Two-Stage Hint–Object Alignment for Text-to-PointCloud Localization

Doyeon Kim[1] and Giseop Kim[1†]

***Abstract*— Localization from natural language descriptions in large-scale environments remains challenging due to the inherent ambiguity and incompleteness of linguistic cues. A system must infer target locations from partially described surroundings without explicit spatial coordinates while distinguishing semantically relevant objects from numerous irrelevant scene elements. Most existing approaches address this task by encoding text and 3D submaps into global descriptors for efficient retrieval. However, such representations aggregate scene objects in a query-independent manner, entangling relevant and irrelevant elements and losing fine-grained hint-object correspondences. To address this limitation, we introduce a two-stage hint-object alignment module that establishes fine-grained correspondences between textual hints and candidate scene objects. Specifically, the first stage performs null-aware alignment, allowing each hint to opt out of matching when no corresponding object exists, while the second stage performs context-refined alignment that enriches the weighted object representations via hint-conditioned cross-attention. Experiments on the KITTI360Pose dataset demonstrate that the proposed approach improves retrieval accuracy over existing baselines, highlighting the effectiveness of hint-object alignment for text-to-pointcloud localization.**

I. Introduction

Localization from natural language descriptions is a crucial capability for autonomous systems operating in human-centered environments. Unlike coordinate-based localization, the system must interpret descriptive cues and identify the corresponding region in the scene. However, grounding language in complex environments is inherently challenging because textual descriptions typically mention only a few salient objects, while each candidate submap may contain many unrelated scene elements. Consequently, reliable localization requires determining which scene objects correspond to each textual hint rather than representing the entire scene with a holistic descriptor.

Most existing approaches address this problem by encoding textual descriptions and candidate 3D submaps into global descriptors and comparing them through embedding similarity. Such representations enable scalable localization in large-scale environments and have demonstrated promising performance in several studies [1, 2, 3]. However, global aggregation treats all scene objects in a query-independent manner, allowing irrelevant elements to influence the matching score and obscure the correspondence between textual hints and scene objects. As illustrated in Fig. 1, this makes it difficult for the system to determine which objects actually ground the given description.

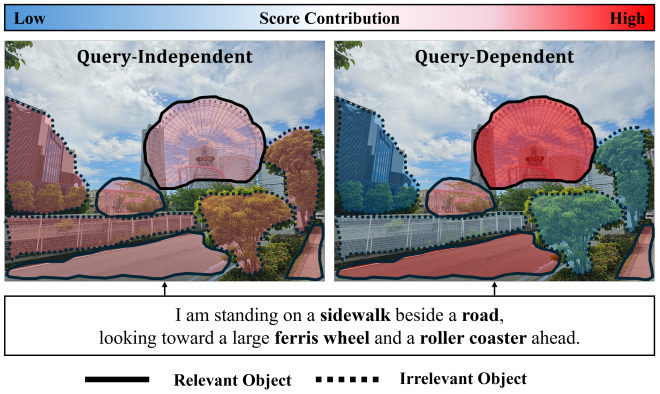

Fig. 1. Conceptual illustration of the fundamental challenge in natural language-based localization. In a query-independent manner (left), all objects in the scene contribute similarly to the representation, causing irrelevant objects (e.g., trees, fences, and roads) to influence the matching score and dilute the discriminative landmark cues. As a result, the representation is dominated by background objects. In contrast, a query-dependent manner (right) selectively emphasizes objects relevant to the textual description (e.g., ferris wheel and roller coaster) while suppressing irrelevant regions. The color overlay indicates the score contribution of each object, ranging from low (blue) to high (red).

Selective aggregation and fine-grained alignment have been widely explored in several related domains. In visual place recognition, recent methods employ selective aggregation strategies to suppress non-informative local features in global representations [4, 5, 6]. In vision-language retrieval [7, 8, 9, 10], attention-based architectures align words with image regions to enable precise cross-modal matching. These observations suggest that explicitly aligning language with relevant scene elements can provide more reliable grounding than holistic scene representations. However, such fine-grained alignment mechanisms remain largely unexplored in text-to-pointcloud localization.

