WHERE IS HERE? SPATIAL FRAME-OF-REFERENCE MISMATCH IN EGOCENTRIC TEXT-TO-POINT-CLOUD LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Textual queries specify target locations through surrounding scene cues when exact coordinates are unavailable. Text-to-point-cloud (T2P) localization estimates the target position in a 3D point-cloud map from such queries. Existing T2P query protocols typically use map-frame directional relations, which can be interpreted as map-space constraints once the referenced objects are grounded. However, localization requests can instead describe local observations relative to an implicit observer-centered reference frame. An observer may know that a tree is on the left while the corresponding map-frame direction remains unknown. Even with correct object grounding, the same egocentric relation can imply different map-space constraints under different reference directions. We formalize this problem as Spatial Frame-of-Reference Mismatch. To study this setting, we introduce EgoT2P, a controlled benchmark with paired map-frame and egocentric queries over matched local maps and target positions. To test whether explicit reference-state modeling helps in this setting, we propose FoRLoc, which predicts a target-centered reference state consisting of a facing anchor and a coarse target-to-anchor bearing, together with auxiliary object correspondences, before estimating the target position. In the paired EgoT2P evaluation, all five evaluated T2P methods yield lower recall with egocentric queries. FoRLoc achieves a 20.6% relative improvement in $5\,\mathrm{m}$ localization recall over the strongest compared method.

1 INTRODUCTION

Text-to-point-cloud (T2P) localization takes a textual query and estimates its corresponding position in an existing 3D point-cloud map (Ao et al., 2021; 2022; 2023), supporting human–robot collaboration in applications such as autonomous delivery (Wu et al., 2026) and language-guided navigation (Kolmet et al., 2022). Existing approaches typically adopt a coarse-to-fine framework (Xia et al., 2024): they first retrieve candidate 3D submaps based on the textual query and then perform fine-grained position estimation within the selected submap. During fine localization, scene objects and their spatial relations provide constraints for estimating the target position within the local map.

However, existing T2P query protocols typically express directional relations relative to the map frame (Kang et al., 2026; Liu et al., 2025a; Xu et al., 2025). For example, "the target is east of the tree" directly specifies a map-frame direction. In contrast, an egocentric query such as "the tree is on my left" expresses the relation relative to an observer-centered frame at the target location (Mei et al., 2026). Without additional reference information, "left" does not specify a map-frame direction.

The issue is not simply replacing "east" with "left". Under map-frame queries, grounding a referenced object anchors it in the map, while a cardinal relation directly constrains the target to a map-space region. This interpretation does not directly carry over to egocentric queries, because the map-frame direction corresponding to a local directional relation is not explicitly specified. As illustrated in Fig. 1, even when the tree in "tree on my left" is correctly grounded, the target can lie in different map regions under different reference directions (Zhang et al., 2025b; Premsri & Kordjamshidi, 2025). Thus, scene grounding alone does not determine the corresponding map-space constraint for an egocentric relation. We formalize this discrepancy between egocentric directional relations and map-space localization constraints as Spatial Frame-of-Reference Mismatch.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

