Hierarchical Graph Alignment for Cross-Modal 3D Scene Grounding

Tianyi Shang
Purdue University, USA
shangt@purdue.edu

Zhenyu Li*
Qilu University of Technology
(Shandong Academy of Sciences), China
lizhenyu@qlu.edu.cn

Abstract

Text-guided point-cloud grounding enables robots to ground natural-language descriptions in 3D environments, making it important for embodied AI and human-robot interaction. Existing coarse-to-fine methods primarily rely on global descriptors for submap retrieval, but these descriptors often compress object semantics, spatial relations, and scene layouts into a single vector, thereby limiting discriminability in large-scale scenes. We propose HiGraLoc, a multi-level cross-modal alignment framework that improves the coarse retrieval stage through three complementary branches. The instance branch uses a Hyperbolic Instance Structure Encoder to model hierarchical object semantics, the relation branch aggregates reliability-weighted pairwise spatial relations, and the global branch employs a Global Spectral Graph Encoder to capture multi-frequency scene structure. A Text2Loc-style coordinate regression stage then refines the location estimate within the retrieved submap. Experiments on KITTI360Pose show that HiGraLoc achieves a 19% improvement in Top-1 recall@10m over existing state-of-the-art methods. Our code and dataset are available at https://github.com/nuozimiaowu/HiGraLoc.

1 Introduction

Text-guided point-cloud localization has become a pivotal capability for embodied AI, offering a robust alternative to traditional Visual Place Recognition (VPR). While VPR often fails under varying illumination or seasonal changes, this paradigm grounds nuanced semantic and spatial descriptions directly into 3D geometric maps. By reducing dependence on viewpoint-specific visual cues, it enables agents to reason about urban environments in a more human-like, intuitive manner.

To tackle the inherent complexity of searching across vast urban areas, existing state-of-the-art methods, including Text2Pos [9], RET [23], Text2Loc [27], and MambaPlace [19], typically adopt a hierarchical coarse-to-fine localization paradigm. In this framework, the system first performs a coarse-level retrieval to identify the most relevant candidate submap from a large-scale database and subsequently executes a fine-grained regression to determine precise coordinates within the retrieved local context. While this two-stage approach effectively narrows down the search space, it places a heavy burden on the discriminative power of the coarse retrieval stage; any ambiguity in global descriptors or failure to capture structural relationships at this level can lead to irreversible localization errors in the final stage.

Despite their effectiveness, current coarse retrieval modules often rely on globally pooled descriptors. This design is compact but tends to collapse heterogeneous cues into a single representation. In text-guided localization, a query may describe object identities, pairwise spatial relations, and holistic scene layouts at the same time. When different places share similar objects, global descriptors may

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

fail to preserve the relational and structural evidence needed to distinguish them. As a result, errors in coarse retrieval can directly limit the subsequent fine-stage coordinate regression.

To address this limitation, we propose HiGraLoc, a multi-level cross-modal alignment framework for text-guided point-cloud localization. HiGraLoc redesigns the coarse stage with three complementary branches.

The instance-level branch introduces a Hyperbolic Instance Structure Encoder (HISE), which maps instance features into a Poincaré ball to model hierarchical object-level structure. The relation-level branch constructs directed pairwise relations between instances and aggregates them with learned reliability weights, improving the alignment of spatial expressions such as "south of" and "north of". The global-level branch adopts a Global Spectral Graph Encoder (GSGE), which applies Chebyshev spectral filtering on an instance graph to capture topology-aware multi-frequency structure, making the global descriptor more discriminative.

These branches capture complementary aspects of a 3D scene: instance-level alignment models what objects are present, relation-level alignment models how objects are arranged, and global-level alignment models the overall scene structure. Their retrieval scores are combined during inference to select a candidate submap, which is then passed to a Text2Loc-style fine stage for coordinate regression. This design preserves the standard coarse-to-fine pipeline while providing a more expressive coarse retrieval signal.

Extensive experiments on KITTI360Pose and our self-built StreetLoc demonstrate that HiGraLoc outperforms state-of-the-art methods. Notably, even when only the global branch is retained, our method still significantly outperforms prior methods that rely on global descriptors. Our contributions are summarized as follows:

- Multi-level alignment framework: HiGraLoc aligns text and 3D point clouds across instance, relation, and global levels for robust city-scale localization.
- Instance-level hyperbolic modeling: HISE leverages hyperbolic geometry in a Poincaré ball to capture hierarchical object-level semantics and category granularities.
- Relation-level spatial reasoning: A reliability-weighted branch aggregates directed spatial relations to improve grounding of complex expressions like "south of" or "north of."
- Global-level spectral descriptors: GSGE extracts topology-aware cues via spectral filtering, outperforming prior methods and providing complementary gains.

2 Related Work

2.1 Visual Place Recognition

Traditional 2D localization techniques generally rely on feature aggregation mechanisms, such as the Vector of Locally Aggregated Descriptors (VLAD) [1] and Generalized Mean (GeM), to extract global descriptors from query frames, facilitating direct 2D feature matching. To bolster system reliability, recent studies have explored the inherent correlations between local descriptors and their corresponding cluster centers [11]. Simultaneously, the field has seen a shift toward end-to-end Transformer-based architectures to produce more discriminative global embeddings [13, 25]. Specialized frameworks have also been developed to tackle environmental extremes, including degraded lighting [12] and drastic shifts in camera perspective [2]. Recent breakthroughs have further pushed the boundaries of VPR, scaling localization capabilities from urban environments to a continental magnitude [15]. Moreover, robust positioning is increasingly achieved through cross-modal retrieval, such as text-to-image [20] or point-cloud-to-image [28] paradigms, which provide critical fallbacks when direct visual input is compromised.

2.2 LiDAR Place Recognition

LiDAR-based Place Recognition (LPR) generates global descriptors either from unstructured point clouds or Bird's-Eye View (BEV) representations [14]. Pioneering works such as PointNetVLAD [21] integrated PointNet [18] with the NetVLAD layer to condense spatial points into compact vectors. Subsequent BEV-centric models, like BevPlace [17] and I2P-Rec [31], have improved rotational robustness via geometric transforms and utilized multimodal projections to gather richer spatial cues.

Alternative LPR strategies focus on the intrinsic topology of point clouds through 3D CNNs or attention-driven graphs. For instance, LPD-Net [16] employs graph structures to represent local geometry, while DH3D [4] and SOENet [26] emphasize high-fidelity pose refinement. To ensure embedding consistency during incremental learning, InCloud [8] was introduced. Furthermore, Transformer-based LPR systems have gained traction for their ability to model long-range context; TransLoc3D [29] utilizes self-attention for global awareness, whereas NDT-Transformer [32] and PPT-Net [6] process NDT-transformed data and multi-scale features through hierarchical structures.

Efficiency in large-scale scenarios is often achieved via sparse 3D convolutions, as demonstrated by MinkLoc3D [10]. More recent advancements include LCDNet [3], which adopts optimal transport theory for superior matching, and LoGG3D-Net [22], which utilizes a local consistency loss to ensure viewpoint-invariant descriptor discriminability.

