sports and recreation venues; Entertainment venues; Resorts and recuperation facilities; Cinemas and theatres; Sports venues; Golf-related facilities.
Public facilities	Public toilets; Public facilities; Public telephones; Newsstands; Emergency shelters.
Companies and enterprises	Companies; Companies and enterprises; Agriculture, forestry, livestock and fishery bases; Factories.
Healthcare	Specialist hospitals; Veterinary facilities; Healthcare facilities; Pharmacies and health-product stores; Emergency centres; Disease-prevention institutions; General hospitals; Clinics.
Commercial and residential properties	Industrial parks; Residential areas; Commercial and residential property services; Buildings.
Place names and addresses	Transport place names; Place names and addresses; General place names; Prominent place names; Natural-feature names; Street-address information.
Indoor facilities	Indoor facilities.
Motorcycle services	Motorcycle-related services; Motorcycle repair; Motorcycle sales.
Government and social organizations	Vehicle and transport administration; Public security, procuratorate and courts; Foreign institutions; Industry, commerce and taxation authorities; Government and social-organization services; Government agencies; Democratic parties; Social organizations.
Automotive services	Used-vehicle trading; Charging stations; Other energy stations; Gas-fuelling stations; Petrol stations; Automobile clubs; Car care and decoration; Roadside assistance; Automotive-related services; Car rental; Auto-parts sales; Car washes.
Automotive repair	Toyota-authorized repair; Audi-authorized repair; BMW-authorized repair; Automotive repair; General automotive repair.
Automotive sales	Dongfeng-authorized sales; Toyota-authorized sales; Volkswagen-authorized sales; BMW-authorized sales; Nissan-authorized sales; Honda-authorized sales; Automotive sales; Hyundai-authorized sales; Hongqi sales; Luxgen sales; Roewe sales; Truck sales; General Motors-authorized sales; Changan sales.

Table 5: Observed POI categories and subcategories in the Xi'an bank (continued).

Category	Subcategories
Daily-life services	Funeral facilities; Agencies; Professional offices; Employment markets; Information consultancy; Ticket offices; Infant-care services; Lottery outlets; Moving companies; Photography and photo-printing shops; Travel agencies; Bathing and massage venues; Laundries; Logistics and courier services; Daily-life service venues; Electricity service offices; Telecommunications service offices; Repair services; Beauty salons and hairdressers; Water service offices; Post offices.
Education, science and culture	Convention and exhibition centres; Media organizations; Museums; Libraries; Training institutions; Planetariums; Schools; Exhibition halls; Cultural palaces; Performing-arts groups; Archives; Science and technology museums; Education, science and culture venues; Research institutions; Art galleries; Driving schools.
Shopping	Specialty shops; Personal-care and cosmetics shops; Sporting-goods shops; Convenience stores; Shopping malls; Home-furnishing and building-material markets; Household-appliance and electronics stores; Stationery and cultural-goods shops; Clothing, footwear, hat and leather-goods shops; Special trading venues; Themed shopping streets; General markets; Flower, bird, fish and pet markets; Shopping-related venues; Supermarkets.
Access facilities	Street-facing compound gates; Building entrances; Virtual gates; Access facilities.
Road ancillary facilities	Toll stations; Warning information; Road ancillary facilities.
Finance and insurance	Insurance companies; ATMs; Securities firms; Finance companies; Financial and insurance institutions; Banks; Bank-related facilities.
Scenic attractions	Parks and squares; Scenic attractions; Attraction-related facilities.
Food and beverage	Chinese restaurants; Casual dining venues; Cold-drink shops; Cafes; Foreign-cuisine restaurants; Fast-food restaurants; Dessert shops; Bakeries; Teahouses; Food-and-beverage-related venues.

B.2 Macro Functional Statistics

For each road, we use a 20 m buffer around its polyline and a distance-decay scale of $\ell_{\mathrm{mac}} = 50\,\mathrm{m}$ in Eq. (2). Category-specific weighted counts are divided by the buffer area. Each density field is log transformed, clipped at its positive-value 99th percentile, and scaled by its positive-value 95th percentile within the city. A field with no positive observations remains zero. These operations reduce the numerical dominance of dense POI categories while retaining differences between roads. Total density and an availability indicator complete the $C + 2$-dimensional input $\mathbf{x}_r^{\mathrm{mac}}$, where C is the number of POI categories; the source representation is $\mathbf{v}_r^{\mathrm{mac}} = \mathrm{MLP}_{\mathrm{mac}}(\mathbf{x}_r^{\mathrm{mac}})$.

