

SPAR: Enhancing Industrial-Scale Generative POI Recommendation via Real-World Spatial Perception

Fangye Wang
AMAP, Alibaba Group
Beijing, China
wangfangye.wfy@alibaba-inc.com

Yunjin Gu*
The Chinese University of Hong
Kong, Shenzhen
Shenzhen, China
yunjingu@link.cuhk.edu.cn

Haowen Lin
AMAP, Alibaba Group
Beijing, China
linhaowen.lhw@taobao.com

Yifang Yuan, Song Yang
AMAP, Alibaba Group
Beijing, China
{yuanyifang,yyf,song.yang}@alibaba-inc.com

Xiaojiang Zhou†
AMAP, Alibaba Group
Beijing, China
zhouxiaojiang.zxj@taobao.com

Pengjie Wang
AMAP, Alibaba Group
Beijing, China
pengjie.wpj@alibaba-inc.com

Abstract

Generative Point-of-Interest (POI) recommendation, autoregressively generating a target POI’s semantic ID (SID), holds great promise for Location-Based Services, where a recommendation helps only if the user can reach it. Yet, existing methods operate within an interest space defined by behavior sequences and collaborative signals, where geography enters only as a textual attribute of the SID, leaving no explicit mechanism to learn or preserve how urban places are related by distance, direction, and reachability; their predictions are thus behaviorally plausible yet far from the user’s real-time location. We argue that such services require injecting real urban spatial knowledge into the interest space, rather than inferring geography from behavior alone. Hence, we propose SPAR, a unified framework whose three synergistic stages jointly construct, cultivate, and preserve urban spatial knowledge: (1) at the tokenization level, Spatially-Intrinsic SID (SI-SID) explicitly encodes longitude–latitude coordinates into a sinusoidal geospatial embedding and fuses it with the textual semantic embedding, producing identifiers via RQ-Kmeans that are simultaneously semantically and geographically consistent; (2) at the cognition level, Multi-Granular Geospatial CPT (MG-CPT) continually pre-trains the base LLM on 25 curated geospatial datasets organized into three tiers of basic attributes, pairwise relations, and city-scale navigation, so that scattered POIs cohere into a connected urban space; and (3) at the adaptation level, Task-Vector Anchored SFT (TV-SFT) anchors the acquired spatial knowledge as a frozen parameter-space task vector to prevent its catastrophic forgetting during behavioral fine-tuning, thereby fusing the two spaces. Extensive quantitative and visualization experiments on two public and four industrial-scale datasets demonstrate the effectiveness of SPAR.

*Work done during the internship at Amap, Alibaba Group
†Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted.

Conference acronym ’XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

CCS Concepts

• Information systems → Information retrieval; Recommender systems.

Keywords

Recommender Systems, Spatial Perception, Large Language Model

ACM Reference Format:

Fangye Wang, Yunjin Gu, Haowen Lin, Yifang Yuan, Song Yang, Xiaojiang Zhou, and Pengjie Wang. 2018. SPAR: Enhancing Industrial-Scale Generative POI Recommendation via Real-World Spatial Perception. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym ’XX)*. ACM, New York, NY, USA, 17 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

With the pervasiveness of Location-Based Services (LBS), next Point-of-Interest (POI) [1, 4, 17, 30, 39] Recommendation has become a cornerstone capability of map platforms such as Kuaishou [3], AMAP [22] and Google [8]. Given a user’s historical behavior and real-time context (e.g., query and location), it recommends the POI the user will visit next. Serving billions of users and POIs, its quality directly shapes users’ decision-making efficiency and experience, making it a central problem for both academia and industry.

Early studies formulate the next-POI recommendation as a sequential prediction task, evolving from Markov chains and matrix factorization to RNN [6, 10] and Transformer-based [14, 20] models that capture behavioral preferences from historical trajectory. Recently, Generative Recommendation Models (GRMs) [31, 39] based on a Large Language Model (LLM) [24, 35] have emerged as a powerful paradigm. Training a GRM typically involves two steps: each POI is first tokenized into a Semantic ID (SID) from its semantic representation via residual quantization (e.g., RQ-VAE [31] or RQ-Kmeans [33]); the LLM is then supervised fine-tuned (SFT) on user behavior sequences rewritten with these SIDs, enabling it to autoregressively generate the SID of the target POI [3, 8, 31]. Since SIDs are newly added tokens, some recent methods further insert a continued pre-training (CPT) step beforehand to align the new SID modality with the LLM’s existing knowledge [8, 32]. Revisiting this pipeline, geographic information enters only as an auxiliary signal in the SID construction step [2, 9, 31], while the subsequent stages

of CPT and SFT are driven by user behavior alone. The pipeline therefore operates largely within an interest space of behavioral co-occurrence, leaving urban spatial knowledge imprecisely encoded at the tokenization stage and largely absent at the training stage.

Imprecise geographic encoding at the tokenization stage. Longitude–latitude coordinates are either serialized as plain textual semantics, whose digit tokens carry no metric structure, or exploited only indirectly through collaborative signals [19, 31] and contrastive objectives [2, 32]. Such signals are themselves largely derived from co-occurrence and thus cannot substitute for explicit geographic measurement. Quantization is therefore dominated by high-variance semantic features while geography degenerates into a weak regularizer, and the resulting identifier space lacks spatial continuity: POIs with adjacent coordinates are not guaranteed adjacent SIDs, while semantically similar yet distant POIs (e.g., two branches of the same chain located tens of kilometers apart in the same city) may share near-identical ones. An effective SID should encode not only what a POI is but also where it lies in the city.

Absent urban spatial knowledge at the training stage. In LBS, existing POI-related GRMs [31, 32] still rely mainly on SFT to model sequences of user behavior, with CPT, when used at all, not aimed at urban spatial knowledge. This overlooks what makes recommendation in such services distinctive: next-POI prediction is essentially movement through real urban space. Users are not merely reproducing co-occurrence patterns but physically traveling, so beyond straight-line proximity they weigh direction and reachability. However, a general LLM’s web-scale pre-training offers little fine-grained knowledge of a specific city, and existing CPT, typically a small-scale warm-up over POI textual attributes, does little to supply it, leaving the road network and its connectivity unlearned. Optimized on behavior alone, SFT captures only the interest side: the model learns which POIs are frequently visited together, but not whether a candidate is accessible from the user’s current location—a POI two kilometers away in a straight line may lie across a river or require a lengthy detour. Its predictions are thus behaviorally plausible, yet ungrounded in real-world geography. Worse still, even injecting spatial knowledge with CPT is not enough: under the massive behavioral data, standard SFT tends to overwrite it, so unless the knowledge is explicitly preserved during adaptation, the model drifts back to behavior-only fitting.

To address these deficiencies, we propose **SPAR**, a unified framework that injects real urban spatial knowledge into the interest space of generative POI recommendation, rather than inferring geography from behavior alone. SPAR realizes this through three synergistic stages that respectively construct, cultivate, and preserve this knowledge. First, at the representation level, **Spatially-Intrinsic SID (SI-SID)** uses an explicit sinusoidal encoder to lift raw longitude–latitude coordinates into a geospatial embedding, fuses it with the POI’s semantic embedding, and discretizes the result via RQ-Kmeans, so that every identifier is anchored to a concrete location. Second, at the cognition level, **Multi-Granular Geospatial CPT (MG-CPT)** continually pre-trains the backbone LLM on 25 datasets that span three tiers of increasing spatial complexity and together comprise millions of instances built from POI, road, and navigation data, internalizing the spatial relations of the whole city. Finally, at the adaptation level, **Task-Vector Anchored SFT (TV-SFT)** fits user behavior while anchoring the acquired

knowledge as a parameter-space task vector to prevent its catastrophic forgetting. Recommendation is thereby no longer a purely behavioral fitting task: user trajectories are grounded in the real urban space, and the next POI is inferred by reasoning over the spatial relations among the visited locations rather than in isolation from the city. The major contributions are summarized as follows:

- We propose SPAR, a unified framework that injects real urban spatial knowledge into the interest space of generative POI recommendation; its three synergistic stages jointly construct, cultivate, and preserve this knowledge.
- SI-SID fuses geospatial and semantic embeddings before residual quantization, making geographic topology an intrinsic property of the identifier codebook.
- MG-CPT continually pre-trains the LLM on 25 multi-tier POI, road, and navigation datasets, internalizing the city’s spatial relations so that POIs become places related by distance, direction, and reachability. TV-SFT then anchors this knowledge as a task vector, grounding behavioral fitting in urban spatial relations.
- Comprehensive experiments validate the effectiveness of SPAR and further confirm that each stage fulfills its intended role. We will release the four industrial-scale POI recommendation datasets, along with a per-city MG-CPT suite of 25 geospatial training datasets and an 18-task spatial cognition benchmark.

2 Related Works

2.1 POI Recommendation.

Traditional next-POI recommendation methods are typically formulated as a sequential prediction task, where the goal is to predict a user’s next visited location based on their historical check-in sequence [12, 14, 27, 31, 38]. With the advent of deep learning, recurrent neural networks (RNNs)—particularly LSTM and GRU variants—were widely adopted to capture sequential preferences [10], followed by transformer-based models that demonstrated superior performance through self-attention mechanisms [12]. For instance, STAN [21] explicitly models pairwise spatio-temporal relationships via interpolated distance encoding, while GeoSAN [15] incorporates geographical influence into self-attention for next-POI recommendation. More recently, LLMs have been integrated to leverage commonsense knowledge for semantic-aware prediction [14, 31]. In addition, graph neural networks gained traction for modeling user-POI interactions [20, 37].

2.2 Generative Recommendation.

Beyond discriminative models, LLM-based GRMs [16, 33, 39] have emerged as a powerful paradigm that directly generates item identifiers, eliminating the multi-stage cascaded pipeline of retrieval and ranking. TIGER [25] pioneers this direction with hierarchical SIDs learned via RQ-VAE [13], while subsequent works adopt RQ-Kmeans for more stable quantization [3]. In industrial practice, OneRec [3] and OneRec-V2 [40] unify retrieval and ranking into end-to-end generative frameworks with reinforcement learning. OneLoc [33] incorporates geographic information into generative recommendation for local lifestyle services with reinforcement learning. GeoGR [32] and GNPR-SID [31] refined SIDs with contrastive learning to better capture POI semantics and collaborative

signals. PLUM [8] employs an LLM to extract features and represent items as SIDs for generative recommendation in YouTube’s recommendation with CPT and SFT [7]. Nevertheless, these GRMs overlook what is specific to next-POI recommendation in LBS: predictions must be linked to the real physical space, rather than fitted from interest co-occurrence and behavior sequences alone. SPAR instead injects real urban spatial knowledge into every stage, grounding behavioral fitting in the city’s geography.

3 Problem Formulation

Let $\mathcal{U} = \{u_1, u_2, \dots, u_M\}$ and $\mathcal{P} = \{p_1, p_2, \dots, p_N\}$ denote the set of M users and N POIs, respectively. Each POI is defined as a tuple $p_i = (\text{lon}_i, \text{lat}_i, \text{addr}_i, \text{ctx}_i)$, capturing geographical coordinates, address, and contextual attributes (e.g., brand, category). Meanwhile, each POI p_i is associated with a SID s_i , generated via the RQ-Kmeans method. For each user $u \in \mathcal{U}$, we record their interaction trajectory (i.e., check-in history) as $T^u = \{r_1^u, r_2^u, \dots, r_n^u\}$, ordered in reverse chronological order. Each interaction record $r_t^u = (u, p, t, \text{con}, a)$ indicates that user u interacted with POI p at timestamp t under condition con with action a . In the AMAP application scenario, action a encompasses behaviors such as navigation, reservation, booking, and collection, while condition con captures real-time context, such as query, current coordinates, and other metadata.

Next POI Recommendation. Given a user u ’s historical trajectory T^u and a real-time condition con_u , the task is to predict the next POI p_{n+1} (i.e., its SID s_{n+1}) by maximizing the conditional probability: $p_{n+1}^* = \arg \max_{p \in \mathcal{P}} \mathbb{P}(p|u, T^u, \text{con}_u)$. Unlike traditional discriminative methods that score all candidates, SPAR formulates this as a generative task [31] where s_{n+1} is autoregressively decoded from the sequence $(s_1, s_2, \dots, s_n)$ conditioned on con_u .

4 Methodology: SPAR

We introduce SPAR, a generative POI recommendation framework with three stages (Figure 1): SI-SID constructs the identifier space, fusing a POI’s semantic embedding with a geospatial encoding of coordinates and quantizing them into spatial semantic tokens. MG-CPT cultivates spatial reasoning, continually pre-training the LLM on multi-granular geospatial data so scattered POIs cohere into a connected urban space. TV-SFT preserves this knowledge, anchoring it as a frozen task vector while the model is fine-tuned on user behavior, fusing the interest space of user trajectories with the real urban space to generate the target POI’s SI-SID.