2.3 Cross Modal Localization

The domain of text-to-point-cloud localization was established by Text2Pos [9], which introduced a multi-stage coarse-to-fine pipeline. It was followed by RET [23], which leveraged attention mechanisms, and Text2Loc [27], which used contrastive learning and a matching-free refinement phase. State space models (SSMs) were recently integrated into this task by MambaPlace [19]. While GOTPR [7] improved efficiency through intermediate graph structures, its utility remains mainly in the coarse retrieval stage.

Current research also addresses practical bottlenecks such as data scarcity and cross-modal ambiguity. IFRP-T2P [24] achieves instance-free positioning using query extractors combined with RowColRPA/RPCA for spatial awareness. Uncertainty modeling has also emerged as an important theme: PMSH [5] handles partially matched submaps by propagating uncertainty into the coarse retrieval metric. Similarly, CMMLoc [30] tackles incomplete textual descriptions through a Cauchy-Mixture-Model (CMM) framework, further refined by cardinal-direction information and spatial consolidation.

3 Method

3.1 Notation Summary

We consider the sample index i for each query–reference pair. The textual query is denoted by $\mathcal{Q}_i$, the paired reference submap by $\mathcal{R}_i$, and the retrieved submap by $\tilde{\mathcal{R}}_i$. The candidate gallery is $\mathcal{G}$. For a submap i, let $\mathbf{X}_i = [\mathbf{x}_{i,1}; \ldots; \mathbf{x}_{i,N_i}] \in \mathbb{R}^{N_i \times D}$ denote the instance feature matrix, where N_i is the number of detected instances and $\mathbf{x}_{i,m} \in \mathbb{R}^D$ is the feature of the m-th instance $\mathbf{s}_{i,m}$. For the textual side, let $\mathbf{T}_i = [\mathbf{t}_{i,1}; \ldots; \mathbf{t}_{i,M_i}] \in \mathbb{R}^{M_i \times D}$ denote the text feature matrix, where M_i is the number of text elements and $\mathbf{t}_{i,j} \in \mathbb{R}^D$ is the j-th text feature.

3.2 Task Formulation

HiGraLoc addresses text-guided point-cloud place recognition via a coarse-to-fine strategy. In the coarse stage of HiGraLoc, we retrieve the most relevant point-cloud submap from the gallery using the textual query:

$$\tilde{\mathcal{R}}_i = \arg\min_{\mathcal{R} \in \mathcal{G}} d\big(f_q(\mathcal{Q}_i), f_r(\mathcal{R})\big), \tag{1}$$

where $f_q(\cdot)$ and $f_r(\cdot)$ encode the query and submap into a shared retrieval space, and $d(\cdot, \cdot)$ is the retrieval metric.

HiGraLoc decomposes this coarse cross-modal alignment into three complementary branches: instance-level alignment for object semantics, relation-level alignment for pairwise spatial structure, and global-level alignment for holistic scene structure.

In the fine stage, HiGraLoc follows the established Text2Loc-style coordinate regression on the retrieved submap $\tilde{\mathcal{R}}_i$.

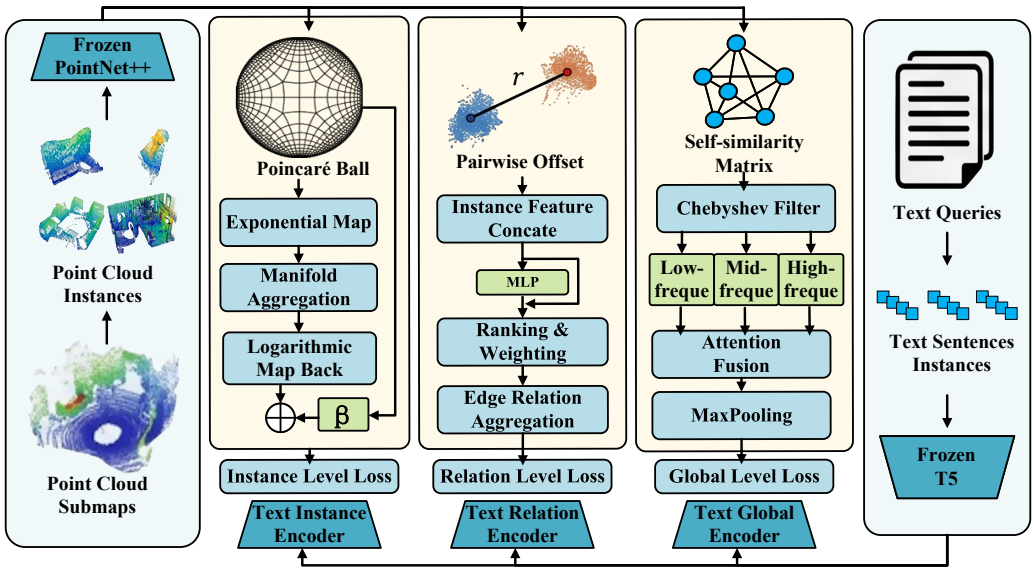

Figure 1: The coarse retrieval stage of the HiGraLoc framework.

3.3 Instance-level Alignment

Instance-level features capture object semantics, but they also inherit the hierarchical organization of the scene. To model this structure, we introduce the Hyperbolic Instance Structure Encoder (HISE), which maps instance features into a Poincaré ball with learnable curvature. Hyperbolic geometry is suitable for nested layouts since its exponential volume growth naturally supports increasingly fine-grained distinctions away from the origin.

We use the Poincaré ball $\mathbb{B}_c^d = \{\mathbf{z} \in \mathbb{R}^d : c\|\mathbf{z}\|^2 < 1\}$ with curvature $-c$, where $c = \mathrm{softplus}(\rho_c) + \epsilon$ is learnable. Instance features are first projected into the tangent space and lifted to the manifold by the exponential map,

$$\mathbf{z}_{i,m} = \exp_{\mathbf{o}}^c(W_x \mathbf{x}_{i,m}), \tag{2}$$

where $\mathbf{o}$ denotes the origin of the Poincaré ball and W_x is a learnable projection matrix. We compute a distance-based manifold aggregation score as

$$\alpha_{i,mn} = \frac{\exp\big(\mathrm{sim}_{\mathbb{H}}(\mathbf{z}_{i,m}, \mathbf{z}_{i,n})/\tau_h\big)}{\sum_{n'=1}^{N_i} \exp\big(\mathrm{sim}_{\mathbb{H}}(\mathbf{z}_{i,m}, \mathbf{z}_{i,n'})/\tau_h\big)}, \tag{3}$$

where τ_h is a temperature parameter, and we define the hyperbolic similarity as

$$\mathrm{sim}_{\mathbb{H}}(\mathbf{z}_{i,m}, \mathbf{z}_{i,n}) = -d_c^2(\mathbf{z}_{i,m}, \mathbf{z}_{i,n}), \tag{4}$$

with $d_c(\cdot,\cdot)$ denoting the geodesic distance in $\mathbb{B}_c^d$.