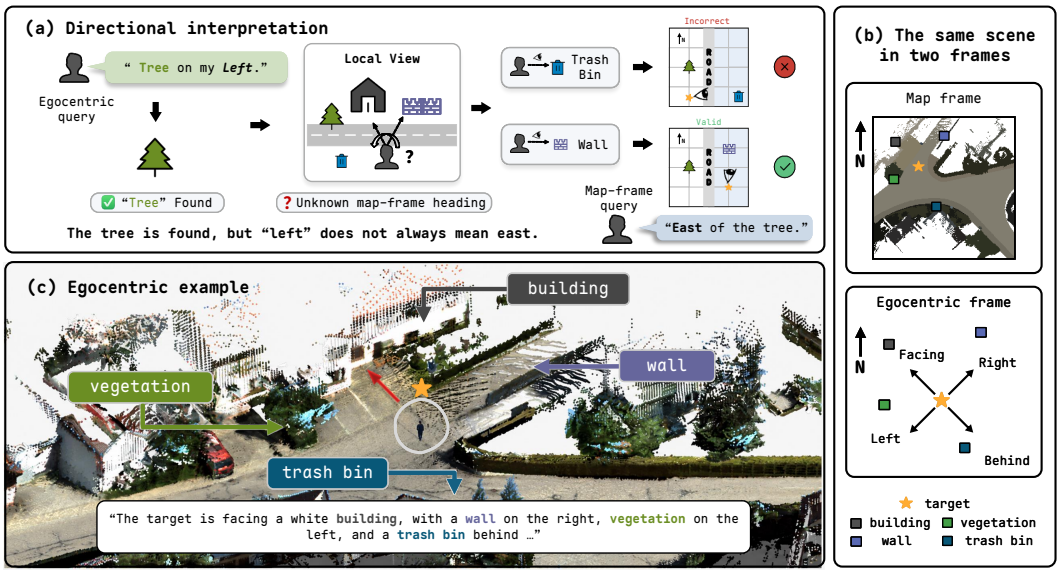


Figure 1: Spatial Frame-of-Reference Mismatch in egocentric T2P fine localization. (a) Grounding the referenced object does not determine the map-frame constraint of an egocentric relation. (b) The same scene in the map and observer-centered frames. (c) An egocentric query describes an unmarked target using surrounding scene relations.

In this work, we study Spatial Frame-of-Reference Mismatch in T2P fine localization. We introduce EgoT2P, a controlled benchmark with template-generated paired map-frame and egocentric queries over matched local maps and target positions, to compare the two query formulations. To examine whether explicit reference-state modeling helps in this setting, we propose Frame-of-Reference-Aware Localization (FoRLoc). FoRLoc predicts a target-centered reference state consisting of a facing anchor and a coarse target-to-anchor bearing in the map frame before estimating the target position. This target-dependent bearing provides coarse geometric information without specifying the final target position. Experiments on EgoT2P show that the five evaluated T2P methods achieve lower recall with paired egocentric queries than with their map-frame counterparts. FoRLoc achieves a 20.6% relative improvement in $R@5\,\mathrm{m}$ localization recall over the strongest compared method. The main contributions are summarized as follows:

- We formulate Spatial Frame-of-Reference Mismatch in T2P fine localization, where object grounding alone does not determine the map-frame interpretation of egocentric directional relations.
- We introduce EgoT2P, a controlled benchmark with template-generated paired map-frame and egocentric queries over matched local maps and target positions, to compare the two query formulations.
- We introduce a target-centered reference state comprising a facing anchor and a coarse target-to-anchor bearing, and develop FoRLoc to explicitly estimate this state before target-position prediction.

2 Related Work

2.1 Language-Guided 3D Localization

Language-guided 3D localization often grounds a referred object or region in an RGB-D or point-cloud scene. ScanRefer and ReferIt3D established benchmarks for this task (Chen et al., 2020; Achlioptas et al., 2020), while subsequent work studies relational disambiguation, viewpoint reasoning, and grounding with vision-language models (VLMs) (Zhao et al., 2021; Wu et al., 2023; Guo et al., 2023; Li et al., 2025b; Liu et al., 2025b; Cheng et al., 2026). CityRefer (Miyanishi et al., 2023) and CityAnchor (Li et al., 2025a) extend object grounding to outdoor and city-scale

point clouds. These works establish language-to-3D grounding for referred entities, while T2P fine localization further uses descriptions of surrounding objects and their spatial relations to infer an unmarked target position.

2.2 Text-to-Point-Cloud Localization

Text-to-point-cloud localization estimates a target position in a map from a description of its surrounding scene. Text2Pos (Kolmet et al., 2022) introduced the task and a coarse-to-fine pipeline consisting of submap retrieval and fine position estimation. RET (Wang et al., 2023) and MambaPlace (Shang et al., 2025) strengthen relation-aware retrieval and cross-modal place recognition, while Text2Loc and Text2Loc++ (Xia et al., 2024; 2026) develop matching-free fine localization and extend it toward more diverse descriptions and cross-environment generalization. MNCL (Liu et al., 2025a) improves contrastive discrimination, whereas CMMLoc (Xu et al., 2025) and PMSH (Feng et al., 2025) model partial correspondence and localization uncertainty. VLM-Loc (Kang et al., 2026) further uses bird's-eye-view (BEV) imagery and scene graphs, together with partial node assignment, to model correspondences between textual hints and map nodes. Existing T2P query protocols commonly express directional relations in map-frame terms, allowing them to constrain target locations once the referenced objects are grounded. For egocentric descriptions, object correspondence alone does not directly specify how these relations should be interpreted as map-space constraints.