When the buffer contains no POIs, the availability indicator is zero. Macro still participates in source fusion, and its biased MLP learns the corresponding missing-environment representation.

B.3 Local Instance Selection

We first retrieve POI candidates within 150 m of a road. For potential associations within 60 m, a normalized ownership score compares the POI's proximity to nearby roads:

$$o_{rp} = \frac{\exp(-d_\perp(r,p)/20)}{\sum_{r':\, d_\perp(r',p)\le 60\,\mathrm{m}} \exp(-d_\perp(r',p)/20)}. \tag{11}$$

A city-specific Local radius is selected from $\{20, 30, 40, 60\}$ m using the criterion that at least 90% of roads have four candidates. The selection prefers the smallest radius meeting this coverage criterion. Up to $K_{\mathrm{loc}} = 4$ instances are retained per road. Subcategory diversity is prioritized, with remaining ties ordered by decreasing ownership score, increasing distance, and POI identifier.

For a retained POI, projection onto the road polyline gives a normalized position $s_{rp} \in [0, 1]$. The geometric input is $\mathbf{g}_{rp} = (\log(1 + d_\perp(r,p)), e_{rp})$, where $e_{rp} = \min(s_{rp}, 1 - s_{rp})$; the two components describe the POI's distance to the road and the proximity of its projection to the road endpoints, respectively. This endpoint-proximity measure is invariant to reversal of the stored polyline direction.

B.4 Landmark Selection and Source Identity

Landmark selection uses the same 150 m retrieval range. A category-covering candidate list of at most 32 POIs is passed to an offline semantic selector, which retains at most $K_{\mathrm{lmk}} = 2$ representative functional anchors. The selector evaluates the functional representativeness of candidate facilities using static road and POI information and runs offline during POI bank construction.

Local uses category, subcategory, road-relative geometry, and a validity mask; Landmark additionally uses the LLM-assigned role and confidence labels stored in its static bank. The encoders have independent parameters, and source-type embeddings identify their different roles at fusion. The LLM invocation settings and prompt template are given in Appendix F.2.

B.5 Instance Encoders and Missing Sources

Category, subcategory, role, and confidence embeddings each have width 32. Local concatenates the category and subcategory embeddings and applies $\mathrm{Linear}(64, d)$–GELU–LayerNorm; Landmark concatenates all four embeddings and applies $\mathrm{Linear}(128, d)$–GELU–LayerNorm. Geometry is encoded by $\mathrm{Linear}(2, d)$–GELU–$\mathrm{Linear}(d, d)$. Let role_{rp} and conf_{rp} denote the categorical role and confidence labels assigned to POI p for road r. The concatenated semantic inputs and instance encodings are

$$\begin{aligned}
\mathbf{a}^{\mathrm{loc}}_{rp} &= \big[\mathrm{Emb}^{\mathrm{loc}}_{\mathrm{cat}}(c_p); \mathrm{Emb}^{\mathrm{loc}}_{\mathrm{sub}}(b_p)\big], \\
\mathbf{a}^{\mathrm{lmk}}_{rp} &= \big[\mathrm{Emb}^{\mathrm{lmk}}_{\mathrm{cat}}(c_p); \mathrm{Emb}^{\mathrm{lmk}}_{\mathrm{sub}}(b_p); \\
&\qquad \mathrm{Emb}_{\mathrm{role}}(\mathrm{role}_{rp}); \mathrm{Emb}_{\mathrm{conf}}(\mathrm{conf}_{rp})\big], \\
\mathbf{u}^{s}_{rp} &= \mathrm{LN}\big(\mathrm{MLP}^{s}_{\mathrm{sem}}(\mathbf{a}^{s}_{rp}) + \mathrm{MLP}^{s}_{\mathrm{geo}}(\mathbf{g}_{rp})\big), \quad s \in \{\mathrm{loc}, \mathrm{lmk}\}.
\end{aligned} \tag{12}$$

The instance-attention query, key, and value projections are bias-free; the output projection has a bias. Invalid slots are masked before the attention softmax. For a fully empty Local or Landmark source, the pooled output is explicitly set to zero *after* the output projection. The empty source, together with its type embedding, is then excluded from the source-level softmax.

Source order is Macro, Local, Landmark. The source-attention projections are bias-free, and the final source MLP is $\mathrm{Linear}(d, d)$–GELU–LayerNorm. Category index zero is a valid category; padding is represented only by the validity mask. Since Macro is always present, the valid source set is never empty.

