

PARC-Loc: Text-to-Point-Cloud Localization with Partial Assignment and Relational Consistency

Shengkai Ma Zhenyu Hou Weihua Cao

Abstract

Text-to-point-cloud localization estimates a position in a city-scale 3D map from descriptions of surrounding objects. Existing coarse-to-fine methods retrieve submaps using aggregate learned compatibility and then localize within a selected submap. However, repetitive or similar urban objects can inflate the embedding similarity between the query and multiple submaps, even when the instance layout within a submap violates the query description. Meanwhile, query-relevant instances often span submap boundaries, leaving the retrieved submap with incomplete contextual evidence. We term these failure modes *layout-inconsistent aliasing* and *boundary evidence incompleteness*, respectively. To address them, we propose **PARC-Loc**, a coarse-to-fine localization framework built on *Partial Assignment with Relational Consistency* (**PARC**). PARC jointly models hint-object compatibility and pairwise spatial relations, allowing unmatched elements while favoring assignments consistent with the queried layout. At the coarse stage, its candidate-level assessment complements neural similarity for layout-consistent submap selection. At the fine stage, the context is expanded with query-relevant instances from adjacent submaps, while PARC yields object-level matching weights that guide cross-modal attention. Extensive experiments on KITTI360Pose and CityLoc show that PARC-Loc outperforms conventional coarse-to-fine baselines. On KITTI360Pose, our method improves Top-1 localization recall at 5 m from 0.50 to 0.67, achieving a 34% relative gain over the strongest baseline.

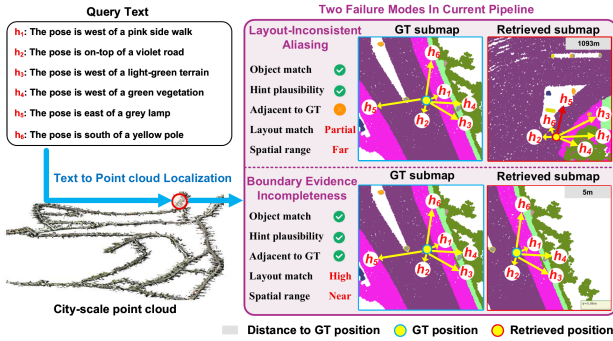


Figure 1: Two failure modes in the standard coarse-to-fine decision structure.

1 Introduction

Text-to-point-cloud (T2P) localization estimates a position in a city-scale 3D map from linguistic descriptions of nearby landmarks (Kolmet et al., 2022). Unlike referred-object grounding, the target does not necessarily correspond to an observed instance. For example, a query such as “west of the green building and east of the gray road” specifies the target through the joint configuration of several surrounding objects. This provides a language-based interface for locating without exact addresses or GPS coordinates, which is promising for applications in autonomous driving and embodied-agent navigation (Anderson et al., 2018; Deruyttere et al., 2019). However, it remains challenging to realize such capabilities: grounding these descriptions in a large-scale point-cloud map requires jointly aligning object semantics and spatial relations across language and point clouds (Chen et al., 2022).

To address this challenging task, most existing systems follow a coarse-to-fine design introduced by Text2Pos (Kolmet et al., 2022): a neural retriever first selects candidate submaps, after which a fine localizer predicts a position within a selected

submap. Subsequent methods advance this pipeline along several dimensions. For spatial structure encoding, RET explicitly models relations at the hint and instance levels, while SpatiaLoc constructs spatial descriptors at instance and global levels (Wang et al., 2023; Shang et al., 2026b). For cross-modal representation alignment, Text2Loc adopts global text-submap contrastive learning, whereas MNCL extends negative contrastive supervision to the global, instance, and relation levels (Xia et al., 2024; Liu et al., 2025). For robustness to partial or ambiguous correspondence, CMMLoc captures semantic interactions under partial relevance, while PMSH propagates localization uncertainty across the coarse and fine stages (Xu et al., 2025; Feng et al., 2025). Despite these advances, these methods generally retain the same decision making: coarse retrieval ranks candidates by an aggregate query-submap similarity, while fine localization treats each selected submap as a self-contained evidence domain. Two limitations exist within this structure: it implicitly equates aggregate similarity with a jointly valid layout and the selected submap with complete localization context.

The first limitation can lead to *layout-inconsistent aliasing*. As illustrated in Fig. 1, common object categories and similar appearances across locations can produce high aggregate query-submap similarity through individually plausible object matches, even when their joint spatial configuration is inconsistent with the relational layout described by the query. Although relation-aware encoders incorporate spatial cues (Wang et al., 2023; Shang et al., 2026b), the final score does not explicitly verify relational consistency at the candidate level. Specifically, these methods do not check whether a query-to-instance configuration satisfies the relational constraints in the query. This motivates an explicit assessment of relational consistency over candidate-specific partial assignments during coarse selection.

The second limitation can lead to *boundary evidence incompleteness*. Cell-based submap construction may distribute query-relevant instances across neighboring submaps, leaving any single submap with incomplete contextual evidence.

This loss of context can make neighboring candidates harder to distinguish and exacerbate the text-submap matching uncertainty studied in prior work (Xu et al., 2025; Feng et al., 2025). Incorporating instances from adjacent submaps can recover cross-boundary evidence. However, it may also introduce distractors whose individual object matches appear plausible while their relational configurations are inconsistent with the query. Fine localization should therefore preserve the selected submap as the coordinate anchor while selectively incorporating query-relevant context from adjacent submaps.