4.1 Spatially-Intrinsic SID (SI-SID)

Recent GRMs derive an item’s SID by encoding its textual descriptions into a semantic embedding via an LLM and then discretizing it through RQ-VAE or RQ-Kmeans. In LBS, however, a user’s interactions are tightly coupled with real-time geographic context, as users tend to choose POIs closer to their current location, which such semantics-driven identifiers fail to reflect. We therefore design SI-SID (Figure 1 stage 1), which incorporates geographic coordinates into the discrete token sequence itself, realized in two steps: explicit geospatial encoding and SI-SID construction with RQ-Kmeans.

4.1.1 Explicit Geospatial Encoding. Raw longitude–latitude scalars are ill-suited as direct LLM inputs: they are dimensionally sparse, related to geographic distance in a nonlinear (spherical) manner,

and cannot express the inherent periodicity of longitude ($-180^\circ \equiv 180^\circ$). We therefore adopt an explicit sinusoidal geospatial encoder inspired by NeRF [23, 26] and Transformer positional encodings [24, 29]. Specifically, given a POI with longitude λ and latitude ϕ , we first normalize each coordinate to $[0, 2\pi]$:

$$\hat{\lambda} = \frac{\lambda + 180}{360} \cdot 2\pi, \quad \hat{\phi} = \frac{\phi + 90}{180} \cdot 2\pi. \quad (1)$$

Each is then projected onto K exponentially spaced frequency bands. Hyperparameter K controls the granularity of the encoding:

$$f_k = 10000^{-k/K}, \quad k = 0, \dots, K-1, \quad (2)$$

yielding the Fourier geospatial features $\gamma(\hat{\lambda}) \in \mathbb{R}^{2K}$:

$$\gamma(\hat{\lambda}) = [\sin(f_0\hat{\lambda}), \cos(f_0\hat{\lambda}), \dots, \sin(f_{K-1}\hat{\lambda}), \cos(f_{K-1}\hat{\lambda})] \quad (3)$$

and analogously $\gamma(\hat{\phi})$. The final geospatial representation $\mathbf{e}_{\text{geo}}$ is obtained by concatenating $\gamma(\hat{\lambda})$, $\gamma(\hat{\phi})$ and passing it through a two-layer MLP with Xavier-uniform [5] initialization:

$$\mathbf{e}_{\text{geo}} = \text{MLP} \left([\gamma(\hat{\lambda}); \gamma(\hat{\phi})] \right) \in \mathbb{R}^{d_g}, \quad d_g = 4K. \quad (4)$$

Spatial properties of the encoding. Let $u \in [0, 2\pi]$ denote either normalized coordinate $\hat{\lambda}$ or $\hat{\phi}$; the following properties hold for each coordinate separately and extend to their concatenation. **Property 1 (Shift-invariant kernel).** By the product-to-sum identity, $\langle \gamma(u_1), \gamma(u_2) \rangle = \sum_{k=0}^{K-1} [\sin(f_k u_1) \sin(f_k u_2) + \cos(f_k u_1) \cos(f_k u_2)] = \sum_{k=0}^{K-1} \cos(f_k(u_1 - u_2))$ depends only on the displacement $u_1 - u_2$ rather than on absolute positions; since $f_k \leq f_0 = 1$, every term is decreasing in $|u_1 - u_2|$ on $[0, \pi]$, a range that covers all intra-city displacements by a wide margin, so embedding similarity decreases monotonically with geographic displacement. **Property 2 (Lipschitz continuity).** Applying $1 - \cos \theta \leq \theta^2/2$ to the squared distance gives $\|\gamma(u_1) - \gamma(u_2)\|_2^2 = \sum_{k=0}^{K-1} 2(1 - \cos(f_k(u_1 - u_2))) \leq (\sum_{k=0}^{K-1} f_k^2)(u_1 - u_2)^2$, so the encoding is Lipschitz-continuous, and the subsequent two-layer MLP, a composition of linear layers and ReLU, preserves this continuity. POIs with adjacent coordinates thus obtain adjacent geospatial embeddings $\mathbf{e}_{\text{geo}}$ and, after the residual quantization described next, tend to share SI-SID prefixes, making geographic proximity an intrinsic property of the identifier space. We refer readers to the supplementary material for a more detailed analysis of these properties and their empirical verification.

4.1.2 SI-SID Construction with RQ-Kmeans. We generate the POI’s semantic embedding $\mathbf{e}_{\text{sem}} \in \mathbb{R}^{d_s}$ from a pretrained text encoder (e.g., Qwen3-embedding-4B [35]) and fuse it with the geospatial embedding via concatenation followed by L2 normalization:

$$\mathbf{x} = [\mathbf{e}_{\text{sem}}; \mathbf{e}_{\text{geo}}] / \|\mathbf{e}_{\text{sem}}; \mathbf{e}_{\text{geo}}\|_2 \in \mathbb{R}^{d_s+d_g}, \quad (5)$$

so that $\mathbf{x}$ lies on the unit hypersphere. This prevents either signal from dominating the clustering due to scale differences and enables cosine-distance-based residual computation.

The fused embedding is then discretized by an L -layer residual K-means quantizer. Let $\mathbf{r}^{(0)} = \mathbf{x}$ for each POI. Each layer’s codebook $C_\ell = \{c_\ell^k\}_{k=1}^{N_c}$ is constructed by applying K-means clustering to the residuals collected from all training POIs at the previous layer:

$$C_\ell = \text{K-Means}(R_\ell, N_c), \quad R_\ell = \{\mathbf{r}_i^{(\ell-1)}\}_{i=1}^N, \quad (6)$$

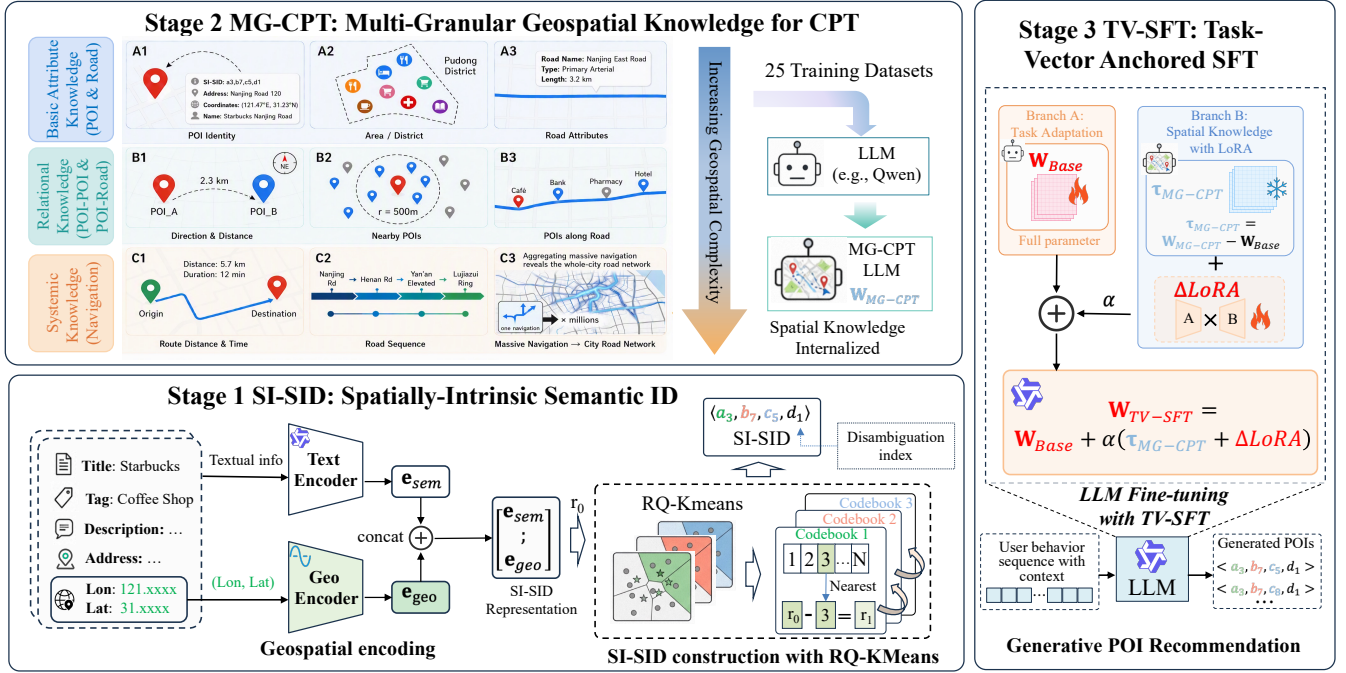


Figure 1: SPAR enhances large-scale generative POI recommendation with urban spatial knowledge. Stage 1: SI-SID constructs a Spatially-Intrinsic SID by fusing textual (e_{sem}) and geospatial (e_{geo}) embeddings and quantizing them via RQ-Kmeans. Stage 2: MG-CPT continually pre-trains the LLM on multi-granular geospatial datasets spanning basic-attribute, relational, and systemic spatial knowledge. Stage 3: TV-SFT jointly performs full-parameter task adaptation (W_{Base}) and anchored spatial-knowledge refinement, where the frozen τ_{MG-CPT} is adjusted by a trainable $\Delta LoRA$, giving W_{TV-SFT} for POI recommendation.

where N_c is the codebook size ($N_c = 512$ for industrial-scale datasets). Given the trained codebooks, each POI is quantized layer by layer:

$$c_\ell^* = \arg \min_{c \in C_\ell} \|r^{(\ell-1)} - c\|_2, \quad (7)$$

$$r^{(\ell)} = (r^{(\ell-1)} - c_\ell^*) / (\|r^{(\ell-1)} - c_\ell^*\|_2 + \epsilon), \quad (8)$$

with $\epsilon = 10^{-8}$ preventing division by zero. We set $L = 4$: the first three layers are learned via RQ-Kmeans, while the fourth is a disambiguation index appended to guarantee SI-SID uniqueness by distinguishing POIs that collide on the first three codewords [31]. We use the tuple (a_*, b_*, c_*, d_*) to represent the SI-SID of a POI, e.g., $\langle a_3, b_4, c_1, d_1 \rangle$. Each SI-SID token serves as the atomic unit of spatial perception in SPAR, the foundational carrier on which MG-CPT builds spatial reasoning and TV-SFT anchors it during adaptation.

4.2 Multi-Granular Geospatial CPT (MG-CPT)

While SI-SID provides a structurally sound carrier for spatial perception, the base foundation model still cannot reason over the geographic relationships encoded in these identifiers. We therefore propose Multi-Granular Geospatial CPT (MG-CPT), which continually pre-trains a foundation LLM on real-world POI, road, and navigation data organized into three tiers of increasing spatial complexity. This is motivated by real LBS behavior: users choose destinations by reasoning about distance, direction, and reachability, and MG-CPT explicitly instills these forms of spatial knowledge. By jointly learning from all three tiers, MG-CPT internalizes the

spatial relations of the whole city into the model's parameters, so that each SI-SID denotes not an isolated POI but a place connected to others through distance, direction, and reachability.

4.2.1 Three-Tier Geospatial Data Taxonomy. As illustrated in Figure 1 stage 2, these tiers are instantiated as training datasets, of which the figure depicts only a few representative ones. In total, we construct 25 training datasets that jointly characterize the real-world profile of POIs, their relationships with the surrounding road network, and a dynamic depiction of the city's road system by aggregating large-scale navigation data, together comprising millions of training instances. The three tiers are summarized below, with full details deferred to the supplementary material.

Tier 1: Basic Attribute Knowledge (POI & Road). This tier grounds each entity in its intrinsic attributes, linking a POI's SI-SID to its name, category, address, district, and coordinates, and a road to attributes such as name, type, and length, so that SI-SID tokens denote concrete geographic entities rather than abstract symbols.

Tier 2: Relational Knowledge (POI-POI & POI-Road). This tier teaches pairwise spatial relations of two complementary kinds. One is *real-world aggregation*: the POIs surrounding a location and those lying along a road, reflecting how places actually cluster in the city. The other is *metric reasoning*: the direction and Euclidean distance between POIs, adding the orientation and proximity computations users rely on when comparing candidate destinations.

Tier 3: Systemic Knowledge (Navigation). Euclidean distance alone can be misleading: two POIs close in straight-line distance may be hard to reach due to rivers, overpasses, or one-way roads. To capture real reachability rather than mere proximity, this tier explicitly introduces navigation data in multiple forms (route distance, travel time, and ordered road sequences), letting the model internalize the connectivity of the whole-city road network instead of memorizing individual routes.

4.2.2 Continued Pre-Training. All datasets are serialized into next-token-prediction instances and jointly used to continually pre-train a base LLM under the standard causal language modeling objective:

$$\mathcal{L}_{\text{MG-CPT}} = - \sum_t \log P(x_{t+1} | x_1, \dots, x_t), \quad (9)$$

where $x_{1:t}$ denotes a training sequence from the mixed corpus. This process extends the base LLM, with parameters $\mathbf{W}_{\text{Base}}$, into an urban-space-aware LLM with parameters $\mathbf{W}_{\text{MG-CPT}}$. As evaluated on an 18-task benchmark in Section 5.7, the trained MG-CPT LLM exhibits substantially enhanced knowledge of urban space, providing a robust foundation for the subsequent TV-SFT stage.