B.6 Expanded Two-Level Attention

Eqs. (3)–(4) use scaled dot-product attention with keys and values stored as columns:

$$\mathrm{Attn}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathbf{V}\,\mathrm{softmax}\left(\frac{\mathbf{K}^{\top}\mathbf{q}}{\sqrt{d}}\right). \tag{13}$$

Queries, keys, and values are projected before entering this operator; the instance-level output projection and source-level MLP remain outside the operator. For a nonempty source s, the columns of $\mathbf{U}^{s}_{r} \in \mathbb{R}^{d \times |\mathcal{P}^{s}_{r}|}$ are valid instance encodings; the columns of $\widetilde{\mathbf{V}}_{r} \in \mathbb{R}^{d \times |\mathcal{S}_{r}|}$ are valid source representations with source-type embeddings added. The two query projections, W^{s}_{q} and W^{src}_{q}, have independent parameters. The corresponding instance and source weights are

$$\begin{aligned}
\alpha^{s}_{rp} &= \mathrm{softmax}_{p \in \mathcal{P}^{s}_{r}}\left(\frac{(W^{s}_{q}\mathbf{q}_{r})^{\top} W^{s}_{k}\mathbf{u}^{s}_{rp}}{\sqrt{d}}\right), \\
\beta_{r,s} &= \mathrm{softmax}_{s \in \mathcal{S}_{r}}\left(\frac{(W^{\mathrm{src}}_{q}\mathbf{q}_{r})^{\top} W^{\mathrm{src}}_{k}\widetilde{\mathbf{v}}^{s}_{r}}{\sqrt{d}}\right).
\end{aligned} \tag{14}$$

C Temporal and Causal Sequence Encoding

C.1 Shared Temporal Basis

The preceding time increment is normalized as $\bar{y}_i = (\log(1 + \Delta_i) - \mu_{\text{train}})/\sigma_{\text{train}}$, where μ_{train} and σ_{train} are the mean and standard deviation of log time increments in the training set. Let s_i be local seconds of day, w_i the weekday, and b_i the time-of-day slot. The local time zone is UTC+8. Seven slots partition the day at hours 0, 6, 9, 12, 14, 17, 20, and 24. The multi-frequency circular representation is

$$\mathbf{c}(s_i) = [\sin(\boldsymbol{\omega}\theta_i); \cos(\boldsymbol{\omega}\theta_i)], \qquad \theta_i = \frac{2\pi s_i}{86400} - \pi. \tag{15}$$

Here $\boldsymbol{\omega} \in \mathbb{R}^{d/2}$ is a fixed frequency vector. The four temporal components are

$$\mathbf{t}_{i,1} = \text{Lin}_\Delta(\bar{y}_i), \quad \mathbf{t}_{i,2} = \text{Emb}_{\text{week}}(w_i), \quad \mathbf{t}_{i,3} = \text{Emb}_{\text{slot}}(b_i), \quad \mathbf{t}_{i,4} = \mathbf{c}(s_i). \tag{16}$$

Each component is scored by a learned linear map, $a_{i,j} = \text{Lin}_j(\mathbf{t}_{i,j})$, with dropout applied to the scores during training. Token-dependent softmax weights combine the components:

$$\boldsymbol{\tau}_i = \text{LN}\left(\sum_{j=1}^{4} \gamma_{i,j}\mathbf{t}_{i,j}\right), \qquad \gamma_{i,j} = \frac{\exp(a_{i,j})}{\sum_{k=1}^{4}\exp(a_{i,k})}. \tag{17}$$

Both branches share this temporal basis, and the Environment branch maps it into its own input space through W_E^{time}. At the first road, the raw increment is zero and is transformed by the *same* normalization as other increments: $\bar{y}_1 = (0 - \mu_{\text{train}})/\sigma_{\text{train}}$.