Although caused by different structural choices, the two failures create the same matching requirement: determining which subset of individually plausible hint-object correspondences forms a coherent explanation of the query layout. At the coarse stage, this assessment differentiates coherent candidates from aliases supported only by isolated coincidences; after neighboring-object aggregation makes cross-boundary evidence available, it separates useful context objects from the accompanying distractors. The required mechanism must therefore allow unsupported hints and irrelevant objects to remain unmatched while jointly evaluating unary attribute compatibility and pairwise layout consistency. We therefore propose PARC-Loc, a text-to-point-cloud localization framework built around *Partial Assignment with Relational Consistency* (PARC). PARC constructs a soft partial hint-object assignment that jointly evaluates attribute compatibility and pairwise spatial consistency while leaving missing or irrelevant elements unmatched. This mechanism is applied to both coarse submap retrieval and fine localization. In the coarse stage, PARC evaluates each neural proposal and combines its explicit candidate-level consistency with the neural retrieval score for submap selection. In the fine stage, the selected submap defines the coordinate frame, while objects collected from neighboring cells form an expanded evidence set. Reapplying PARC to this set produces object-level matching weights that bias cross-attention toward query-relevant context.

Taken together, these two stage-specific applications of PARC yield substantial gains. On the KITTI360Pose test set, PARC-Loc improves Top-1 @5m localization recall from 0.50 to 0.67, a 34% relative improvement over MNCL. We further conduct transfer and zero-shot experiments on CityLoc-K and CityLoc-C, where PARC-Loc achieves competitive performance. Module-wise ablations support the individual contributions of our method. In a controlled single-scene stress test, the PARC selector retains 96.7% of its Top-1 retrieval recall as the gallery grows from 819 to 2,549 submaps, compared with 83.8–89.7% for reported baselines.

Our main contributions are summarized as follows:

- We thoroughly analyze the standard coarse-to-fine pipeline adopted in T2P localization and diagnose two structural failure modes: *layout-inconsistent aliasing* during coarse submap retrieval and *boundary evidence incompleteness* during fine localization.
- We propose PARC-Loc, a novel language-based localization framework. Within this framework, *Layout-Consistent Submap Selection* and *Cross-Boundary Context Localization* specifically mitigate the two structural failure modes

through *Partial Assignment with Relational Consistency* (PARC), a soft hint-object assignment that jointly evaluates semantic and spatial compatibility under partial correspondence.

- Extensive experiments on the KITTI360Pose dataset demonstrate that our method outperforms state-of-the-art methods. Experiments on the CityLoc-K and CityLoc-C datasets and under controlled gallery growth further assess the transfer and robustness of our method.

2 Related Work

Text-to-point-cloud localization. Text2Pos establishes the prevailing coarse-to-fine formulation on KITTI360Pose, retrieving candidate submaps before estimating a target offset from hint-instance associations (Kolmet et al., 2022). Subsequent methods address different sources of cross-modal ambiguity: RET models instance relations, Text2Loc adopts stronger language representations and matching-free localization, and MNCL introduces multi-level negative contrastive learning over global, instance, and relation cues (Wang et al., 2023; Xia et al., 2024; Liu et al., 2025). CMMLoc and PMSH model partial correspondence and uncertainty (Xu et al., 2025; Feng et al., 2025), while SpatiaLoc and SympLoc strengthen spatial descriptors and alignment geometry (Shang et al., 2026b; Shang and Li, 2026). Vehicle-SceneInteraction studies text-driven 3D LiDAR place recognition for autonomous driving (Shang et al., 2026a), whereas VLM-Loc performs localization through BEV maps and scene graphs (Kang et al., 2026). These methods improve the representation, training, or retrieval of submap compatibility. PARC-Loc instead validates joint layouts during coarse selection and prevents the selected submap boundary from limiting fine-stage evidence.

Language-guided 3D spatial reasoning. Language-guided 3D grounding identifies objects or regions from textual descriptions (Huang et al., 2022; Jain et al., 2022). ScanRefer and ReferIt3D establish indoor benchmarks (Chen et al., 2020; Achlioptas et al., 2020), followed by methods for relation modeling, spatial-language alignment, progressive point selection, and dense alignment (He et al., 2021; Zhao et al., 2021; Roh et al., 2022; Luo et al., 2022; Wu et al., 2023). CityRefer extends grounding to city-scale point clouds (Miyanishi et al., 2023). Unlike referred-object grounding, the target in T2P localization may not correspond to an observed instance; it must be inferred from the joint configuration of surrounding objects. This distinction motivates validating hint-object correspondences jointly rather than grounding each hint independently.

3 Preliminaries and Pipeline Diagnosis

We formulate text-to-point-cloud localization and diagnose two limitations of its standard coarse-to-fine pipeline.

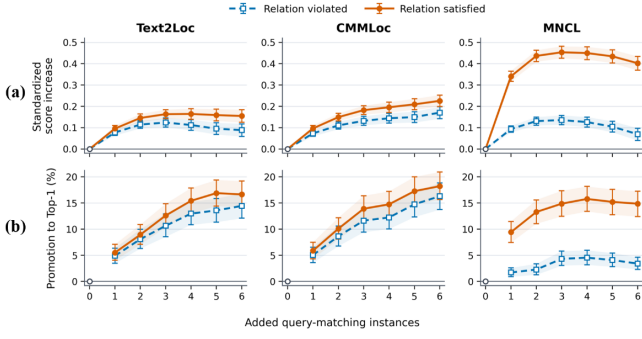


Figure 2: Effect of added query-matching instances on candidate scores. Top: standardized score change; bottom: promotion rate among initially non-Top-1 candidates. Curves compare relation-satisfied and relation-violated placements; error bars denote 95% cluster-bootstrap confidence intervals.