In this work, we argue that text-to-pointcloud localization relies on aligning linguistic hints with scene objects rather than comparing global scene descriptors. Instead of treating each submap as a holistic representation, a localization system should explicitly identify which objects correspond to each linguistic hint and which objects should be ignored. This perspective naturally motivates a fine-grained alignment framework that associates linguistic hints with relevant scene objects while suppressing irrelevant ones.

To this end, we propose a two-stage hint-object alignment framework for text-to-pointcloud localization. The first stage

†Corresponding author.
[1]Doyeon Kim and Giseop Kim are with the Department of Robotics and Mechatronics Engineering, DGIST, Daegu, Republic of Korea [doyeon.kim, gsk]@dgist.ac.kr
This work was supported by the InnoCORE program of the Ministry of Science and ICT (26-InnoCORE-01).

performs *null-aware alignment*, allowing each linguistic hint to abstain from matching when no corresponding object exists in a candidate submap. The second stage performs *context-refined alignment*, enriching the weighted object representations through hint-conditioned cross-attention.

Our contributions are summarized as follows:

- We identify the limitations of global descriptor retrieval in text-to-pointcloud localization and reformulate the task as a hint-object alignment problem.
- We propose a two-stage hint-object alignment framework consisting of null-aware alignment that allows hints to abstain from matching absent objects, and context-refined alignment that enriches weighted object representations via hint-conditioned cross-attention.
- Experiments on the KITTI360Pose dataset demonstrate that the proposed approach improves retrieval accuracy over existing baselines.

II. Related Works

A. Text-to-PointCloud Localization

Text-to-pointcloud localization aims to estimate a target location in a 3D map from natural language descriptions. Text2Pos [1] first introduced this task together with a benchmark dataset and baseline models. Subsequent works have focused on improving cross-modal representation learning between textual descriptions and 3D scenes.

RET [2] models relationships among scene objects to align textual relations with scene structure. Text2Loc [3] introduces a hierarchical transformer encoder for coarse localization and a matching-free strategy for fine localization. MNCL [11] improves representation learning through multi-level negative contrastive learning and performs retrieval using multi-level similarity matching. CMMLoc [12] employs a Cauchy Mixture Model–based Transformer to mitigate the influence of irrelevant objects during cross-modal matching. PMSH [13] addresses inconsistencies in textual descriptions through partially described data augmentation, while MambaPlace [14] improves efficiency and scalability using Mamba-based sequence modeling.

Despite these advances, many methods still compress a candidate submap into a single global representation for retrieval. Such query-independent aggregation entangles signals from both relevant and irrelevant objects, making it difficult to capture fine-grained correspondences between textual hints and scene objects.

B. Selective Alignment for Retrieval

In visual place recognition, aggregating local features into a compact global descriptor is a core component of retrieval pipelines. NetVLAD [4] performs soft assignment of local descriptors to learnable clusters. More recent methods introduce mechanisms to suppress non-informative features during aggregation: SALAD [5] employs optimal transport with a dustbin cluster to discard uninformative descriptors, and BoQ [6] uses learnable global queries to selectively attend to relevant features.

In image-text retrieval, fine-grained alignment between visual regions and textual words has proven more effective than global embedding comparison. SCAN [7] introduced stacked cross-attention to discover latent alignments between image regions and words, aggregating pairwise similarities to infer overall image-text correspondence. SGRAF [8] extended this line by introducing similarity graph reasoning to capture relationships among local alignments, along with an attention filtration module that suppresses less meaningful correspondences. CHAN [9] further observed that cross-attention produces redundant region-word pairs, and proposed hard alignment to retain only the most relevant correspondences while discarding the rest. More recently, SFAN [10] introduced modality-specific filter modules to remove irrelevant features within each modality before alignment, combined with a state-space model-based module for selective cross-modal matching.

Our work extends this principle to text-to-pointcloud localization, where candidate submaps similarly contain many objects unrelated to the query.