C.2 Internal Structure of the Causal Transformer

Within each branch's causal Transformer, the early layers use rotary positional encoding to better capture relative spatiotemporal relationships. The positional coordinates comprise road-midpoint longitude and latitude and local seconds of day. The destination token occupies sequence position zero and uses the destination midpoint and a zero-reference temporal coordinate; it participates in attention as a known condition, while history prediction positions start at one. Rotary positional encoding (Su et al., 2024) acts on attention queries and keys. Writing $\text{Rot}(\boldsymbol{\xi}_i)$ for the block-diagonal rotation associated with the spatial and temporal coordinates $\boldsymbol{\xi}_i$, an attention head computes

$$A_{ij} = \text{softmax}_j\left(\frac{(\text{Rot}(\boldsymbol{\xi}_i)\mathbf{q}_i)^\top(\text{Rot}(\boldsymbol{\xi}_j)\mathbf{k}_j)}{\sqrt{d_h}} + M_{ij}\right), \qquad \mathbf{o}_i = \sum_j A_{ij}\mathbf{v}_j. \tag{18}$$

Here $\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i$ are the query, key, and value projections of a sequence attention head, and d_h is the head width. Each coordinate is first encoded by 64-dimensional circular features; a shared linear projection of their concatenation produces the rotation phases.

The causal mask satisfies $M_{ij} = 0$ for valid $j \le i$ and $M_{ij} = -\infty$ otherwise, including padding. Causal attention is used throughout each branch's Transformer. Each history position can therefore use the destination condition and trajectory history up to the current position.

The Road and Environment causal Transformers begin with two and one Post-LN layers using this spatiotemporal encoding, respectively, followed by four and two Pre-LN layers (Xiong et al., 2020). The two groups of layers are connected by a $\text{Linear}(d, d)$–GELU–Dropout–$\text{Linear}(d, d)$ projection. Learned sequence-position embeddings are added before the Pre-LN layers to encode order. A Pre-LN layer has the form

$$\mathbf{U} = \mathbf{X} + \text{MHA}(\text{LN}(\mathbf{X}); M), \qquad \mathbf{Y} = \mathbf{U} + \text{FFN}(\text{LN}(\mathbf{U})). \tag{19}$$

The feed-forward width is 512 and dropout is 0.1. The two branches have independent Transformer parameters. Shared road indices and a shared temporal basis synchronize input positions across the two branches.

D Training and Generation Procedures

D.1 Supervision and Time Recovery

For a minibatch, let $\mathcal{B}_{\text{all}}$ contain all valid prediction positions, including terminal labels, and let $\mathcal{B}_{\text{move}}$ contain actual road transitions. The two classification losses are

$$\mathcal{L}_{\text{road}} = -\frac{1}{|\mathcal{B}_{\text{all}}|} \sum_{(n,i)\in\mathcal{B}_{\text{all}}} \log P^R_{n,i}(y_{n,i}), \qquad \mathcal{L}_{\text{env}} = -\frac{1}{|\mathcal{B}_{\text{all}}|} \sum_{(n,i)\in\mathcal{B}_{\text{all}}} \log P^E_{n,i}(y_{n,i}). \tag{20}$$

The time head reads the fused Road state and the Road token of a nonterminal candidate c. With $\mathbf{h}^{\text{time}}_i = W_{\text{time}}\widetilde{\mathbf{H}}^R_i$, the candidate-conditioned prediction is

$$\begin{aligned} \boldsymbol{\eta}_c &= 2\,\text{sigmoid}(\text{RMSNorm}(W_c\mathbf{z}^R_c)) - \mathbf{1}, \\ \widehat{\bar{y}}_{i+1,c} &= \mathbf{w}_\Delta^\top\big(\mathbf{h}^{\text{time}}_i \odot (\mathbf{1}+\boldsymbol{\eta}_c)\big) + b_\Delta. \end{aligned} \tag{21}$$

Candidate modulation uses RMSNorm (Zhang & Sennrich, 2019). Since W_c is initialized to zero, $\boldsymbol{\eta}_c$ is initially zero and the multiplicative factor $\mathbf{1}+\boldsymbol{\eta}_c$ is initially one. Time predictions use the same scale as the standardized log increments in Appendix C.1. The transition-time loss is evaluated on the true next road:

$$\mathcal{L}_{\text{time}} = \frac{1}{|\mathcal{B}_{\text{move}}|} \sum_{(n,i)\in\mathcal{B}_{\text{move}}} \text{SmoothL1}\Big(\widehat{\bar{y}}_{n,i+1,r_{n,i+1}}, \bar{y}_{n,i+1}\Big). \tag{22}$$

The transition threshold of SmoothL1 is 1. $\mathcal{B}_{\text{move}}$ contains only valid road transitions; if a minibatch has no such transitions, $\mathcal{L}_{\text{time}} = 0$.