2.3 Frame-of-Reference Reasoning

Spatial descriptions depend on the reference frame, and the same directional expression may correspond to different spatial configurations under different frames (Song et al., 2025). Existing benchmarks and VLM studies investigate spatial relations, orientation, and reference-frame ambiguity (Yang et al., 2019; Liu et al., 2023; Kamath et al., 2023; Zhang et al., 2025a;b; Premsri & Kordjamshidi, 2025), while other works explore grounded spatial reasoning, reference-frame transformation, and viewpoint-aware understanding (Chen et al., 2024; Cheng et al., 2024; Yang et al., 2025; Zhu et al., 2025; Lee et al., 2025; Jin et al., 2023; Jeong et al., 2026). We study reference-frame interpretation in T2P fine localization, where a local point-cloud map is given and an unmarked target position must be inferred from egocentric relational descriptions.

3 Method

3.1 Task Definition

T2P fine localization is defined over a local point-cloud map M containing the target position. Given an egocentric query $T = (h_k)_{k=1}^{m}$, the goal is to predict the target position $\mathbf{p} = (x, y)$ on the ground plane. The target is the observer's current position, and $\mathbf{p}$ is expressed in the map frame. Each hint h_k contains a spatial relation and an object phrase. We denote an instance by $(M, T, \mathbf{p})$ and its prediction by $\hat{\mathbf{p}}$. Unlike map-frame queries, egocentric queries express relations with respect to an implicit observer-centered frame whose reference state is not explicitly provided.

3.2 Framework Overview

Fig. 2 illustrates FoRLoc. We convert the local map M into a BEV image I and a structured scene representation $\mathcal{H}$. Stage 1 predicts an intermediate target-centered reference state F, consisting of a facing anchor and a coarse target-to-anchor bearing in the map frame, together with partial node assignments A for the query's object mentions. Stage 2 conditions on these predictions, the query, and local map evidence to estimate the target position. The two stages use separate predictors with stage-specific system prompts s_1 and s_2.

3.3 Map Evidence Representation

The BEV image provides visual evidence from the local map. To make individual object instances explicit, including candidates that share the same semantic label, and to distinguish their spatial-

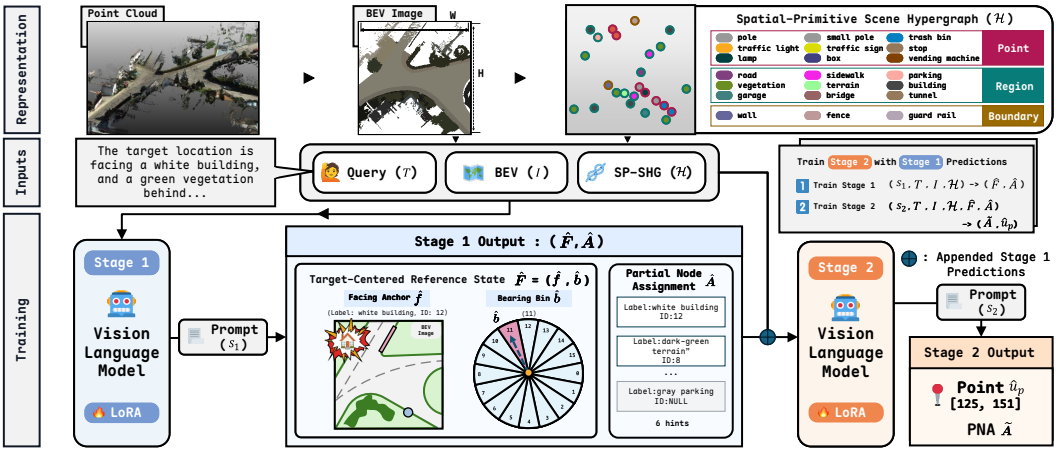


Figure 2: Framework of FoRLoc. The local point-cloud map is represented by a BEV image and an SP-SHG; the latter panel shows the fixed semantic-to-primitive mapping. Stage 1 predicts a target-centered reference state and partial node assignments, and Stage 2 uses these predictions with the query and map evidence to estimate the target position and auxiliary assignments.