III. Methodology

A. Preliminaries

Following the hierarchical coarse-to-fine paradigm introduced in [1], we consider the coarse retrieval stage for text-to-pointcloud localization. Given a large-scale 3D map $\mathcal{M} = \{m_i\}_{i=1}^{N}$, each submap m_i contains a set of object instances $\{o_{ij}\}_{j=1}^{J_i}$. Each object o_{ij} is represented by its pointcloud $\mathbf{P}_{ij} \in \mathbb{R}^{n_{ij}\times 3}$ and semantic label $l_{ij} \in \mathcal{C}$. Additional attributes such as color, spatial position, and point count are also considered. For simplicity, we omit the submap index and denote a submap as $m = \{o_j\}_{j=1}^{J}$.

Given a query text description $\mathcal{T} = \{h_k\}_{k=1}^{K}$, where each hint h_k describes a spatial relationship between the target location and a nearby object, the goal of coarse retrieval is to identify the most relevant submap:

$$m^* = \arg\max_{m_i \in \mathcal{M}} S(\mathcal{T}, m_i) \tag{1}$$

where $S(\mathcal{T}, m)$ denotes the text–submap matching score.

B. Hint–Object Feature Extraction

Following the baseline feature extraction pipeline in [3], we employ the same text and submap encoders to extract hint and object features. Unlike the baseline, which aggregates individual embeddings into global descriptors, we retain hint-level and object-level embeddings.

Given a query text $\mathcal{T}$ with K hints and a submap m with J objects, the encoders produce hint embeddings $\mathbf{H} = \{\mathbf{h}_k\}_{k=1}^{K}$ and object embeddings $\mathbf{O} = \{\mathbf{o}_j\}_{j=1}^{J}$, where $\mathbf{h}_k, \mathbf{o}_j \in \mathbb{R}^D$. The resulting features are projected into a shared embedding space and ℓ_2-normalized before computing similarity.

C. Two-stage Hint–Object Alignment

Given the normalized hint embeddings $\mathbf{H}$ and object embeddings $\mathbf{O}$, we compute the text–submap matching score through two sequential stages.