(a) Prevalence of Boundary-Truncated Evidence			
Analysis unit	Affected	Total	Rate
Boundary-truncated instances	7,907	19,629	40.28%
Hints referencing them	11,905	18,459	64.49%
Queries	3,179	3,187	99.75%

(b) Submap Retrieval Outcomes				
Model	Top-1 Recall		Top-1 Retrieval Errors	
	Top-1	Close@1	Close non-GT	Other
Text2Loc (2024)	33.0%	64.3%	46.8%	53.2%
CMMLoc (2025)	34.7%	66.3%	48.3%	51.7%
MNCL (2025)	48.4%	79.0%	59.2%	40.8%

Table 1: Boundary-related evidence on the KITTI360Pose validation split. (a) Prevalence of boundary-truncated instances and affected hints and queries. (b) Exact Top-1 recall, Close@1 recall within 15 m, and the partition of Top-1 retrieval errors into close non-GT and other submaps.

3.1 Text-to-Point-Cloud Localization

Let $M = \{M_n\}_{n=1}^N$ denote a city-scale point-cloud map partitioned into overlapping submaps, where $M_n = \{P_{n,j}\}_{j=1}^{O_n}$ contains O_n segmented instances. Given a query $q = \{h_i\}_{i=1}^H$, whose hints describe target-to-object relations, the task is to estimate the target position (x, y) from the joint configuration of the referenced objects. Most existing methods first rank submaps by query-submap similarity and then predict the position from q and the instances in the selected submap $M_{\hat{n}}$ (Kolmet et al., 2022; Xia et al., 2024; Liu et al., 2025; Xu et al., 2025).

Despite its practical effectiveness, this pipeline leaves two structural gaps: (1) a submap is retrieved through an aggregate embedding comparison without explicitly verifying the uniqueness and full consistency between the query and the instance layout. (2) fine localization infers the target position within a single submap, ignoring that the context is often incomplete because of each submap’s fixed spatial extent.

3.2 Diagnosis of Pipeline Limitations

We analyze the two gaps using released Text2Loc, CMMLoc, and MNCL models (Xia et al., 2024; Liu et al., 2025; Xu et al., 2025). All analyses are conducted with their official checkpoints on the KITTI360Pose dataset (Kolmet et al., 2022).

Partial Matches Elevate Inconsistent Candidates. Urban environments typically contain abundant similar or repetitive objects (Torii et al., 2013). Repeated objects allow a candidate to contain individually plausible but jointly inconsistent matches, which can raise the aggregate submap score when no explicit joint-layout verification is imposed. To measure this effect, we test 911 partially matched candidates by replacing unrelated instances with one to six exact semantic matches. Paired variants have identical geometry and semantic attributes but differ in whether their placement satisfies the queried relation. Figure 2 shows that both candidate scores and Top-1 promotion rates increase even under relation-violated placements, while satisfying the relation provides an additional gain. These results demonstrate that partial matches may mislead models without explicit joint layout verification, resulting in *layout-inconsistent aliasing*.

Submap Boundaries Truncate Query Evidence. Large-scale objects such as sidewalks and roads often span multiple submaps and are therefore truncated at cell boundaries. Table 1(a) shows that 40.28% of instances are boundary-truncated, 64.49% of hints refer to them, and 99.75% of queries contain at least one such hint. The resulting fragments make neighboring submaps similar while reducing the evidence available within any single submap. Across the three baselines, nearby non-ground-truth submaps account for 46.8–59.2% of Top-1 retrieval errors (Table 1(b)), indicating that neighboring candidates often share relevant evidence while omitting different fragments. Thus, even when the coarse stage retrieves the submap containing the most relevant evidence, the fine stage may still operate on incomplete context, resulting in *boundary evidence incompleteness*.

Together, these diagnoses define two requirements: coarse selection should validate matched instances as a joint layout, whereas fine localization should preserve the selected coordinate frame while admitting relevant cross-boundary evidence. Both require semantic and spatial compatibility to be evaluated under partial correspondence. This motivates PARC’s candidate-level assessment for submap selection and object-level weights for context routing.

4 Method

4.1 Overview

Given a query q , the coarse retrieval stage first uses a neural retriever to rank the submaps in M by neural similarity scores $S_{\text{sim}}(q, M_n)$ and retains the top- K candidates. For each candidate M_n , PARC evaluates whether the objects form a partial hint-object assignment that is explicitly consistent in

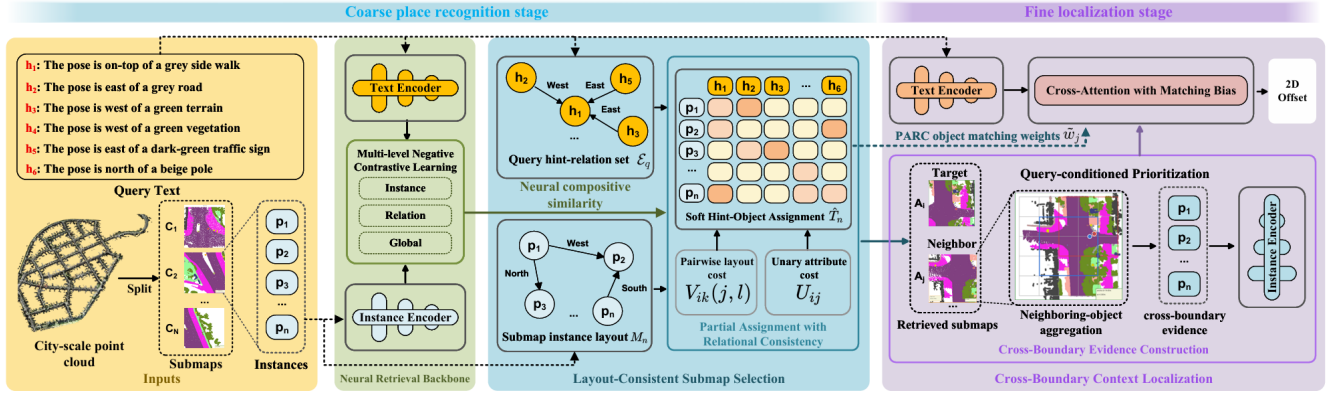


Figure 3: Overview of PARC-Loc. In the coarse stage, the relational consistency between hints and each submap is explicitly assessed through partial assignment to produce a reliable submap selection. In the fine stage, neighboring-object aggregation first makes cross-boundary context available; PARC is then reapplied to produce object weights that bias cross-attention before offset prediction.

both attributes and pairwise layout. Combining its candidate-level score with neural similarity, our approach identifies the best-matching submap $M_{\hat{n}}$.