primitive types, we additionally construct a spatial-primitive scene hypergraph (SP-SHG), $\mathcal{H} = (V, \mathcal{E})$. This query-independent representation is constructed from the semantic-instance annotations provided with each local map.

The node set $V = \{v_0, \ldots, v_{N-1}\}$ contains the map's object instances. Each node $v_i = (\ell_i, \tau_i, \mathbf{u}_i)$ records its semantic label ℓ_i, spatial-primitive type $\tau_i \in \{\mathrm{P}, \mathrm{R}, \mathrm{B}\}$, and centroid pixel coordinate $\mathbf{u}_i$ in the BEV image. The three types denote point-like landmarks, region-like areas, and boundary-like structures, respectively, following the fixed mapping in Fig. 2.

The hyperedge set $\mathcal{E}$ groups nodes sharing the same semantic label or primitive type. We serialize $\mathcal{H}$ as an indexed node list followed by these semantic and primitive groups. The SP-SHG contains no target coordinates, query-derived relations, reference-state labels, or ground-truth node assignments. Further construction and serialization details are provided in Appendix C.1.

3.4 Target-Centered Reference State

Egocentric queries express directional relations whose interpretation depends on an implicit reference state. The corresponding direction in the map frame is not explicitly given by the query. Thus, directional cues such as left and right cannot be directly used as map-space constraints without additional reference information. We represent the target-centered reference state as

$$F = (f, b), \tag{1}$$

where $f = (i_f, \ell_{i_f})$, with $i_f \in \{0, \ldots, N-1\}$, identifies the facing anchor by its node index and semantic label, and $b \in \{0, \ldots, N_{\mathrm{bin}} - 1\}$ specifies the coarse map-frame direction from the target position to this anchor.

Let $\mathbf{u}_p$ denote the BEV pixel coordinate corresponding to $\mathbf{p}$, and let $\theta_f \in [0, 2\pi)$ denote the direction from $\mathbf{u}_p$ to the facing anchor's centroid in the fixed BEV coordinate system, measured clockwise from the rightward horizontal axis. We quantize this direction into N_{bin} uniformly spaced bins as

$$b = \left\lfloor \frac{N_{\mathrm{bin}}\theta_f}{2\pi} + \frac{1}{2} \right\rfloor \bmod N_{\mathrm{bin}}. \tag{2}$$

A direction exactly halfway between two adjacent bin centers is assigned to the next clockwise bin.

The facing anchor f identifies the scene object associated with the facing relation, while b specifies the map-frame direction associated with this relation, which is used to interpret the remaining egocentric relations. Because b is defined by the target-to-anchor bearing, F contains coarse target geometry as well as reference information. However, it does not specify the target-to-anchor distance,

and the same reference state may therefore correspond to multiple target positions. Thus, F serves as an intermediate geometric state rather than a complete localization solution. The reference-state representation is further illustrated in Appendix C.2.

3.5 Frame-of-Reference-Aware Localization

FoRLoc uses the target-centered reference state F as its main intermediate representation, together with mention-level object correspondences, to estimate the target position. Following the partial node assignment (PNA) strategy of VLM-Loc (Kang et al., 2026), we represent these correspondences as

$$A = (a_k)_{k=1}^{m}, \qquad a_k \in \{0, \dots, N-1\} \cup \{\varnothing\}. \tag{3}$$

Here, $a_k = j$ associates hint h_k with scene node v_j, whereas $a_k = \varnothing$ indicates that no corresponding node exists in the local map. The assignments provide auxiliary object-level evidence for localization.