4.3 Task-Vector Anchored SFT (TV-SFT)

Because MG-CPT has internalized the spatial relations of the whole city, next-POI recommendation can be framed as projecting a user’s behavioral trajectory onto real urban space and inferring the next location from the spatial relations among visited POIs, rather than fitting behavior sequences alone. Realizing this, however, requires fine-tuning on large-scale user-behavior data, which creates a dilemma: fine-tuning directly from $\mathbf{W}_{\text{MG-CPT}}$ gradually erodes the spatial knowledge just acquired (catastrophic forgetting), whereas fine-tuning from the base model, with parameters $\mathbf{W}_{\text{base}}$, leaves behavioral fitting ungrounded in urban space. To couple these two spaces rather than trade one off against the other, we draw inspiration from recent advances in model editing and knowledge injection [11, 18, 34], which show that a specific capability acquired by an LLM can be represented and manipulated as a direction in parameter space. Building on this insight, we propose TV-SFT, which decouples behavioral adaptation from spatial preservation into two parallel branches (Figure 1 Stage 3). First, we isolate the spatial knowledge acquired by MG-CPT as a task vector:

$$\tau_{\text{MG-CPT}} = \mathbf{W}_{\text{MG-CPT}} - \mathbf{W}_{\text{base}}. \quad (10)$$

This task vector captures the parameter change induced by learning spatial knowledge during MG-CPT; since this change arises solely from the injected spatial knowledge, $\tau_{\text{MG-CPT}}$ serves as an explicit parametric representation of the spatial knowledge increment. TV-SFT then trains the LLM through two parallel branches, one for each space. **(i) Task Adaptation** learns the interest space: the base LLM $\mathbf{W}_{\text{Base}}$ undergoes full-parameter SFT to fit user behavior, which is the only branch existing GRMs employ. **(ii) Spatial Knowledge** preserves the urban space: $\tau_{\text{MG-CPT}}$ is kept frozen as an anchor, while a low-rank LoRA adapter $\Delta\text{LoRA} = \mathbf{BA}$ is applied to the Linear-layer weights, where $\mathbf{B} \in \mathbb{R}^{d_{\text{lin}} \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{lin}}}$, and the rank $r \ll d_{\text{lin}}$ is a hyperparameter. Accordingly, $\mathbf{W}_{\text{Base}}$ and ΔLoRA are the only trainable parameters of the TV-SFT stage, while $\tau_{\text{MG-CPT}}$

Table 1: Statistics of the processed datasets.

Type	Dataset	#Users	#POIs	#Instances	Sparsity
Public	NYC	2,083	5,135	104,074	98.10%
	TKY	2,293	7,873	361,430	98.98%
Industry	Beijing	5,431,217	678,819	20,980,626	99.99%
	Shanghai	4,819,962	919,594	22,333,204	99.99%
	Tianjin	2,552,323	499,972	15,182,776	99.99%
	Zhejiang	11,252,453	3,281,512	28,068,268	99.99%

stays frozen. In this way, the frozen $\tau_{\text{MG-CPT}}$ keeps the acquired spatial knowledge intact, while ΔLoRA grants it just enough flexibility to meet the demands of the generative POI recommendation task.

The two branches are combined into the final fine-tuned model:

$$\mathbf{W}_{\text{TV-SFT}} = \mathbf{W}_{\text{Base}} + \alpha (\tau_{\text{MG-CPT}} + \Delta\text{LoRA}). \quad (11)$$

Hyperparameter α controls the strength of the spatial knowledge, set to 1 by default. Since $\mathbf{W}_{\text{Base}} + \tau_{\text{MG-CPT}} = \mathbf{W}_{\text{MG-CPT}}$, this formulation recovers two regimes as special cases: $\alpha = 1$ amounts to fine-tuning based on MG-CPT LLM $\mathbf{W}_{\text{MG-CPT}}$, through the full-parameter branch together with ΔLoRA ; $\alpha = 0$ discards the spatial knowledge, recovering full-parameter SFT of the base LLM $\mathbf{W}_{\text{Base}}$.

Training objective. Given a prompt composed of the user’s historical behavior sequence and contextual signals, the fine-tuned LLM learns to autoregressively generate the SI-SID of the target POI $\mathbf{Q}_t = \langle a_*, b_*, c_*, d_* \rangle$ by minimizing the negative log-likelihood:

$$\begin{aligned} \mathcal{L}_{\text{TV-SFT}} &= -\log P_{\mathbf{W}_{\text{TV-SFT}}}(\mathbf{Q}_t | \text{prompt}) \\ &= -\sum_{i=1}^4 \log P_{\mathbf{W}_{\text{TV-SFT}}}(q_i | \text{prompt}, q_{<i}), \end{aligned} \quad (12)$$

where q_i denotes the i -th token of $\mathbf{Q}_t$ and $q_{<i}$ the preceding tokens.

5 Experiments

5.1 Experimental Setting

Datasets. We evaluate SPAR on **two public benchmarks**, NYC and TKY [31, 36], and **four industrial-scale datasets** collected from AMAP platform, a prominent navigation and mapping platform serving billions of users and POIs. The public datasets are pre-processed following previous works [31]: after sorting interactions chronologically, we allocate 80% for training, 10% for validation, and 10% for testing. For industrial dataset, we collect and clean one month of user trajectories and interaction logs. Every interaction therein is treated as one prediction target, for which we trace back the user’s preceding behaviors to construct its historical sequence, so that each (history, target) pair forms one complete training instance. The interactions on the last day are reserved as the test set and all earlier ones are used for training. Dataset statistics are shown in Table 1. We will release the four industrial datasets.

Evaluation Metrics. We adopt Recall@K, NDCG@K and MRR@K with $K \in \{5, 10, 20\}$, which measure top-K hit coverage, ranking quality, and the reciprocal rank of the first correct prediction, respectively. We abbreviate them as R@K, N@K, and M@K where space is limited in some tables.

Comparison Methods. We compare SPAR against three categories of baselines: (1) *Traditional sequential models*: SASRec [12], BERT4Rec [27], GRU4Rec [10], Caser [28], and S³-Rec [41]; (2) *Transformer based POI recommendation methods*: TPG [20] and

Table 2: Offline performance on NYC and TKY. "Rela.Imp" indicates the relative improvement compared to the best baselines.

Dataset	Metric	Traditional sequential models					POI related		GRMs			SPAR-4B (Rela.Imp)	SPAR-8B (Rela.Imp)
		SASRec	BERT4Rec	GRU4Rec	Caser	S ³ -Rec	TPG	Rotan	TIGER	GNPR-SID	PLUM		
NYC	R@5	0.3151	0.2857	0.1977	0.2883	0.3071	0.3551	0.4448	0.4965	0.5311	0.5619	0.6213 (10.57%)	0.6556 (16.68%)
	R@10	0.3896	0.3564	0.2460	0.3570	0.3854	0.4441	0.5223	0.5514	0.5942	0.6143	0.6686 (8.84%)	0.6996 (13.89%)
	R@20	0.4506	0.4130	0.2889	0.4135	0.4503	0.5121	0.5834	0.6001	0.6455	0.6642	0.7120 (7.20%)	0.7649 (15.16%)
	N@5	0.2224	0.2074	0.1442	0.2044	0.2235	0.2464	0.3471	0.4131	0.4430	0.4672	0.5116 (9.50%)	0.5415 (15.90%)
	N@10	0.2467	0.2304	0.1599	0.2267	0.2489	0.2755	0.3723	0.4276	0.4634	0.4870	0.5363 (10.12%)	0.5651 (16.04%)
	N@20	0.2622	0.2448	0.1708	0.2410	0.2654	0.2927	0.3878	0.4443	0.4766	0.4951	0.5608 (13.27%)	0.5906 (19.29%)
TKY	R@5	0.3450	0.2649	0.2514	0.3257	0.3365	0.3725	0.4333	0.5031	0.5354	0.5526	0.6108 (10.53%)	0.6474 (17.16%)
	R@10	0.4284	0.3326	0.3106	0.4067	0.4115	0.4601	0.5113	0.5808	0.6130	0.6255	0.6712 (7.31%)	0.7098 (13.48%)
	R@20	0.4976	0.3943	0.3651	0.4758	0.4739	0.5291	0.5894	0.6431	0.6675	0.6812	0.7217 (5.95%)	0.7623 (11.91%)
	N@5	0.2384	0.1907	0.1833	0.2273	0.2423	0.2591	0.3293	0.4003	0.4437	0.4659	0.4994 (7.19%)	0.5342 (14.66%)
	N@10	0.2655	0.2127	0.2025	0.2535	0.2666	0.2881	0.3568	0.4251	0.4623	0.4843	0.5116 (5.64%)	0.5473 (13.01%)
	N@20	0.2831	0.2284	0.2163	0.2711	0.2825	0.3051	0.3739	0.4401	0.4788	0.4936	0.5331 (8.00%)	0.5712 (15.72%)

Table 3: Performance comparison on four industrial AMAP Datasets. Ave.Imp denotes the average improvement over PLUM.

Dataset	Beijing			Shanghai			Tianjin			Zhejiang			Ave.Imp
	R@5	N@5	M@5	R@5	N@5	M@5	R@5	N@5	M@5	R@5	N@5	M@5	
TIGER (8B)	0.4112	0.2959	0.2452	0.3426	0.2328	0.1924	0.3362	0.2373	0.2029	0.4358	0.2808	0.2432	-17.54%
GNPR-SID (8B)	0.4243	0.3272	0.2840	0.3593	0.2501	0.2215	0.3624	0.2590	0.2241	0.4404	0.3209	0.2843	-6.19%
PLUM (8B)	0.4514	0.3403	0.3017	0.3704	0.2712	0.2396	0.3910	0.2761	0.2408	0.4665	0.3518	0.3006	(Base)
SPAR-0.6B	0.4729	0.3546	0.3193	0.4062	0.2975	0.2469	0.4310	0.3083	0.2631	0.4892	0.3618	0.3165	9.92%
SPAR-4B	0.5708	0.4325	0.3866	0.4910	0.3656	0.3241	0.5051	0.3775	0.3352	0.5424	0.4079	0.3633	28.54%
SPAR-8B	0.6176	0.4643	0.4113	0.5353	0.3968	0.3502	0.5528	0.4122	0.3654	0.5708	0.4266	0.3825	38.32%
R.I (SPAR 8B to 4B)	8.20%	7.35%	6.39%	9.02%	8.53%	8.05%	9.44%	9.19%	9.00%	5.23%	4.58%	5.28%	-

Rotan [4]; (3) *Generative recommendation models (GRMs)*: TIGER [25], GNPR-SID [31] and PLUM [8].

Implementation Details. All experiments are conducted on 64 NVIDIA A100 GPUs. We use Qwen3-0.6B, -4B, and -8B as the backbone LLMs for MG-CPT and TV-SFT, with their corresponding embedding models producing POI representations. Unless otherwise stated, we use the 4B backbone by default for a balance of effectiveness and efficiency. The SI-SID module uses $L = 4$ levels (three RQ-Kmeans codebooks and one disambiguation index), with $N_c = 32$ for NYC and TKY and $N_c = 512$ for the AMAP datasets. For a fair comparison, all GRMs adopt four-level SIDs, whose last level serves as a disambiguation index. All models are trained with a batch size of 32 and a learning rate of 2×10^{-5} .

5.2 Overall Performance

5.2.1 Results on Public Benchmarks. As shown in Table 2, SPAR achieves state-of-the-art(SOTA) performance across all metrics on NYC and TKY. Two critical observations emerge:

GRMs significantly outperform other methods. Across all six metrics, GRMs (TIGER, GNPR-SID, PLUM, SPAR) consistently surpass traditional and POI-related methods: on NYC R@5, the best GRM (0.6556) exceeds the best traditional one (0.3071) by 113.5% and the best POI-related one (0.4448) by 47.4%. Through autoregressive decoding, the generative paradigm more comprehensively models user-POI dependencies, avoiding the representation bottleneck of the candidate-scoring stage in discriminative approaches.

SPAR significantly outperforms existing generative baselines. SPAR-8B surpasses the strongest baseline PLUM (8B) across all metrics, with relative improvements of 11.91%–19.29%, confirming the efficacy of reinforcing spatial perception throughout the training lifecycle: SI-SID grounds identifiers topologically, MG-CPT

cultivates spatial-relationship understanding, and TV-SFT preserves it during adaptation. Notably, even SPAR-4B outperforms all baselines on every metric, showing that spatial-perception enhancement yields greater gains than merely scaling parameters.