During training, time increments are recovered from predictions for the true next road:

$$\widehat{\Delta}_{i+1} = \exp\Big(\min\Big\{b_{\max},\ \max\Big\{0,\ \mu_{\text{train}} + \sigma_{\text{train}}\widehat{\bar{y}}_{i+1,r_{i+1}}\Big\}\Big\}\Big) - 1. \tag{23}$$

The inverse transform clips the unstandardized log value to the interval from 0 to $b_{\max} = 15$; generation applies the same transform and bounds to the sampled road. The trip-level loss compares the logarithms of total predicted and observed durations:

$$\mathcal{L}_{\text{trip}}(\tau) = \text{SmoothL1}\Big(\log(1+\textstyle\sum_{i=1}^{L-1}\widehat{\Delta}_{i+1}), \log(1+\sum_{i=1}^{L-1}\Delta_{i+1})\Big). \tag{24}$$

This loss is averaged over trajectories and is zero for a trajectory without real transitions. Training retains continuous recovered seconds for the trip-duration loss; generation truncates the recovered values to nonnegative integer seconds. For epoch e starting at one,

$$\lambda_{\text{trip}}(e) = 0.5\min(e/5, 1). \tag{25}$$

The trip objective uses time predictions for the true next roads under teacher forcing.

D.2 End-to-End Training

1. Read a minibatch of road sequences, timestamps, destinations, and validity masks. Construct next-road labels ending in EOS and time targets for real transitions.
2. Encode road attributes, frozen topology vectors, and regional fields to produce Road tokens. Project the attribute query $\mathbf{q}_r$ separately at the instance and source levels and aggregate the corresponding POI banks through two-level attention to form Environment tokens.
3. Construct shared temporal features, prepend the destination condition to each stream, and obtain history states through each branch's causal sequence encoder.
4. Gate Environment hidden states into Road hidden states. Score legal next-road candidates in the two vocabularies and predict candidate-conditioned time increments from the fused Road states.
5. Evaluate the two next-road losses, the real-transition time loss, and the trip-duration loss with the current warm-up weight.
6. Update the trainable tokenizers, temporal modules, causal sequence encoders, and prediction heads. Keep raw spatial banks and pretrained topology vectors fixed.

The Road loss and both time losses propagate through the environmental gate into the Environment branch. The auxiliary Environment loss directly trains its sequence states and candidate environmental representations. These losses can all update the encoding parameters of the attribute query $\mathbf{q}_r$ through environmental tokenization.

D.3 Autoregressive Generation

1. Compute the static Road and Environment vocabularies once for the trained model. Initialize the generated trajectory with (O, t_1) and prepend the destination token to both streams.
2. Encode the current synchronized road and time prefix, apply asymmetric coupling, and obtain the Road probabilities over $\mathrm{Succ}(r_i) \cup \{\mathrm{EOS}\}$.
3. Sample a candidate. If it is EOS, stop without producing a time increment.
4. For a real road c, recover the nonnegative time increment from the same candidate's prediction $\widehat{\widetilde{y}}_{i+1,c}$ using Eq. (23), truncate it to integer seconds, and append $(c, t_i + \widehat{\Delta}_{i+1})$.
5. Feed the appended road and time into both branches and repeat until termination.

Generation also stops at a dead end, an unreachable destination, or the configured length cap. Environmental candidate probabilities support auxiliary training; generated roads are always sampled from the Road probability distribution. Generation uses temperature 0.8, top-$k = 4$, and top-$p = 0.9$, with a length cap of 400 and three generated trajectories per condition.

E Evaluation Definitions

The main comparison uses the four macro-level distributions defined below; supplementary trajectory-level metrics are reported in Appendix G. We compute statistics for real and generated trajectory collections within each city. Coordinate-space outputs use the geometric measurements and region assignment described in Appendix H.2.

Distance and Radius. If $\ell(r)$ is segment length, then

$$\mathrm{Distance}(\tau) = \sum_{i=1}^{L} \ell(r_i). \tag{26}$$

For road trajectories, spatial extent is measured independently of road segmentation density by resampling the geometry at $J = 200$ equal-arc-length locations $\mathbf{u}_j$. With geographic center $\bar{\mathbf{u}} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{u}_j$,

$$\mathrm{Radius}(\tau) = \frac{1}{J}\sum_{j=1}^{J} d_{\mathrm{geo}}(\mathbf{u}_j, \bar{\mathbf{u}}), \tag{27}$$

where d_{geo} is Haversine distance. This is a mean spatial radius, not a root-mean-square radius of gyration.