Stage 1: Reference-State and Assignment Estimation. Given the egocentric query T, BEV image I, SP-SHG $\mathcal{H}$, and system prompt s_1, Stage 1 predicts

$$(\hat{F}, \hat{A}) = g_{\theta_1}(s_1, T, I, \mathcal{H}), \tag{4}$$

where g_{θ_1} is an autoregressive VLM. The predicted $\hat{F}$ specifies the facing anchor and coarse target-to-anchor bearing in the map frame, while $\hat{A}$ associates query mentions with available map nodes. Stage 1 does not output the final target coordinate, but provides coarse geometric and correspondence evidence for Stage 2.

Stage 2: Reference-Guided Position Estimation. Stage 2 conditions on the original inputs and the Stage 1 predictions:

$$(\tilde{A}, \hat{\mathbf{u}}_p) = g_{\theta_2}(s_2, T, I, \mathcal{H}, \hat{F}, \hat{A}), \tag{5}$$

where g_{θ_2} is a separate autoregressive VLM, $\tilde{A}$ is an auxiliary assignment output, and $\hat{\mathbf{u}}_p$ is the predicted target position in BEV pixel coordinates. The BEV raster has an upper-left pixel origin, with coordinates increasing rightward and downward. We convert $\hat{\mathbf{u}}_p$ to the map-frame position $\hat{\mathbf{p}}$ using the known raster-to-map transform of the local map for evaluation. At inference, Stage 2 receives the predicted $\hat{F}$ and $\hat{A}$, rather than their ground-truth values.

3.6 Loss Functions

During training, the annotated target position and query construction provide the reference-state label $F = (f, b)$ and partial node assignments A. At inference, both are unknown and must be predicted from the query and local map. The two stages are trained sequentially with autoregressive, stage-specific target sequences.

The Stage 1 target sequence Y_{ref} serializes (F, A), yielding

$$\mathcal{L}_{\text{ref}} = -\sum_{q} \log p_{\theta_1}\left(y_q^{\text{ref}} \mid y_{<q}^{\text{ref}}, s_1, T, I, \mathcal{H}\right). \tag{6}$$

After fitting Stage 1 on the training split, we generate its predictions $(\hat{F}, \hat{A})$ offline on that split for the Stage 2 training inputs. Conditioned on these predictions, the Stage 2 target sequence Y_{loc} serializes the auxiliary assignments and target BEV pixel coordinate $(A, \mathbf{u}_p)$, yielding

$$\mathcal{L}_{\text{loc}} = -\sum_{q} \log p_{\theta_2}\left(y_q^{\text{loc}} \mid y_{<q}^{\text{loc}}, s_2, T, I, \mathcal{H}, \hat{F}, \hat{A}\right), \tag{7}$$

where y_q^{ref} and y_q^{loc} denote the q-th tokens of the respective target sequences. Stage 2 is therefore trained with predicted intermediate inputs rather than ground-truth F and A, matching the type of intermediate information available at inference.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

EgoT2P Benchmark. To evaluate fine localization with egocentric directional evidence, we introduce EgoT2P, a benchmark in which local evidence is expressed in an egocentric observation frame. Each query comprises one *facing* hint, which defines an implicit observation heading, and five support hints that describe surrounding evidence relative to this heading. A representative query reads *"The target location is facing a white building; vegetation on the left, parking on the right, and a traffic light behind..."*

EgoT2P is built on the KITTI-360 dataset (Liao et al., 2021), following the CityLoc-K data split and using its $50\,\mathrm{m} \times 50\,\mathrm{m}$ local maps rasterized into 224×224 BEV images (Kang et al., 2026). The target pool combines retained target locations from CityLoc-K and KITTI360Pose (Kolmet et al., 2022) with additional samples drawn from region interiors and stable locations adjacent to boundary elements. Queries are generated from controlled templates, with support relations sampled from continuous, potentially overlapping target-centered relation regions rather than fixed disjoint angular bins (Zhang et al., 2025b). The resulting benchmark contains 14,377 training, 1,181 validation, and 5,632 test queries.

