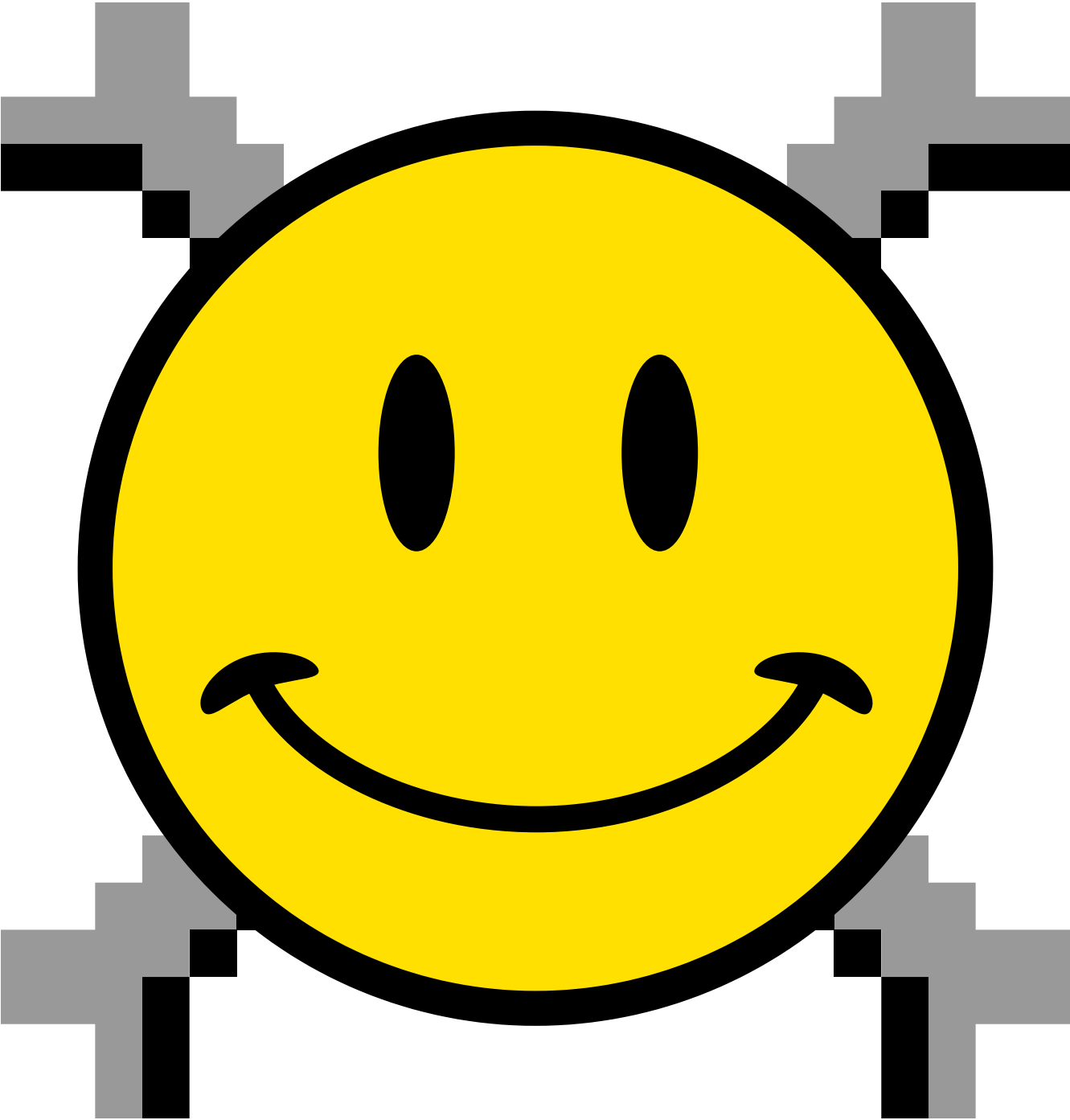




arXiv is now an independent nonprofit! [Learn more \(https://info.arxiv.org/about\)](https://info.arxiv.org/about) ×



Why HTML?
(https://info.arxiv.org/about/accessible_HTML.html)



Report Issue



Back to Abstract



Download PDF







License: arXiv.org perpetual non-exclusive license (<https://info.arxiv.org/help/license/index.html#licenses-available>)

arXiv:2608.30479v1 [cs.LR] 31 Aug 2026

HF-SID: High-Fidelity Semantic IDs for Generative Retrieval in Location-Based Services

†

Haowen Lin

These authors contributed equally to this research.

linhaowen.lhw@taobao.com (<mailto:linhaowen.lhw@taobao.com>)

AMAP, Alibaba Group, Beijing, China

Jing Li

lijingsgy@mail.ustc.edu.cn (<mailto:lijingsgy@mail.ustc.edu.cn>)

University of Science and Technology of
China, Hefei, China

Zhibin Hao

hzb24@mails.tsinghua.edu.cn (<mailto:hzb24@mails.tsinghua.edu.cn>)

Tsinghua University, Beijing, China

Fangye Wang

wangfangye.wfy@alibaba-inc.com (<mailto:wangfangye.wfy@alibaba-inc.com>)

AMAP, Alibaba Group, Beijing, China

Lihui Su

slh@pku.edu.cn (<mailto:slh@pku.edu.cn>)

AMAP, Alibaba Group, Beijing, China

Song Yang

song.yang@alibaba-inc.com (<mailto:song.yang@alibaba-inc.com>)

AMAP, Alibaba Group, Beijing, China

Xiaojiang Zhou

Corresponding author.

zhouxiaojiang.zxj@taobao.com (<mailto:zhouxiaojiang.zxj@taobao.com>)

Pengjie Wang

pengjie.wpj@alibaba-inc.com (<mailto:pengjie.wpj@alibaba-inc.com>)

AMAP, Alibaba Group, Beijing, China

AMAP, Alibaba Group, Beijing, China

2025

Abstract.

Generative retrieval has attracted increasing attention in Location-Based Services (LBS), where each Point-of-Interest (POI) is represented as a Semantic ID (SID). As the SID is the only channel through which POI information reaches the generative model, whatever it fails to preserve is irrecoverable at decoding time, and LBS retrieval is especially sensitive to the fine-grained differences that existing SIDs blur. Specifically, (1) LLMs embed continuous coordinates discontinuously, so their numeric differences do not reflect true geographic distance; (2) dynamic numerical attributes differ vastly in scale, so an identical gap may be decisive for one attribute yet negligible for another; and (3) short text cannot convey hierarchical affiliation, as text-similar POIs may belong to different hierarchies. We therefore propose HF-SID, which restores geographic, numerical, and structural fidelity at the representation stage, before any information is committed to a discrete code. It transforms coordinates into a continuous 3D Cartesian form and encodes each numerical value as a single unit, consolidated inside the LLM by Geo-CPT and Num-CPT with type-aware embeddings; a Structure-based Contrastive Learning objective, applied only to the last-layer residual, then separates co-located POIs that share a coarse tag but differ at the fine level. Because these mechanisms enrich the representation rather than lengthen the identifier, HF-SID uses a 3-token SID at no extra decoding cost. On a large-scale industrial dataset it reduces the average intra-SID geographic distance by 95.9% and improves Hit@200 by 3.66 points over the strongest baseline, with online A/B gains of +6.74% PV_CVR and +6.03% UV_CVR. We further release **AMap-S**, a large-scale real-world POI dataset.

Keywords: : Semantic ID, Generative Retrieval, Location-Based Services, Large Language Model

1. Introduction

Large language models (LLMs) ([Naveed et al., 2025](#); [Minaee et al., 2024](#)) have achieved remarkable success across many tasks, driving the emergence of Generative Retrieval (GR) ([Li et al., 2025b](#); [Pang et al., 2025](#); [Chen et al., 2025](#)). A central challenge in GR is transforming diverse item attributes into Semantic IDs (SIDs) ([Rajput et al., 2023](#); [Deng et al., 2025](#); [Zhou et al., 2025](#)), since the SID is the only channel through which item information reaches the generative model: whatever it fails to preserve is permanently invisible at decoding time. This bottleneck is acute in Location-Based Services (LBS) ([Zhang et al., 2025a](#); [Wang et al., 2026](#); [Lin et al., 2025](#)), where retrieval outcomes hinge on fine-grained POI attributes that mainstream LLMs cannot perceive precisely: they fragment coordinates into digit sub-tokens ([Liu et al., 2026](#)), are insensitive to the scale of numerical attributes ([Li et al., 2025a](#); [Ni, 2026](#)), and infer structure only from surface text ([Lv et al., 2026](#)). The central question for LBS generative retrieval is therefore not how to *organize* SIDs, but how to make them *high-fidelity*. Existing POI-oriented SID methods pursue this goal but intervene too late. They append discrete spatial tokens such as Geohash or S2 Cell IDs as geographic prefixes ([Chen et al., 2026](#); [Lv et al., 2026](#)), inject geographic offsets during codebook quantization ([He et al., 2026](#)), or fine-tune the embedding model with co-visit contrastive signals ([Wang et al., 2026](#)). All of them operate on representations whose precision is already lost, since it vanishes the moment the tokenizer splits a coordinate into digit fragments; no downstream grid, rotation, or clustering can recover a value that was never continuous. Moreover, they address geography alone, leaving numerical attributes and structural information to plain text, and pay for their gains with longer identifiers that increase autoregressive decoding cost. Three challenges must therefore be resolved before the identifier is formed.

C1: Geographic coordinates. In LBS retrieval a difference of merely 0.1° , roughly 10 km, already carries substantial business impact, yet LLMs mishandle coordinates in three ways (Figure 1(a)). Their numeric encoding is discontinuous ([Jin et al., 2026](#); [Yang et al., 2025b](#); [Li et al., 2025a](#)), so numerically close coordinates may land in distant embedding regions. Discrete grid encodings such as GeoHash ([Liu et al., 2014](#)) and S2 Cell ([Veach et al., 2017](#)), which some methods directly adopt as SID prefixes ([Chen et al., 2026](#); [Lv et al., 2026](#)), impose hard

cell boundaries where two geographically adjacent POIs can be assigned entirely unrelated codes simply because a boundary falls between them. And coordinates are angular: the same longitude difference spans about 111 km at the equator but far less at high latitudes.

C2: Dynamic numerical attributes. POIs also carry ratings, average prices, favorite counts and so on, whose scales differ by orders of magnitude, so an identical gap means entirely different things (Figure 1(b)): 0.5 in rating (range 0–5) separates a well-regarded POI from a mediocre one, whereas 0.5 in visit count (range 0–120K) is indistinguishable from noise ([Guo et al., 2024](#)). Prior work ([He et al., 2026](#); [Jiang et al., 2025](#); [Wang et al., 2026](#)) serializes them as plain text without type awareness, forcing one shared numeric subspace to absorb all of them.