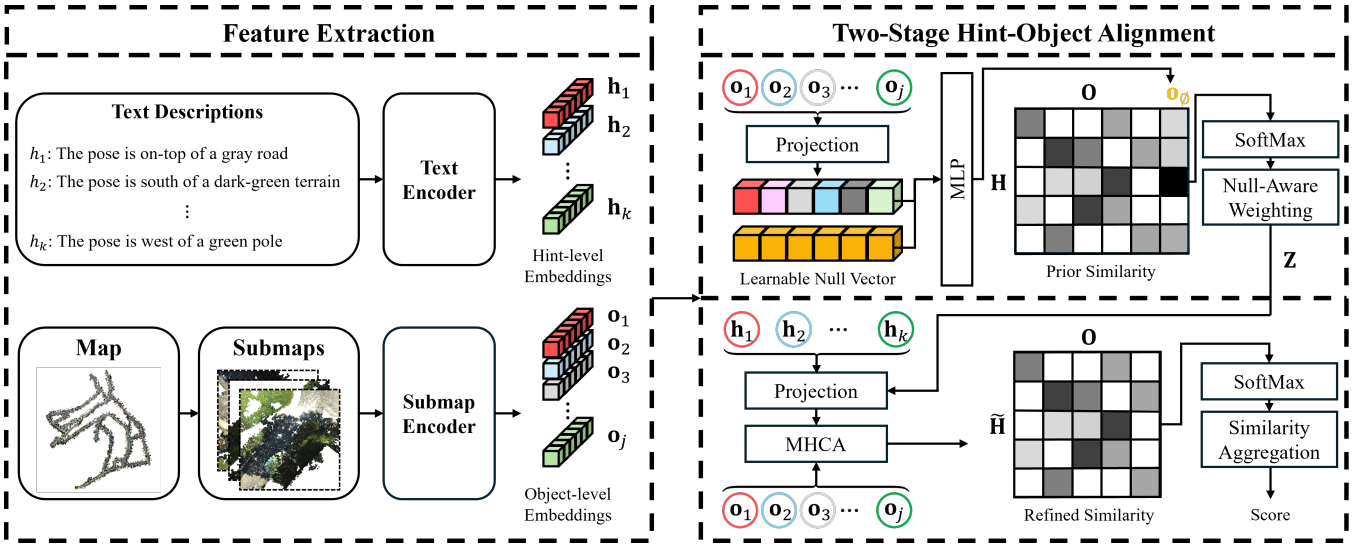

Fig. 2. Overview of the proposed two-stage hint–object alignment pipeline. Text descriptions are encoded into hint-level embeddings, while each submap is represented by object-level embeddings. In Stage 1, null-aware alignment estimates hint–object alignment while accounting for missing objects via a submap-conditioned null token. In Stage 2, context-refined alignment refines these alignments through hint-conditioned attention over scene objects. Finally, similarity aggregation combines the refined hint–object similarities to produce the query–submap matching score.

1) Null-Aware Alignment: A candidate submap may be spatially close to the ground-truth location while not containing all objects mentioned in the query. Forcing every hint to match an object in such cases introduces false correspondences. To handle this, we augment the object set with a learnable *null token* $\mathbf{o}_\varnothing$ that accounts for ungrounded hints, forming the augmented set $\mathbf{O}^+ = \{\mathbf{o}_1, \ldots, \mathbf{o}_J, \mathbf{o}_\varnothing\}$.

We compute hint–object similarities over $\mathbf{O}^+$, apply a learnable temperature scaling and an additional learnable bias to the null token, and obtain alignment weights α_{kj} via softmax. The per-hint score is then:

$$s_k = \gamma_k \sum_{j=1}^{J} \alpha_{kj} (\mathbf{h}_k^\top \mathbf{o}_j) \tag{2}$$

where $\gamma_k = 1 - \alpha_{k,J+1}$ is the support, measuring how strongly hint k is grounded in the candidate submap. When no matching object exists, the null token receives most of the alignment weight, driving γ_k toward zero and suppressing the contribution of that hint.

The null token itself is submap-conditioned: we concatenate the mean-pooled object representation $\bar{\mathbf{o}}$ with a learnable base vector $\mathbf{n} \in \mathbb{R}^D$, pass the result through a two-layer MLP, and ℓ_2-normalize the output. This allows the null token's representation to adapt to each candidate submap.

2) Context-Refined Alignment: The alignment scores from Stage 1 rely solely on independent pairwise similarities between hints and objects. However, objects in a scene often exhibit contextual co-occurrence patterns that provide additional cues for resolving ambiguous matches. Rather than encoding explicit spatial relations, the second stage captures co-occurrence patterns implicitly through hint-conditioned attention. Each hint guides the attention to focus on objects that contextually co-occur with it.

For each hint k, we first compute the Stage 1 weighted object representation $\mathbf{z}_k = \sum_{j=1}^{J} \alpha_{kj}\, \mathbf{o}_j$, which aggregates object embeddings using the alignment weights from Stage 1 and serves as the contextual signal connecting the two stages. We then form a query $\mathbf{q}_k = \mathbf{z}_k + W_p \mathbf{h}_k$ and compute attention over the object set:

$$\mathbf{c}_k = \mathrm{softmax}\left(\frac{(W_Q \mathbf{q}_k)(W_K \mathbf{O})^\top}{\sqrt{d_h}} \right) W_V \mathbf{O} \tag{3}$$

where W_Q, W_K, and W_V are learnable projection matrices and d_h denotes the head dimension.