For each egocentric query, we construct a map-frame counterpart by expressing directional relations with the cardinal directions used in existing T2P benchmarks. Within each pair, only the directional expressions are changed; all other scene and query content remains unchanged. The complete benchmark construction procedure, relation-region definitions, and dataset statistics are provided in Appendix B.

Baselines and Protocol. We compare FoRLoc with five representative T2P baselines: Text2Pos (Kolmet et al., 2022), Text2Loc (Xia et al., 2024), MNCL (Liu et al., 2025a), CMMLoc (Xu et al., 2025), and VLM-Loc (Kang et al., 2026). All baselines are retrained from their official implementations under a unified data split and evaluation protocol. All methods receive the same ground-truth local submap containing the target. VLM-based methods use the same backbone and training settings where applicable.

Evaluation Metric. We employ Recall@$K\,\mathrm{m}$, which measures the percentage of queries whose predicted 2D positions lie within K meters of the ground-truth (GT) targets. We report results for $K \in \{5, 10, 15\}$, with all values presented as percentages.

Implementation Details. We use Qwen3-VL-8B-Instruct (Bai et al., 2025) as the shared frozen backbone and optimize an independent LoRA adapter (Hu, 2022) for each stage with rank $r = 8$ and scaling factor $\alpha = 16$. The reference state is parameterized by a facing anchor and a target-to-anchor bearing quantized into $N_{\mathrm{bin}} = 16$ bins. Each stage is trained for one epoch on a single NVIDIA GeForce RTX 5090 GPU using fused AdamW (Loshchilov & Hutter, 2017), batch size 1, gradient accumulation 8, learning rate 1×10^{-4}, warmup ratio 0.05, and BF16 precision. Additional training and inference details are provided in Appendix C.4.

4.2 RESULTS

Controlled Query-Frame Analysis. To compare fine-localization performance under map-frame and egocentric directional descriptions, we retrain five representative T2P methods separately on paired map-frame and egocentric versions of EgoT2P using the same data split and training protocol. As shown in Table 1, all five methods achieve lower recall on the egocentric versions. At $R@10\,\mathrm{m}$, the performance decreases by 6.89 to 18.13 percentage points across the four other methods, while VLM-Loc decreases from 86.33% to 59.02%. Since each query pair shares the same local map, target position, and referenced objects, this comparison characterizes the localization gap between map-frame and egocentric directional formulations under the controlled EgoT2P protocol.

Localization Accuracy. We next compare complete localization systems under the egocentric query formulation. Table 2 shows that FoRLoc achieves the best performance across all validation and test thresholds. Compared with the strongest baseline, VLM-Loc, FoRLoc improves $R@5\,\mathrm{m}$

Table 1: Performance comparison of T2P methods under paired map-frame and egocentric queries on EgoT2P. Each entry reports $R@5/10/15\,\mathrm{m}$ recall (%); Δ denotes the decrease.

Query	Text2Pos	Text2Loc	MNCL	CMMLoc	VLM-Loc
Map-frame	25.16/59.59/81.73	18.36/49.75/75.71	18.15/51.05/77.61	15.06/46.64/74.41	59.61/86.33/94.98
Egocentric	12.55/41.46/67.93	12.64/39.93/65.96	12.07/38.51/64.91	11.70/39.75/67.03	26.01/59.02/77.43
Δ Drop	12.61/18.13/13.80	5.72/9.82/9.75	6.08/12.54/12.70	3.36/6.89/7.38	33.60/27.31/17.55

Table 2: Localization results on EgoT2P. Best results are **bold**; second-best results are underlined. Parentheses denote absolute improvements in percentage points over the strongest baseline.