C3 Hierarchical structure. Each POI carries a system-annotated two-level tag such as “Shopping/Consumer Electronics”, and sits inside a geographic unit such as a business district ([Gao et al., 2022](#)), neither of which is recoverable from surface text. Tags may agree at the coarse level yet diverge at the fine level: “Apple Store” and “Apple Market” are textually almost identical and share the coarse tag “Shopping”, yet belong to “Consumer Electronics” and “Super Market” (Figure 1(c)). A business district likewise hosts many heterogeneous shops that proximity cannot separate, and stronger geographic modeling only aggravates this: the more tightly co-located POIs are grouped, the less distinguishable they become.

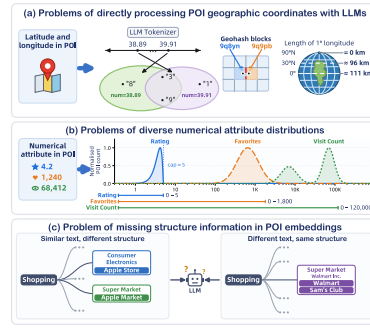


Figure 1. Three key challenges in encoding POI information into Semantic IDs: geographic coordinates, dynamic numerical attributes, and hierarchical structure.

To address these challenges, we propose HF-SID, which restores fidelity at the representation stage, before any information is committed to a discrete code. For C1, we encode each coordinate as a single continuous unit through a dedicated numerical encoder ([Jin et al., 2026](#)) and transform spherical coordinates into a 3D Cartesian form, which replaces hard grid boundaries with a continuous space in which Euclidean distance is monotonic in true geographic distance; Geo-CPT then consolidates this geometry inside the LLM through pairwise-distance and nearest-pair reasoning samples. For C2, we give each attribute type an independent Type Embedding that projects it into its own subspace, and Num-CPT trains the model on cross-attribute comparison samples so that a magnitude is always read relative to the attribute it belongs to. For C3, Structure-based Contrastive Learning treats two POIs as positives only when both tag levels match exactly, making coarse-level-only matches hard negatives. We apply it solely to the residual before the last quantization layer, since the first two layers have by then collapsed co-located POIs together while the fidelity they established remains untouched. As these mechanisms enrich the representation rather than append tokens to the identifier, HF-SID achieves this with a 3-token SID, adding no decoding step over the plain quantization baseline. It cuts the average intra-SID geographic distance by 95.9% and lifts Hit@200 by 3.66 points on a large-scale industrial dataset, with consistent online gains after deployment.

In summary, our main contributions are as follows:

- We identify that existing POI SID methods lose geographic, numerical, and structural fidelity *before* quantization, and propose HF-SID, which restores all three at the representation level while keeping the SID length unchanged.
- We introduce a unified numerical representation that encodes coordinates and dynamic attributes as single continuous units, combining a 3D Cartesian transformation with a numerical encoder and per-type type embeddings, and align it with the LLM through two-stage CPT, namely Geo-CPT for spatial reasoning and Num-CPT for scale-aware attribute comparison.
- We design Structure-based Contrastive Learning, which separates POIs that are co-located or share only a coarse tag but differ at the fine-grained level, and apply it only to the residual of the last quantization layer so that this refinement does not disturb the geographic signal encoded at coarser levels.
- We open-source AMap-S, a large-scale real-world dataset of POIs and trajectories, and validate HF-SID with large-scale offline experiments and a production deployment on AMap.

2. Related Work

Generative Retrieval reformulates item retrieval as a sequence-to-sequence task in which a language model directly generates structured item identifiers ([Tay et al., 2022](#); [Wang et al., 2023](#)). Its central design choice is the Semantic ID (SID), a short discrete code that must be both generatable by the LLM and semantically faithful to the item. SIDs are commonly obtained by Residual-Quantized Variational AutoEncoders or hierarchical K-Means over item embeddings ([Rajput et al., 2023](#); [Zheng et al., 2024](#); [Luo et al., 2024](#)), yielding a coarse-to-fine codebook sequence whose prefixes are shared by similar items, which lets constrained beam search traverse a billion-scale item space and underpins industrial deployments ([Deng et al., 2025](#); [Zhou et al., 2025](#)). A direct consequence is that whatever an item embedding fails to encode is irreversibly discarded once quantization is applied, so embedding fidelity, rather than codebook design, is the decisive factor. POI scenarios break down precisely here:

text-driven SIDs assign nearly identical codes to semantically similar but spatially distant POIs, while adjacent POIs of different categories share no prefix at all, causing geographical hallucination during generation.

Attribute Enhanced Semantic ID. To inject attribute awareness into POI-oriented SIDs, existing work follows three routes. *Discrete spatial tokenization* prepends quantized location codes to the semantic code, using S2 Cell encodings as hierarchical geographic prefixes ([Lv et al., 2026](#)), Geohash-based Geographic IDs under trie-constrained decoding ([Chen et al., 2026](#)), or hierarchical geographic tokenization with residual quantization ([Jiang et al., 2025](#)). *Token-level continuous injection* instead perturbs embeddings, through the Geographic Position Embedding of GenPOI ([Chen et al., 2026](#)) and the geo-centroid Geo-RoPE frame of Pro-GEO ([He et al., 2026](#)). *Behavior-driven alignment* reshapes the embedding space with log-mined supervision, contrasting spatiotemporal co-visit pairs with EM-style ([Dempster et al., 1977](#)) SID refinement ([Wang et al., 2026](#)), or adding a diversity loss to RQ-VAE ([Lee et al., 2022](#)) against semantic collapse ([Wang et al., 2025](#)). All of them act after coordinates have already been serialized as text, and all treat geography as the only structured signal worth modeling. Numerical attributes fare no better: sub-word tokenization destroys numeric continuity and leaves LLMs unreliable at magnitude and comparison ([Yang et al., 2025b](#); [Li et al., 2025a](#); [Ni, 2026](#)), and although dedicated encoders that embed a scalar as a single continuous unit largely restore this ability ([Jin et al., 2026](#)), they have been evaluated on reasoning tasks rather than identifier construction. Closest to us, ReSID ([Liang et al., 2026](#)) and HiD-VAE ([Fang et al., 2025](#)) quantize structured recommendation fields and multi-granularity semantics, yet both target general recommendation and neither confronts the heterogeneous attribute scales and spatial hierarchy of POIs.

3. Methods

Given a POI p with its textual description, HF-SID encodes it into a Semantic ID SID_p , guided by a single principle: restore geographic, numerical, and structural fidelity at the representation level, before quantization commits the POI to a discrete code. As illustrated in Figure 2, the framework consists of four stages. (1) *Coordinate and Numerical Representation* (§3.1): coordinates are converted from

spherical to 3D Cartesian form, and every numerical value in the description, coordinates included, is decomposed into a polar representation of its sign, per-digit magnitude, and an angular encoding of its fractional part, with these positions marked as numerical tokens. (2) *Numerical Encoder* (§3.2): a dedicated encoder maps each polar representation to a continuous embedding, pretrained to reconstruct the sign, integer digits, and fractional part through three decoder heads, and then frozen and reused by all later stages. (3) *Integrating the Numerical Encoder into the LLM* (§3.3): each numerical embedding is projected into the LLM input space, offset by a Type Embedding that gives every attribute type its own subspace, and substituted at its original position to form a unified sequence together with text token embeddings; the LLM is then continually pre-trained by Geo-CPT and Num-CPT under a hybrid objective that applies cross-entropy at text positions and the numerical multi-task loss at numerical positions, and the hidden state of its last token becomes the POI representation. (4) *HF-SID Generation with Structure-based Contrastive Learning* (§3.4): the POI representations are hierarchically quantized, where the first two layers cluster on geographic and numerical semantics and therefore give nearby POIs a shared prefix, while Structure-based Contrastive Learning refines the second-order residual before the third layer, so that POIs collapsed together by the coarser levels are separated by their tags in the final SID token.

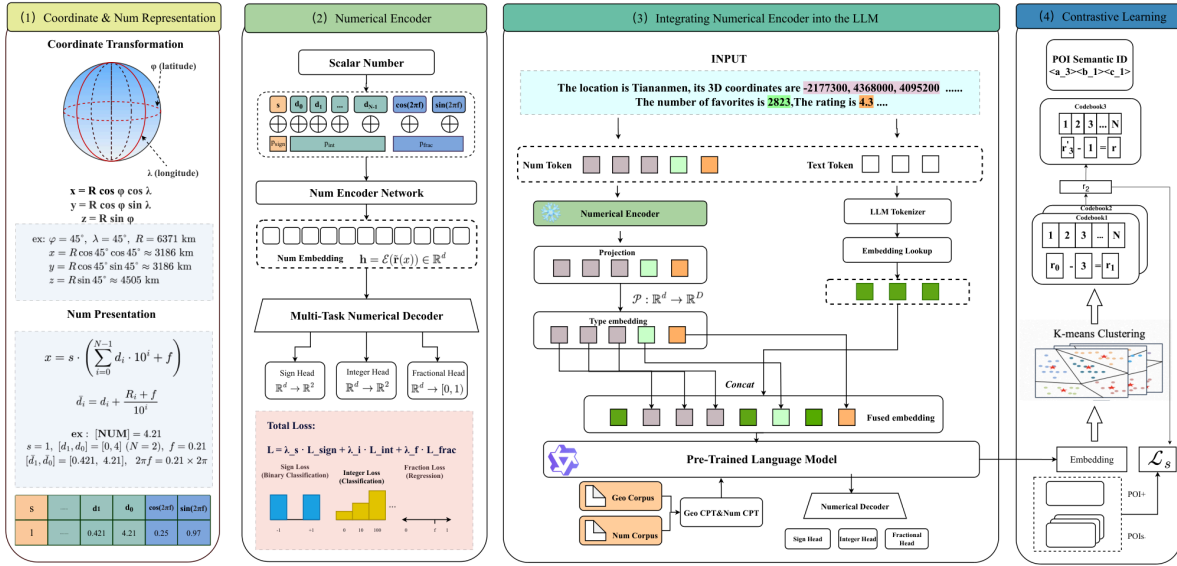


Figure 2. Overview of the HF-SID framework.

3.1. Coordinate and Numerical Representation

POIs carry two categories of numerical information: geographic coordinates and dynamic numerical attributes. Standard LLM tokenizers fragment both into sub-tokens but destroy numerical continuity. And coordinates suffer from an additional problem, as raw latitude and longitude differences do not scale linearly with true distance. We therefore introduce two representations: a Cartesian transformation applied to the coordinates, and a unified polar decomposition applied to every numerical value, coordinates included.