At the fine localization stage, the selected submap $M_{\hat{n}}$ serves as the coordinate anchor. Objects from its neighborhood are expressed in its coordinate frame, and PARC is reapplied to compute query-conditioned object-matching weights. These weights bias hint-to-object cross-attention before the model predicts a location offset.

4.2 Partial Assignment with Relational Consistency (PARC)

At both stages, a query refers to only a subset of map objects, while some referenced objects may be absent or truncated and unrelated objects may be present. PARC models this partial correspondence while requiring the assigned objects to preserve the relations expressed by the query. It combines unary attribute compatibility, pairwise layout consistency (Mémoli, 2011; Peyré et al., 2016), and partial assignment constraints in one objective.

Consider a query q with H hints and an input instance set $P = \{P_j\}_{j=1}^O$. Hint i specifies a category ℓ_i , an appearance attribute κ_i , and a target-to-object direction. Instance P_j has category ℓ_j , appearance κ_j , and a two-dimensional ground-plane center p_j . We represent their soft correspondence by a nonnegative matrix $T \in \mathbb{R}^{H \times O}$, where T_{ij} is the assignment mass between hint i and instance P_j .

Unary attribute compatibility. Let $\delta(x, y)$ be zero if $x = y$ and one otherwise. The unary cost of an individual correspondence is

$$U_{ij} = \lambda_{\text{cat}} \delta(\ell_i, \ell_j) + \lambda_{\text{app}} \delta(\kappa_i, \kappa_j), \quad (1)$$

where λ_{cat} and λ_{app} specify mismatch cost weights. Unary compatibility identifies plausible pairs but, as diagnosed in Sec. 3, cannot determine whether they form one valid layout.

Pairwise layout consistency. We therefore evaluate the relations among assigned objects. From the target-to-object directions in q , we derive a set $\mathcal{E}_q$ of valid hint pairs. Each $(i, k) \in \mathcal{E}_q$ implies an object-to-object relation r_{ik}^q . For example, if the target is west of one referenced object and east of another, the first object should lie east of the second. For $(i, k) \in \mathcal{E}_q$, the cost $V_{ik}(j, l)$ penalizes assigning the hint pair (i, k) to the object pair (j, l) when the ground-plane relation between instances P_j and P_l is inconsistent with the object-to-object relation r_{ik}^q implied by hints h_i and h_k . Following the fused transport principle of combining node attributes and pairwise structure (Vayer et al., 2019), we define

$$\mathcal{J}_{q,P}(T) = \sum_{i,j} U_{ij} T_{ij} + \frac{\alpha_{\text{rel}}}{Z_q} \sum_{(i,k) \in \mathcal{E}_q} \sum_{j,l} V_{ik}(j, l) T_{ij} T_{kl}, \quad (2)$$

where $Z_q = \max(1, |\mathcal{E}_q|)$ normalizes the number of valid query relations and α_{rel} controls the relative penalty for layout violations.

Partial correspondence. A query mentions only a subset of map objects, and boundary truncation may remove referenced objects. A complete one-to-one assignment would force missing hints or unrelated objects into incorrect correspondences. We therefore impose partial rather than complete matching (Chapel et al., 2020; Sarlin et al., 2020):

$$\begin{cases} T_{ij} \geq 0, & \sum_j T_{ij} \leq 1, & \sum_i T_{ij} \leq 1, \\ \sum_{i,j} T_{ij} = \mu, & \mu = \min(\rho H, O). \end{cases} \quad (3)$$

where $\rho \in [0, 1]$ specifies the matched fraction of hint mass. The capacities limit repeated use of a hint or object, while the partial mass μ leaves irrelevant objects and missing references unmatched. We define $\hat{T}(q, P) = \text{PARC}(q, P)$ as a feasible plan obtained by approximately minimizing the objective in Eq. (2) subject to the constraints in Eq. (3). Its normalized objective value provides the candidate-level consistency used

in the coarse stage, while its column mass provides the object matching weights used in fine localization.

4.3 Coarse Stage: Layout-Consistent Submap Selection

Neural retrieval efficiently narrows the city-scale search space (Uy and Lee, 2018; Komorowski, 2021), while PARC explicitly validates whether the matched objects form a coherent query layout in each of the K candidate submaps. For each candidate M_n , PARC returns $\hat{T}_n = \text{PARC}(q, M_n)$. Its normalized score is

$$S_{\text{parc}}(q, M_n) = -\frac{\mathcal{J}_{q, M_n}(\hat{T}_n)}{\max(\|\hat{T}_n\|_1, \varepsilon_0)}, \quad (4)$$

where $\varepsilon_0 > 0$ is a numerical stabilizer. A high score requires low-cost unary matches that also preserve the joint layout. Because PARC complements rather than replaces learned neural similarity, we combine its candidate-level consistency S_{parc} with the neural similarity S_{sim} to select the submap. This design effectively suppresses layout-inconsistent aliases. The selected submap defines the coordinate frame for fine localization, but not its complete evidence domain.

4.4 Fine Stage: Cross-Boundary Context Localization

Even a correct or nearby selected submap may omit query-relevant instances across its boundary. The fine stage therefore first expands the available context and then uses PARC to determine which objects support the joint query layout before routing them through cross-attention.