5.2.2 Results on Industrial Datasets. Since GRMs are systematically superior, we compare SPAR with three SOTA GRMs. As shown in Table 3, SPAR maintains substantial advantages:

SPAR maintains significant advantages at industrial scale. SPAR-8B outperforms PLUM on all four datasets, with an average relative gain of 38.32% across R@5/N@5/M@5; SPAR-4B and SPAR-0.6B reach 28.54% and 9.92%, respectively. Notably, even the 0.6B variant surpasses every 8B baseline on all twelve metrics. This comparison against PLUM is particularly informative, as PLUM instantiates a generic SID-CPT-SFT pipeline [8], whereas SPAR reorganizes all three stages around urban spatial knowledge. The consistent gains across the four cities thus stem from making each stage spatially aware rather than from the multi-stage pipeline itself, showing that such knowledge is especially valuable under the noise, sparsity, and distribution shifts of production LBS data.

Model scaling continues to yield consistent gains. Performance grows monotonically with model scale, with the average gain over PLUM rising from 9.92% (0.6B) to 28.54% (4B) and 38.32% (8B); as reported in the last row (R.I), scaling from 4B to 8B still adds 4.58%–9.44% per metric. These results indicate that larger LLMs store and understand urban spatial knowledge better, which translates into sustained gains in recommendation accuracy.

5.3 Ablation Study

To validate the individual contributions and synergistic effects of SPAR’s core components, we conduct a systematic ablation study

Table 4: Ablation study on industrial datasets. R.I(TV-SFT) is the relative improvement contributed by the TV-SFT module.

Datasets	Models	Recall@5	Recall@10	Recall@20	NDCG@5	NDCG@10	NDCG@20	MRR@5	MRR@10	MRR@20	Ave.Imp
Beijing	TextOnlyGR	0.4391	0.5423	0.6455	0.3059	0.3392	0.3654	0.2552	0.2689	0.2761	-
	SPAR-NoCPT	0.4656	0.5790	0.6768	0.3566	0.3933	0.4181	0.3138	0.3291	0.3359	14.62%
	SPAR-FullSFT	0.5324	0.6429	0.7321	0.3975	0.4334	0.4560	0.3527	0.3676	0.3739	27.34%
	SPAR	0.5708	0.6813	0.7702	0.4325	0.4684	0.4910	0.3866	0.4015	0.4078	37.47%
	R.I(TV-SFT)	7.21%	5.97%	5.20%	8.80%	8.07%	7.67%	9.61%	9.22%	9.06%	-
ShangHai	TextOnlyGR	0.3706	0.4794	0.5861	0.2693	0.3045	0.3315	0.2292	0.2437	0.2512	-
	SPAR-NoCPT	0.4056	0.5165	0.6164	0.2907	0.3267	0.3520	0.2528	0.2677	0.2746	8.14%
	SPAR-FullSFT	0.4376	0.5486	0.6537	0.3204	0.3564	0.3830	0.2817	0.2966	0.3039	17.91%
	SPAR	0.4910	0.5989	0.6927	0.3656	0.4007	0.4245	0.3241	0.3386	0.3452	32.09%
	R.I(TV-SFT)	12.20%	9.17%	5.97%	14.11%	12.43%	10.84%	15.05%	14.16%	13.59%	-
Tianjin	TextOnlyGR	0.3801	0.4881	0.5961	0.2646	0.2996	0.3269	0.2331	0.2475	0.2551	-
	SPAR-NoCPT	0.4050	0.5149	0.6184	0.2948	0.3304	0.3567	0.2585	0.2732	0.2804	8.64%
	SPAR-FullSFT	0.4380	0.5516	0.6571	0.3225	0.3593	0.3861	0.2843	0.2996	0.3070	17.97%
	SPAR	0.5051	0.6191	0.7171	0.3775	0.4145	0.4394	0.3352	0.3506	0.3575	35.67%
	R.I(TV-SFT)	15.32%	12.24%	9.13%	17.05%	15.36%	13.80%	17.90%	17.02%	16.45%	-
Zhejiang	TextOnlyGR	0.4719	0.5920	0.6992	0.3240	0.3628	0.3901	0.2884	0.3044	0.3120	-
	SPAR-NoCPT	0.4995	0.6135	0.7096	0.3695	0.4065	0.4309	0.3265	0.3418	0.3486	11.53%
	SPAR-FullSFT	0.5129	0.6269	0.7224	0.3797	0.4167	0.4409	0.3356	0.3509	0.3577	14.88%
	SPAR	0.5424	0.6506	0.7379	0.4079	0.4431	0.4653	0.3633	0.3779	0.3841	23.29%
	R.I(TV-SFT)	5.74%	3.77%	2.15%	7.43%	6.33%	5.54%	8.25%	7.70%	7.37%	-

across four industrial datasets. As shown in Table 4, we compare four variants that progressively incorporate additional components: TextOnlyGR (text-based SID + FullSFT), SPAR-NoCPT (SI-SID + FullSFT), SPAR-FullSFT (SI-SID + MG-CPT + FullSFT), and finally SPAR (SI-SID + MG-CPT + TV-SFT), observing the stepwise performance evolution. R.I(TV-SFT) denotes the relative improvement of SPAR over SPAR-FullSFT, isolating the contribution of TV-SFT. Based on Table 4, we draw three key conclusions:

SI-SID and MG-CPT enable hierarchical spatial perception. The average improvement from TextOnlyGR to SPAR-NoCPT is 10.76%, confirming that SI-SID encodes geographic proximity directly into identifier structure via RQ-Kmeans, providing a foundational topological carrier that enables the model to distinguish spatially plausible candidates from semantically similar but distant ones. Further introducing MG-CPT (SPAR-NoCPT → SPAR-FullSFT) yields an additional 15.38% gain, demonstrating that continued pre-training on multi-granular spatial data successfully teaches the model POI locations and spatial relationships between POIs and roads. The consistent and non-overlapping gains across both stages validate that SI-SID solves the "where" grounding problem while MG-CPT solves the "spatial relationship understanding" problem, confirming the effectiveness of SPAR's "representation construction → spatial relationship learning" design pathway.

TV-SFT achieves better performance with spatial knowledge preservation. The R.I(TV-SFT) row shows relative improvements ranging from 2.15% to 17.90% across the four datasets, with Tianjin exhibiting the most significant gain (15.32% R@5, 17.05% N@5 and 17.90% MRR@5). This result demonstrates that while MG-CPT followed by Full-SFT already improves over SI-SID-only baselines, the massive volume of user behavior sequences in SFT gradually overwhelms spatial knowledge learned during CPT, causing the model to degenerate toward a purely behavioral recommender. The proposed TV-SFT prevents this dilution by keeping τ_{MG-CPT} frozen, ensuring spatial relationships remain immune to parameter updates, while LoRA provides lightweight refinement that adapts spatial knowledge to co-occurring behavioral patterns. This dual mechanism achieves superior POI retrieval performance by maintaining an optimal balance between spatial knowledge

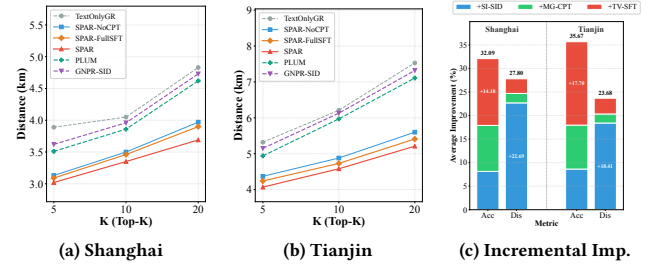


Figure 2: (a)/(b): distance comparison across Top-K recommended POIs to user's location; (c): Incremental contribution to distance and accuracy average improvement.

preservation and task-specific fitting that conventional Full-SFT cannot sustain under large-scale behavioral data pressure.

Cumulative advantage of the full SPAR framework. SPAR achieves 23.29%–37.47% improvement over TextOnlyGR across four datasets, far exceeding the simple sum of individual component contributions, indicating positive synergy among SI-SID, MG-CPT, and TV-SFT: high-quality spatial identifiers are prerequisite for effective continued pre-training, and task-vector anchored fine-tuning ensures complete release of pre-trained spatial knowledge in downstream POI recommendation. This validates that systematic spatial reinforcement throughout the training lifecycle, from representation construction to relationship learning to knowledge-preserving adaptation, is essential for practical POI recommendation.

5.4 Spatial Proximity and Evolution Analysis

We examine how SI-SID affects spatial proximity and how each stage contributes to SPAR. Figure 2(a)/(b) plot the average haversine distance between the Top-K recommended POIs and the user's real-time location (represented by Dis@K), while Figure 2(c) decomposes the incremental accuracy improvement (Ave.Imp over TextOnlyGR, from Table 4) contributed by sequentially adding SI-SID, MG-CPT, and TV-SFT. We draw three key conclusions:

Table 5: Summary of cold- and warm-start performance.

Datasets		Shanghai		Zhejiang	
Task	Metrics	TextOnlyGR	SPAR	TextOnlyGR	SPAR
Cold-start	R@5	0.2821	0.4164	0.3033	0.4490
	N@5	0.2107	0.3157	0.2442	0.3286
	M@5	0.1705	0.2608	0.1981	0.2735
Warm-start	R@5	0.3908	0.5106	0.4922	0.5654
	N@5	0.2950	0.3773	0.3512	0.4217
	M@5	0.2593	0.3483	0.2911	0.3661

(1) **SI-SID establishes the spatial foundation by pulling recommended POIs closer to the user.** TextOnlyGR, GNPR-SID, and PLUM all derive identifiers from textual descriptions containing coordinates and addresses, further reinforced by contrastive objectives and collaborative signals, yet their Dis@5 remains 3.51–3.89 km on Shanghai and 4.94–5.32 km on Tianjin. SI-SID instead fuses normalized geospatial features into the representation before quantization, making proximity a structural property of the codebook. SPAR-NoCPT, which differs from these baselines mainly in adopting SI-SID, already lowers Dis@5 to 3.13 and 4.37 km, with accuracy gains of 8.14% and 8.64%. SI-SID thus provides the spatial foundation on which MG-CPT and TV-SFT further build.

(2) **MG-CPT and TV-SFT are the main drivers of accuracy improvement.** Adding MG-CPT (SPAR-NoCPT → SPAR-FullSFT) lifts accuracy from 8.14% to 17.91% on Shanghai and 8.64% to 17.97% on Tianjin, and the final TV-SFT stage brings the largest gains (32.09% and 35.67%), while distance keeps decreasing (to 27.80% and 23.68% cumulative reduction). Explicit spatial-reasoning cultivation and knowledge-preserving adaptation are thus the core mechanisms that turn spatial awareness into recommendation quality.

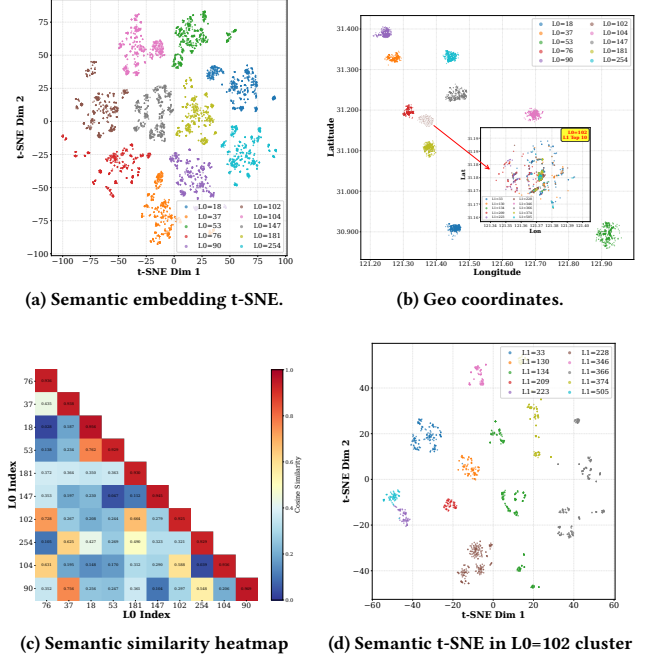
(3) **The stepwise evolution validates the SPAR design.** Distance reduction is led by SI-SID while accuracy gains are led by MG-CPT and TV-SFT: SI-SID embeds spatial topology into the identifier vocabulary, upon which MG-CPT and TV-SFT cultivate and preserve spatial reasoning. This division is consistent with our observation: coordinate-infused SI-SID is the foundation and consistently reinforcing real urban spatial knowledge is the core driver of effective generative POI recommendation.

5.5 Cold-Start Performance

To evaluate SPAR’s robustness under data-sparse conditions, we split each dataset by user behavior sequence length into cold-start (< 8) and warm-start (≥ 16) subsets; note that query terms and real-time location remain available in both, and only historical interactions are sparse. Table 5 summarizes the performance.

Sparse behavioral histories substantially degrade retrieval. TextOnlyGR suffers substantial performance degradation when behavior histories are sparse, with Recall@5 falling 27.9% on Shanghai (0.3908 → 0.2821) and 38.3% on Zhejiang (0.4922 → 0.3033), with similar declines in NDCG@5 and MRR@5, confirming that behavior sequences are a critical signal for POI recommendation.