The contextual signal is integrated through a gated residual connection using $g_k = \sigma(W_g[\mathbf{h}_k; \mathbf{z}_k])$, yielding the enriched representation:

$$\mathbf{e}_k = \mathrm{LayerNorm}(\mathbf{q}_k + g_k \cdot \mathbf{c}_k) \tag{4}$$

We then recompute alignment weights α'_{kj} between the ℓ_2-normalized $\mathbf{e}_k$ and the object set, and aggregate the refined hint–object similarities into the overall text–submap matching score:

$$S(\mathcal{T}, m) = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha'_{kj} (\mathbf{e}_k^\top \mathbf{o}_j) \tag{5}$$

D. Loss Function

We adopt the symmetric cross-modal contrastive loss in [3]. Given a mini-batch of B paired text–submap samples, the model computes pairwise matching scores S_{ij} for all text–submap pairs. The training objective is defined as:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^{B} \left(\log \frac{\exp(S_{ii}/\tau_\mathcal{L})}{\sum_j \exp(S_{ij}/\tau_\mathcal{L})} + \log \frac{\exp(S_{ii}/\tau_\mathcal{L})}{\sum_j \exp(S_{ji}/\tau_\mathcal{L})} \right) \tag{6}$$

where $\tau_\mathcal{L}$ is a predefined temperature parameter.