Neighboring-object aggregation. We collect objects from the neighborhood centered at $M_{\hat{n}}$ and express every object in the selected-submap frame:

$$\tilde{p}_j = \frac{p_j - o_{\hat{n}}}{L_{\hat{n}}}, \quad (5)$$

where $o_{\hat{n}}$ is the two-dimensional origin and $L_{\hat{n}}$ is the side length of $M_{\hat{n}}$. This preserves a single coordinate system across submap boundaries. We retain the most relevant objects using a prioritization strategy that considers their compatibility with the query and the surrounding spatial context. This deterministic capacity control makes cross-boundary evidence available without asserting that every retained object is relevant.

Object matching weights. Because the expanded context also introduces irrelevant objects, we reapply PARC to the retained instance set P_{ctx} , obtaining $\hat{T}_{\text{ctx}} = \text{PARC}(q, P_{\text{ctx}})$. Its object-wise column mass and normalized weight are

$$w_j = \sum_i (\hat{T}_{\text{ctx}})_{ij}, \quad \tilde{w}_j = \frac{w_j}{\max_l w_l + \varepsilon_0}. \quad (6)$$

The normalized weight quantifies the contribution of instance P_j to a low-cost joint explanation. These quantities are matching weights, which enable the fine localizer to route the available context toward query-relevant objects.

Cross-attention with matching bias. Let q_i and k_j denote the d_k -dimensional projected representation of hint h_i and the key projection of context instance P_j , respectively. We incorporate the PARC-derived weight as an additive log bias in hint-to-object attention:

$$A_{ij} = \text{softmax}_j \left(\frac{q_i^\top k_j}{\sqrt{d_k}} + \beta \log(\tilde{w}_j + \varepsilon_0) \right), \quad (7)$$

where $\beta \geq 0$ controls the strength of the matching bias. When $\beta = 0$, Eq. (7) reduces to standard content-based attention (Vaswani et al., 2017). Increasing β gives more attention to objects supported by a relationally consistent partial assignment while retaining the flexibility of learned cross-modal interaction.

The fine decoder produces a normalized two-dimensional offset $\hat{d}$ in the selected-submap frame, which is converted back to map coordinates by

$$\hat{p} = o_{\hat{n}} + L_{\hat{n}} \hat{d}. \quad (8)$$

The neural retriever retains its standard training objective, while the fine decoder is trained with the fine-localization loss $\mathcal{L}_{\text{fine}} = \|\hat{d} - d^*\|_2^2$, where $d^* = (p^* - o_{\hat{n}})/L_{\hat{n}}$ is the ground-truth normalized offset and p^* is the ground-truth map position. At inference, PARC first refines the neural candidate ranking and is then reapplied to route cross-boundary evidence during offset prediction.

5 Experiments

5.1 Experimental Setup

Dataset and protocol. Experiments are conducted on KITTI360Pose (Kolmet et al., 2022; Liao et al., 2022) and CityLoc (Kang et al., 2026), two benchmarks for city-scale T2P localization. Following the official KITTI360Pose protocol, we use five scenes for training, one for validation, and three for testing, with 30 m cubic submaps extracted at a 10 m stride. CityLoc comprises CityLoc-K for training and in-domain evaluation and CityLoc-C (Hu et al., 2022) for zero-shot cross-domain testing. Each 50×50 m map is paired with six textual hints. KITTI360Pose contains 3,187 validation queries and 11,505 test queries, corresponding to 1,116 and 3,932 distinct target-position clusters, respectively.

Evaluation metrics. Submap retrieval is evaluated by top-1/3/5 recall of the annotated ground-truth submap. Fine localization uses the top- k candidates, $k \in \{1, 5, 10\}$, and reports recall within $\tau \in \{5, 10, 15\}$ m (Kolmet et al., 2022; Liu et al., 2025).

Implementation. PARC-Loc adopts the neural backbone of MNCL (Liu et al., 2025). For hyper-parameters, we set $\alpha_{\text{rel}} = 0.8$, $\lambda_{\text{cat}} = 1$, and $\lambda_{\text{app}} = 0.5$. The benchmark experiments were run on an NVIDIA RTX A6000 GPU; the matched-seed analysis below used an NVIDIA A800 80GB. We make no runtime comparison across GPU types.

Method	Venue	Localization Recall ($\tau = 5/10/15$ m) $\uparrow$					
		Validation Set			Test Set		
		$k = 1$	$k = 5$	$k = 10$	$k = 1$	$k = 5$	$k = 10$
Text2Pos (2022)	CVPR'22	0.14/0.25/0.31	0.36/0.55/0.61	0.48/0.68/0.74	0.13/0.21/0.25	0.33/0.48/0.52	0.43/0.61/0.65
RET (2023)	AAAI'23	0.19/0.30/0.37	0.44/0.62/0.67	0.52/0.72/0.78	0.16/0.25/0.29	0.35/0.51/0.56	0.46/0.65/0.71
Text2Loc (2024)	CVPR'24	0.37/0.57/0.63	0.68/0.85/0.87	0.77/0.91/0.93	0.33/0.48/0.52	0.61/0.75/0.78	0.71/0.84/0.86
CMMLoc (2025)	CVPR'25	0.44/0.62/0.68	0.75/0.88/0.90	0.83/0.93/0.95	0.39/0.53/0.56	0.67/0.80/0.82	0.77/0.87/0.89
PMSH (2025)	ICCV'25	0.42/0.62/0.68	0.75/0.89/0.90	0.83/0.94/0.95	0.39/0.55/0.59	0.68/0.81/0.83	0.78/0.89/0.90
MNCL (2025)	AAAI'25	<u>0.54/0.75/0.79</u>	<u>0.83/0.94/0.95</u>	<u>0.89/0.97/0.98</u>	<u>0.50/0.66/0.68</u>	<u>0.76/0.87/0.89</u>	<u>0.83/0.93/0.94</u>
Des4Pos (2026a)	T-ITS'26	<u>0.53/0.75/0.80</u>	<u>0.81/0.95/0.96</u>	<u>0.86/0.97/0.98</u>	<u>0.48/0.67/0.71</u>	<u>0.74/0.89/0.90</u>	<u>0.81/0.94/0.95</u>
PARC-Loc	Ours	0.71/0.85/0.87	0.91/0.97/0.98	0.94/0.98/0.99	0.67/0.80/0.82	0.87/0.94/0.95	0.90/0.96/0.97

Table 2: Full localization results on KITTI360Pose. Each entry reports Top- k recall at 5/10/15 m. Best results are in **bold**, and second-best results are underlined.