3.1.1. Coordinate Transformation

We convert geographic coordinates from spherical to Cartesian form before they are embedded:

$$(1) \quad x = R \cos \phi \cos \lambda, \quad y = R \cos \phi \sin \lambda, \quad z = R \sin \phi$$

where ϕ and λ denote latitude and longitude, and R is the mean Earth radius, so that $x, y, z \in \mathbb{R}$ are expressed in meters. Two properties follow. The representation is continuous across the antimeridian, since $\lambda = +179^\circ$ and $\lambda = -179^\circ$ now

yield adjacent points. And the Euclidean distance between two POIs becomes $2R \sin(d / 2R)$ for a great-circle distance d , a strictly monotonic and, at city scale, nearly linear function of true distance, which is far easier to model than raw latitude–longitude differences.

3.1.2. Polar Representation of Numerical Values

For every numerical value in the POI text, coordinates and dynamic attributes alike, we adopt a unified decomposition that treats each scalar as a single unit rather than a digit string. We call it polar because the fractional part is represented by its angle on the unit circle.

Formally, any scalar x is decomposed into **a sign, N integer digits, and a fractional part**:

$$(2) \quad x = s \cdot \left(\sum_{i=0}^{N-1} d_i \cdot 10^i + f \right)$$

where $s \in \{-1, +1\}$, $d_i \in \{0, 1, \dots, 9\}$ is the i -th integer digit, and $f \in [0, 1)$ is the fractional part. We set N to 9, which covers the metric coordinates produced by Eq. (1) as well as all dynamic attributes, and pad shorter values with leading zeros.

Instead of feeding the raw digits d_i , we augment each integer position with the magnitude that follows it:

$$(3) \quad \bar{d}_i = (|x| / 10^i) \bmod 10$$

so that $\bar{d}_i \in [d_i, d_i + 1)$ carries both the digit itself and a fractional summary of all lower positions. This makes integer carries continuous: $399 \rightarrow 400$ moves $\bar{d}_2$ smoothly from 3.99 to 4.00 instead of changing three digits at once. Symmetrically, encoding f as $(\cos 2\pi f, \sin 2\pi f)$ makes the fractional carry continuous, since $f = 0.99$ and $f = 0.01$ are adjacent on the unit circle.

Concatenating all components yields the polar vector

$$(4) \quad \mathbf{r}(x) = [s, \bar{d}_0, \bar{d}_1, \dots, \bar{d}_{N-1}, \cos 2\pi f, \sin 2\pi f]^\top \in \mathbb{R}^{N+3}.$$

Since the entries of $\mathbf{r}(x)$ play different semantic roles, we lift each of them into a d_e -dim vector and add a learnable field embedding:

$$(5) \quad \tilde{\mathbf{r}}_j(x) = r_j(x) \mathbf{w} + \mathbf{b} + \mathbf{e}_{\varphi(j)} \in \mathbb{R}^{d_e}, \quad j = 0, \dots, N+2$$

where $r_j(x)$ is the j -th entry of $\mathbf{r}(x)$, $\mathbf{w}, \mathbf{b} \in \mathbb{R}^{d_e}$ are shared learnable parameters, and $\varphi(j) \in \{\text{sign}, \text{int}, \text{frac}\}$ returns the field type of position j . All integer positions share a single embedding $\mathbf{e}_{\text{int}}$, because their order is already encoded in $\tilde{d}_j$. We write $\tilde{\mathbf{r}}(x) = \text{vec}([\tilde{\mathbf{r}}_0(x), \dots, \tilde{\mathbf{r}}_{N+2}(x)]) \in \mathbb{R}^{(N+3)d_e}$ for the resulting representation. Together, the two carry-smoothing encodings and the field embeddings ensure that numerically proximate values are mapped to geometrically adjacent points in the input space.

3.2. Numerical Encoder

We pass the field-augmented polar representation $\tilde{\mathbf{r}}(x)$ through the Numerical Encoder $\mathcal{E}(\cdot)$, a multi-layer perceptron (MLP), to obtain a numerical embedding:

$$(6) \quad \mathbf{h} = \mathcal{E}(\tilde{\mathbf{r}}(x)) \in \mathbb{R}^d$$

where d is the encoder embedding dimension. Note that d is far larger than the dimension of $\tilde{\mathbf{r}}(x)$, so $\mathcal{E}(\cdot)$ lifts a single scalar into a high-dimensional space rather than compressing it.

To make this space numerically structured, we pretrain $\mathcal{E}(\cdot)$ with a multi-task objective requiring every component of the decomposition to be recoverable from $\mathbf{h}$ alone. The sign and the N integer positions are predicted by softmax classifiers,

$$(7) \quad \mathbf{p}_{\text{sign}} = \text{softmax}(\mathbf{W}_{\text{sign}} \mathbf{h} + \mathbf{b}_{\text{sign}}) \in \mathbb{R}^2,$$

$$(8) \quad \mathbf{p}_{\text{int}}^{(i)} = \text{softmax}(\mathbf{W}_{\text{int}}^{(i)} \mathbf{h} + \mathbf{b}_{\text{int}}^{(i)}) \in \mathbb{R}^{10},$$

for $i = 0, 1, \dots, N-1$, while the fractional part is predicted by regression through a sigmoid-activated linear layer:

$$(9) \quad \hat{f}(\mathbf{h}) = \sigma(\mathbf{W}_f \mathbf{h} + \mathbf{b}_f) \in (0, 1)$$

This objective is not a trivial reconstruction. Since $\mathcal{E}(\cdot)$ increases dimensionality, no information is at risk of being discarded; what the objective constrains is *how* that information is arranged. A single $\mathbf{h}$ must simultaneously support N independent 10-way classifications, one per digit position, together with a sign classification and a fractional regression, which forces the digit-wise magnitude structure of x to be linearly decodable from separate directions of the embedding. This is also what later allows the same three heads to be attached to the LLM output (§3.3), since they define what counts as a valid numerical representation.

For a single scalar x , the pretraining loss combines three heads:

$$(10) \quad \ell_{num}(x) = \lambda_s \mathcal{L}_{sign} + \lambda_i \sum_{i=0}^{N-1} \omega_i \mathcal{L}_{int}^{(i)} + \lambda_f \mathcal{L}_{frac}$$

where $\mathcal{L}_{sign}$ and $\mathcal{L}_{int}^{(i)}$ are cross-entropy losses for the sign and the i -th integer digit, $\mathcal{L}_{frac}$ is the mean squared error of the fractional regression, and $\lambda_s, \lambda_i, \lambda_f \geq 0$ balance the three terms. The positional weight $\omega_i = 1 + 0.2i$ mildly favours higher-order digits, whose errors shift the value by larger amounts, while keeping the low-order digits supervised; a steeper weighting would let the model ignore the trailing digits, which for attributes such as ratings carry most of the discriminative signal.

After pretraining, $\mathcal{E}(\cdot)$ is frozen and reused across all downstream stages of HF-SID, providing one consistent embedding space for geographic coordinates and heterogeneous numerical attributes alike.

3.3. Integrating Numerical Encoder into the LLM

3.3.1. Type Embedding

Since the Numerical Encoder is frozen, we bridge it to the LLM with two light-weight trainable components. A linear projection first aligns its output with the LLM input space:

$$(11) \quad \mathbf{h}_k^{proj} = \mathbf{W}_{proj} \mathbf{h}_k + \mathbf{b}_{proj}$$

where $\mathbf{h}_k$ is the encoder output for the k -th numerical token, $\mathbf{W}_{proj} \in \mathbb{R}^{d_{llm} \times d}$ and $\mathbf{b}_{proj} \in \mathbb{R}^{d_{llm}}$ are learnable parameters, and d_{llm} is the input embedding dimension of the LLM.

A single shared projection, however, maps every attribute into one common numeric subspace, in which a value of 0.5 is indistinguishable whether it denotes a rating or a visit count. We therefore add a learnable type embedding, drawn from a matrix $\mathbf{M}_{\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}| \times d_{llm}}$ indexed by a finite set of attribute types $\mathcal{T}$, to each numerical embedding according to its own type $\tau_k \in \mathcal{T}$:

$$(12) \quad \hat{\mathbf{h}}_k = \mathbf{h}_k^{proj} + \mathbf{M}_{\mathcal{T}}[\tau_k]$$

Because the offset is constant within a type, magnitude relations among values of the same attribute are preserved exactly, while different attributes are translated into separate regions of the input space. This is exactly the behavior we want: the model can compare two ratings on a common scale, yet it can no longer confuse a rating with a visit count. The mechanism is analogous to positional embeddings in Transformers ([Vaswani et al., 2023](#)), which likewise inject a prior about the role of a position without modifying its content.

3.3.2. Fused Embedding Construction

When a POI is serialized, every numerical value is replaced by a single placeholder token **[NUM]**, so that one scalar occupies exactly one position irrespective of how many digits it has; as a side effect this also shortens the input sequence, since a multi-digit number no longer expands into several sub-tokens. We then substitute the type-aware numerical embeddings at these positions and keep ordinary token embeddings elsewhere, before feeding the sequence to the pre-trained LLM:

$$(13) \quad \mathbf{E}_{fused}[k] = \begin{cases} \hat{\mathbf{h}}_k & \text{if } k \in \mathcal{J}_{num} \\ \text{Emb}(w_k) & \text{otherwise} \end{cases} \quad \mathbf{E}_{out} = \text{LLM}(\mathbf{E}_{fused})$$

where $\mathcal{J}_{num}$ denotes the **[NUM]** positions of the input sequence, $\text{Emb}(\cdot)$ is the LLM token embedding lookup, w_k is the k -th text token, and $\mathbf{E}_{fused}, \mathbf{E}_{out} \in \mathbb{R}^{T \times d_{llm}}$ with T the total sequence length. Following common decoder-based embedding models ([Zhang et al., 2025b](#)), we take the hidden state of the last token as the

POI representation $\mathbf{E}_{out}(p) \in \mathbb{R}^{d_{lm}}$, serving as the input to HF-SID generation (§3.4).

3.3.3. Two-Stage Continual Pre-Training

Training Data Construction.

We construct two categories of training samples, one for geographic reasoning and one for numerical reasoning, to explicitly supervise the LLM on both dimensions. In every sample, all numerical quantities appear as **[NUM]** tokens and are encoded by the frozen Numerical Encoder, both in the input and in the target sequence.