Using these weights, we aggregate instances in the local tangent space and map the result back to the manifold,

$$\tilde{\mathbf{z}}_{i,m} = \exp_{\mathbf{z}_{i,m}}^c\left(\sum_{n=1}^{N_i} \alpha_{i,mn} \log_{\mathbf{z}_{i,m}}^c(\mathbf{z}_{i,n})\right). \tag{5}$$

The enhanced manifold features are then converted back to Euclidean space,

$$\mathbf{v}_{i,m}^{\mathrm{HISE}} = W_o \log_{\mathbf{o}}^c(\tilde{\mathbf{z}}_{i,m}), \qquad \mathbf{V}_i^{\mathrm{HISE}} = [\mathbf{v}_{i,1}^{\mathrm{HISE}}, \ldots, \mathbf{v}_{i,N_i}^{\mathrm{HISE}}]^\top. \tag{6}$$

Finally, we combine them with the original instance features through a gated residual connection,

$$\mathbf{P}_i^{\mathrm{Inst}} = \beta_{\mathrm{inst}} \mathbf{V}_i^{\mathrm{HISE}} + (1 - \beta_{\mathrm{inst}})\mathbf{X}_i, \qquad \beta_{\mathrm{inst}} = \sigma(\rho_{\mathrm{inst}}). \tag{7}$$

For numerical stability, manifold outputs are projected back to $\mathbb{B}_c^d$ after exponential mapping. For the textual modality, we intentionally use a lightweight MLP to generate instance-level textual descriptors $\mathbf{T}_i^{\mathrm{Inst}}$, since the language encoder already provides strong semantic token representations.

3.4 Relation-level Alignment

To model discriminative spatial context beyond object-level semantics, we construct relation-aware representations over point-cloud instances. For the i-th submap, let $\mathbf{X}_i = [\mathbf{x}_{i,1}; \ldots; \mathbf{x}_{i,N_i}] \in \mathbb{R}^{N_i \times D}$ denote the instance feature matrix, where $\mathbf{x}_{i,m}$ is the semantic feature of the m-th detected instance $\mathbf{s}_{i,m}$. The instance features are extracted by a frozen PointNet++ encoder.
For the point-cloud modality, we first compute the pairwise geometric offset between two instances $\mathbf{s}_{i,m}$ and $\mathbf{s}_{i,n}$ based on their centroids:

$$\mathbf{o}_{i,mn} = \mathrm{centroid}(\mathbf{s}_{i,m}) - \mathrm{centroid}(\mathbf{s}_{i,n}). \tag{8}$$

This offset encodes the relative spatial displacement from instance $\mathbf{s}_{i,n}$ to instance $\mathbf{s}_{i,m}$, providing geometric cues for modeling directed pairwise relations. We then project the geometric offset into the latent feature space by $\mathbf{g}_{i,mn} = f_{\mathrm{geo}}(\mathbf{o}_{i,mn})$, where $f_{\mathrm{geo}}(\cdot)$ is a learnable geometric embedding function.

To obtain an initial relation representation, we fuse the semantic features of the two instances with their geometric embedding as $\mathbf{e}^p_{i,mn} = f_{\mathrm{edge}}(\mathrm{concat}(\mathbf{x}_{i,m}, \mathbf{x}_{i,n}, \mathbf{g}_{i,mn}))$, where $f_{\mathrm{edge}}(\cdot)$ maps the concatenated semantic-geometric feature into a relation edge representation. The resulting $\mathbf{e}^p_{i,mn}$ describes the directed relation from instance $\mathbf{s}_{i,n}$ to instance $\mathbf{s}_{i,m}$.

Since not all pairwise relations are equally reliable or discriminative, we further transform each edge into a relation-aware latent representation. Instead of directly using the raw edge feature, we employ a residual relation projection:

$$\boldsymbol{\theta}^p_{i,mn} = \mathrm{LN}\left(\mathbf{e}^p_{i,mn} + \phi_r\left(\mathbf{e}^p_{i,mn}\right)\right), \tag{9}$$

where $\phi_r(\cdot)$ is a lightweight MLP and $\mathrm{LN}(\cdot)$ denotes layer normalization. This projection preserves the original edge information while enhancing relation-specific patterns for subsequent reliability estimation.

We then convert the projected relation feature into a scalar reliability score $\rho^p_{i,mn} = w_r^\top \boldsymbol{\theta}^p_{i,mn}$, where w_r is a learnable vector. The score $\rho^p_{i,mn}$ measures the reliability of the relation between instances $\mathbf{s}_{i,m}$ and $\mathbf{s}_{i,n}$. For each instance $\mathbf{s}_{i,m}$, we normalize the reliability scores over all instances in the same submap using $\alpha^p_{i,mn} = \exp(\rho^p_{i,mn}) / \sum_{n'=1}^{N_i} \exp(\rho^p_{i,mn'})$, where $\alpha^p_{i,mn}$ represents the normalized importance of the relation from instance $\mathbf{s}_{i,n}$ to instance $\mathbf{s}_{i,m}$.

Finally, we aggregate the relation features associated with each instance $\mathbf{s}_{i,m}$ to obtain the relation-aware point-cloud descriptor set of the submap:

$$\mathbf{P}^{\mathrm{Rela}}_i = \left\{\mathbf{p}^{\mathrm{Rela}}_{i,m} = \sum_{n=1}^{N_i} \alpha^p_{i,mn} \boldsymbol{\theta}^p_{i,mn}\right\}_{m=1}^{N_i} \in \mathbb{R}^{N_i \times 256}. \tag{10}$$

The resulting descriptor set $\mathbf{P}^{\mathrm{Rela}}_i$ encodes both semantic and pairwise spatial context, enabling the model to capture discriminative relations such as "next to", "behind", and "across from".

Since the textual descriptions already contain explicit directional terms, we directly apply an Attention layer followed by an MLP to encode $\mathbf{T}_i$, obtaining the textual relation descriptor $\mathbf{T}^{\mathrm{Rela}}_i$.

3.5 Global-level Alignment

We propose a Global Spectral Graph Encoder (GSGE) to capture global-level structural invariants from point-cloud submaps. GSGE builds a self-similarity graph over object instances and applies Chebyshev spectral filtering to obtain multi-frequency global representations.

Given the point-cloud instance matrix $\mathbf{X}_i \in \mathbb{R}^{N_i \times D}$, we construct a self-similarity adjacency matrix $\mathbf{A}_i \in \mathbb{R}^{N_i \times N_i}$ from pairwise feature distances,

$$A_{i,mn} = \exp\left(-\frac{\|\mathbf{x}_{i,m} - \mathbf{x}_{i,n}\|_2^2}{\tau}\right) + \delta_{mn}, \tag{11}$$

where the learnable parameter $\tau > 0$ controls the sensitivity of edge weights to feature distances.