Spatial knowledge substantially optimizes cold-start performance. SPAR improves over TextOnlyGR in both settings, with disproportionately larger cold-start gains: Recall@5 rises 47.6% on Shanghai and 48.0% on Zhejiang, versus 30.7% and 14.9% in warm-start. That the gains are larger when behavioral signals are scarce

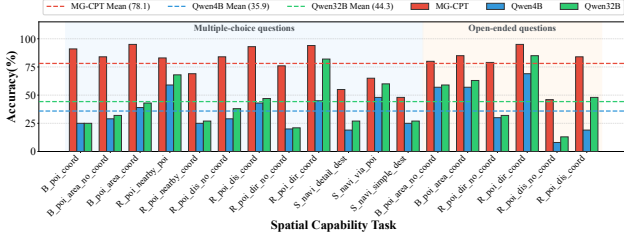
**Figure 3: Multi-view analysis of SI-SID clusters reveals hierarchical semantic and geographic structure.**

attributes this advantage to SPAR’s spatial mechanisms: SI-SID’s topologically-grounded identifiers and MG-CPT’s learned spatial relationships supply ranking signals that compensate for missing history. Accordingly, SPAR narrows the cold-to-warm Recall@5 gap from 27.9%–38.3% (TextOnlyGR) to 18.5%–20.6%, showing that spatial knowledge also strengthens robustness against data sparsity.

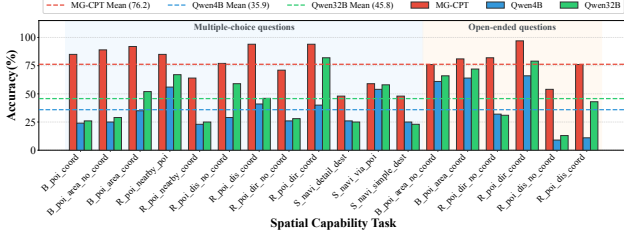
5.6 Hierarchical Visualization of SI-SID

To understand how SI-SID organizes POIs in a hierarchical semantic-geographic space, we conduct a multi-view analysis of its quantization structure: Figure 3(a) and (b) show the semantic t-SNE and geographic coordinates of the top-10 L0 clusters, Figure 3(c) quantifies their intra- and inter-cluster semantic similarity, and Figure 3(d) zooms into L0=102 to visualize its top-10 L1 sub-clusters (geographic distribution in the inset of Figure 3(b)). We highlight two important conclusions:

L0 clusters exhibit both semantic consistency and geographic distinguishability. Figure 3(a) shows that POIs within the same L0 cluster form compact groups in the semantic t-SNE space, while Figure 3(c) confirms high intra-cluster semantic similarity. Simultaneously, Figure 3(b) reveals that different L0 clusters occupy distinct geographic regions with minimal spatial overlap. This indicates that the first-level quantization primarily partitions POIs based on geographic proximity: nearby venues tend to fall into the same L0 cluster, and their semantic consistency likely arises because adjacent urban regions often share similar business compositions or functional roles.



(a) Spatial cognition evaluation on Shanghai Dataset



(b) Spatial cognition evaluation on Beijing Dataset

Figure 4: Spatial cognition evaluation across 18 tasks.

L1 sub-clusters capture fine-grained semantic distinctions beyond geographic proximity. Within $L0=102$, the top-10 L1 sub-clusters separate into distinct semantic groups (Figure 3(d)) yet overlap geographically (inset of Figure 3(b)). Since POIs in an $L0$ cluster are already geographically proximate, the geographic dimension is largely exhausted at the first level; the residual passed to L1 is therefore dominated by semantic variation, and L1 naturally refines these POIs by semantic category rather than location. This coarse-to-fine progression, geographic partitioning at $L0$ followed by semantic refinement at L1, confirms the effectiveness of SI-SID tokenization, allowing SPAR to combine the coarse geographic constraint from $L0$ with fine-grained semantic preferences for retrieval conditioned on user location and query intent.

5.7 Spatial Cognition Evaluation for MG-CPT

To assess whether MG-CPT effectively acquires spatial cognition, we construct one benchmark for each industrial dataset, each consisting of 18 tasks with 2,000 questions, spanning multiple-choice and open-ended formats; Figure 4 reports results on Shanghai and Beijing. Most numbers in the subsection are accuracies in percent, averaged over the 18 spatial tasks of each benchmark. Each task name combines a category prefix—B, R, S for basic-attribute, relational, and systemic questions—with an ability suffix: *dis/dir* for distance/direction reasoning, *area/nearby* for region/surroundings queries, and *navi* for navigation. The *coord/no_coord* suffix indicates whether coordinates are provided or omitted, with the latter forcing reliance on internalized geographic knowledge. Based on these results, we draw several key conclusions:

Geospatial cognition is instilled by MG-CPT rather than acquired through scale. MG-CPT raises the average accuracy (%) of a 4B backbone (Qwen3-4B) from 36.8 to 76.1 (a 39.3-point absolute gain, more than $2\times$ its base accuracy) and further surpasses the $8\times$

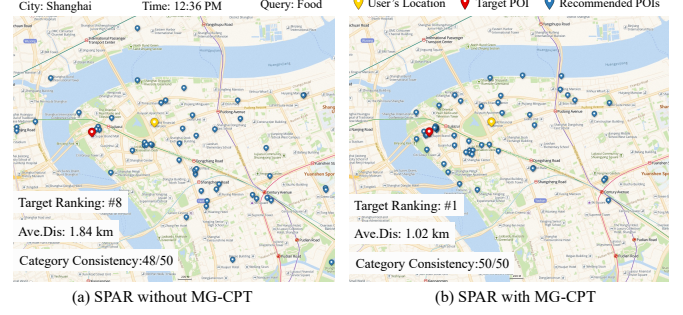


Figure 5: Visualization of recommended POIs distribution.

larger Qwen3-32B (45.3) by 30.8 points. That a 4B LLM post-MG-CPT outperforms a general-purpose 32B model by such a margin shows that geospatial competence stems from targeted CPT on 25 structured per-city spatial datasets, not from parameter count.

Scaling a general LLM yields only marginal spatial gains. Enlarging the general backbone from 4B to 32B improves the average accuracy by merely 8.5 ($36.8 \rightarrow 45.3$), an order of magnitude smaller than the gain delivered by MG-CPT. This confirms that industrial-scale pre-training alone, however large, does not endow a model with the fine-grained geographic knowledge of a specific city, which must instead be explicitly cultivated.

MG-CPT internalizes the city’s geographic layout rather than computing over given coordinates. Base LLMs depend heavily on explicit coordinates: Qwen3-32B, for example, jumps from 21 to 82 on directional questions once coordinates are provided. In contrast, MG-CPT remains strong even without coordinates (76 and 79 on directional Choice and Open questions), exceeding the coordinate-free Qwen3-32B by more than $3\times$. This indicates that MG-CPT encodes the geospatial configuration of the corresponding city into its parameters, enabling metric and directional reasoning without coordinates supplied at inference time.

Performance follows the three-tier complexity gradient. Accuracy is highest on Basic-attribute tasks (80–95), moderate on Relational tasks, and lowest on Systemic navigation tasks (48–65), mirroring the increasing spatial complexity by which the training data are organized. Although navigation remains the hardest tier, MG-CPT still substantially outperforms both baselines on it, showing that even city-scale connectivity is partially internalized.

5.8 Case Study: Proximity versus Reachability

Figure 5 visualizes the top-50 POIs recommended by SPAR **with and without MG-CPT** for a user located in Shanghai, together with the average distance to the user’s location, the category consistency with the target, and the target’s ranking.

As shown in Figure 5(a), without MG-CPT (i.e., urban spatial knowledge), the recommendations are already semantically sound, with 48 of the 50 POIs matching the target category, yet spatially scattered: the average distance reaches 1.84 km and the target is ranked only #8. Many candidates even fall on the opposite bank of the river, close in straight-line distance and correct in category but reachable only through a bridge or tunnel detour.

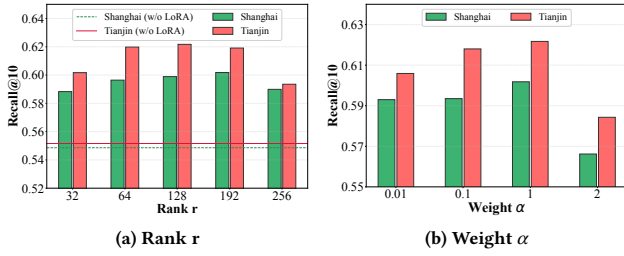


Figure 6: Analysis of two hyper-parameters in TV-SFT.

On the contrary, with MG-CPT (Figure 5(b)), the recommendations concentrate on the user’s own side of the river along the surrounding roads, cutting the average distance to 1.02 km (a 44.6% reduction), raising category consistency to 50 of 50, and promoting the target to #1. As the two variants are comparable in category consistency, the improvement stems from spatial grounding rather than better semantics. In conclusion, this case shows that injecting urban spatial knowledge through MG-CPT makes recommendations not only geographically close but also practically reachable, corroborating the distance and ablation analyses above.

5.9 Hyper-Parameters Analysis of TV-SFT

As shown in Figure 6, with all other settings fixed, we investigate two key hyperparameters of the TV-SFT stage: LoRA rank $r \in \{32, 64, 128, 192, 256\}$ and fusion weight $\alpha \in \{0.01, 0.1, 1, 2\}$.

LoRA Rank r . The LoRA adapter refines the spatial knowledge carried by τ_{MG-CPT} so that it better fits the full-parameter branch of TV-SFT. Adding LoRA consistently outperforms the variant without it (horizontal reference line in Figure 6(a)), and although the optimal rank differs slightly across datasets ($r=192$ in Shanghai, $r=128$ in Tianjin), performance varies within a narrow band over all tested ranks, indicating that the low-rank constraint confines spatial refinement to a robust subspace.

Fusion Weight α . The α controls how strongly τ_{MG-CPT} contributes to the full-parameter fine-tuned model. As shown in Figure 6(b), both datasets peak at $\alpha = 1$ and degrade on either side, matching the design rationale: since τ_{MG-CPT} is exactly the parameter shift acquired during MG-CPT, $\alpha = 1$ integrates the preserved spatial knowledge in full, whereas $\alpha < 1$ underutilizes it and $\alpha > 1$ over-amplifies it at the cost of behavioral adaptation.

6 Conclusion

In this paper, we argue that generative POI recommendation in LBS should ground the interest space of user behavior in the real urban space, rather than inferring geography from behavior alone. We therefore propose SPAR, a unified framework that injects real urban spatial knowledge through three synergistic stages. SI-SID constructs the geographic foundation by encoding coordinates directly into the semantic identifier space; building on it, MG-CPT injects urban spatial knowledge at scale, continually pre-training the backbone LLM on 25 self-constructed geospatial datasets so that scattered POIs cohere into a connected urban space; and TV-SFT preserves this knowledge as a frozen task vector while adapting to user behavior, fusing the interest space with the real urban space. Extensive quantitative and qualitative experiments on two public

benchmarks and four industrial datasets show that SPAR achieves SOTA performance, each stage (SI-SID, MG-CPT, and TV-SFT) fulfills its intended role, and recommendations become geographically closer and practically reachable, validating that grounding interest modeling in real urban spatial knowledge is key to accurate generative POI recommendation.

References

- [1] Ben Chen, Xian Guo, Siyuan Wang, Zihan Liang, Yue Lv, Yufei Ma, Xinlong Xiao, Bowen Xue, Xuxin Zhang, Ying Yang, et al. 2025. Onesearch: A preliminary exploration of the unified end-to-end generative framework for e-commerce search. *arXiv preprint arXiv:2509.03236* (2025).
- [2] Dong Chen, Shuai Zheng, Haoyang Shao, Hongsheng Wu, Muhao Xu, Yeyu Yan, Ruifang Li, and Zhenfeng Zhu. 2026. Revisiting General Map Search via Generative Point-of-Interest Retrieval. *arXiv preprint arXiv:2605.03397* (2026).
- [3] Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965* (2025).
- [4] Shanshan Feng, Feiyu Meng, Lisi Chen, Shuo Shang, and Yew Soon Ong. 2024. Rotan: A rotation-based temporal attention network for time-specific next poi recommendation. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 759–770.
- [5] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.
- [6] Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.
- [7] Yulong Gu, Lixin Zou, and Chenliang Li. 2026. Deep Learning to Rank in Industrial Search Engines, Recommender Systems, and Online Advertising: An Overview and New Perspectives. *ACM Transactions on Information Systems* 44, 4 (2026), 1–52.
- [8] Ruining He, Lukasz Heldt, Lichan Hong, Raghunandan Keshavan, Shifan Mao, Nikhil Mehta, Zhengyang Su, Alicia Tsai, Yueqi Wang, Shao-Chuan Wang, et al. 2026. Plum: Adapting pre-trained language models for industrial-scale generative recommendations. In *Proceedings of the ACM Web Conference 2026*. 8093–8104.
- [9] Tian He, Chen Yang, Jiawei Zhang, Lin Guo, Wei Lin, and Zhuqing Jiang. 2026. Birds of a Feather Cluster Nearby: a Proximity-Aware Geo-Codebook for Local Service Recommendation. *arXiv preprint arXiv:2604.23156* (2026).
- [10] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).
- [11] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089* (2022).
- [12] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.
- [13] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11523–11532.
- [14] Peibo Li, Maarten de Rijke, Hao Xue, Shuang Ao, Yang Song, and Flora D Salim. 2024. Large language models for next point-of-interest recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1463–1472.
- [15] Defu Lian, Yongji Wu, Yong Ge, Xing Xie, and Enhong Chen. 2020. Geography-aware sequential location recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2009–2019.
- [16] Enze Liu, Bowen Zheng, Cheng Ling, Lantao Hu, Han Li, and Wayne Xin Zhao. 2024. End-to-end learnable item tokenization for generative recommendation. *arXiv e-prints* (2024), arXiv:2409.2409.
- [17] Shuai Liu, Ning Cao, Yile Chen, Yue Jiang, and Gao Cong. 2024. nextlocilm: next location prediction using LLMs. *arXiv preprint arXiv:2410.09129* (2024).
- [18] Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. 2025. PAVE: Patching and adapting video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3306–3317.
- [19] Zhanyu Liu, Shiyao Wang, Xingmei Wang, Rongzhou Zhang, Jiaxin Deng, Honghui Bao, Jinghao Zhang, Wuchao Li, Pengfei Zheng, Xiangyu Wu, et al. 2025. Onerec-think: In-text reasoning for generative recommendation. *arXiv preprint arXiv:2510.11639* (2025).
- [20] Yan Luo, Haoyi Duan, Ye Liu, and Fu-Lai Chung. 2023. Timestamps as prompts for geography-aware location recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1697–1706.