Duration and LocFreq. Trip duration is the elapsed time between the first and last record:

$$\mathrm{Duration}(\tau) = t_L - t_1 = \sum_{i=1}^{L-1} \Delta_{i+1}. \tag{28}$$

For trajectory collection $\mathcal{T}$, let $n_z(\mathcal{T})$ count observations assigned to region z. The visitation distribution is

$$\mathrm{LocFreq}_z(\mathcal{T}) = \frac{n_z(\mathcal{T})}{\sum_{z' \in \mathcal{Z}} n_{z'}(\mathcal{T})}. \tag{29}$$

Real and generated counts use the same region index.

Distribution comparison. Distance, Radius, and Duration are converted into probability vectors with common histogram edges. The convention uses 100 finite bins determined from training reference data, plus underflow and overflow bins. LocFreq is already a categorical distribution over regions. For real distribution P, generated distribution Q, and $M = (P + Q)/2$, the reported distance is

$$\mathrm{JSD}(P, Q) = \sqrt{\tfrac{1}{2}\,\mathrm{KL}(P\|M) + \tfrac{1}{2}\,\mathrm{KL}(Q\|M)}. \tag{30}$$

Logarithms are natural and zero-mass terms contribute zero. The square root is part of the definition used here.

Density visualization. Figure 4 uses shared histogram bins and a one-bin Gaussian display kernel, with densities normalized over all trips. Its central ranges retain approximately 99% or more of spatial values and 97.5% or more of durations in each population.

F Experimental Configuration

F.1 Model and Training Settings

Table 6 summarizes the model and training settings. Training uses four NVIDIA RTX 4090 GPUs with distributed data parallelism, one process per GPU. The four-GPU launcher uses 32 trajectories per GPU and one gradient-accumulation step, giving an effective global batch size of $4 \times 32 \times 1 = 128$.

Table 6: Model and training settings.

Setting	Value
Hidden dimension / attention heads	128 / 4
Road Transformer layers (Post-LN / Pre-LN)	2 / 4
Environment Transformer layers (Post-LN / Pre-LN)	1 / 2
Transformer feed-forward width / dropout	512 / 0.1
Road and region field width	16
POI category / subcategory embedding widths	32 / 32
Landmark role / confidence embedding widths	32 / 32
Local / Landmark instance capacity	4 / 2
Macro buffer / decay scale	20 m / 50 m
POI candidate radius / Landmark candidate cap	150 m / 32
Optimizer / learning rate / weight decay	AdamW / 5×10^{-4} / 0.02
Maximum epochs / selection	50 / validation performance
Road / Environment / time loss weights	1 / 0.1 / 1
Trip-loss weight at epoch $e \geq 1$	$0.5 \min(e/5, 1)$
Log-increment recovery bounds	0 to 15
Local time zone / number of day slots	UTC+8 / 7
Training GPUs / distributed processes	4 NVIDIA RTX 4090 / 4
Training batch per GPU / effective global batch	32 / 128
Gradient accumulation steps / random seed	1 / 42
Data-loader workers per process / mixed precision	4 / enabled

F.2 Offline LLM Configuration and Prompt

Landmark selection invokes `gpt-5.6-sol` through the Codex CLI, processing one road per request with at most 32 candidate POIs and returning at most two functional anchors. Each request uses an ephemeral read-only session with JSON-Schema-constrained output. Temperature, top-p, maximum output tokens, and reasoning effort are not explicitly overridden in the invocation script. The request timeout is 900 s, with at most 12 retries for transient failures, a retry-wait base of 15 s, and a cap of 300 s; the default interval between successful requests is 0.2 s. Completed roads are cached, and exact request contexts and structured responses are stored separately for verification.

The prompt restricts selection to supplied candidate IDs and requires explanations grounded in the supplied name, type, category, distance, and address. Inputs include the city, road midpoint, road

attributes, and each candidate's name, two-level categories, full type, district, coordinates, road distance, and address. Output fields are `poi_id`, `role`, `confidence`, and `reason`. Role and confidence are LLM-generated categorical labels, distinct from the provider's categories and subcategories, and are embedded as additional Landmark inputs. Zero or one anchor may be returned without filling the two-anchor capacity. LLM selection runs only during POI bank construction.