We compute the normalized Laplacian and use its rescaled form for Chebyshev filtering:

$$\mathbf{L}_i = \mathbf{I} - \mathbf{D}_i^{-1/2} \mathbf{A}_i \mathbf{D}_i^{-1/2}, \qquad \widetilde{\mathbf{L}}_i = \mathbf{L}_i - \mathbf{I}, \tag{12}$$

where $\mathbf{D}_i$ is the degree matrix of $\mathbf{A}_i$.

On the constructed instance graph, spectral graph filtering is used to extract structural patterns. To avoid the costly eigen-decomposition required by conventional spectral methods, we approximate the filters with Chebyshev polynomials $T_k(\widetilde{\mathbf{L}}_i)$ for efficient graph information propagation. The Chebyshev polynomials are defined recursively as

$$T_0(\widetilde{\mathbf{L}}_i) = \mathbf{I}, \qquad T_1(\widetilde{\mathbf{L}}_i) = \widetilde{\mathbf{L}}_i, \qquad T_{k+1}(\widetilde{\mathbf{L}}_i) = 2\widetilde{\mathbf{L}}_i T_k(\widetilde{\mathbf{L}}_i) - T_{k-1}(\widetilde{\mathbf{L}}_i). \tag{13}$$

We employ three fixed Chebyshev filters to extract low-, mid-, and high-frequency graph features:

$$\mathbf{Y}_i^r = \sum_{k=0}^{K-1} \beta_k^r \, T_k(\widetilde{\mathbf{L}}_i)\mathbf{X}_i, \quad r \in \{l, m, h\}. \tag{14}$$

where l, m, and h denote the low-, mid-, and high-frequency branches, respectively. K is the polynomial order, and $\boldsymbol{\beta}^r = \{\beta_k^r\}_{k=0}^{K-1}$denotes the predefined coefficient set of the (r)-th Chebyshev filter. The coefficients are fixed throughout training and are used only to emphasize the corresponding low-frequency, mid-frequency, or high-frequency graph components.

The low frequency branch captures smooth global layouts, the medium frequency branch models local object compositions, and the high frequency branch emphasizes fine-grained instance differences. Then, the low-frequency, medium-frequency, and high-frequency representations are fused by a lightweight attention-based aggregation block and then max-pooled to obtain the global point-cloud descriptor $\mathbf{P}_i^{\mathrm{Glob}}$. In parallel, the textual global descriptor $\mathbf{T}_i^{\mathrm{Glob}}$ is obtained by applying an attention operation over the language features followed by max pooling.

3.6 Loss Function

Textual descriptions are inherently ambiguous compared to precise geometric cues. To avoid over-constraining positive pairs with incomplete textual hints, we adopt an asymmetric repulsion strategy. Language is treated primarily as exclusionary evidence to optimize the separation of negative pairs. For feature sets $\mathcal{U}$ and $\mathcal{V}$, the loss is defined as:

$$\mathcal{L}(\mathcal{U}, \mathcal{V}) = -\frac{1}{B} \sum_{i,j=1}^{B} \mathbb{I}_{ij} \log\big(1 - \frac{1}{2}\mathcal{S}(\mathcal{U}_i, \mathcal{V}_j)\big) \tag{15}$$

where $\mathbb{I}_{ij}$ is an indicator function (1 for negative pairs, 0 otherwise). The bidirectional similarity $\mathcal{S}$ is:

$$\mathcal{S}(\mathcal{U}_i, \mathcal{V}_j) = \frac{\exp(\lambda_{ij}^{\mathcal{U}\to\mathcal{V}}/\gamma)}{\sum_{b=1}^{B}\exp(\lambda_{ib}^{\mathcal{U}\to\mathcal{V}}/\gamma)} + \frac{\exp(\lambda_{ji}^{\mathcal{V}\to\mathcal{U}}/\gamma)}{\sum_{b=1}^{B}\exp(\lambda_{jb}^{\mathcal{V}\to\mathcal{U}}/\gamma)} \tag{16}$$

where γ is a temperature parameter and λ denotes the matching score.

We apply the same asymmetric alignment objective to all three branches independently, using their respective feature granularities: $\mathcal{L}_{\mathrm{Instance}} = \mathcal{L}(\mathbf{P}_i^{\mathrm{Inst}}, \mathbf{T}_i^{\mathrm{Inst}})$, $\mathcal{L}_{\mathrm{Relation}} = \mathcal{L}(\mathbf{P}_i^{\mathrm{Rela}}, \mathbf{T}_i^{\mathrm{Rela}})$, and $\mathcal{L}_{\mathrm{Global}} = \mathcal{L}(\mathbf{P}_i^{\mathrm{Glob}}, \mathbf{T}_i^{\mathrm{Glob}})$, where the inputs represent the collections of instance, relation, and global features within a batch, respectively.

During training, HiGraLoc optimizes the three coarse branches independently; during inference, it sums their similarity scores to produce the final cross-modal similarity.

3.7 Fine Stage

For the final coordinate regression within the retrieved submap $\tilde{\mathcal{R}}_i$, we follow the fine-stage design of Text2Loc. Specifically, a cascaded cross-attention module fuses geometry-enhanced point-cloud features with aligned text features, and an MLP regressor predicts the relative offset to the submap anchor point. The absolute location is then recovered directly from this offset. The key point is that a better coarse retrieval provides a more accurate candidate submap, which in turn gives the fine stage a better starting point and improves the final localization accuracy.

Table 1: A comprehensive comparison between HiGraLoc and existing state-of-the-art methods on the KITTI360Pose (test set) and StreetLoc dataset.

Methods	Localization Recall ($\epsilon < 5/10m$) ↑					
	KITTI360Pose			StreetLoc		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
Text2Pos (CVPR'22)	0.13/0.20	0.33/0.42	0.43/0.61	0.14/0.17	0.27/0.31	0.36/0.52
RET (AAAI'23)	0.16/0.25	0.35/0.51	0.46/0.65	0.13/0.21	0.32/0.47	0.40/0.58
Text2Loc (CVPR'24)	0.33/0.48	0.60/0.75	0.70/0.84	0.29/0.42	0.47/0.66	0.55/0.72
IFRP-T2P (ACM MM'24)	0.22/0.40	0.47/0.68	0.58/0.78	0.20/0.33	0.45/0.63	0.53/0.68
MambaPlace (IROS'25)	0.38/0.52	0.66/0.79	0.76/0.87	0.32/0.42	0.50/0.70	0.58/0.76
CMMLoc (CVPR'25)	0.39/0.53	0.67/0.80	0.77/0.87	0.31/0.44	0.50/0.68	0.57/0.78
PMSH (ICCV'25)	0.39/0.55	0.68/0.81	0.78/0.89	0.30/0.43	0.52/0.70	0.59/0.77
HiGraLoc (Ours)	**0.52/0.74**	**0.78/0.93**	**0.84/0.96**	**0.44/0.56**	**0.67/0.82**	**0.76/0.90**

4 Experiment

4.1 Dataset Preparation and Experiment Setup

We evaluate HiGraLoc on the KITTI360Pose benchmark, which covers a 15.51 km^2 urban area across nine sequences. Following the standard protocol, the dataset is divided into training, validation, and testing sets with 28,689, 3,187, and 11,505 query-submap pairs, respectively.