Method	Val			Test		
	$R@5\,\mathrm{m}\uparrow$	$R@10\,\mathrm{m}\uparrow$	$R@15\,\mathrm{m}\uparrow$	$R@5\,\mathrm{m}\uparrow$	$R@10\,\mathrm{m}\uparrow$	$R@15\,\mathrm{m}\uparrow$
Text2Pos (CVPR'22)	13.80	44.96	72.57	12.55	41.46	67.93
Text2Loc (CVPR'24)	14.82	48.26	74.01	12.64	39.93	65.96
MNCL (AAAI'25)	15.16	47.76	74.43	12.07	38.51	64.91
CMMLoc (CVPR'25)	15.16	47.08	75.70	11.70	39.75	67.03
VLM-Loc (CVPR'26)	25.91	58.76	81.20	26.01	59.02	77.43
FoRLoc (Ours)	**28.45** (+2.54)	**64.01** (+5.25)	**86.62** (+5.42)	**31.37** (+5.36)	**65.11** (+6.09)	**82.28** (+4.85)

from 25.91% to 28.45% on the validation split and from 26.01% to 31.37% on the test split, corresponding to relative improvements of 9.8% and 20.6%, respectively. The improvement is also observed at larger distance thresholds, with test $R@10\,\mathrm{m}$ increasing from 59.02% to 65.11% and remaining consistent at $R@15\,\mathrm{m}$. Qualitative examples in Fig. 3 further illustrate these improvements. Further results and analyses are provided in Appendix D.7.

Query Rewrite Evaluation. To test sensitivity to the fixed query templates, we construct a scene-stratified subset of 200 EgoT2P test queries by random sampling within each scene and manually rewrite each query. The rewrites alter wording, sentence structure, and relation ordering while preserving the referenced objects, attributes, and spatial relations; each rewrite is manually checked against its original for semantic consistency. As shown in Table 3, FoRLoc's $R@5\,\mathrm{m}$ decreases by 4.50 percentage points after rewriting, compared with 11.00 points for VLM-Loc. FoRLoc achieves higher recall than VLM-Loc on the rewritten queries. Examples are provided in Appendix D.2.

Native-Submap Fine Localization. We further evaluate fine localization on native KITTI360Pose submaps (Kolmet et al., 2022). All methods receive the same ground-truth submap containing the target, ensuring that the comparison focuses on fine localization rather than coarse retrieval. We generate egocentric queries following the EgoT2P protocol and train all models under the same native egocentric split. Compared with the enlarged EgoT2P maps, these native submaps preserve the original $30\,\mathrm{m}$ extent of KITTI360Pose and provide a smaller local context. Additional protocol details are provided in Appendix D.1.

As shown in Table 4, FoRLoc improves $R@10\,\mathrm{m}$ from 54.71% to 58.94% while maintaining higher recall at all evaluated thresholds. These results show that FoRLoc retains its advantage in a separately trained fine-localization setting with the original KITTI360Pose submap extent. Because both settings are derived from KITTI-360, this experiment does not establish cross-dataset generalization.

Egocentric Condition Analysis. To assess performance under challenging egocentric evidence, we evaluate three overlapping EgoT2P test subsets: Ambiguous Reference, Partial Context, and Weak Directional Support. They capture repeated facing-anchor candidates, mentions without a local-map match, and incomplete directional support, respectively; their definitions are given in Appendix D.3. These condition subsets are not mutually exclusive; a test sample may belong to multiple subsets.

As shown in Fig. 4, Partial Context yields the lowest $R@10\,\mathrm{m}$ recall for every evaluated method and remains among the most difficult subsets at $R@5\,\mathrm{m}$ and $R@15\,\mathrm{m}$. FoRLoc achieves the highest recall across all three subsets at each threshold, consistently outperforming VLM-Loc. These results