Geographic-enhanced Continual Pre-Training (Geo-CPT). To train the model to reason about spatial relationships between POIs, we construct two types of geographic samples: **Pairwise Distance Computation.** Given two POIs p_A and p_B , whose Cartesian coordinates (x_A, y_A, z_A) and (x_B, y_B, z_B) are supplied as numerical tokens, the model is asked to predict the distance between them at a single **[NUM]** position of the target sequence. The ground truth is the great-circle distance $d(p_A, p_B)$ in kilometers, computed by the Haversine formula. Note that the quantity directly available from the Cartesian inputs is the chord length $2R \sin(d / 2R)$ rather than d itself; the two agree to within $10^{-3}\%$ for POI pairs closer than 100 km, so at the scale relevant to LBS retrieval this target is effectively the Euclidean distance in the transformed space, while remaining the geographically meaningful quantity. **Nearest Pair Identification.** Given three POIs p_A , p_B , and p_C , the model is asked to identify the closest pair among (p_A, p_B) , (p_A, p_C) , and (p_B, p_C) , and to emit it as text tokens. This requires jointly reasoning over three pairwise distances rather than computing a single one.

Numerical-enhanced Continual Pre-Training (Num-CPT). We construct samples in which two POIs p_A and p_B are described with multiple numerical attributes, and the model is asked which of them has the larger value of a queried attribute τ . The target is emitted as text tokens, and we define it over the two candidate labels $\{A, B\}$:

$$(14) \quad y^{(\tau)} = \begin{cases} A & \text{if } v_A^{(\tau)} > v_B^{(\tau)} \\ B & \text{if } v_A^{(\tau)} < v_B^{(\tau)} \end{cases}$$

where $v_A^{(\tau)}$ and $v_B^{(\tau)}$ are the values of the two POIs under attribute type τ . Pairs with $v_A^{(\tau)} = v_B^{(\tau)}$ are discarded, since ties carry no comparative signal and would otherwise train a spurious preference for one of the two positions; we also balance the two labels so that the answer cannot be inferred from position alone. Instantiated across attributes, this formulation forces the model to locate the value belonging to the queried attribute and to compare it against its counterpart of the same type, even though different types differ by orders of magnitude in scale.

Hybrid Training Objective

Because the target sequences contain both discrete text tokens and numerical tokens, we adopt a hybrid objective that applies a different loss at each kind of position. Let $\mathcal{J}_{text}$ and $\mathcal{J}_{num}$ denote the text and numerical positions of the target sequence, and let $\mathbf{E}_{<k}$ denote the fused prefix of position k , which may contain both text token embeddings and injected numerical embeddings. At text positions we apply the standard cross-entropy next-token prediction (NTP) loss:

$$(15) \quad \mathcal{L}_{CE} = - \frac{1}{|\mathcal{J}_{text}|} \sum_{k \in \mathcal{J}_{text}} \log P(w_k | \mathbf{E}_{<k})$$

where w_k is the ground-truth text token at position k . Conditioning on $\mathbf{E}_{<k}$ rather than on $w_{<k}$ is what allows a numerical value to act as context for the tokens that follow it.

At numerical positions we reuse the multi-task loss ℓ_{num} , applying the same three decoder heads to the LLM output representation:

$$(16) \quad \mathcal{L}_{num}^{NTP} = \frac{1}{|\mathcal{J}_{num}|} \sum_{k \in \mathcal{J}_{num}} \ell_{num}(x_k)$$

where x_k is the target value at position k , and its sign, integer, and fractional losses are computed from the LLM hidden state at that position. Both terms are averaged over their own positions so that their relative weight does not drift with the text-to-number ratio of a sample. And each of the two stages is trained with

$$(17) \quad \mathcal{L}_{NTP} = \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{num}^{NTP}$$

where $\alpha \geq 0$ balances the numerical alignment loss against next-token prediction. We run Geo-CPT first and Num-CPT second, since geographic fidelity is the dominant factor for LBS retrieval and therefore benefits from being established on the unmodified backbone.

3.4. HF-SID Generation with Structure-based Contrastive Learning

We obtain Semantic IDs by residual-quantizing the POI representations of the entire corpus with three layers of K-means. Writing $\mathbf{R}_p^{(0)} = \mathbf{E}_{out}(p) \in \mathbb{R}^{d_{lm}}$, layer l fits a codebook $\mathcal{C}^{(l)} = \{\mathbf{c}_1^{(l)}, \dots, \mathbf{c}_{N_c}^{(l)}\}$ of size N_c over the l -th order residuals of all POIs, assigns each POI to its nearest centroid, and forwards the remaining residual:

$$(18) \quad s_p^l = \arg \min_j \|\mathbf{R}_p^{(l-1)} - \mathbf{c}_j^{(l)}\|_2, \quad \mathbf{R}_p^{(l)} = \mathbf{R}_p^{(l-1)} - \mathbf{c}_{s_p^l}^{(l)},$$

where $\mathbf{E}_{out}(p)$ is the last-token embedding of the LLM and every layer uses the same codebook size N_c . The first two layers are applied in standard form and yield the coarse tokens s_p^1 and s_p^2 , which reflect the geographic and numerical semantics.

Precisely because these two layers are geographically compact, POIs located in the same business district are largely collapsed into a shared prefix $\langle s_p^1, s_p^2 \rangle$ irrespective of what they actually are. We therefore refine the second-order residual $\mathbf{R}_p^{(2)}$ before the final quantization, supervised by the two-level tag of each POI. Two POIs form a positive pair only when both levels of their tags coincide, and every other POI in the mini-batch serves as a negative, so that POIs agreeing on the coarse level alone act as hard negatives. The residual is linearly projected,

$$(19) \quad \mathbf{z}_p = \mathbf{W}_s \mathbf{R}_p^{(2)} + \mathbf{b}_s$$

and optimized with an InfoNCE objective over inner products:

$$(20) \quad \mathcal{L}_s(p_i, p_j) = -\log \frac{\exp(\mathbf{z}_{p_i}^\top \mathbf{z}_{p_j} / \gamma)}{\exp(\mathbf{z}_{p_i}^\top \mathbf{z}_{p_j} / \gamma) + \sum_{p_k \in \mathcal{B}^-} \exp(\mathbf{z}_{p_i}^\top \mathbf{z}_{p_k} / \gamma)}$$

where $\mathbf{W}_s \in \mathbb{R}^{d_z \times d_{llm}}$ and $\mathbf{b}_s \in \mathbb{R}^{d_z}$ are learnable projection parameters with output dimension d_z and the only parameters updated at this stage, $\mathcal{B}^-$ collects the in-batch POIs whose tag differs from that of p_i at either level, $\gamma > 0$ is a temperature hyperparameter, and the loss is averaged over all positive pairs in the batch. Since the LLM and the first two codebooks stay fixed, $\mathbf{R}_p^{(2)}$ is a constant input and the refinement never invalidates the coarse tokens.

The refined residual is then quantized by the third K-means layer into s_p^3 , so that POIs sharing a fine-grained tag tend to fall into the same third-layer cluster, and each POI is finally represented as

$$(21) \quad \mathbf{HF-SID}_p = \langle s_p^1, s_p^2, s_p^3 \rangle$$

Note that this last layer operates in the projected space $\mathbb{R}^{d_z}$ rather than in $\mathbb{R}^{d_{llm}}$, and therefore no longer reconstructs $\mathbf{R}_p^{(2)}$ in the strict residual-quantization sense. This is deliberate: an SID serves only as an identifier to be generated autoregressively and is never used to reconstruct the POI representation, so the final layer can trade reconstruction error for discriminative power. Restricting the projection to the last layer is what keeps the geographic and numerical fidelity of $\langle s_p^1, s_p^2 \rangle$ intact, as the analysis in §4.4 confirms.

The resulting sequence retains the prefix-sharing property of residual quantization: POIs with similar representations agree on their leading tokens, while the last token separates them by fine-grained tag. Candidate generation can thus be performed by beam search constrained to the prefix tree of valid SIDs.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate on two industrial datasets from the AMap platform¹ [<https://amap.com> (<https://amap.com>)] and two public Foursquare check-in datasets (NYC, TKY). **AMap-L** is built from one full day of anonymized user interaction logs. For data quality, we remove POIs with no associated interaction sequence and exclude sessions without authenticated user identifiers. **AMap-S** is a publicly released subset from the same platform: smaller in scale but retaining the same attributes, it can serve as a standardized benchmark for future work. Statistics are summarized in Table 1.

Baselines. We compare our HF-SID against the following SID construction methods. RQ-KMeans ([Luo et al., 2024](#)), RQ-OPQ ([Chen et al., 2025](#)), GNPR-SID ([Wang et al., 2025](#)), Cosine-RQ ([Zhou et al., 2026](#)), Pro-GEO ([He et al., 2026](#)), GenPOI ([Chen et al., 2026](#)), GeoGR ([Wang et al., 2026](#)). Detailed information can be found in the given references. For fair comparison, all baselines are re-implemented under the same configuration as HF-SID: codebook depth $L = 3$, per-layer codebook size $N_c = 4096$ on AMap-L and $N_c = 512$ on AMap-S, and identical dataset splits.

Evaluation Metrics. We evaluate both the quality of HF-SID and downstream performance through Next POI recommendation task ([He et al., 2026](#); [Wang et al., 2025](#)). **(1) Quantization Metrics of SID.** Follow prior work ([He et al., 2026](#)), we use 4 metrics: ICR, Avg.Dist, p90 Dist, p95 Dist. ICR (Independent Codeword Rate, $\uparrow$) measures the degree to which codewords are assigned uniquely across items. Avg.Dist ($\downarrow$) computes the mean Euclidean distance from each POI’s position to the centroid of its quantification SID, providing a measure of intra-group spatial compactness. Additionally, we report p90 Dist and p95 Dist, representing the maximum distance to the SID centroid within the top 90% and 95% of POIs. **(2) Offline Recommendation Metrics.** Hit@200 and Hit@300 ($\uparrow$) measure the proportion of ground-truth target POIs appearing within the top-200 and top-300 generated candidates, respectively, jointly reflecting retrieval precision and recall coverage. **(3) Online Recommendation**

Metrics. We evaluate online performance using five metrics: WinRate ($\uparrow$), a composite metric that quantifies overall user engagement gains by weighting strategic actions such as clicks, route navigation, and bookings according to their value, reflecting holistic improvements in both user satisfaction and platform utility. PV_CTR and UV_CTR measure click-through rates at the exposure and user levels, respectively. PV_CVR and UV_CVR measure conversion rates at the exposure and user levels, respectively.

Table 1. Dataset statistics.