We further construct a new street-level text-to-point-cloud localization dataset, termed StreetLoc, to evaluate the generalization capability of our model in real-world urban scenarios. StreetLoc is collected along an urban street of approximately $4.7\,\mathrm{km}$ in length, where high-precision 3D point clouds are acquired using a 64-beam LiDAR sensor, and global pose information is recorded with a GPS navigation system. Based on the collected street-level point clouds, we partition the scene into point-cloud submaps following the same protocol as KITTI360Pose and construct query-submap pairs accordingly. In total, StreetLoc contains 1,600 query-submap pairs. To support fine-grained semantic relation modeling, we manually annotate the objects in the scenes, covering 19 semantic categories.

All models are implemented in PyTorch and run on a workstation with a 128-core AMD EPYC CPU and an NVIDIA RTX V100 32GB GPU. HiGraLoc is trained for 20 epochs in the coarse retrieval stage and 100 epochs in the fine localization stage, using the same hyperparameter settings as the baselines for fair comparison.

4.2 Overall Performance

We first compare HiGraLoc with state-of-the-art methods on the KITTI360Pose benchmark. As shown in Table 1, HiGraLoc achieves the best performance across the evaluated metrics. On the test set, HiGraLoc reaches a Top-1 localization recall of 0.74 at the 10m threshold, outperforming the strongest baseline PMSH by an absolute margin of 0.19. Under the stricter 5m threshold, HiGraLoc also improves the Top-1 recall from 0.39 to 0.52. These results indicate that the proposed multi-level alignment strategy enables more accurate localization than methods relying mainly on global-level representations.

We further evaluate all methods on our newly constructed StreetLoc dataset. HiGraLoc again obtains the best results across all evaluation settings, achieving 0.44/0.56 Top-1 recall under the 5m/10m thresholds. Compared with the best competing methods, HiGraLoc brings clear improvements, especially at larger retrieval ranks, reaching 0.67/0.82 for Top-5 and 0.76/0.90 for Top-10. This suggests the effectiveness and generalization ability of HiGraLoc in real-world street-level scenarios. The coarse-stage submap retrieval results presented in Table 2 reveal the underlying factors behind the success of HiGraLoc. By decomposing the retrieval task into instance, relation, and global alignment, HiGraLoc captures scene structures across multiple granularities, leading to a substantial improvement in the quality of candidate selection. On the test set, HiGraLoc achieves a Top-1 retrieval recall of 0.51 and a Top-5 recall of 0.82, representing absolute gains of 0.17 and 0.17,

Table 2: Coarse-stage submap retrieval accuracy on KITTI360Pose (test set). HiGraLoc (G) denotes the variant of our method that only utilizes the global branch in the coarse stage.

Methods	Submap Retrieval Recall ↑		
	$k=1$	$k=3$	$k=5$
Text2Loc	0.28	0.49	0.58
IFRP-T2P	0.23	0.39	0.48
MambaPlace	0.31	0.53	0.62
CMMLoc	0.32	0.53	0.63
PMSH	0.34	0.56	0.65
HiGraLoc (G)	<u>0.39</u>	<u>0.62</u>	<u>0.71</u>
HiGraLoc	**0.51**	**0.75**	**0.82**

Table 3: Ablation study of coarse retrieval branches on KITTI360Pose.

Branches			Recall ↑		
Rel.	Inst.	Glo.	$k=1$	$k=3$	$k=5$
✓			0.46	0.70	0.78
	✓		0.48	0.72	0.79
		✓	0.39	0.62	0.71
✓	✓		0.50	0.74	0.81
✓		✓	0.48	0.72	0.80
	✓	✓	0.49	0.73	0.80
✓	✓	✓	**0.51**	**0.75**	**0.82**

respectively, over the previous state-of-the-art method PMSH (0.34/0.65). These results suggest that the coarse retrieval stage remains a critical bottleneck in text-to-point-cloud localization, and that our multi-level structural alignment helps mitigate the information loss inherent in methods that rely solely on global descriptors.

To validate the global design, we compare HiGraLoc (G) with global-descriptor methods in Table 2. HiGraLoc (G) achieves 0.39/0.62/0.71 Recall@1/3/5, outperforming PMSH by 0.05/0.06/0.06. These results indicate that GSGE is already effective on its own, even without the instance and relation branches. In particular, the multi-frequency spectral filtering and fusion may help preserve scene-level topology and capture complementary structural patterns across different spatial scales, making the learned global descriptor more discriminative for semantically similar submaps.

4.3 Ablation for Three Branches

To verify each branch's independent contribution, we progressively add branches and measure retrieval recall. As shown in Table 3, each single branch provides meaningful gains (Top-1: 0.39–0.48), two-branch combinations further improve performance, and the full three-branch model achieves the best Top-1 recall of 0.51, surpassing single branches by 3–12 points and two-branch combinations by 1–3 points.

The results show that the instance branch performs best alone (Top-1: 0.48), suggesting that the hierarchical object structure captured by HISE is the most discriminative signal for cross-modal matching. The relation branch (0.46) outperforms the global branch (0.39), suggesting that pairwise spatial relationships may be more informative than global structural cues for distinguishing semantically similar candidates. Although weakest alone, the global branch remains complementary: the full model improves over the relation-plus-instance setting by 1 point at Top-1, suggesting that the multi-frequency scene structure captured by GSGE is complementary to hierarchical and relational cues.

Overall, the additive gains support HiGraLoc's design of optimizing three alignment objectives and aggregating their similarity scores. The diminishing gain from the third branch suggests some cross-level redundancy, but each branch still captures distinct geometric or semantic information that cannot be fully replaced by the others.

4.4 Ablation for Key Components

4.4.1 Analysis of Instance-level Alignment

Table 4 ablates three core HISE components: w/o Poincaré ball replaces the hyperbolic manifold with Euclidean space; w/o Manifold Aggregation replaces tangent-space fusion with a standard MLP; and w/o Gated Residual removes adaptive residual fusion with the original instance features.

Replacing the Poincaré ball with Euclidean space lowers Recall@1 from 0.48 to 0.44, suggesting that hyperbolic geometry better models the hierarchical and nested structure of 3D scenes. Removing manifold aggregation yields the largest drop, reducing Recall@1 to 0.37, which suggests that tangent-

Table 4: Ablation study of Instance-level Alignment.

Method / Variation	Recall@k (↑)		
	1	3	5
w/o Poincaré ball	0.44	0.66	0.76
w/o Manifold Agg.	0.37	0.62	0.72
w/o Gated Residual	0.46	0.71	0.78
Instance (Full)	**0.48**	**0.72**	**0.79**

Table 5: Ablation study of Relation-level Alignment.