Method	Submap Retrieval Recall $\uparrow$					
	Validation Set			Test Set		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
Text2Pos (2022)	0.14	0.28	0.37	0.12	0.25	0.33
Text2Loc (2024)	0.32	0.56	0.67	0.28	0.49	0.58
CMMLoc (2025)	0.35	0.61	0.73	0.32	0.53	0.63
PMSH (2025)	0.37	0.63	0.73	0.34	0.56	0.65
MNCL (2025)	<u>0.50</u>	0.75	<u>0.84</u>	0.44	0.67	0.75
Des4Pos (2026a)	<u>0.50</u>	<u>0.76</u>	<u>0.84</u>	<u>0.45</u>	<u>0.69</u>	<u>0.77</u>
PARC-Loc	0.54	0.81	0.89	0.51	0.76	0.83

Table 3: Coarse submap retrieval recall on KITTI360Pose. Best results are in **bold**, and second-best results are underlined.

For KITTI360Pose, neural retrieval proposes 20 candidates for PARC reranking. The fine decoder uses at most 64 object tokens from the selected-submap-centered 3×3 neighborhood. Exact category–appearance matches are retained first, category matches second, and remaining objects third, with deterministic source order and distance used to break ties. This ordering controls the token budget; PARC subsequently computes matching weights over all retained objects.

5.2 Quantitative Analysis

End-to-end localization. As shown in Table 2, PARC-Loc obtains the best recall across all metrics on both KITTI360Pose splits. Under the strict Top-1@5 m metric, it reaches 0.71 on validation and 0.67 on test, exceeding MNCL by 17 percentage points on both splits. Its advantage persists across larger thresholds and candidate ranks, although the margin narrows as recall approaches saturation. These results demonstrate the overall effectiveness of our method. The analyses below further disentangle the sources of the observed gains.

Submap retrieval. Table 3 shows that PARC-Loc outperforms the strongest baseline across all retrieval metrics, by 4–5 percentage points on validation and 6–7 percentage points on test. By construction, this metric credits only the annotated

Variant	Validation Set			Test Set		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
Neural retriever	0.49	0.75	0.83	0.46	0.70	0.78
PARC w/o unary	0.53	0.80	0.89	0.51	0.76	0.82
PARC w/o relation	0.52	0.79	0.87	0.50	0.74	0.82
PARC w/o partial	0.51	0.77	0.85	0.48	0.72	0.79
Full PARC	0.54	0.81	0.89	0.51	0.76	0.83

Table 4: Retrieval-level ablation of PARC on KITTI360Pose. All variants share our MNCL-style neural retriever. The full model is highlighted in **bold**.

ground-truth submap. A neighboring non-GT submap can still support a correct final prediction once cross-boundary context is included. This retrieval metric is used to isolate coarse-stage submap selection, while final localization remains the primary measure.

5.3 Ablation Studies

We evaluate the PARC terms and assess the complete pipeline by cumulatively enabling its coarse and fine stage components.

Ablation on Submap Retrieval. All variants in Table 4 start from the same neural backbone and differ only in the objective used in PARC. Full PARC raises Top-1 retrieval recall from 0.49 to 0.54 on validation and from 0.46 to 0.51 on test. Replacing partial assignment causes the largest test degradation, reducing Top-1 recall to 0.48. Removing relational consistency lowers Top-1 recall to 0.50, while removing unary compatibility ties full PARC at Top-1 but reduces Top-5 recall from 0.83 to 0.82. The degradation of the relation and partial variants further supports assessing a relationally consistent partial layout rather than relying only on individually compatible objects. These results demonstrate that explicit assessment of relational consistency through partial assignment effectively constrains the joint layout of candidates, achieving considerable improvements in submap retrieval.

Cumulative variant	Localization Recall $\uparrow$	
	Validation	Test
Neural pipeline	0.54/0.82/0.89	0.51/0.78/0.85
+ Coarse PARC	0.62/0.89/0.93	0.61/0.84/0.88
+ Neighbor context	0.66/0.90/0.93	0.64/0.85/0.89
+ Query-conditioned priority	0.69/0.90/0.93	0.66/0.85/0.90
Full PARC-Loc (+ bias)	0.71/0.91/0.94	0.67/0.87/0.90

Table 5: Cumulative end-to-end ablation on KITTI360Pose. Each entry reports Top-1/5/10 localization recall within 5 m. The full model is highlighted in **bold**.

Table 6: Validation Top-1 localization recall (%) under controlled context construction. Each column gives the distance threshold; higher is better.

Context construction	@5m	@10m	@15m
Single selected submap	62.22	82.65	86.41
Five-cell context	67.74	84.00	86.57
Distance-prioritized 3×3	66.68	83.68	86.32

Ablation on End-to-End Localization. The first row of Table 5 is the controlled neural baseline, which combines the MNCL-style neural retriever with the original single-submap regressor. Coarse PARC raises test Top-1@5 m recall from 0.51 to 0.61. With coarse PARC fixed, neighboring context, query-conditioned prioritization, and matching bias increase it successively to 0.64, 0.66, and 0.67. These order-conditioned increments show that the coarse-stage PARC provides the largest gain, while context construction and routing contribute a further 6 percentage points. The two tables therefore show the effectiveness of the shared PARC formulation and its two stage-specific readouts: candidate-level energy for coarse assessment and object-level assignment mass for fine-stage routing.