Dataset	#Users	#POIs	#Inter.	Avg.Len
NYC	1,083	5,135	104,074	136
TKY	2,293	7,873	361,430	195
AMap-L	11.0M	48.1M	16.3M	305
AMap-S	0.90M	4.58M	0.97M	453

Table 2. Performance comparison on AMap-L and AMap-S datasets. **Bold** denotes the best result within each dataset.

Dataset	Method	SID Length	Quantization	
			ICR($\uparrow$)	Avg.Dist. ($\downarrow$)
AMap-L	RQ-KMeans (Luo et al., 2024)	3	80.92%	6.07 km
	RQ-OPQ (Chen et al., 2025)	3	98.93%	7.27 km
	GNPR-SID (Wang et al., 2025)	3	77.57%	10.67 km
	Cosine-RQ (Zhou et al., 2026)	3	82.63%	5.26 km
	Pro-GEO (He et al., 2026)	3	81.18%	7.02 km
	GeoGR (Wang et al., 2026)	3	54.69%	0.49 km
	GenPOI (Chen et al., 2026)	6(3+3)	83.80%	0.97 km
	HF-SID (Ours)	3	84.22%	0.25 km
AMap-S	RQ-KMeans (Luo et al., 2024)	3	61.34%	37.91 km
	RQ-OPQ (Chen et al., 2025)	3	98.71%	41.20 km
	GNPR-SID (Wang et al., 2025)	3	47.52%	77.82 km
	Cosine-RQ (Zhou et al., 2026)	3	63.68%	36.32 km
	Pro-GEO (He et al., 2026)	3	64.64%	24.25 km
	GeoGR (Wang et al., 2026)	3	37.83%	8.94 km
	GenPOI (Chen et al., 2026)	5(2+3)	62.12%	4.84 km
	GenPOI (Chen et al., 2026)	6(3+3)	65.97%	3.22 km
	HF-SID (Ours)	3	59.79%	3.99 km

Implementation Details. All experiments are implemented in PyTorch. Following prior work ([Jin et al., 2026](#)), we implement the Numerical Encoder. The embedding model’s backbone is initialized from Qwen3–0.6B ([Zhang et al., 2025b](#)) for POI representation learning and SID generation, while the downstream next–POI prediction task uses Qwen3–4B ([Yang et al., 2025a](#)) with full–parameter supervised fine–tuning (SFT). Beam search is performed with a width of 300 to generate the final list of POIs. Continual pre–training runs for 1 epoch with AdamW ([Loshchilov and Hutter, 2019](#)) at learning rate 1×10^{-3} . SFT also runs for 1 epoch with a cosine scheduler, learning rate 2×10^{-5} , per–device batch size 8, and gradient accumulation of 4 steps, using DeepSpeed ZeRO–2 ([Rajbhandari et al., 2020](#)) with bf16 mixed precision. For RQ–KMeans we set $L = 3$ layers, consistent with prior work ([Deng et al., 2025](#)). All experiments run on 64 NVIDIA A100 GPUs.

4.2. Overall Performance

We examine whether the representation–level fidelity that HF–SID enforces actually translates into better SIDs and better retrieval performance. Table 2 compares HF–SID against seven baselines covering the three ways of building POI SIDs: text–only quantization, post–hoc geographic injection, and embedding–space alignment.

HF–SID Quality. On both datasets the compactness of the SID space improves monotonically with how early geography enters the pipeline, and HF–SID, which intervenes earliest, is the most compact of all. Methods that quantize text–only embeddings stay at 5.26–10.67 km of Avg. Dist. on AMap–L and 36.32–77.82 km on AMap–S, regardless of how their codebooks are refined: orthogonal product quantization ([Chen et al., 2025](#)), a codeword diversity loss ([Wang et al., 2025](#)), and stronger text representations ([Zhou et al., 2026](#)) all leave the geographic error essentially untouched, because none of them puts geographic information into the embedding to begin with. Injecting geography afterwards helps (7.02 km for Pro–GEO, 0.97 km for the 6–token GenPOI), and reshaping the embedding space with behavioral supervision helps further (0.49 km for GeoGR), but HF–SID reaches 0.25 km and 3.99 km, cutting Avg. Dist. by 95.9% and 89.5% relative to plain RQ–KMeans and by a further 48.98% and 55.4% relative to GeoGR. Its advantage

is largest in the tail, which is the regime that produces visibly wrong candidates: p90 Dist. drops by 83.8% and 61.0% and p95 Dist. by 42.0% and 75.9% against the same baseline. Notably, on AMap-S HF-SID halves the p90 Dist. of the 6-token GenPOI (11.30 $\rightarrow$ 5.49 km) and reduces its p95 Dist. by 38.9% (18.19 $\rightarrow$ 11.12 km) using half the identifier length, although GenPOI retains a slightly lower mean of 3.22 km under that doubled token budget.

Two baselines further show that codeword uniqueness on its own is not a meaningful objective. RQ-OPQ attains a near-saturated ICR of 98.93% and 98.71%, yet its Avg. Dist. is worse than that of plain RQ-KMeans on both datasets and its Hit@300 is the lowest of all methods on both (55.31% and 26.71%): spreading items evenly over the codebook says nothing about whether the code carries usable structure. GeoGR fails in the opposite direction, buying geographic compactness with the lowest ICR of the table (54.69% and 37.83%), because contrastive fine-tuning on co-visit pairs collapses behaviorally similar POIs onto shared codewords. HF-SID avoids both failure modes because it encodes the coordinate itself rather than a proxy of it, and with an ICR of 84.22% on AMap-L it is the most collision-resistant among all methods whose Avg. Dist. stays below 1 km.

Retrieval Performance. The gains in SID quality carry over to downstream retrieval in full, and the ordering of methods by Hit mirrors their ordering by SID compactness almost exactly. All methods are fine-tuned under identical settings and differ only in their SIDs. HF-SID attains Hit@200 of 81.24% and 48.98% and Hit@300 of 83.36% and 57.92%, exceeding the strongest baseline on each dataset by 3.66 and 6.78 points on Hit@200 and by 2.94 and 11.01 points on Hit@300, the latter amounting to a 23.48% relative gain on AMap-S while using half the tokens of the 6-token GenPOI it surpasses. Across the eight AMap-L systems, Hit@200 rises from the 54–58% range for the four text-only methods, to 65.69% and 70.61% once geography is injected post hoc, to 77.58% and 81.24% once the embedding itself becomes geography-aware. That the two orderings coincide is the central evidence of this paper: the retrieval gain is not an artefact of a particular codebook design but a direct consequence of how much geographic and numerical fidelity survives into the identifier.

Table 3. Ablation study on AMap-L.

Method	Hit@200 ($\uparrow$)	Δ Hit@200	Hit@300 ($\uparrow$)	Δ Hit@300
HF-SID (Full)	81.24%	—	83.36%	—
w/o Num	77.73%	−4.32%	80.65%	−3.25%
w/o Struct	71.88%	−11.52%	78.35%	−6.01%
w/o Geo	68.34%	−15.88%	71.11%	−14.69%
w/o All (Baseline)	54.47%	−32.95%	56.02%	−32.79%

4.3. Ablation Study

We ablate HF-SID on AMap-L in a leave-one-out fashion. No variant removes any *information*: when a component is disabled, the corresponding attribute stays in the POI description as plain text. **w/o Geo** drops the Cartesian transformation, the Numerical Encoder on coordinates, and Geo-CPT; **w/o Num** drops the Numerical Encoder on dynamic attributes, the Type Embedding, and Num-CPT; **w/o Struct** drops the contrastive refinement, letting the third layer quantize $\mathbf{R}_p^{(2)}$ directly; **w/o All** reduces to plain RQ-KMeans. All variants therefore see identical content and differ only in how faithfully it is represented, so the drops below measure the value of representation fidelity rather than of information availability.

Every component contributes, and the size of its contribution follows the level of the SID hierarchy at which it acts. Removing the geographic pathway costs the most, with Hit@200 falling from 81.24% to 68.34% (−15.88%) and Hit@300 from 83.36% to 71.11% (−14.69%): geographic fidelity is established before quantization and thus shapes the first two codebook layers, that is, the entire prefix tree over which generation is constrained, so degrading it corrupts every subsequent decoding decision. Structure-based Contrastive Learning acts on only one token out of three, and its cost accordingly halves from Hit@200 (−11.52%) to Hit@300 (−6.01%), indicating that tag purity matters mainly for ranking the correct POI early rather than for reaching it at all. The numerical pathway is the smallest (−4.32% and −3.25%), as ratings and visit counts refine the ordering

among already plausible candidates instead of deciding which region of the identifier space is explored.

The entire gap to the baseline is attributable to the representation alone. With all three components disabled the model returns to plain RQ-KMeans at 54.47% and 56.02%, so HF-SID’s 26.77-point gain on Hit@200 is obtained without changing the quantization algorithm, the codebook size, or the identifier length. The same procedure that ranks last in Table 2 ranks first once the representation it quantizes is made geography- and numeracy-aware.

4.4. Visualization Analysis of HF-SID

Through detailed experiments, we explore how much geographic, numerical, and structural information HF-SID carries.

4.4.1. Visualization of Geographic Information

We sample 1,554 POIs across three administrative levels (Province, City, and District), embed them with t-SNE using the geographic label as color, and report the NMI before and after Geo-CPT. As shown in Figure 3, the baseline embeddings remain largely overlapped at every granularity, whereas HF-SID forms tight and well-separated clusters, with NMI rising from 0.1993 to 0.7454 at Province, 0.2615 to 0.9796 at City, and 0.3972 to 0.9399 at District. This embedding-space geometry is exactly what produces the SID-space compactness of Table 2: fidelity established in the representation survives into the code.