Method / Variation	Recall@k (↑)		
	1	3	5
w/o Instance Semantics	0.38	0.69	0.70
w/o Geometric Offsets	0.41	0.66	0.75
w/o Reliable Agg.	0.43	0.67	0.74
Relation (Full)	**0.46**	**0.70**	**0.78**

space instance interaction via logarithmic and exponential maps is crucial for preserving geometric integrity during global context fusion. Removing the gated residual slightly degrades performance, indicating that it helps retain original semantic information and produces a more robust instance representation.

4.4.2 Analysis of Relation-level Alignment

Table 5 evaluates the relation-level branch with three variants: w/o Instance Semantics removes object instance features, w/o Geometric Offsets removes relative position information, and w/o Reliable Aggregation replaces learnable weighting with standard max-pooling. Removing either instance semantics or geometric offsets causes a substantial performance drop, showing that discriminative relation modeling depends on the joint use of object semantics and relative spatial offsets between instances.

Reliable aggregation is also important: replacing the learnable reliability score with max-pooling reduces Recall@1 by 3 percent. This suggests that pairwise relations have unequal reliability, and the proposed scoring mechanism helps emphasize stable, discriminative spatial configurations while suppressing noisy or ambiguous ones.

4.4.3 Analysis of Global-level Alignment

Table 6: Ablation study of Global-level Alignment on KITTI360Pose.

Method / Variation	Recall@k (↑)		
	1	3	5
w/o Self-similarity Graph	0.32	0.55	0.65
w/o Low-frequency Comp.	0.37	0.61	0.69
w/o Mid-frequency Comp.	0.38	0.60	0.70
w/o High-frequency Comp.	**0.39**	0.59	0.70
Global (Full)	**0.39**	**0.62**	**0.71**

The self-similarity graph appears to be the most critical component, as its removal triggers the sharpest decline in Recall@1 from 0.39 to 0.32. This suggests that establishing a graph-based topological foundation is important for capturing holistic structural information and providing a robust geometric signature that resists local instance variations. Furthermore, removing any frequency branch degrades performance on at least one recall metric, supporting the effectiveness of low-frequency (global layouts), mid-frequency (structural patterns), and high-frequency (fine-grained details) features.

5 Conclusion

In this paper, we presented HiGraLoc, a hierarchical graph alignment framework for text-guided point-cloud localization. By decomposing coarse retrieval into instance-level hyperbolic modeling, relation-level spatial reasoning, and global-level spectral representation learning, HiGraLoc captures complementary semantic, relational, and structural cues in 3D scenes. This multi-level design improves candidate submap retrieval and provides a stronger starting point for the subsequent Text2Loc-style fine-stage regression. Extensive experiments on KITTI360Pose and StreetLoc datasets show that HiGraLoc consistently outperforms existing state-of-the-art methods. These results suggest that hierarchical cross-modal alignment is a promising direction for scalable text-guided 3D localization in complex urban environments.

Acknowledgments and Disclosure of Funding

This work was supported by the Natural Science Foundation of Shandong Province (ZR2024QF284).

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.

[2] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11080–11090, 2023.

[3] Daniele Cattaneo, Matteo Vaghi, and Abhinav Valada. Lcdnet: Deep loop closure detection and point cloud registration for lidar slam. *IEEE Transactions on Robotics*, 38(4):2074–2093, 2022.

[4] Juan Du, Rui Wang, and Daniel Cremers. Dh3d: Deep hierarchical 3d descriptors for robust large-scale 6dof relocalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 744–762. Springer, 2020.

[5] Mingtao Feng, Longlong Mei, Zijie Wu, Jianqiao Luo, Fenghao Tian, Jie Feng, Weisheng Dong, and Yaonan Wang. Partially matching submap helps: Uncertainty modeling and propagation for text to point cloud localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8296–8305, 2025.

[6] Le Hui, Hang Yang, Mingmei Cheng, Jin Xie, and Jian Yang. Pyramid point cloud transformer for large-scale place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6098–6107, 2021.

[7] Donghwi Jung, Keonwoo Kim, and Seong-Woo Kim. Gotpr: General outdoor text-based place recognition using scene graph retrieval with openstreetmap. *IEEE Robotics and Automation Letters*, 2025.

[8] Joshua Knights, Peyman Moghadam, Milad Ramezani, Sridha Sridharan, and Clinton Fookes. Incloud: Incremental learning for point cloud place recognition. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8559–8566. IEEE, 2022.

[9] Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, and Laura Leal-Taixé. Text2pos: Text-to-point-cloud cross-modal localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6687–6696, 2022.

[10] Jacek Komorowski. Minkloc3d: Point cloud based large-scale place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1790–1799, 2021.

[11] Zhenyu Li and Tianyi Shang. a^2gc: A symmetric a ggregation with geometric constraints for locally aggregated descriptors. *arXiv preprint arXiv:2511.14109*, 2025.

[12] Zhenyu Li, Tianyi Shang, Pengjie Xu, Zhaojun Deng, and Ruirui Zhang. Toward robust visual place recognition for mobile robots with an end-to-end dark-enhanced net. *IEEE Transactions on Industrial Informatics*, 2024.

[13] Zhenyu Li, Pengjie Xu, Zhenbiao Dong, Ruirui Zhang, and Zhaojun Deng. Feature-level knowledge distillation for place recognition based on soft-hard labels teaching paradigm. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[14] Baihong Lin, Zhengxia Zou, and Zhenwei Shi. Rsbev: Multi-view collaborative segmentation of 3d remote sensing scenes with bird's-eye-view representation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[15] Philipp Lindenberger, Paul-Edouard Sarlin, Jan Hosang, Matteo Balice, Marc Pollefeys, Simon Lynen, and Eduard Trulls. Scaling image geo-localization to continent level. *arXiv preprint arXiv:2510.26795*, 2025.

[16] Zhe Liu, Shunbo Zhou, Chuanzhe Suo, Peng Yin, Wen Chen, Hesheng Wang, Haoang Li, and Yun-Hui Liu. Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2831–2840, 2019.

[17] Lun Luo, Shuhang Zheng, Yixuan Li, Yongzhi Fan, Beinan Yu, Si-Yuan Cao, Junwei Li, and Hui-Liang Shen. Bevplace: Learning lidar-based place recognition using bird's eye view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8700–8709, 2023.

[18] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[19] Tianyi Shang, Zhenyu Li, Wenhao Pei, Pengjie Xu, ZhaoJun Deng, and Fanchen Kong. Mambaplace: Text-to-point-cloud cross-modal place recognition with attention mamba mechanisms. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[20] Tianyi Shang, Zhenyu Li, Pengjie Xu, Jinwei Qiao, Gang Chen, Zihan Ruan, and Weijun Hu. Bridging text and vision: A multi-view text-vision registration approach for cross-modal place recognition. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[21] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4470–4479, 2018.