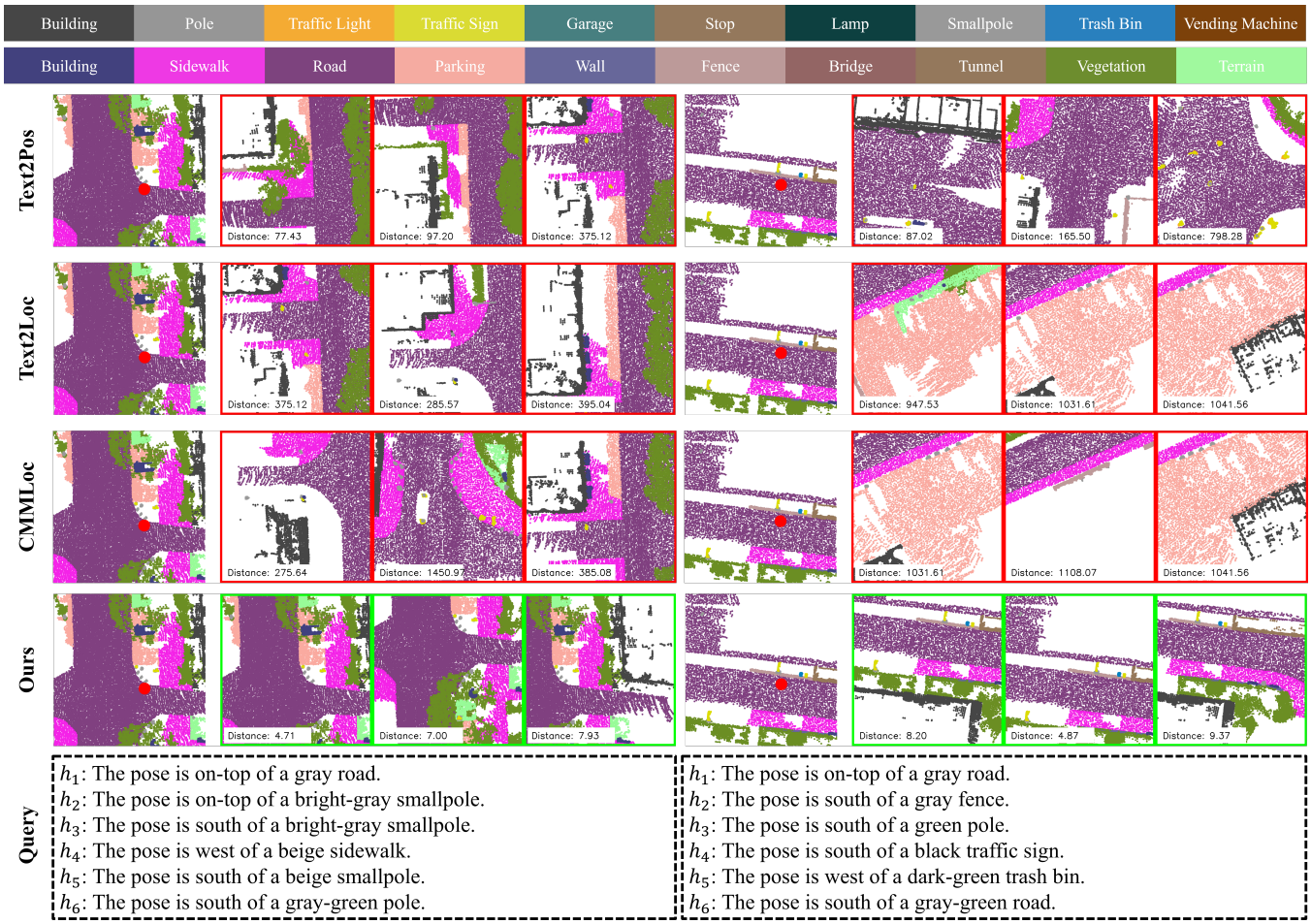

Fig. 3. Qualitative retrieval results on the KITTI360Pose dataset. Each query consists of a natural language description, and the top-3 retrieved submaps are shown with their distances to the ground-truth location. Retrievals within 10 m are marked in green and others in red.

IV. Experimental Results

1) Implementation Details: The model is trained for 20 epochs using the AdamW optimizer with a batch size of 32. The initial learning rate is set to 5×10^{-4}, and a step learning rate scheduler is applied with a decay step of 7 epochs and a decay factor of 0.4. The softmax temperature is initialized to 0.1. All experiments are conducted on a single NVIDIA RTX 5080.

2) Benchmark Dataset: We conduct experiments on the KITTI360Pose benchmark [1], which provides language descriptions aligned with large-scale urban 3D maps. The dataset contains pointclouds from nine urban scenes and includes 43,381 text–location query pairs covering a total area of 15.51 km^2. The scenes are split into five training scenes, one validation scene, and three testing scenes. The 3D map is partitioned into overlapping submaps by sliding a cubic window of size 30 m with a stride of 10 m. This preprocessing results in 11,259, 1,434, and 4,308 submaps for the training, validation, and testing sets, respectively, with 17,001 submaps in total.

3) Evaluation Metrics: We evaluate retrieval performance using Recall@k with $k \in \{1, 3, 5\}$, which measures whether the ground-truth submap appears within the top-k retrieved candidates. To analyze the relative strengths of different methods, we define two complementary statistics based on Top-1 retrieval results. The *Recovery Ratio* measures the fraction of queries where a baseline fails but our method succeeds at Recall@1, while the *Reverse Ratio* measures the fraction of queries where our method fails but the baseline succeeds.

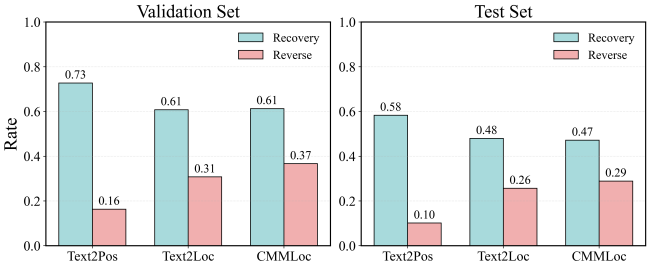

Fig. 4. Recovery and Reverse ratios comparing our method with baseline models.

TABLE I. Submap retrieval accuracy measured by Recall@k.

Method	Validation Set			Test Set			Time (ms)
	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$	
Text2Pos [1]	0.14	0.28	0.37	0.12	0.25	0.33	1
RET [2]	0.18	0.34	0.44	–	–	–	–
Text2Loc [3]	0.32	0.56	0.67	0.28	0.49	0.58	8
MNCL [11]	0.50	0.75	0.84	0.44	0.67	0.75	46
CMMLoc [12]	0.35	0.61	0.73	0.32	0.53	0.63	9
PMSH [13]	0.37	0.63	0.73	0.34	0.56	0.65	–
MambaPlace [14]	0.35	0.61	0.72	0.31	0.53	0.62	10
Ours	0.48	0.76	0.84	0.44	0.69	0.77	12

* Baseline results are reported from their original papers.
* Time denotes the average retrieval latency per query after map encoding.