- [21] Yingtao Luo, Qiang Liu, and Zhaocheng Liu. 2021. Stan: Spatio-temporal attention network for next location recommendation. In *Proceedings of the web conference 2021*. 2177–2185.
- [22] Dongyi Lv, Qiuyu Ding, Heng-Da Xu, Zhaoxu Sun, Zhi Wang, Feng Xiong, and Mu Xu. 2026. Reasoning Over Space: Enabling Geographic Reasoning for LLM-Based Generative Next POI Recommendation. *arXiv preprint arXiv:2601.04562* (2026).
- [23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- [24] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology* 16, 5 (2025), 1–72.
- [25] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 10299–10315.
- [26] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. 2024. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *International Conference on Learning Representations*, Vol. 2024. 1746–1759.
- [27] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.
- [28] Jiaxi Tang and Ke Wang. 2018. Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding. In *ACM International Conference on Web Search and Data Mining*.
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [30] Dongsheng Wang, Shen Gao, Chengrui Huang, Yuxi Huang, Ruixiang Feng, and Shuo Shang. 2025. TOOL4POI: A Tool-Augmented LLM Framework for Next POI Recommendation. *arXiv preprint arXiv:2511.06405* (2025).
- [31] Dongsheng Wang, Yuxi Huang, Shen Gao, Yifan Wang, Chengrui Huang, and Shuo Shang. 2025. Generative next poi recommendation with semantic id. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 2904–2914.
- [32] Fangye Wang, Haowen Lin, Yifang Yuan, Siyuan Wang, Xiaojiang Zhou, Song Yang, and Pengjie Wang. 2026. GeoGR: A Generative Retrieval Framework for Spatio-Temporal Aware POI Recommendation. *arXiv preprint arXiv:2602.10411* (2026).
- [33] Zhipeng Wei, Kuo Cai, Junda She, Jie Chen, Minghao Chen, Yang Zeng, Qiang Luo, Wencong Zeng, Ruiming Tang, Kun Gai, et al. 2025. Oneloc: Geo-aware generative recommender systems for local life service. *arXiv preprint arXiv:2508.14646* (2025).
- [34] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems* 36 (2023), 7093–7115.
- [35] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
- [36] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2014. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 1 (2014), 129–142.
- [37] Song Yang, Jiamou Liu, and Kaiqi Zhao. 2022. Getnext: trajectory flow map enhanced transformer for next poi recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*. 1144–1153.
- [38] Penglong Zhai, Jie Li, Fanyi Di, Yue Liu, Yifang Yuan, Jie Huang, Peng Wu, Sicong Wang, Mingyang Yin, Tingting Hu, et al. 2025. Cognitive-Aligned Spatio-Temporal Large Language Models For Next Point-of-Interest Prediction. *arXiv preprint arXiv:2510.14702* (2025).
- [39] Qianru Zhang, Peng Yang, Junliang Yu, Haixin Wang, Xingwei He, Siu-Ming Yiu, and Hongzhi Yin. 2025. A survey on point-of-interest recommendation: Models, architectures, and security. *IEEE Transactions on Knowledge and Data Engineering* (2025).
- [40] Guorui Zhou, Hengrui Hu, Hongtao Cheng, Huanjie Wang, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Lu Ren, Liao Yu, et al. 2025. Onerec-v2 technical report. *arXiv preprint arXiv:2508.20900* (2025).
- [41] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1893–1902.

Appendices

A Theoretical Analysis of the Geospatial Encoder

This appendix complements Section 4.1 Spatially-Intrinsic SID by analyzing why the sinusoidal geospatial encoder produces embeddings whose geometry faithfully mirrors geographic topology, the property that SI-SID inherits through residual quantization. We first state two exact properties of the Fourier features $\gamma(\cdot)$, then empirically verify that both properties survive the learned projection.

A.1 Exact Properties

In what follows, let $u \in [0, 2\pi]$ denote a generic normalized coordinate (either $\hat{\lambda}$ or $\hat{\phi}$ from Section 4.1); the two properties hold for each coordinate separately and extend to their concatenation.

Property 1 (Shift-invariant kernel). For any two values u_1, u_2 , the product-to-sum identity gives

$$\begin{aligned} \langle \gamma(u_1), \gamma(u_2) \rangle &= \sum_{k=0}^{K-1} [\sin(f_k u_1) \sin(f_k u_2) + \cos(f_k u_1) \cos(f_k u_2)] \\ &= \sum_{k=0}^{K-1} \cos(f_k (u_1 - u_2)), \end{aligned} \quad (13)$$

which depends only on the difference $u_1 - u_2$ rather than on absolute positions. Moreover, since $f_k \leq f_0 = 1$, every term $\cos(f_k \Delta)$ is decreasing in Δ on $[0, \pi]$, so the inner product is strictly decreasing in $|u_1 - u_2|$ whenever $|u_1 - u_2| \leq \pi$, a range that covers all intra-city displacements by a wide margin. Embedding similarity therefore behaves as a monotonically decreasing function of coordinate displacement. Note that although longitude is periodic ($-180^\circ \equiv 180^\circ$), all POIs in our setting lie within a single city, so coordinate displacements satisfy $|u_1 - u_2| \ll \pi$ and the wrap-around case does not arise in practice.

Property 2 (Lipschitz continuity). Expanding the squared distance between two feature vectors,

$$\begin{aligned} \|\gamma(u_1) - \gamma(u_2)\|_2^2 &= \sum_{k=0}^{K-1} 2(1 - \cos(f_k (u_1 - u_2))) \\ &\leq \left(\sum_{k=0}^{K-1} f_k^2 \right) (u_1 - u_2)^2, \end{aligned} \quad (14)$$

where the inequality follows from $1 - \cos \theta \leq \theta^2/2$. The encoding is thus Lipschitz-continuous: nearby coordinates are guaranteed nearby features. The subsequent MLP, a composition of Lipschitz maps (linear layers and ReLU), preserves this continuity, so POIs with adjacent coordinates enter RQ-Kmeans with adjacent geospatial embeddings $\mathbf{e}_{\text{geo}}$, the prerequisite for adjacent SI-SIDs. The frequencies $f_k = 10000^{-k/K}$ are exponentially spaced, so different bands respond to coordinate differences at different rates, and the MLP learns a weighted combination across bands during end-to-end training.

Remark. Both properties are stated for the Fourier features γ . The MLP preserves continuity exactly (Property 2), whereas the distance-similarity monotonicity of Property 1 is preserved qualitatively after the learned projection and verified empirically (Figure 7).

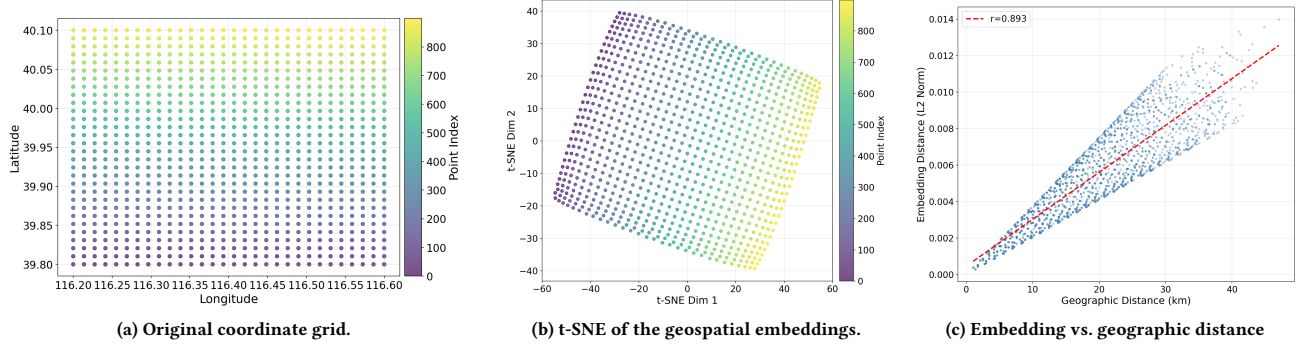


Figure 7: Empirical verification of the geospatial encoder on a 30×30 coordinate grid over Beijing. (a) The original grid, color-coded by point index. (b) The t-SNE projection of the embeddings preserves the grid topology: geographically adjacent points remain adjacent. (c) The embedding distance grows monotonically with the geographic distance (Pearson $r = 0.893$), confirming that the city’s metric structure is faithfully carried into the embedding space.

In addition, the kernel depends on differences of normalized coordinates rather than geodesic distance; within a single city the two are approximately proportional, so the stated monotonicity carries over to real distances at the scale relevant to next-POI recommendation.

A.2 Empirical Verification

We further verify that the two exact properties of γ survive the MLP projection, i.e., that they hold for the final geospatial embedding $\mathbf{e}_{\text{geo}}(\mathbf{p})$ of a location $\mathbf{p} = (\lambda, \phi)$. To this end, we encode a 30×30 uniform coordinate grid over an urban region of Beijing ($116.2^\circ - 116.6^\circ\text{E}$, $39.8^\circ - 40.1^\circ\text{N}$; 900 points covering roughly $44\text{ km} \times 33\text{ km}$, Figure 7(a)) and inspect the geometry of the resulting embeddings, as summarized in Figure 7.

First, shift-invariance manifests itself in the regularity of the embedded lattice. As shown in Figure 7(b), the t-SNE projection of the 900 embeddings still forms an evenly spaced grid that mirrors the original one: the same coordinate step induces a nearly identical embedding change wherever it is applied, and no region of the city is stretched or compressed. This position-independent response is exactly the behavior predicted by Property 1.

Second, continuity and metric preservation are confirmed by the distance relation in Figure 7(c): over sampled point pairs of the grid, the embedding distance grows monotonically and near-linearly with the geographic distance (Pearson $r = 0.893$). Nearby coordinates thus receive nearby embeddings, while farther places are consistently mapped farther apart, matching Property 2.

Together, these results confirm that POIs enter SI-SID construction with a geometry-faithful geospatial embedding. Notably, this encoder-level property is corroborated by the identifier-level structure observed in the main text: because the fused representation carries an explicit metric structure over coordinates, the first-level quantization inherits it and partitions POIs primarily by geographic proximity, with the top first-level clusters occupying compact and mutually distinct urban regions. Once the first-level codeword absorbs this shared geographic component, the residual passed to deeper levels is dominated by semantic variation. Accordingly, the

second-level sub-clusters within a single first-level cluster overlap geographically yet separate cleanly by semantic category. The geometry verified here at the encoder level and the hierarchy observed at the identifier level thus validate each other, closing the chain from coordinate encoding to spatially continuous identifiers.

B Details of the MG-CPT Training Corpus

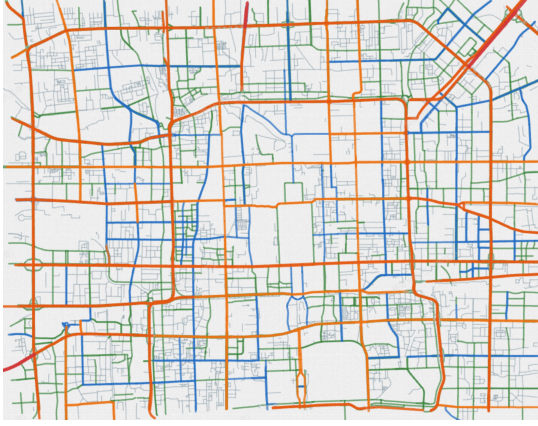
The goal of MG-CPT is to internalize the spatial relations of the whole city into the LLM, so that POIs become interconnected places rather than isolated tokens. To this end, we build a training corpus that spans points, lines, and flows over urban space, comprising **25 datasets**, one per task type, each serialized into (instruction, input, output) triples. Specifically, the instruction assigns a city-geography expert role and states the requirement, the input is the question, and the output is the answer. For all open-ended reasoning tasks, the output follows an “answer-first, derivation-second” format written in plain language, without any special reasoning tags. The corpus is derived from four kinds of geospatial data, summarized in Table 6.