[22] Kavisha Vidanapathirana, Milad Ramezani, Peyman Moghadam, Sridha Sridharan, and Clinton Fookes. Logg3d-net: Locally guided global descriptor learning for 3d place recognition. In *International Conference on Robotics and Automation (ICRA)*, pages 2215–2221. IEEE, 2022.

[23] Guangzhi Wang, Hehe Fan, and Mohan Kankanhalli. Text to point cloud localization with relation-enhanced transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2501–2509, 2023.

[24] Lichao Wang, Zhihao Yuan, Jinke Ren, Shuguang Cui, and Zhen Li. Instance-free text to point cloud localization with relative position awareness. *arXiv preprint arXiv:2404.17845*, 2024.

[25] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. Transvpr: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13648–13657, 2022.

[26] Yan Xia, Yusheng Xu, Shuang Li, Rui Wang, Juan Du, Daniel Cremers, and Uwe Stilla. Soe-net: A self-attention and orientation encoding network for point cloud based place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11348–11357, 2021.

[27] Yan Xia, Letian Shi, Zifeng Ding, Joao F Henriques, and Daniel Cremers. Text2loc: 3d point cloud localization from natural language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14958–14967, 2024.

[28] Huaiyuan Xu, Huaping Liu, Shoudong Huang, and Yuxiang Sun. C2l-pr: Cross-modal camera-to-lidar place recognition via modality alignment and orientation voting. *IEEE Transactions on Intelligent Vehicles*, 2024.

[29] Tian-Xing Xu, Yuan-Chen Guo, Zhiqiang Li, Ge Yu, Yu-Kun Lai, and Song-Hai Zhang. Transloc3d: Point cloud based large-scale place recognition using adaptive receptive fields. *arXiv preprint arXiv:2105.11605*, 2021.

[30] Yanlong Xu, Haoxuan Qu, Jun Liu, Wenxiao Zhang, and Xun Yang. Cmmloc: Advancing text-to-pointcloud localization with cauchy-mixture-model based framework. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6637–6647, 2025.

[31] Shuhang Zheng, Yixuan Li, Zhu Yu, Beinan Yu, Si-Yuan Cao, Minhang Wang, Jintao Xu, Rui Ai, Weihao Gu, Lun Luo, et al. I2p-rec: Recognizing images on large-scale point cloud maps through bird's eye view projections. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1395–1400. IEEE, 2023.

[32] Zhicheng Zhou, Cheng Zhao, Daniel Adolfsson, Songzhi Su, Yang Gao, Tom Duckett, and Li Sun. Ndt-transformer: Large-scale 3d point cloud localisation using the normal distribution transform representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5654–5660. IEEE, 2021.

A Technical appendices and supplementary material

A.1 Visualization Analysis

In this section, we further investigate the performance of HiGraLoc in 3D scene grounding through a visual presentation of prediction results on the KITTI360Pose dataset (see Figure 2 and Figure 3). These qualitative cases not only demonstrate the model's efficacy in decoding intricate linguistic instructions but also highlight its robustness in achieving semantically consistent mapping within expansive search spaces. The panels provide a structured layout of the original text queries, ground-truth references, and the corresponding Top-1 to Top-3 candidate submaps. We employ red dots to denote the actual coordinates (L_{gt}) and blue dots to mark the model's estimated points ($\boldsymbol{\eta}$). For qualitative assessment, candidate submaps where retrieval is successful and the error remains within 10 meters are highlighted with green borders; otherwise, they are marked with red borders to indicate significant deviations.

As observed from the representative samples, the integration of hierarchical spatial logic modeling enables HiGraLoc to successfully constrain the target within the top three candidates in the vast majority of scenarios. This superior retrieval recall underscores the critical role of our multi-branch alignment mechanism in bridging the cross-modal gap between natural language and 3D point cloud features. However, an attribution analysis of the rare failure cases indicates that grounding accuracy is primarily hindered by "perceptual aliasing"—where geographically distant areas (e.g., exceeding 1000 meters) with highly similar geometric topologies lead to retrieval confusion. This suggests that the discriminative capability of the coarse retrieval phase remains the primary bottleneck in large-scale urban environments. Future efforts should focus on developing more distinctive local topological representations to further enhance the reliability of global spatial anchoring.

A.2 Computational Resources

To further analyze the computational overhead of the proposed method, we compare HiGraLoc with the current state-of-the-art method PMSH in the coarse retrieval stage. Specifically, we evaluate the computational complexity of the two methods in terms of model parameters and inference runtime.

In the coarse retrieval stage, PMSH contains 372M parameters in total, including 336.39M backbone parameters and 35.61M non-backbone parameters, with an inference runtime of approximately 39 ms. In comparison, HiGraLoc contains 381.85M parameters in total, where the backbone also accounts for 336.39M parameters and the non-backbone components account for 45.45M parameters, with an inference runtime of approximately 44 ms.

It can be observed that HiGraLoc only introduces an additional 9.85M non-backbone parameters compared with PMSH, while increasing the inference runtime by only about 5 ms. This marginal overhead mainly comes from the proposed multi-level spatial feature alignment module, which is designed to more effectively model the structured correspondence between textual descriptions and point-cloud scenes. Meanwhile, since both methods adopt the same T5 text encoder and PointNet++ point-cloud encoder, the overall runtime is dominated by the backbone components. Therefore, HiGraLoc maintains a computational overhead comparable to that of PMSH.

More importantly, HiGraLoc achieves substantial performance improvements in the coarse retrieval stage without introducing significant additional computational cost. This demonstrates that the proposed structured alignment mechanism is both parameter-efficient and computationally efficient, enabling more effective cross-modal feature matching while preserving inference efficiency, and thereby leading to superior text-to-point-cloud localization performance.



Figure 2: The first set of qualitative visualization results produced by HiGraLoc.