A. Text-to-PointCloud Localization Performance

We evaluate the proposed method on the KITTI360Pose benchmark and report submap retrieval performance in terms of Recall@k. Compared with existing baselines, our method consistently achieves competitive retrieval performance across both validation and test sets. As shown in Table I, the proposed approach attains the best Recall@3 and Recall@5 on both splits, while achieving comparable performance to the state-of-the-art method on Recall@1. In addition, our method requires only 12 ms per query, compared with 46 ms for MNCL, while reducing latency by 3.8×.

Fig. 3 presents qualitative retrieval examples. Notably, our method successfully retrieves correct submaps even when many hints refer to objects that occupy only a small portion of the submap. Because such objects contribute weakly to global scene descriptors, baseline methods often fail to exploit them effectively. In contrast, our method treats them as explicit alignment targets, enabling more reliable retrieval.

Finally, Fig. 4 compares the Recovery and Reverse ratios between our method and baseline models. Across all comparisons, the Recovery ratio consistently exceeds the Reverse ratio. On the validation set, the Recovery-to-Reverse ratios are approximately $4.6\times$, $2.0\times$, and $1.6\times$ when compared with Text2Pos [1], Text2Loc [3], and CMMLoc [12], respectively. On the test set, these ratios further increase to $5.8\times$, $1.8\times$, and $1.6\times$. This asymmetry indicates that our method is substantially more effective at recovering baseline failures than introducing additional retrieval errors.

V. Conclusion

We presented a two-stage hint–object alignment framework for text-to-pointcloud localization that establishes correspondences between textual hints and scene objects without relying on global scene descriptors. The first stage introduces a null-aware alignment mechanism that allows hints to abstain from matching when no corresponding object exists, while the second stage refines the resulting object representations through hint-conditioned cross-attention.

Experiments on the KITTI360Pose dataset demonstrate that the proposed approach improves text-to-pointcloud retrieval performance over existing baselines. Recovery and Reverse analysis further indicates that these improvements are broadly distributed across queries rather than concentrated on a small subset.

Overall, our results highlight the effectiveness of explicitly aligning linguistic hints with scene objects for robust text-to-pointcloud localization.

References

[1] M. Kolmet, Q. Zhou, A. Ošep, and L. Leal-Taixé, "Text2pos: Text-to-point-cloud cross-modal localization," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6677–6686, 2022.

[2] G. Wang, H. Fan, and M. Kankanhalli, "Text to point cloud localization with relation-enhanced transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 2501–2509, 2023.

[3] Y. Xia, L. Shi, Z. Ding, J. F. Henriques, and D. Cremers, "Text2loc: 3d point cloud localization from natural language," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14958–14967, 2024.

[4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016.

[5] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17658–17668, 2024.

[6] A. Ali-bey, B. Chaib-draa, and P. Giguère, "BoQ: A place is worth a bag of learnable queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17794–17803, June 2024.

[7] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 201–216, 2018.

[8] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 1218–1226, 2021.

[9] Z. Pan, F. Wu, and B. Zhang, "Fine-grained image-text matching by cross-modal hard aligning network," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19275–19284, 2023.

[10] Y. Huang, Z. Liu, S. Sun, N. Cui, and J. Li, "Sfan: Selective filter and alignment network for cross-modal retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 10, pp. 18792–18804, 2025.

[11] D. Liu, S. Huang, W. Li, S. Shen, and C. Wang, "Text to point cloud localization with multi-level negative contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 5397–5405, 2025.

[12] Y. Xu, H. Qu, J. Liu, W. Zhang, and X. Yang, "Cmmloc: Advancing text-to-pointcloud localization with cauchy-mixture-model based framework," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6637–6647, 2025.

[13] M. Feng, L. Mei, Z. Wu, J. Luo, F. Tian, J. Feng, W. Dong, and Y. 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*, pp. 8296–8305, 2025.

[14] T. Shang, Z. Li, P. Xu, and J. Qiao, "Mambaplace: Text-to-point-cloud cross-modal place recognition with attention mamba mechanisms," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 16985–16992, 2025.