B.1 Three Tiers of Datasets

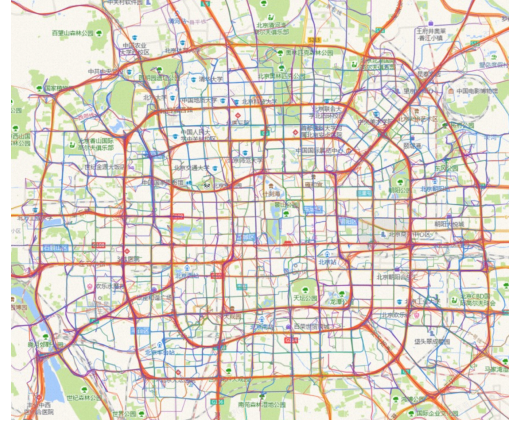
The 25 training datasets instantiate the three tiers of MG-CPT and together assemble a virtual city street map that the backbone can internalize. **Basic-attribute (B, 13 datasets)** describe the static attributes of a single object: twelve POI datasets establish the bidirectional mapping among a POI’s SI-SID, coordinates, and attributes (category, district, address), while one road dataset characterizes a single road (class, length, lanes, speed limit). This tier is the foundation on which all other spatial abilities rest. **Relational (R, 9 datasets)** target pairwise geometry; the eight distance/direction datasets form a fully orthogonal $2 \times 2 \times 2$ design over coordinate visibility (coord/nocoord), reasoning target (dis/dir), and answer format (choice/open), and the contrast between the coord and nocoord groups directly quantifies how much positional knowledge has been internalized, since the latter provides no numeric hint. **Systemic (S, 3 datasets)** move from points to the road network, connecting an origin, a path, and a destination into a complete

Table 6: The four data sources of the MG-CPT training corpus.

Source	Information provided	Granularity
POI snapshot	SI-SID, name, coordinates, address, category, district, ...	point (static)
Road feature	road class, length, chain/link count, lanes, speed limit, ...	line (static)
POI-pair relation	distance and direction of a POI pair, with derivations, ...	point–point
Navigation route	driving distance, duration, road sequence, turn-by-turn steps, ...	point–line–point (dynamic)



(a) Mainly static road network rendered from road features (Beijing).



(b) Dynamic flow aggregated from 2,000 navigation trajectories; each trajectory is a POI–road–POI connection.

Figure 8: Two complementary views of Beijing: the static road skeleton and the dynamic movement flow.

movement through three progressively finer questions: how far, which roads, and how to drive step by step.

B.2 Static Road Network and Dynamic Flow

Internalizing the spatial relations of the whole city relies on two complementary signals: the static road skeleton and the dynamic flow of movement. The road-feature tasks capture the static skeleton—the class, orientation, and connectivity of arterials, ring roads, and expressways—which alone outlines the spatial framework of a city, as visualized in Figure 8(a). However, the static skeleton does not reveal how people actually move through the city. Navigation data encode real movement in the form of an origin POI, a traversed road sequence, and a destination POI, so that each route is a POI–road–POI connection. As shown in Figure 8(b), even 2,000 navigation trajectories (the training set uses about 50,000) already reveal the main flow arteries of the city; when a large number of such connections are aggregated, isolated POIs are woven by real travel paths into a dynamic virtual street map of Beijing. The contrast between the two figures makes the value of this design concrete. Figure 8(a) alone tells the model where the roads are, but the streets remain empty: it cannot say which places people actually travel between, or by what route. Figure 8(b) fills these streets with real movement: every trajectory threads an origin POI, a sequence of roads, and a destination POI into a single line, and as thousands of them overlay, the busy corridors of the city light up while previously isolated POIs turn into nodes on a living transportation graph. By continually pre-training on both views, MG-CPT internalizes not only the straight-line distance between two POIs, but also how one is

reached from another through the road network. This directly benefits downstream next-POI recommendation: rather than merely fitting behavioral co-occurrence, the model can favor destinations that are genuinely reachable from the user’s current location, so that its recommendations respect real urban geography instead of historical transition patterns alone.

B.3 Distance and Direction Annotation

Spherical distance (Haversine). Open-ended distance tasks provide a full derivation. For two points (λ_1, φ_1) and (λ_2, φ_2) ,

$$a = \sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos \varphi_1 \cos \varphi_2 \sin^2\left(\frac{\Delta\lambda}{2}\right), \quad d = 2R \arcsin \sqrt{a}, \quad (15)$$

with $R = 6371$ km. **Planar approximation.** For direction judgment and some tasks we adopt a city-scale planar approximation $d \approx \sqrt{(\Delta\lambda \cdot 85)^2 + (\Delta\varphi \cdot 111)^2}$ km, where 111 is the kilometers per degree of latitude and 85 that per degree of longitude at the city’s latitude. **Eight-way direction.** The sign of $\Delta\lambda$ and $\Delta\varphi$ determines east/west and north/south, and comparing the east–west displacement $|\Delta\lambda| \cdot 85$ with the north–south displacement $|\Delta\varphi| \cdot 111$ decides between a principal and a composite direction, over the set $\{N, NE, E, SE, S, SW, W, NW\}$.

B.4 Option and Distractor Construction

The position of the correct option is determined by a hash of the primary key so that A–D are uniformly distributed and reproducible. Numeric distractors scale the ground truth ($\times 1.5, \times 0.6, \times 2.0, \times 0.4$

for distances) or progressively offset the coordinates, while categorical distractors are drawn from a circular list with offsets coprime to its length (e.g., eight directions or sixteen districts), which prevents duplicated or trivially eliminable options. Open-ended distance and direction tasks additionally include explicit derivations, and the three POI-description tasks each traverse the full POI set, differing not in which POIs they cover but in which attributes serve as the question and which as the answer.

Table 7 lists the templates of all 25 training tasks. To avoid leaking proprietary data, in the following demonstrations we denote a POI's SI-SID by the placeholder $\{\text{SI-SID}_i\}$ and a coordinate by (lon_i, lat_i) ; other public information (names, addresses, road names) is kept for readability.

B.5 Dataset Demonstrations

We show a full card for every one of the 25 training tasks, each consisting of the (translated) instruction, the question, and the gold answer. To avoid leaking proprietary data, SI-SIDs, coordinates, POI names, addresses, and road names are all replaced by placeholders, $\{\text{SI-SID}_i\}$, (lon_i, lat_i) , $\{\text{name}_i\}$, $\{\text{addr}_i\}$, and $\{\text{road}_i\}$, respectively, while generic attributes such as category and district are kept for readability.

B.5.1 Basic-attribute: POI point (12 tasks).

- **poi_coord_qa**
Instruction: You are a geography expert of Shanghai, familiar with the coordinates, districts, and addresses of all local POIs. Accurately provide the coordinates and district of the given POI.
Question: What are the coordinates of $\{\text{SI-SID}_0\}$?
Answer: (lon_0, lat_0) .
- **poi_coord_reverse**
Instruction: You are a geography expert of Shanghai. Given a coordinate, accurately identify the corresponding POI, its name, and its district.
Question: Which POI is located at (lon_0, lat_0) ?
Answer: $\{\text{name}_0\}$ ($\{\text{SI-SID}_0\}$), in Huangpu District.
- **poi_describe_name**
Instruction: You are a geography expert of Shanghai. Given a POI's name or SID, accurately provide its coordinates, district, and address.
Question: Describe the basic geographic information of $\{\text{SI-SID}_0\}$.
Answer: $\{\text{SI-SID}_0\}$ is a shopping mall in Huangpu District, at $\{\text{addr}_0\}$, with coordinates (lon_0, lat_0) .
- **poi_describe_coord**
Instruction: You are a geography expert of Shanghai. Given a coordinate, provide the POI located there together with its category, district, and address.
Question: What is located at (lon_0, lat_0) ?
Answer: $\{\text{SI-SID}_0\}$, a shopping mall in Huangpu District, at $\{\text{addr}_0\}$.
- **poi_describe_reverse**
Instruction: You are a geography expert of Shanghai. Given address information, accurately provide the corresponding POI name, coordinates, and district.
Question: Which POI is a shopping mall at $\{\text{addr}_0\}$, Huangpu District?
Answer: This is $\{\text{name}_0\}$ ($\{\text{SI-SID}_0\}$), with coordinates (lon_0, lat_0) .

- **qa_category**
Instruction: You are a geography expert of Shanghai. Answer the question concisely and accurately.
Question: What is the category of $\{\text{SI-SID}_0\}$?
Answer: Shopping mall.
- **qa_coord_choice**
Instruction: You are a geography expert of Shanghai. For multiple-choice questions, directly output the correct option letter.
Question: What are the coordinates of $\{\text{SI-SID}_0\}$? A. (lon_1, lat_1) B. (lon_2, lat_2) C. (lon_3, lat_3) D. (lon_4, lat_4)
Answer: B.
- **qa_coord_open**
Instruction: You are a geography expert of Shanghai. Answer the question concisely and accurately.
Question: Give the coordinates of $\{\text{SI-SID}_0\}$.
Answer: (lon_0, lat_0) .
- **qa_coord_open_full**
Instruction: You are a geography expert of Shanghai. Given the category, district, and address, provide the coordinates and SI-SID of the POI.
Question: Where is the shopping mall at $\{\text{addr}_0\}$, Huangpu District?
Answer: (lon_0, lat_0) , i.e., $\{\text{SI-SID}_0\}$.
- **qa_district_choice**
Instruction: You are a geography expert of Shanghai. For multiple-choice questions, directly output the correct option letter.
Question: Which district of Shanghai is $\{\text{SI-SID}_0\}$ in? A. Songjiang B. Chongming C. Huangpu D. Jiading
Answer: C.
- **qa_district_open**
Instruction: You are a geography expert of Shanghai. Answer the question concisely and accurately.
Question: Which district of Shanghai is $\{\text{SI-SID}_0\}$ in?
Answer: Huangpu District.
- **qa_nearby**
Instruction: You are a geography expert of Shanghai. Answer the question concisely and accurately.
Question: Which POI is near the coordinate (lon_0, lat_0) ?
Answer: Near this coordinate is $\{\text{SI-SID}_0\}$, a shopping mall in Huangpu District, at $\{\text{addr}_0\}$.

B.5.2 Basic-attribute: line (1 task).

- **road_describe**
Instruction: You are a geography expert of Beijing, familiar with the detailed attributes of the city's road network. Accurately describe the class, length, lanes, and speed limit of the given road.
Question: Describe the basic road information of $\{\text{road}_0\}$, Beijing.
Answer: $\{\text{road}_0\}$ is a secondary road about 31.4 km long, consisting of 26 directional chains and 361 links, mostly 6 lanes, with a speed limit of 70 km/h.

B.5.3 Relational: point to point (9 tasks).

- **poi_nearby_to_poi**
Instruction: You are a geography expert of Shanghai. For multiple-choice questions, directly output the correct option letter.

Table 7: The 25 training datasets of the MG-CPT training corpus (from the Shanghai dataset). “Num(POI)” and “Num(Road)” denote the numbers of POIs and roads, respectively. Additionally, “MC” and “Open” denote the multiple-choice and open-ended template types. For brevity, the “Input → Output” column lists only representative attributes rather than the full set involved in each dataset.

Tier	Task type	Format	Input → Output	Scale
Basic (B)	poi_coord_qa	Open	SI-SID → coordinates	Num(POI)
	poi_coord_reverse	Open	coordinates → SI-SID, name, district	Num(POI)
	poi_describe_name	Open	name, SI-SID → category, district, address, coordinates	Num(POI)
	poi_describe_coord	Open	coordinates → SI-SID, category, district, address	Num(POI)
	poi_describe_reverse	Open	name, district, address, category → SI-SID, coordinates	Num(POI)
	qa_category	Open	SI-SID → category	20,000
	qa_coord_choice	MC	SI-SID + 4 coordinate options → option	20,000
	qa_coord_open	Open	SI-SID → coordinates	20,000
	qa_coord_open_full	Open	category + district + address → coordinates, SI-SID	20,000
	qa_district_choice	MC	SI-SID + 4 district options → option	20,000
Relational (R)	qa_district_open	Open	SI-SID → district	20,000
	qa_nearby	Open	coordinates → SI-SID, category, district, address	20,000
	road_describe	Open	road name → class, length, chain/link count, lanes, speed limit	Num(Road)
	poi_nearby_to_poi	MC	reference coordinates + 4 candidates → nearest option (with distances)	20,000
	dis_coord_choice	MC	two SI-SIDs + coordinates → distance option	20,000
	dis_coord_open	Open	two SI-SIDs + coordinates → distance, with Haversine derivation	20,000
	dir_coord_choice	MC	two SI-SIDs + coordinates → direction option	20,000
	dir_coord_open	Open	two SI-SIDs + coordinates → direction, with derivation	20,000
	dis_nocoord_choice	MC	two SI-SIDs only → distance option	20,000
	dis_nocoord_open	Open	two SI-SIDs only → distance, recalling coordinates first	20,000
Systemic (S)	dir_nocoord_choice	MC	two SI-SIDs only → direction option	20,000
	dir_nocoord_open	Open	two SI-SIDs only → direction, recalling coordinates first	20,000
	route_dist	Open	origin, destination → driving distance, duration, districts	50,000
	route_roads	Open	origin, destination → distance, duration, road sequence	50,000
	route_navi	Open	origin, destination → turn-by-turn navigation	50,000

Question: Which of the following coordinates is closest to (lon_0, lat_0) ?