A Static instruction

You are a careful urban-geography analyst for __CITY__. Select only supplied POI IDs that are functional environment anchors for this specific road in a trajectory-generation study.
An anchor is a stable, named place whose presence can add place/function identity beyond the road topology and a coarse POI-category histogram, and can therefore help distinguish the surrounding environment state in a travel sequence. Strong examples are a scenic/cultural or religious site, metro/rail/coach/airport hub, university, hospital, government/civic institution, named major market/mall/industrial or office complex, or a clearly named hotel complex.
Do NOT select ordinary road intersections, ordinary bus stops, parking lots, ATMs, toilets, charging stations, generic entrances/exits, or ordinary shops/restaurants/services: these are either already road-topology information, too generic, or too volatile. A named entrance may be selected only when its name unambiguously identifies a major parent place and it is the relevant road-side access point. A transport POI must be a hub-scale facility, not an ordinary bus stop.
Use real geographic knowledge of __CITY__ only to interpret a supplied name and coordinates. Never invent a POI, address, popularity fact, star rating, traffic-flow fact, or claim that a place is famous/important. The reason must cite only supplied name/type/category/distance/address evidence. Do NOT decide traffic routing. Return at most the requested number of anchors; return zero or one whenever a second candidate is not independently convincing.

Output JSON only in this schema:

```
{"anchors": [{"poi_id": "supplied id",
"role":
"tourism|transport|education|healthcare|civic|commercial|other",
"confidence": "high|medium|low",
"reason": "explanation"}]}
```

B Per-road context: request.json

```
{
"task": "Select at most 2 anchor
POIs. JSON output is required.",
"city": <city>,
"road_id": <road ID>,
"road_location": {
  "district_hint": <district or
unknown>,
  "mid_longitude_wgs84":
<longitude>,
  "mid_latitude_wgs84":
<latitude>},
"road_properties": {
  "length_m": <length>,
  "highway_code": <highway code>,
  "lanes": <lane count>,
  "oneway": <boolean>,
  "in_degree": <in-degree>,
  "out_degree": <out-degree>},
"candidate_pois": [{
  "poi_id": <supplied ID>,
  "name": <name>,
  "category": <category>,
  "subcategory": <subcategory>,
  "type": <full type string>,
  "district": <district>,
  "longitude_wgs84": <longitude>,
  "latitude_wgs84": <latitude>,
  "distance_to_road_m":
<distance>,
  "address": <address>}, ...]
}
```

Figure 6: Prompt template for offline Landmark selection. The static instruction is followed by the per-road JSON context. Placeholders are filled from the city, road, and candidate POI records.

G OD-Conditioned Micro-Level Evaluation and Route Examples

G.1 Destination Attainment and Route Similarity

We evaluate MUSE at the trajectory level using destination accuracy (Dest) and normalized longest common subsequence similarity (LCSS). Since generation starts from the given origin road, Dest measures whether the final generated road equals the target destination. LCSS matches road IDs exactly and divides the longest common subsequence length by the shorter of the generated and reference sequence lengths.

For each test condition, Mean averages the scores of the three generated trajectories, while Best takes their maximum; these per-condition scores are then aggregated over the evaluation set. For Dest, Best records whether at least one sample reaches the destination.

Table 7: OD-conditioned micro-level results for MUSE. Mean and Best summarize three samples per test condition; higher is better.

City	Dest ↑		LCSS ↑	
	Mean	Best	Mean	Best
Beijing	0.9947	0.9987	0.7232	0.8177
Chengdu	0.9996	1.0000	0.8451	0.9133
Xi'an	0.9991	0.9999	0.8281	0.9033

Mean destination accuracy ranges from 99.47% to 99.96% across the three cities, indicating consistent destination attainment under the supplied OD conditions. LCSS complements this endpoint measure by quantifying agreement with the reference road order.

G.2 Qualitative Route Examples

Figure 7 provides one selected OD example per city for qualitative inspection of route geometry under shared endpoints.

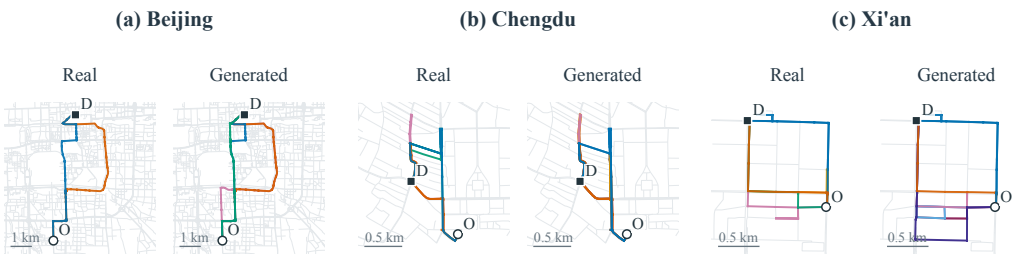

Figure 7: Real and generated paths for three selected ODs. Each city shows real paths (left) and generated paths (right), with a shared map extent and endpoints within each city. Colors distinguish major route corridors, with spatially corresponding real and generated corridors sharing a color. All distinct paths are retained; circles and squares mark origins and destinations.