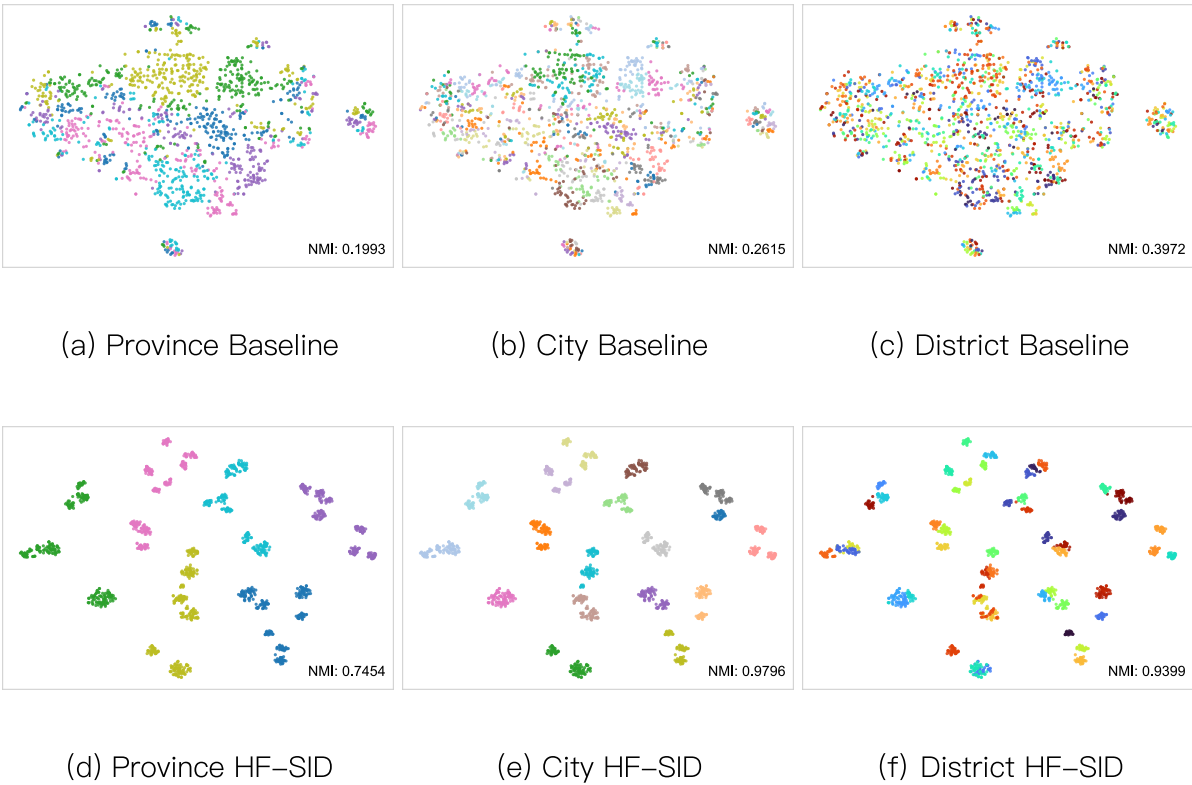


Figure 3. t-SNE visualization of POI embeddings.

Table 4. Intra-SID numerical attribute dispersion comparison across methods. **Bold** denotes the best result.

Method	Rating ↓			POI Visited ↓			Avg.St
	Avg.Std	Avg.CV	Avg.MAD	Avg.Std	Avg.CV	Avg.MAD	
GeoGR	0.2474	0.1001	0.2880	1.92×10^{11}	1.2474	1.92×10^{11}	0.612
Pro- GEO	0.1831	0.0918	0.2610	4.39×10^{11}	1.0661	4.39×10^{11}	0.489
GenPOI	0.2027	0.0947	0.2731	253.178	1.0600	236.968	0.513
RQ- KMeans	0.2138	0.0982	0.2808	222.486	1.0753	208.249	0.512
HF-SID	0.1817	0.0916	0.2608	49.955	1.0297	48.180	0.443

4.4.2. Analysis of Numerical Awareness

To assess numerical awareness, we measure the intra-HF-SID dispersion of five heterogeneous attributes: Rating, POI Visited, POI Index, Collect UV, and Week View UV. For every HF-SID cluster with more than one POI we compute the Standard Deviation (Std), Coefficient of Variation (CV), and Mean Absolute Deviation (MAD), and report their macro-averages, where lower values indicate greater numerical coherence (Table 4).

HF-SID attains the best dispersion on all five attributes, and the margin scales with how demanding the attribute is. On Rating, which is confined to a narrow 0–5 range, the improvement over the second-best baseline Pro-GEO is small but consistent (Avg.Std 0.1831 → 0.1817, Avg.CV 0.0918 → 0.0916, Avg.MAD 0.2610 → 0.2608). The advantage widens sharply on heavy-tailed attributes spanning several orders of magnitude: on POI Visited, HF-SID cuts Avg.Std by 77.5% (222.486 → 49.955) and Avg.MAD by 76.9% (208.249 → 48.180) against the second-best baseline on each metric, and also lowers Avg.CV from 1.0600 to 1.0297, showing that relative dispersion is reduced even under extreme scale. The same pattern holds on POI Index (Avg.Std −9.3%, Avg.CV −7.2%, Avg.MAD −9.5% over Pro-GEO) and on Collect UV and Week View UV (Avg.Std −38.2% and −42.4%, Avg.MAD −29.5% and −38.2% over their second-best baselines). That a single frozen encoder handles all five distributions confirms that the numerical pathway generalizes across scales rather than fitting any one attribute.

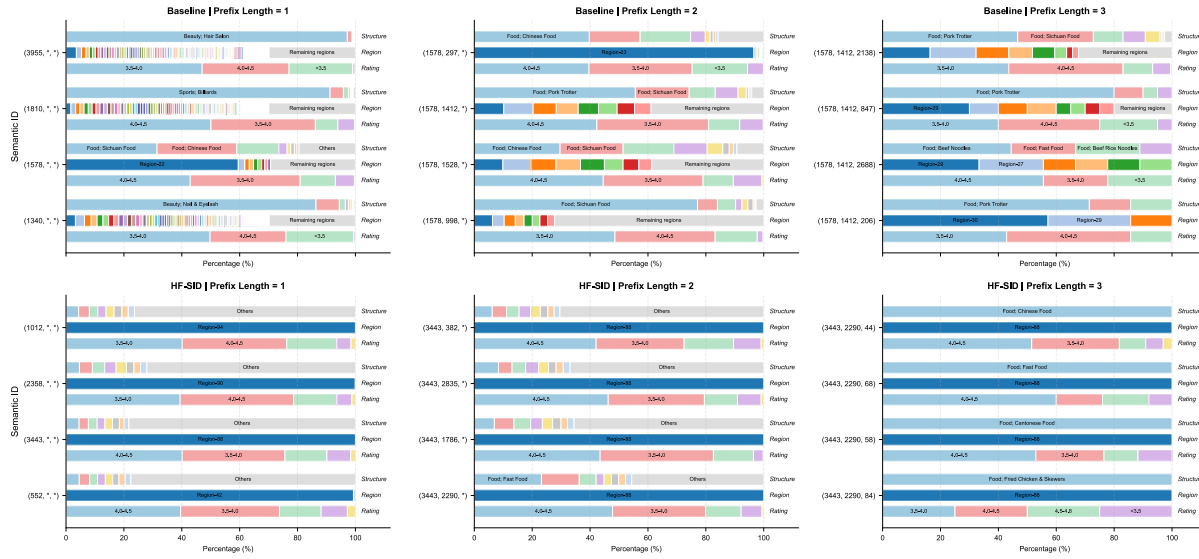


Figure 4. Hierarchical semantic consistency of SID prefixes. For prefixes of increasing length $(a,*,*)$, $(a,b,*)$, (a,b,c) , we visualize the tag, region, and rating distributions under each. HF-SID shows stronger geographic locality and more concentrated ratings at deeper levels; under a shared parent $(3443,2290,*)$, its L3 children are nearly tag-pure, each a distinct fine-grained category, reflecting Structure-based Contrastive Learning on the third-level residual.

4.4.3. Visualization of Structural Awareness

Here each POI's structure is its two-level tag of the form "parent/child", e.g., "Food/Chinese Food". We compare HF-SID with RQ-KMeans by visualizing the tag, region, and rating distributions induced by SID prefixes of increasing length, namely $(a,*)$, $(a,b,*)$, and (a,b,c) (Figure 4). HF-SID shows better geographic and rating organization at all three prefix lengths, while its structural advantage emerges specifically at the third token: under a shared parent prefix $(3443,2290,*)$, its L3 children are almost tag-pure, each corresponding to a distinct fine-grained food category. This is by design, as Structure-based Contrastive Learning acts only on the third-level residual. The first two levels are left to encode geographic, numerical, and semantic information, geography above all; applying the contrastive objective there would compete with the geographic signal that dominates retrieval, which is precisely the failure the last-layer-only design avoids.

Table 5. Performance comparison on public datasets.

Dataset	Method	ICR ($\uparrow$)	Avg.Dist. ($\downarrow$)	Hit@5 ($\uparrow$)
NYC	RQ-KMeans	52.13%	6.96 km	41.23%
	GNPR-SID	19.74%	5.52 km	44.34%
	Cosine-RQ	16.57%	4.61 km	42.87%
	Pro-GEO	55.29%	6.10 km	53.00%
	HF-SID	56.12%	3.99 km	59.83%
TKY	RQ-KMeans	49.12%	7.21 km	30.00%
	GNPR-SID	18.50%	5.67 km	33.68%
	Cosine-RQ	13.69%	5.66 km	35.27%
	Pro-GEO	45.60%	6.19 km	49.27%
	HF-SID	54.46%	4.41 km	57.37%

4.5. Experiments on Public Datasets

In this section, we evaluate on two public Foursquare check-in datasets, NYC and TKY. Unlike the AMap dataset, these datasets provide only coordinates and check-in records, without the rich dynamic numerical attributes or fine-grained category structure that Num-CPT and Structure-based Contrastive Learning rely on. We therefore apply only Geo-CPT and report results in Table 5.

HF-SID achieves the best performance on both datasets across all three metrics. On the SID quality side, Avg.Dist. drops from 6.96 km to 3.99 km on NYC (−42.7%) and from 7.21 km to 4.41 km on TKY (−38.8%), while ICR improves to 56.12% and 54.46% respectively—both the highest among all methods. Notably, while GNPR-SID and Cosine-RQ achieve better Avg.Dist. than RQ-KMeans on NYC (5.52 km and 4.61 km), their ICR collapses to below 20%, indicating severe codeword collision at the cost of geographic coherence. HF-SID is the only method that improves both metrics simultaneously. On the retrieval side, HF-SID achieves Hit@5 of 59.83% on NYC and 57.37% on TKY, outperforming the strongest baseline Pro-GEO by 6.83 and 8.10 points respectively, demonstrating that finer geographic SIDs directly translate to better retrieval accuracy.

The relative Avg.Dist. reduction is smaller than on AMap-L (−42.7% vs. −95.9%), which is expected: public datasets contain far fewer POIs, so each codebook cluster spans a larger and more geographically dispersed set of POIs, making fine-grained spatial aggregation inherently harder. Even so, consistent gains across two independent cities confirm that Geo-CPT’s Cartesian coordinate encoding is robust and platform-agnostic.

Table 6. Online A/B experiments.