Context availability. To examine how much evidence is lost at a submap boundary, we compare three context constructions on a frozen validation candidate stream and checkpoint (Table 6). This diagnostic complements the cumulative pipeline ablation above; its rows are not intended as a row-wise reconstruction of Table 5.

Both expanded contexts improve the strict 5m endpoint over the single-submap input, by 5.52 and 4.46 percentage points, respectively. However, the larger distance-prioritized neighborhood does not outperform the five-cell context. The comparison therefore supports making neighboring evidence available, without implying that a larger context alone is always better.

Stability across training seeds. We next assess whether the matching-bias benefit persists across training randomness. Table 7 reports matched epoch-3 runs with seeds 2027, 2028, and 2029. The $\beta = 0.3$ condition has higher mean recall at every reported candidate rank. With only three seeds, these results provide a limited stability check, not a population-level

Table 7: Matched-seed localization recall within 5m (%). Values are mean $\pm$ sample standard deviation over seeds 2027–2029 at epoch 3. Higher is better.

Candidate rank	$\beta = 0$	$\beta = 0.3$
Top-1	68.92 $\pm$ 0.73	71.10 $\pm$ 0.49
Top-3	84.96 $\pm$ 0.51	85.90 $\pm$ 0.79
Top-5	89.11 $\pm$ 0.54	90.97 $\pm$ 0.68
Top-10	92.15 $\pm$ 0.46	93.95 $\pm$ 0.50

Method	CityLoc-K Val.			CityLoc-K Test		
	R@5m	R@10m	R@15m	R@5m	R@10m	R@15m
Text2Pos	16.48	40.69	62.92	14.62	38.27	59.55
Text2Loc	18.91	45.26	64.28	17.97	41.22	61.50
MNCL	19.30	45.94	64.50	18.76	42.63	62.58
CMMLoc	20.77	48.65	67.89	21.71	46.67	66.00
VLM-Loc	36.23	63.66	<u>77.77</u>	35.91	63.81	<u>76.79</u>
PARC-Loc	<u>32.84</u>	<u>61.12</u>	80.47	<u>31.04</u>	<u>59.31</u>	77.64

Table 8: Localization on CityLoc-K. Recalls are reported as percentages under 5 m, 10 m, and 15 m thresholds. Best results are in **bold**, and second-best results are underlined. VLM-Loc (2026) is included for reference and uses vision-language reasoning over BEV and scene-graph inputs, whereas the remaining methods follow the conventional T2P setting.

significance claim.

5.4 Further Analysis

Generalization on CityLoc. Table 8 reports in-domain validation and test results on CityLoc-K. PARC-Loc obtains the best R@15 m, reaching 80.47% on validation and 77.64% on test, while VLM-Loc remains stronger at the stricter R@5 m and R@10 m thresholds.

We further evaluate zero-shot cross-domain transfer from CityLoc-K to CityLoc-C without training or threshold tuning on CityLoc-C. As shown in Table 9, PARC-Loc obtains the best R@15 m at 73.57% and remains within 0.16 percentage points of VLM-Loc at R@10m, while VLM-Loc leads under the strict R@5m criterion. Given that VLM-Loc is tailored to vision-language reasoning over BEV and scene-graph inputs (Kang et al., 2026), this competitive performance supports the cross-domain transferability of PARC-Loc without implying uniform superiority across all localization thresholds.

Robustness to gallery growth. Following MNCL (Liu et al., 2025), we fix 2,373 query descriptions within the 200 m region and progressively expand the gallery around the same center. As shown in Fig. 5, PARC-Loc retains 96.7% of its Top-1 recall at 500 m, decreasing from 54.11% to 52.34%. In comparison, MNCL, Text2Loc, and CMMLoc retain only 89.7%, 83.8%, and 84.4%, respectively. This smaller degradation indicates that explicit layout validation improves robustness to distractors arising from repeated instances and partial layout matches.