A. (lon_1, lat_1) B. (lon_2, lat_2) C. (lon_3, lat_3) D. (lon_4, lat_4)

Answer: B. (The answer computes the distance to each option).

- **dis_coord_choice**

Instruction: You are a geography expert. Based on the two coordinates, judge the straight-line distance. For multiple-choice questions, output the option letter.

Question: Distance between $\{SI-SID_0\}$ (lon_0, lat_0) and $\{SI-SID_1\}$ (lon_1, lat_1) ? A. about 4.7 km B. about 7.9 km C. about 15.8 km D. about 3.2 km

Answer: B.

- **dis_coord_open**

Instruction: You are a geography expert. Based on the two coordinates, compute the straight-line distance; give the answer first, then the derivation.

Question: What is the straight-line distance between $\{SI-SID_0\}$ (lon_0, lat_0) and $\{SI-SID_1\}$ (lon_1, lat_1) ?

Answer: About 7.9 km. {Followed by the Haversine derivation.}

- **dir_coord_choice**

Instruction: You are a geography expert. Based on the two coordinates, judge the direction of A relative to B. For multiple-choice questions, output the option letter.

Question: Direction of $\{SI-SID_0\}$ (lon_0, lat_0) relative to $\{SI-SID_1\}$ (lon_1, lat_1) ? A. Northeast B. Southeast C. Southwest D. Northwest

Answer: C.

- **dir_coord_open**

Instruction: You are a geography expert. Based on the two coordinates, reason about the direction of A relative to B; give the answer first, then the reasoning.

Question: Reason about the direction of $\{SI-SID_0\}$ (lon_0, lat_0) relative to $\{SI-SID_1\}$ (lon_1, lat_1) .

Answer: Southwest. {With a component-wise analysis.}

- **dis_nocoord_choice**

Instruction: You are a geography expert. Based only on the two POIs' SIDs, judge the straight-line distance. For multiple-choice questions, output the option letter.

Question: Distance between $\{SI-SID_0\}$ and $\{SI-SID_1\}$? A. about 4.7 km B. about 7.9 km C. about 15.8 km D. about 3.2 km

Answer: B.

- **dis_nocoord_open**

Instruction: You are a geography expert. Based only on the two POIs' SIDs, judge the straight-line distance; give the answer first, then the analysis.

Question: Approximately how far apart are $\{SI-SID_0\}$ and $\{SI-SID_1\}$?

Answer: About 7.9 km; the answer first recalls the two coordinates, then derives the distance.

- **dir_nocoord_choice**

Instruction: You are a geography expert. Based only on the two POIs' SIDs, judge the direction. For multiple-choice questions, output the option letter.

Question: Direction of $\{SI-SID_0\}$ relative to $\{SI-SID_1\}$? A. Northeast B. Southeast C. Southwest D. Northwest

Answer: C.

- **dir_nocoord_open**

Instruction: You are a geography expert. Based only on the two POIs' SIDs, reason about the direction; give the answer first, then the analysis.

Question: Reason about the direction of {SI-SID₀} relative to {SI-SID₁}.

Answer: Southwest; the answer recalls the coordinates, then analyzes the components.

B.5.4 Systemic: point–line–point, navigation (3 tasks).

- **route_dist**

Instruction: You are a geography expert of Shanghai. From the origin and destination POI SIDs, you can obtain approximate navigation information, including distance and time.

Question: How far is it from {SI-SID₀} to {SI-SID₁}, and how long does it take?

Answer: About 11.3 km, roughly 28 min by car (a road-network distance, distinct from the 7.9 km straight-line distance).

- **route_roads**

Instruction: You are a geography expert of Shanghai. Based on the origin and destination SIDs, list the main roads passed by this trip.

Question: Which roads should one take from {SI-SID₀} to {SI-SID₁}?

Answer: About 11.3 km, roughly 28 min, mainly via {road₁}; {road₂}; {road₃}; ...

- **route_navi**

Instruction: You are a geography expert of Shanghai. From the origin and destination POI SIDs, you can obtain detailed navigation.

Question: Give the detailed navigation from {SI-SID₀} to {SI-SID₁}.

Answer: Head north on {road₁} for 180 m, turn right onto {road₂}; ...; arrive at the destination.

system prompt, which assigns the model the role of a city geography expert and specifies the expected output format. The final evaluation input thus follows the template “[system prompt] + [question]”. The system prompts for the two formats are instantiated as follows:

- **Multiple-choice.** “You are a geography expert of City, familiar with its road network and POI locations. Please answer the question accurately. For multiple-choice questions (with options A/B/C/D), directly output the option letter.”
- **Open-ended.** “You are a geography expert of City, familiar with its POI locations. Please answer the question accurately. For open-ended questions, directly output the answer (e.g., the direction, distance, or district).”

C Details of the Spatial Cognition Evaluation Benchmark

To comprehensively evaluate the spatial cognition acquired by MG-CPT, we construct a dedicated benchmark for each industrial dataset, consisting of 18 tasks with 2,000 questions each. All questions are automatically generated from the corresponding city’s POI, road, and navigation data, and are strictly disjoint from the MG-CPT training corpus. The benchmark follows the three-tier organization of MG-CPT: B (Basic-attribute) tasks query the intrinsic properties of a single POI, such as its coordinates or the area it belongs to; R (Relational) tasks examine pairwise spatial relations between POIs, covering distance (dis), direction (dir), and surroundings (nearby); and S (Systemic) tasks assess city-scale route understanding (navi), such as inferring the destination or the passed-by POIs of a navigation route. Each task is instantiated in a multiple-choice format, and six representative tasks are additionally instantiated in an open-ended format to test generation beyond option matching. Moreover, the coord/no_coord suffix controls whether the coordinates are provided in the question: the former tests coordinate-grounded computation, whereas the latter removes all numeric hints and forces the model to rely solely on its internalized geographic knowledge. Table 8 summarizes all tasks along with representative examples. In addition to the question and the ground-truth answer, each task is equipped with a task-specific

Table 8: Details and demonstrations of the 18 spatial cognition evaluation tasks for MG-CPT (examples from the Shanghai dataset). B/R/S denote Basic-attribute, Relational, and Systemic categories; coord/no_coord indicates whether coordinates are provided. {SI-SID_i} is a placeholder for the four-token SI-SID of a POI. For brevity, the “Example” column shows abridged questions, with long navigation steps and road sequences truncated by “...”; the system prompt prepended to each question at evaluation time is omitted.

Format	Task	Description	Example	Answer
MC	B_poi_coord	Identify the coordinates of a given POI.	What are the coordinates of {SI-SID ₀ }? A. (<i>lon</i> ₁ , <i>lat</i> ₁) B. (<i>lon</i> ₂ , <i>lat</i> ₂) C. (<i>lon</i> ₃ , <i>lat</i> ₃) D. (<i>lon</i> ₄ , <i>lat</i> ₄)	A
	B_poi_area_no_coord	Identify the district a POI belongs to, without coordinates.	Which district of Shanghai is {SI-SID ₀ } located in? A. Qingpu B. Jiading C. Xuhui D. Hongkou	B
	B_poi_area_coord	Identify the district a POI belongs to, with coordinates.	Which district of Shanghai is {SI-SID ₀ } (<i>lon</i> ₀ , <i>lat</i> ₀) located in? A. Qingpu B. Jiading C. Xuhui D. Hongkou	B
	R_poi_nearby_poi	Select the POI nearest to a given POI.	Which POI is closest to the coordinate (<i>lon</i> ₀ , <i>lat</i> ₀)? A. {SI-SID ₀ } B. {SI-SID ₁ } C. {SI-SID ₂ } D. {SI-SID ₃ }	D
	R_poi_nearby_coord	Select the coordinate nearest to a given coordinate.	Which of the following coordinates is nearest to (<i>lon</i> ₀ , <i>lat</i> ₀)? A. (<i>lon</i> ₁ , <i>lat</i> ₁) B. (<i>lon</i> ₂ , <i>lat</i> ₂) C. (<i>lon</i> ₃ , <i>lat</i> ₃) D. (<i>lon</i> ₄ , <i>lat</i> ₄)	C
	R_poi_dis_no_coord	Estimate the distance between two POIs, without coordinates.	What is the approximate distance between {SI-SID ₀ } and {SI-SID ₁ }? A. about 18.9 km B. about 37.8 km C. about 75.6 km D. about 113.4 km	B
	R_poi_dis_coord	Estimate the distance between two POIs, with coordinates.	What is the approximate distance between {SI-SID ₀ } (<i>lon</i> ₀ , <i>lat</i> ₀) and {SI-SID ₁ } (<i>lon</i> ₁ , <i>lat</i> ₁)? A. about 18.9 km B. about 37.8 km C. about 75.6 km D. about 113.4 km	B
	R_poi_dir_no_coord	Judge the direction between two POIs, without coordinates.	In which direction is {SI-SID ₀ } relative to {SI-SID ₁ }? A. North B. East C. West D. South	D
	R_poi_dir_coord	Judge the direction between two POIs, with coordinates.	In which direction is {SI-SID ₀ } (<i>lon</i> ₀ , <i>lat</i> ₀) relative to {SI-SID ₁ } (<i>lon</i> ₁ , <i>lat</i> ₁)? A. North B. East C. West D. South	D
	S_navi_detail_dest	Infer the destination from a detailed step-by-step navigation route.	Starting from {SI-SID ₀ } (<i>lon</i> ₀ , <i>lat</i> ₀), follow the detailed navigation: 1. drive west for 55 m and turn right; 2. drive north for 21 m and turn right onto the main road; ...; 20. along the pedestrian street for 119 m to the destination (37.2 km, 68 min in total). Which POI is the most likely destination? A. {SI-SID ₁ } B. {SI-SID ₂ } C. {SI-SID ₃ } D. {SI-SID ₄ }	B
Open	S_navi_via_poi	Judge which stopover POI is more on the way of a trip.	I depart from Zixing Tobacco Store for Gongyuan Pipeline and want to stop by either the waste recycling station (22.6 km, 43 min) or the Linjiang Road Ferry Terminal (23.7 km, 61 min). Which stopover is more on the way? A. via the recycling station B. via the ferry terminal C. equally convenient D. cannot be determined	A
	S_navi_simple_dest	Infer the destination from a road-level navigation route.	Starting from {SI-SID ₀ } (<i>lon</i> ₀ , <i>lat</i> ₀), drive along Longqi Rd; Longteng Ave; Xumei Rd; Humin Interchange; G60 Expressway exit; Jiamin Elevated Rd; ...; Pedestrian St (37.2 km, 68 min in total). Which POI is the most likely destination? A. {SI-SID ₁ } B. {SI-SID ₂ } C. {SI-SID ₃ } D. {SI-SID ₄ }	B
	B_poi_area_no_coord	State the district a POI belongs to, without coordinates.	Which district of Shanghai is {SI-SID ₀ } located in?	Jiading
	B_poi_area_coord	State the district a POI belongs to, with coordinates.	Which district of Shanghai is {SI-SID ₀ } (<i>lon</i> ₀ , <i>lat</i> ₀) located in?	Jiading
	R_poi_dir_no_coord	State the direction between two POIs, without coordinates.	In which direction is {SI-SID ₀ } relative to {SI-SID ₁ }?	East
	R_poi_dir_coord	State the direction between two POIs, with coordinates.	In which direction is {SI-SID ₀ } (<i>lon</i> ₀ , <i>lat</i> ₀) relative to {SI-SID ₁ } (<i>lon</i> ₁ , <i>lat</i> ₁)?	East
	R_poi_dis_no_coord	Estimate the distance between two POIs, without coordinates.	What is the approximate distance between {SI-SID ₀ } and {SI-SID ₁ }?	37.8 km
	R_poi_dis_coord	Estimate the distance between two POIs, with coordinates.	What is the approximate distance between {SI-SID ₀ } (<i>lon</i> ₀ , <i>lat</i> ₀) and {SI-SID ₁ } (<i>lon</i> ₁ , <i>lat</i> ₁)?	37.8 km