H Baseline Methods and Experimental Protocols

H.1 Shared Experimental Protocol

All learned generators use the same city-specific training, validation, and test split: 80/10/10 in fixed file order. Models are trained separately for each city and evaluated on the same held-out trips. We retain each baseline's original architecture, training objective, and recommended training and sampling settings from its paper or public implementation. The origin, destination, and departure-time information of each test trip is supplied through the method's conditioning interface.

Spatial evaluation uses the same city-level region partition and distribution-comparison protocol. Macro-level distributions aggregate the generated samples, while the Mean and Best summaries in Appendix G.1 describe MUSE's trajectory-level results. Duration is reported for VOLUNTEER, HOSER, and MUSE, whose evaluated outputs include temporal information; the remaining entries are marked with a dash in Table 2. The following subsections describe coordinate-space adaptation and each method's training and generation mechanism.

H.2 Adaptation of Coordinate-Based Baselines

DiffTraj, ControlTraj, and TrajFlow generate continuous coordinate sequences. Their training trajectories are constructed from the same road trajectories used by the road-network-based methods: we concatenate the full geographic polylines of the visited road segments in traversal order, remove consecutive duplicate coordinates at shared endpoints, and linearly interpolate the resulting polyline into a fixed-length coordinate sequence. Training and sampling then proceed in continuous coordinate space using each method's own model.

Generated GPS sequences are evaluated directly, without map matching. Distance sums the geographic distances between consecutive generated points; Radius measures their mean distance to the geographic center; and LocFreq assigns the generated points to the same region partition used for road-based evaluation. These baselines provide coordinate sequences without comparable per-step travel-time predictions, so Duration is not reported for them. This interface preserves their coordinate-space outputs while using shared spatial distributional measures.

H.3 Model-Specific Training and Generation

Dijkstra. Dijkstra finds a minimum-cost path between a given origin and destination on a road network with nonnegative edge weights, providing a deterministic routing baseline.

MoveSim. MoveSim (Feng et al., 2020) uses adversarial learning with a self-attention sequence network to model location transitions, incorporating urban structural priors based on geographic distance, POI functional similarity, and historical transitions.

VOLUNTEER. VOLUNTEER (Long et al., 2023) uses a user VAE and a trajectory VAE to model population distributions and individual mobility patterns, respectively. Its trajectory generator incorporates temporal point processes and models travel time separately from dwell time.

TS-TrajGen. TS-TrajGen (Jiang et al., 2023) combines adversarial learning with A* search to model movement conditioned on the partial trajectory and destination. It first generates a regional route and then uses that route to guide the generation of connected road-segment sequences.

HOSER. HOSER (Cao et al., 2025) integrates road, regional, and trajectory semantics through a Transformer trajectory encoder and a destination-oriented navigator to predict the next road and timestamp. Generation searches for the highest-probability path between the given endpoints.

MobilityGPT. MobilityGPT (Haydari et al., 2026) generates road sequences autoregressively using a decoder-only GPT, with a road connectivity matrix constraining transitions. Its training incorporates gravity-based sampling and fine-tuning with trajectory feedback.

DiffTraj. DiffTraj (Zhu et al., 2023) generates trajectories in continuous GPS coordinate space, progressively denoising Gaussian noise through a conditional Traj-UNet.

ControlTraj. ControlTraj (Zhu et al., 2024) encodes road topology constraints with RoadMAE and combines the resulting road embeddings with trip attributes to condition GeoUNet diffusion for GPS trajectory generation.

Diff-RNTraj. Diff-RNTraj (Wei et al., 2024) embeds discrete road segments and continuous within-segment moving ratios into a continuous space for diffusion generation. A trajectory decoder recovers the road segments and moving ratios, while a spatial-validity loss constrains generation.

TrajFlow. TrajFlow (Li et al., 2026) uses conditional flow matching to learn a continuous vector field that transforms Gaussian noise into GPS trajectories. It combines trajectory simplification and OD-conditioned normalization to harmonize trajectory representations, then reconstructs geographic coordinates after sampling with an ordinary differential equation solver.