Online Metrics	WinRate	PV_CTR	UV_CTR	PV_CVR	UV_CVR
Restaurant	+0.35%	+0.24%	+0.14%	+5.89%	+5.63%
Life Services	+1.54%	+0.79%	+0.23%	+2.81%	+2.68%
Tourist Attraction	+0.87%	+0.97%	+0.50%	+11.51%	+9.79%
Ave. Imp.	+0.92%	+0.67%	+0.29%	+6.74%	+6.03%

4.6. Online A/B Test

HF-SID has been deployed on the AMap homepage POI recommendation pipeline, replacing a previously deployed generative retrieval method that already encodes geographic information into its SIDs. The inference pipeline is unchanged: a prompt is assembled from personalized preferences, historical behaviour trajectories, and real-time context including current location, timestamp, and search query; candidates are then generated autoregressively via beam search and passed to the downstream fine-ranking and re-ranking modules. Since the two systems differ only in the SID generation model, the comparison isolates the contribution of HF-SID itself. We ran a one-week A/B test across three representative scenarios, namely Restaurant, Life Services, and Tourist Attraction.

As shown in Table 6, HF-SID improves every metric in every scenario, and the gains concentrate where fidelity is decisive. Averaged over the three scenarios, PV_CVR rises by 6.74% and UV_CVR by 6.03%, an order of magnitude beyond the

corresponding CTR gains of 0.67% and 0.29%. The asymmetry is expected: clicks are driven largely by surface content such as the POI title and cover image, whereas conversion depends on whether the POI is genuinely reachable and genuinely of the intended category, which is exactly what a higher-fidelity identifier preserves. Consistently, Tourist Attraction benefits most with PV_CVR up 11.51%, as it combines the widest geographic span of candidates with dense co-located POIs inside a single scenic area. These gains are obtained over a production system that was already geography-aware.

5. Conclusion

In this paper, we present HF-SID, a high-fidelity Semantic ID framework for generative POI retrieval in LBS. HF-SID converts coordinates into a continuous 3D Cartesian form and encodes every numerical value as a single unit through a dedicated Numerical Encoder, fuses these embeddings into the LLM with a Type Embedding that gives each attribute type its own subspace, and aligns the model via Geo-CPT and Num-CPT. A Structure-based Contrastive Learning objective, applied only to the residual before the last quantization layer, then separates co-located POIs that share a coarse tag but differ at the fine level. All mechanisms act at the embedding level and add no token overhead. Deployed on the AMap platform, HF-SID consistently improves both offline SID quality and online business metrics over the previous system, and we release AMap-S, a large-scale real-world POI benchmark.

References

- Chen et al. (2025) B. Chen, X. Guo, S. Wang, Z. Liang, Y. Lv, Y. Ma, X. Xiao, B. Xue, X. Zhang, Y. Yang, et al. Onesearch: a preliminary exploration of the unified end-to-end generative framework for e-commerce search. *arXiv preprint arXiv:2509.03236*.
Cited by: [§1](#), [§4.1](#), [§4.2](#), [Table 2](#), [Table 2](#).

- Chen et al. (2026) D. Chen, S. Zheng, H. Shao, H. Wu, M. Xu, Y. Yan, R. Li, and Z. Zhu
Revisiting general map search via generative point-of-interest retrieval.
 External Links: 2605.03397
 Cited by: [§1](#), [§1](#), [§2](#), [§4.1](#), [Table 2](#), [Table 2](#), [Table 2](#).
- Dempster et al. (1977) A. P. Dempster, N. M. Laird, and D. B. Rubin
 Maximum likelihood from incomplete data via the em algorithm.
Journal of the royal statistical society: series B (methodological) **39** (1), pp. 1–22.
 Cited by: [§2](#).
- Deng et al. (2025) J. Deng, S. Wang, K. Cai, L. Ren, Q. Hu, W. Ding, Q. Luo, and G. Zhou
 Onerec: unifying retrieve and rank with generative recommender and iterative preference alignment.
arXiv preprint arXiv:2502.18965.
 Cited by: [§1](#), [§2](#), [§4.1](#).
- Fang et al. (2025) D. Fang, J. Gao, C. Zhu, Y. Li, X. Zhao, and Y. Chang
 Hid-vae: interpretable generative recommendation via hierarchical and disentangled semantic ids.
arXiv preprint arXiv:2508.04618.
 Cited by: [§2](#).
- Gao et al. (2022) Y. Gao, Y. Xiong, S. Wang, and H. Wang
 GeoBERT: pre-training geospatial representation learning on point-of-interest.
Applied Sciences **12** (24), pp. 12942.
 Cited by: [§1](#).
- Guo et al. (2024) X. Guo, J. Pan, X. Wang, B. Chen, J. Jiang, and M. Long

- On the embedding collapse when scaling up recommendation models.*
External Links: 2310.04400
Cited by: [§1](#).
- He et al. (2026) T. He, C. Yang, J. Zhang, L. Guo, W. Lin, and Z. Jiang
Birds of a feather cluster nearby: a proximity-aware geo-codebook for local service recommendation.
External Links: 2604.23156
Cited by: [§1](#), [§1](#), [§2](#), [§4.1](#), [§4.1](#), [Table 2](#), [Table 2](#).
- Jiang et al. (2025) H. Jiang, G. Wang, D. Zhou, S. Yu, Y. Zeng, W. Zeng, K. Gai, and G. Zhou
LLM-aligned geographic item tokenization for local-life recommendation.
External Links: 2511.14221
Cited by: [§1](#), [§2](#).
- Jin et al. (2026) S. Jin, T. Chen, C. Gao, and J. Han
GeoNum: bridging numerical continuity and language semantics via geometric embedding.
In Proceedings of the AAAI Conference on Artificial Intelligence,
Vol. **40**, pp. 22426–22434.
Cited by: [§1](#), [§1](#), [§2](#), [§4.1](#).
- Lee et al. (2022) D. Lee, C. Kim, S. Kim, M. Cho, and W. Han
Autoregressive image generation using residual quantization.
In 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR),
pp. 11513–11522.
Cited by: [§2](#).
- Li et al. (2025a) H. Li, X. Chen, Z. Xu, D. Li, N. Hu, F. Teng, Y. Li, L. Qiu, C. J. Zhang, L. Qing, et al.
Exposing numeracy gaps: a benchmark to evaluate fundamental numerical abilities in large

language models.

In Findings of the Association for Computational Linguistics: ACL 2025, pp. 20004–20026.

Cited by: [§1](#), [§1](#), [§2](#).

Li et al. (2025b)

X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, and Z. Dou

From matching to generation: a survey on generative information retrieval.

ACM Transactions on Information Systems **43** (3), pp. 1–62.

Cited by: [§1](#).

Liang et al. (2026)

Y. Liang, Z. Zhang, Y. Zhu, K. Zhang, Z. Guo, W. Zhou, Z. Yang, K. Wu, Y. Ni, A. Zeng, et al.

Rethinking generative recommender tokenizer: recsys-native encoding and semantic quantization beyond llms.

arXiv preprint arXiv:2602.02338.

Cited by: [§2](#).

Lin et al. (2025)

H. Lin, Z. Yang, J. Xue, Z. Zhang, L. Wang, Y. Gu, Y. Xu, and X. Li

Spacetime-gr: a spacetime-aware generative model for large scale online poi recommendation.

arXiv preprint arXiv:2508.16126.

Cited by: [§1](#).

Liu et al. (2014)

J. Liu, H. Li, Y. Gao, H. Yu, and D. Jiang

A geohash-based index for spatial data management in distributed memory.

In 2014 22Nd international conference on geoinformatics, pp. 1–4.

Cited by: [§1](#).

Liu et al. (2026)

W. Liu, Z. Liu, M. Xie, H. Zhu, J. Yu, J. Yin, and W. Lee

- Geography-aware large language models for next poi recommendation.*
In 2026 IEEE 42nd International Conference on Data Engineering (ICDE),
pp. 2264–2277.
Cited by: [§1](#).
- Loshchilov and Hutter (2019) I. Loshchilov and F. Hutter
Decoupled weight decay regularization.
External Links: 1711.05101
Cited by: [§4.1](#).
- Luo et al. (2024) X. Luo, J. Cao, T. Sun, J. Yu, R. Huang, W. Yuan, H. Lin, Y. Zheng, S. Wang, Q. Hu, C. Qiu, J. Zhang, X. Zhang, Z. Yan, J. Zhang, S. Zhang, M. Wen, Z. Liu, K. Gai, and G. Zhou
QARM: quantitative alignment multi-modal recommendation at kuaishou.
External Links: 2411.11739
Cited by: [§2](#), [§4.1](#), [Table 2](#), [Table 2](#).
- Lv et al. (2026) D. Lv, Q. Ding, H. Xu, Z. Sun, Z. Wang, F. Xiong, and M. Xu
Reasoning over space: enabling geographic reasoning for llm-based generative next poi recommendation.
In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),
pp. 7322–7336.
Cited by: [§1](#), [§1](#), [§2](#).
- Minaee et al. (2024) S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao
Large language models: a survey.
arXiv preprint arXiv:2402.06196.
Cited by: [§1](#).

- Naveed et al. (2025) H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian
A comprehensive overview of large language models.
ACM Transactions on Intelligent Systems and Technology **16** (5), pp. 1–72.
Cited by: [§1](#).
- Ni (2026) A. Ni
Numeracy in large language models: fundamental limitations and paths to improvement.
arXiv preprint arXiv:2608.13129.
Cited by: [§1](#), [§2](#).
- Pang et al. (2025) M. Pang, C. Yuan, X. He, Z. Fang, D. Xie, F. Qu, X. Jiang, C. Peng, Z. Lin, Z. Luo, et al.
Generative retrieval and alignment model: a new paradigm for e-commerce retrieval.
In Companion Proceedings of the ACM on Web Conference 2025,
pp. 413–421.
Cited by: [§1](#).
- Rajbhandari et al. (2020) S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He
ZeRO: memory optimizations toward training trillion parameter models.
External Links: 1910.02054
Cited by: [§4.1](#).
- Rajput et al. (2023) S. Rajput, N. Mehta, A. Singh, R. Hulikal Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Tran, J. Samost, et al.
Recommender systems with generative retrieval.
Advances in Neural Information Processing Systems **36**, pp. 10299–10315.
Cited by: [§1](#), [§2](#).