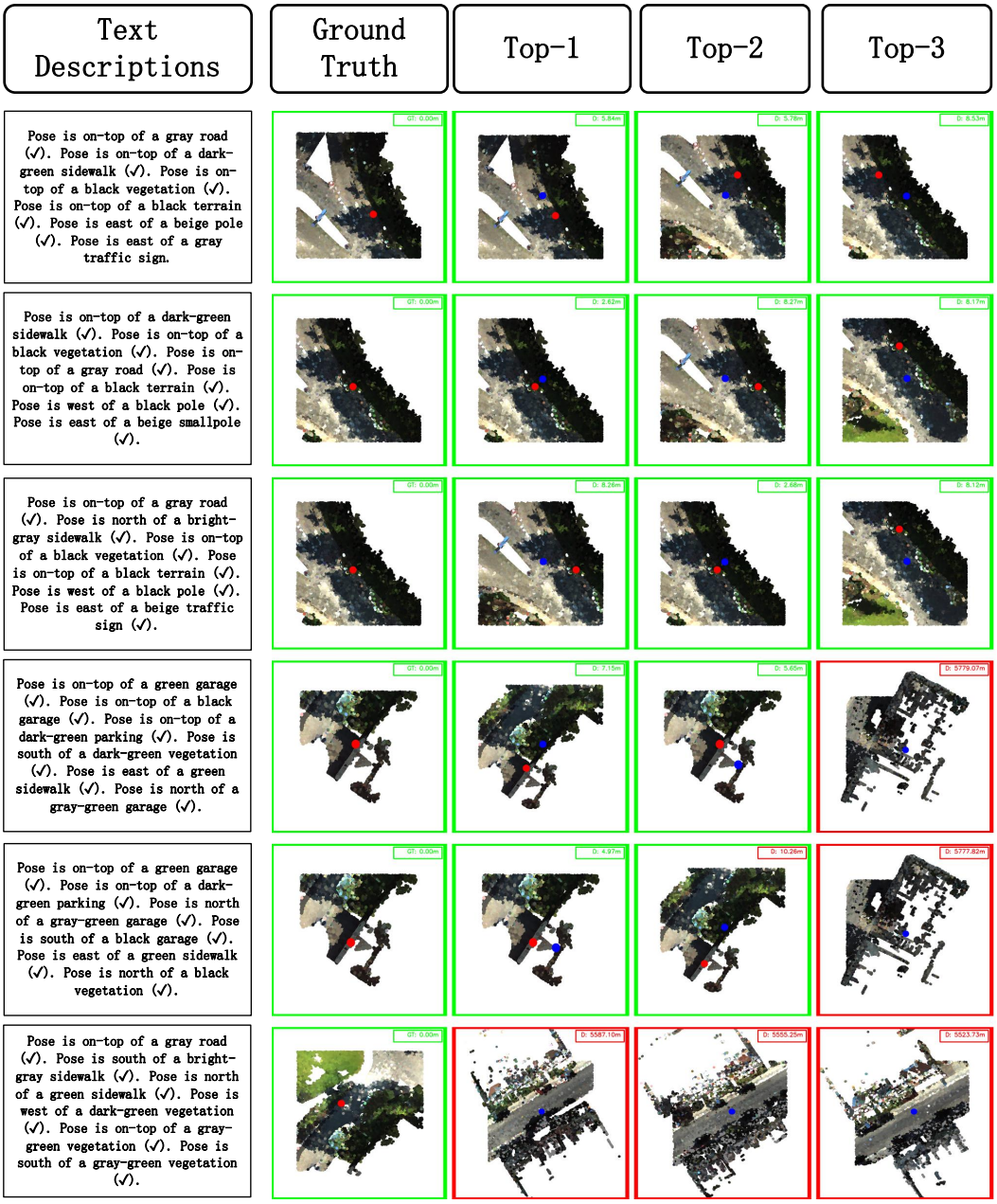

Figure 3: The second set of qualitative visualization results produced by HiGraLoc.

A.3 Limitations and Future Work

Despite the robust grounding performance of HiGraLoc, certain constraints suggest avenues for subsequent refinement. A primary limitation lies in the framework's current reliance on a closed-set semantic taxonomy for instance extraction and alignment. Real-world urban landscapes, however, are characterized by an almost infinite variety of physical entities, paired with the inherent ambiguity and unconstrained nature of natural language. Consequently, the precision of scene association may be compromised when the system encounters out-of-distribution objects or highly idiosyncratic linguistic patterns that deviate from the training corpus.

To mitigate these issues, our future research will emphasize the integration of Large Language Models (LLMs) and open-vocabulary 3D perception strategies. By harnessing the extensive common-sense reasoning and semantic elasticity of foundation models, we seek to decouple the grounding process from fixed category definitions. This transition will enable the architecture to dynamically associate arbitrary, free-form textual queries with novel 3D instances, thereby significantly enhancing the scalability and generalization of cross-modal interaction in unstructured and heterogeneous environments.

A.4 Broader Impacts

The introduction of HiGraLoc establishes a novel technical pathway for the active perception and spatial cognition of autonomous agents within complex urban environments. By integrating the logical depth of natural language with the structural information of 3D point clouds, this research not only bolsters the operational resilience of autonomous delivery and navigation systems in extreme conditions—such as low-light or smoke-filled scenarios where traditional visual sensors often fail—but also enables the semantic management of large-scale urban infrastructure. Such a language-guided scene anchoring mechanism significantly diminishes the reliance of automated systems on high-cost, high-bandwidth sensor data, thereby propelling the evolution of intelligent robotics toward greater energy efficiency and reliability.

On a societal level, this work substantially enhances the inclusivity and intuitiveness of human-robot interaction. By enabling non-technical users to direct service robots through natural language instructions, the technology effectively bridges the digital divide, offering more accessible and convenient assistive living solutions for the elderly and the visually impaired. Furthermore, compared to conventional video-based navigation, the semantic association paradigm between text and point clouds holds a distinct advantage in privacy preservation for public space applications. It establishes spatial mappings through abstract topological structures rather than sensitive imagery, thereby better safeguarding individual privacy boundaries while maintaining public safety.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are consistent with the proposed method and experimental evidence. The paper introduces HiGraLoc with instance-level, relation-level, and global-level alignment branches, and the reported results on KITTI360Pose and StreetLoc support the claimed improvements over prior methods.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a "Limitations and Future Works" section (Appendix A.3) discussing the reliance on a closed-set semantic taxonomy and possible performance degradation under out-of-distribution objects or unconstrained language expressions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include formal theoretical results such as theorems, lemmas, or propositions that require proofs. The equations are used to define the proposed model components and training objective.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides the main model design, datasets, splits, training epochs, hardware, evaluation metrics in Section 4.1. We also provide open-source code.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The abstract provides an anonymized repository link. The repository includes the source code, data access and preparation instructions, environment configuration, and training/evaluation scripts for reproducing the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the datasets, train/validation/test splits, evaluation metrics, training stages, number of epochs, hardware configuration, and baseline settings in Section 4.1. Additional implementation details are provided in the anonymized repository.

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The current paper reports point estimates but does not include error bars, confidence intervals, or statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the workstation configuration, including a 128-core AMD EPYC CPU and an NVIDIA RTX V100 32GB GPU in Section 4.1. The Appendix A.2 further reports model parameter counts and inference runtime for the coarse retrieval stage.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work does not involve human-subject experiments or high-risk release, and the research is intended for 3D localization and embodied AI. The authors should ensure the data collection and release comply with relevant laws and institutional requirements.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Section A.4 in Appendix discusses potential societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk generative models, pretrained language models, or scraped internet datasets. The released assets are intended for text-guided 3D localization research.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses existing benchmark datasets and methods with proper citations, including KITTI360Pose and prior text-to-point-cloud localization baselines such as Text2Pos, RET, Text2Loc, MambaPlace, CMMLoc, and PMSH. The datasets and baselines are credited through citations.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have uploaded the details of all assets to our open-source repository. The repository provides documentation for data preparation, code usage, training, and evaluation to support reproducibility.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper is purely computational/theoretical and does not involve any human subjects, or crowdsourcing experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not use LLMs as an important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.