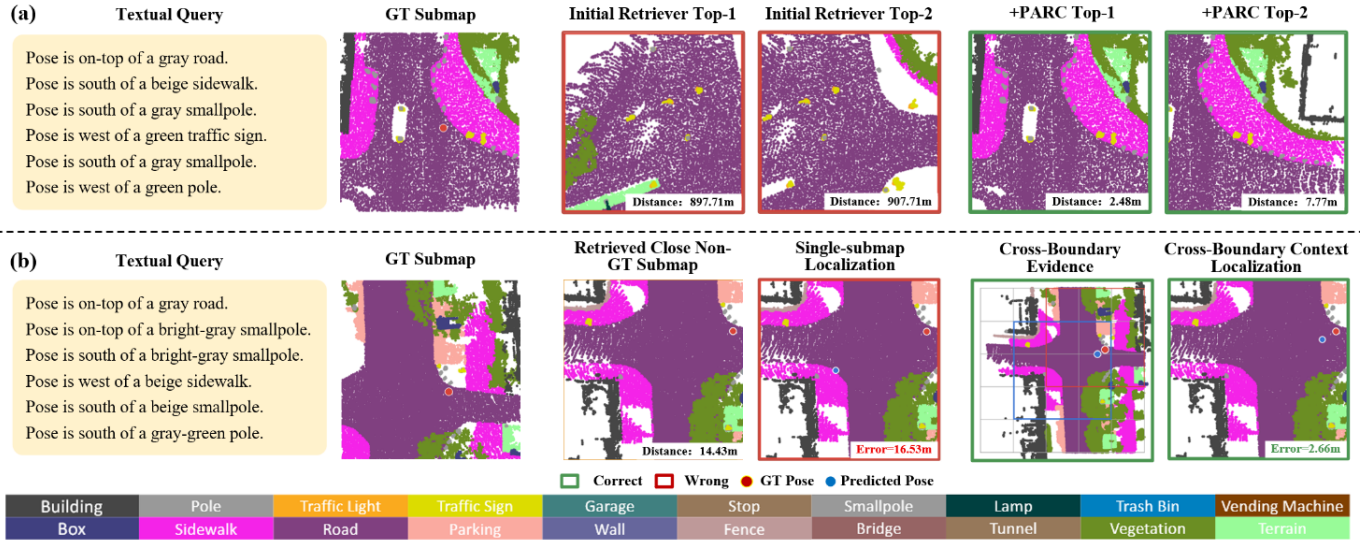


Figure 4: Qualitative examples of the two stage-specific mechanisms. Query boxes reproduce the relations while omitting the repeated subject “Pose is.” (a) PARC replaces the MNCL neural Top-1 distant alias with a candidate within 15m. (b) A close non-GT selected submap omits neighboring evidence; the 3×3 context reduces localization error. Red and green outlines mark incorrect and correct results, respectively. The yellow marker denotes the ground-truth pose, and the blue marker denotes the predicted pose.

Method	R@5m $\uparrow$	R@10m $\uparrow$	R@15m $\uparrow$
Text2Pos	8.11	27.21	50.01
Text2Loc	9.45	29.44	51.17
MNCL	13.68	35.93	53.78
CMMLoc	11.68	34.79	54.71
VLM-Loc	21.37	49.12	68.26
PARC-Loc	<u>16.56</u>	<u>48.96</u>	<u>73.57</u>

Table 9: Zero-shot cross-domain localization on CityLoc-C. Recalls are reported as percentages. PARC-Loc is trained on CityLoc-K and evaluated on CityLoc-C. Best results are in **bold**, and second-best results are underlined.

5.5 Qualitative Analysis

The quantitative analyses measure aggregate performance; we now examine two fixed cases to illustrate the two stage-specific mechanisms of PARC-Loc (Figure 4). In the first case, the neural Top-1 candidate is a distant layout-inconsistent alias, and PARC promotes a candidate within 15m of the target. In the second, a nearby selected submap is not the annotated ground-truth cell and omits neighboring evidence; cross-boundary context reduces the final localization error. Together, these cases illustrate why candidate verification and context expansion address different failure modes. They are case-level evidence, not an estimate of how frequently either mechanism corrects an error.

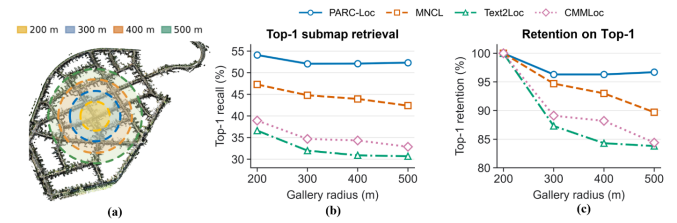


Figure 5: Robustness to controlled gallery growth. (a) Gallery expansion. (b) Top-1 submap retrieval recall at each radius. (c) Recall retention relative to the 200m gallery.

6 Conclusion

We presented PARC-Loc to address layout-inconsistent aliasing and boundary evidence incompleteness in coarse-to-fine text-to-point-cloud localization. PARC evaluates attribute and relational compatibility under partial correspondence: its candidate-level assessment complements neural similarity for coarse selection, while reapplying PARC to cross-boundary context yields object-level weights for localization in the selected-submap frame. PARC-Loc achieves 0.67 Top-1@5m test recall on KITTI360Pose, while ablations, CityLoc transfer, and gallery-growth tests support its effectiveness, transferability, and distractor robustness.

References

- Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Proceedings of the European Conference on Computer Vision*, pages 422–440, 2020.
- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.
- Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 2903–2913, 2020.
- Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision*, pages 202–221, 2020.
- Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *Advances in Neural Information Processing Systems*, volume 35, pages 20522–20535, 2022.
- Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2088–2098, 2019.
- 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.
- Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021.
- Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. SensatUrban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, 130(2):316–343, 2022.
- Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022.
- Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *Proceedings of the European Conference on Computer Vision*, pages 417–433, 2022.
- Shuhao Kang, Youqi Liao, Peijie Wang, Wenlong Liao, Qilin Zhang, Benjamin Busam, Xieyuanli Chen, and Yun Liu. Vlm-loc: Localization in point cloud maps via vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41365–41375, 2026.
- 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.
- 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.
- Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.
- Dunqiang Liu, Shujun Huang, Wen Li, Siqi Shen, and Cheng Wang. Text to point cloud localization with multi-level negative contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5397–5405, 2025.
- Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022.
- Facundo Mémoli. Gromov-wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.
- Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. Cityrefer: Geography-aware 3d visual grounding dataset on city-scale point cloud data. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proceedings of the International Conference on Machine Learning*, pages 2664–2672, 2016.
- Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Proceedings of the Conference on Robot Learning*, pages 1046–1056, 2022.

- Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020.
- Tianyi Shang and Zhenyu Li. Riemannian and symplectic geometry for hierarchical text-driven place recognition. *arXiv preprint arXiv:2604.01598*, 2026.
- Tianyi Shang, Zhenyu Li, Pengjie Xu, and Zhaojun Deng. Vehicle-sceneinteraction: A text-driven 3-d lidar place recognition method for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–13, 2026a. doi: 10.1109/TITS.2026.3676077.
- Tianyi Shang, Pengjie Xu, Zhaojun Deng, Zhenyu Li, Zhicong Chen, and Lijun Wu. Spatialloc: Leveraging multi-level spatial enhanced descriptors for cross-modal localization. *arXiv preprint arXiv:2601.03579*, 2026b.
- Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 883–890, 2013.
- 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.
- Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *Proceedings of the International Conference on Machine Learning*, pages 6275–6284, 2019.
- 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.
- Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19231–19242, 2023.
- Yan Xia, Letian Shi, Zifeng Ding, João 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.
- 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6637–6647, 2025.
- Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021.