- Tay et al. (2022) Y. Tay, V. Q. Tran, M. Dehghani, J. Ni, D. Bahri, H. Mehta, Z. Qin, K. Hui, Z. Zhao, J. Gupta, T. Schuster, W. W. Cohen, and D. Metzler
Transformer memory as a differentiable search index.
External Links: 2202.06991
Cited by: [§2](#).
- Vaswani et al. (2023) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin
Attention is all you need.
External Links: 1706.03762
Cited by: [§3.3.1](#).
- Veach et al. (2017) E. Veach, J. Rosenstock, E. Engle, and T. Manshreck
S2 geometry library: computational geometry and spatial indexing on the sphere.
Software.
Cited by: [§1](#).
- Wang et al. (2025) D. Wang, Y. Huang, S. Gao, Y. Wang, C. Huang, and S. Shang
Generative next poi recommendation with semantic id.
External Links: 2506.01375
Cited by: [§2](#), [§4.1](#), [§4.1](#), [§4.2](#), [Table 2](#), [Table 2](#).
- Wang et al. (2026) F. Wang, H. Lin, Y. Yuan, S. Wang, X. Zhou, S. Yang, and P. Wang
GeoGR: a generative retrieval framework for spatio-temporal aware poi recommendation.
External Links: 2602.10411
Cited by: [§1](#), [§1](#), [§2](#), [§4.1](#), [Table 2](#), [Table 2](#).
- Wang et al. (2023) Y. Wang, Y. Hou, H. Wang, Z. Miao, S. Wu, H. Sun, Q. Chen, Y. Xia, C. Chi, G. Zhao, Z. Liu, X. Xie, H. A. Sun, W. Deng, Q. Zhang, and M. Yang
A neural corpus indexer for document retrieval.

External Links: 2206.02743

Cited by: [§2](#).

Yang et al. (2025a)

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu

Qwen3 technical report.

External Links: 2505.09388

Cited by: [§4.1](#).

Yang et al. (2025b)

H. Yang, Y. Hu, S. Kang, Z. Lin, and M. Zhang
Number cookbook: number understanding of language models and how to improve it.

External Links: 2411.03766

Cited by: [§1](#), [§2](#).

Zhang et al. (2025a)

F. Zhang, X. Liu, X. Jia, Y. Zhang, Z. Xia, F. Jiang, F. Zhuang, W. Lin, and Z. Zhang
HierGR: hierarchical semantic representation enhancement for generative retrieval in food delivery search.

In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pp. 444–455.

Cited by: [§1](#).

Zhang et al. (2025b)

Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou

Qwen3 embedding: advancing text embedding and reranking through foundation models.

External Links: 2506.05176

Cited by: [§3.3.2](#), [§4.1](#).

Zheng et al. (2024)

B. Zheng, Y. Hou, H. Lu, Y. Chen, W. X. Zhao, M. Chen, and J. Wen

Adapting large language models by integrating collaborative semantics for recommendation.

External Links: 2311.09049

Cited by: [§2](#).

Zhou et al. (2025)

G. Zhou, H. Hu, H. Cheng, H. Wang, J. Deng, J. Zhang, K. Cai, L. Ren, L. Ren, L. Yu, et al.

Onerec-v2 technical report.

arXiv preprint arXiv:2508.20900.

Cited by: [§1](#), [§2](#).

Zhou et al. (2026)

J. Zhou, C. Chen, K. Zuo, M. Xu, Z. Fu, Y. Chen, X. Tang, and Y. Hu

HyMiRec: a hybrid multi-interest learning framework for llm-based sequential recommendation.

In Proceedings of the ACM Web Conference 2026,

pp. 5998–6008.

Cited by: [§4.1](#), [§4.2](#), [Table 2](#), [Table 2](#).

Appendix A Appendix

A.1. Case Study

We further provide a case study to illustrate how the SID quality improvements translate into concrete retrieval behavior. Figure [5](#) compares the generated POI candidates of the baseline and HF-SID under the same user context.

The baseline retrieves candidates that are only partially category-consistent but geographically scattered: although 29 out of the top-30 candidates share the target category, their average distance from the target reaches 9.22 km, and the ground-truth target is missed. In contrast, HF-SID concentrates the retrieved candidates around the target region, reduces the average distance to 2.40 km, improves category consistency to 30 / 30, and ranks the target POI at the first position. This example is consistent with the aggregate results in Table 2: preserving geographic and structural fidelity in the representation yields SIDs whose candidate sets are both spatially compact and semantically reliable.

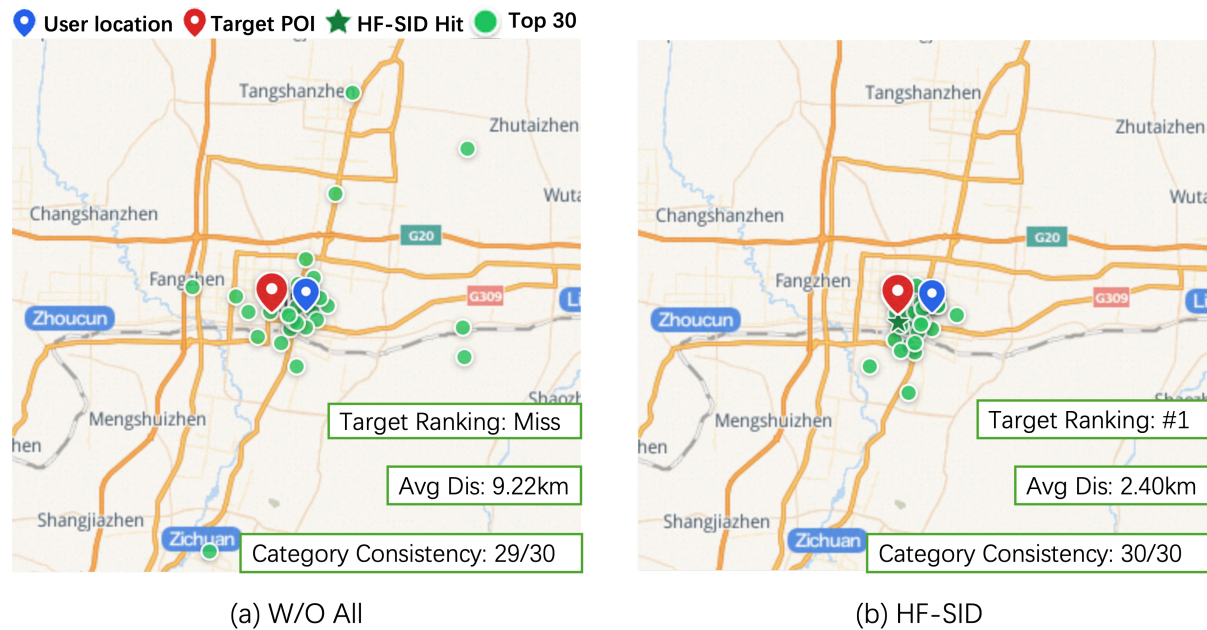


Figure 5. Case study of generated POI candidates. The baseline misses the target and produces a geographically dispersed candidate set, whereas HF-SID ranks the target first while achieving lower average distance and higher category consistency among the top-30 candidates.

A.2. CPT Prompt Templates

Table [7](#) summarizes the prompt templates used for continual pre-training. We design two groups of templates corresponding to Geo-CPT and Num-CPT. Geo-CPT focuses on spatial representation and geographic reasoning, including POI spatial information serialization, pairwise distance prediction, and nearest-location identification. Num-CPT focuses on heterogeneous numerical attribute comparison, where the model is required to compare values under a specified attribute type rather than relying on raw numerical magnitude alone.

Table 7. Prompt templates for Geo-CPT and Num-CPT.

Template Type	Prompt Template
Prompt 1 Pairwise Geographic Distance Prediction	The 3D spatial coordinates of {POI A} are < start_num_0 >{X_A}< end_num_0 >, < start_num_0 >{Y_A}< end_num_0 >, and < start_num_0 >{Z_A}< end_num_0 >; its address is {Address A}. The 3D spatial coordinates of {POI B} are < start_num_0 >{X_B}< end_num_0 >, < start_num_0 >{Y_B}< end_num_0 >, and < start_num_0 >{Z_B}< end_num_0 >; its address is {Address B}. The spherical distance between the two locations is < start_num_1 >{Distance}< end_num_1 > meters.
Prompt 2 Nearest-Location Identification	The 3D spatial coordinates of {POI A} are < start_num_0 >{X_A}< end_num_0 >, < start_num_0 >{Y_A}< end_num_0 >, and < start_num_0 >{Z_A}< end_num_0 >. The 3D spatial coordinates of {POI B} are < start_num_0 >{X_B}< end_num_0 >, < start_num_0 >{Y_B}< end_num_0 >, and < start_num_0 >{Z_B}< end_num_0 >. The 3D spatial coordinates of {POI C} are < start_num_0 >{X_C}< end_num_0 >, < start_num_0 >{Y_C}< end_num_0 >, and < start_num_0 >{Z_C}< end_num_0 >. The location closer to {Query POI} is {Answer POI}.
Prompt 3 Attribute Comparison	{POI A} has rating < start_num_2 >{Rating_A}< end_num_2 >, weekly views < start_num_6 >{WeekView_A}< end_num_6 >, and collected users < start_num_5 >{Collect_A}< end_num_5 >. {POI B} has rating < start_num_2 >{Rating_B}< end_num_2 >, weekly views < start_num_6 >{WeekView_B}< end_num_6 >, and collected users < start_num_5 >{Collect_B}< end_num_5 >. {POI A} has a higher {Queried Attribute} than {POI B}.

Experimental support, please [view the build logs](#) for errors. Generated by [L^AT_EX_{ML} 1.1](https://math.nist.gov/~BMiller/LaTeXML/) (<https://math.nist.gov/~BMiller/LaTeXML/>).

Instructions for reporting errors

We are continuing to improve HTML versions of papers, and your feedback helps enhance accessibility and mobile support. To report errors in the HTML that will help us improve conversion and rendering, choose any of the methods listed below:

- Click the "Report Issue" (🐛) button, located in the page header.

Tip: You can select the relevant text first, to include it in your report.

Our team has already identified [the following issues](https://github.com/arXiv/html_feedback/issues) (https://github.com/arXiv/html_feedback/issues). We appreciate your time reviewing and reporting rendering errors we may not have found yet. Your efforts will help us improve the HTML versions for all readers, because disability should not be a barrier to accessing research. Thank you for your continued support in championing open access for all.

Have a free development cycle? Help support accessibility at arXiv! Our collaborators at LaTeXML maintain a [list of packages that need conversion](https://github.com/bruceMiller/LaTeXML/wiki/Porting-LaTeX-packages-for-LaTeXML) (<https://github.com/bruceMiller/LaTeXML/wiki/Porting-LaTeX-packages-for-LaTeXML>), and welcome [developer contributions](https://github.com/bruceMiller/LaTeXML/issues) (<https://github.com/bruceMiller/LaTeXML/issues>).