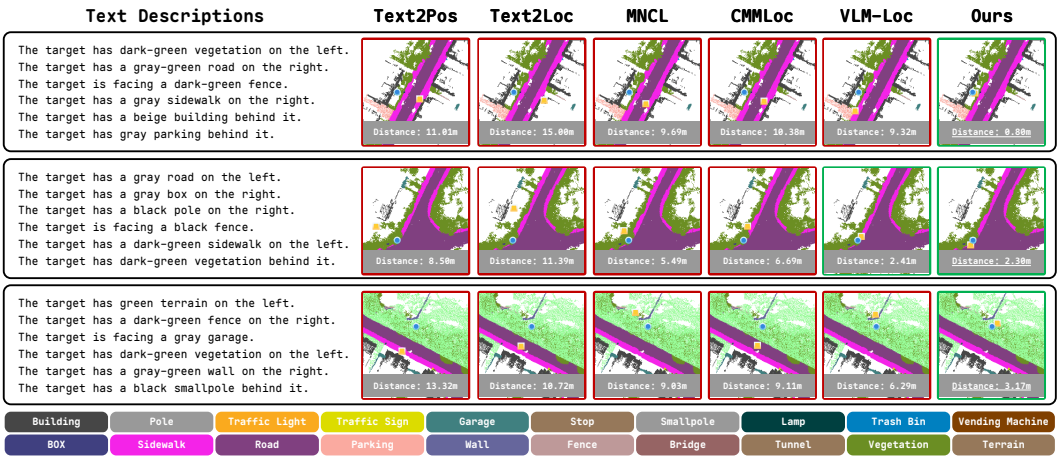


Figure 3: Qualitative localization results on EgoT2P. Blue circles and yellow squares mark Ground-truth and predicted positions; green/red borders indicate errors below/above 5 m.

Table 3: Performance on a scene-stratified random subset of 200 original–rewrite query pairs. Rewrites preserve referenced objects, attributes, and spatial relations.

Method	Query	R@5 m	R@10 m	R@15 m
VLM-Loc	Original	27.00	61.50	78.50
	Rewritten	16.00	48.00	70.50
FoRLoc	Original	31.00	67.00	82.50
	Rewritten	26.50	58.00	74.00

Table 4: Fine-localization performance on native KITTI360Pose submaps.

Method	R@5 m	R@10 m	R@15 m
Text2Pos	11.84	36.61	60.05
Text2Loc	13.17	40.93	63.88
MNCL	12.81	40.17	64.06
CMMLoc	13.42	42.36	66.19
VLM-Loc	24.36	54.71	73.29
FoRLoc (Ours)	**29.02**	**58.94**	**79.18**

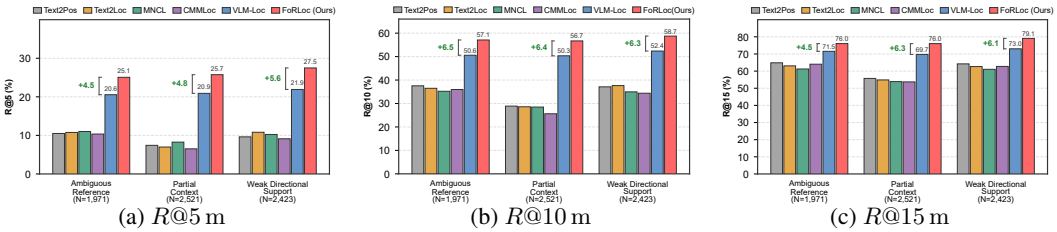
show that the complete method retains its advantage in the evaluated conditions with repeated facing-anchor candidates or incomplete support evidence.

4.3 Ablation Study

Component Analysis. Table 5 examines the role of intermediate information in FoRLoc. One-Stage Direct achieves 24.52% at R@5 m, and Two-Stage Direct achieves 23.08%, compared with 31.37% for complete FoRLoc. Thus, adding a localization stage alone does not reproduce FoRLoc's gain in the evaluated control. Within the FoRLoc variants, providing both $\hat{F}$ and $\hat{A}$ yields the highest recall at all three thresholds, and withholding $\hat{F}$ produces a larger decrease than withholding $\hat{A}$.

Table 6 examines how Stage 2 localization responds to test-time substitutions of its intermediate inputs. Jointly shuffling the predicted $\hat{F}$ and $\hat{A}$ across samples reduces recall, whereas replacing either prediction with ground truth improves it, with a larger gain from replacing F at all three thresholds. This pattern is consistent with Stage 2 using sample-specific intermediate information. When both inputs are replaced by ground truth, R@5 m reaches 36.42%. Ground-truth F contains a bearing derived from the true target, but the final target position is still predicted by Stage 2; this setting therefore serves as an oracle-input diagnostic rather than a localization upper bound. Additional ablations are reported in Appendix D.5.

Reference-State Analysis. We examine how reference-state prediction quality relates to localization by grouping test samples according to facing-anchor correctness and circular bearing-bin error, defined as $e_b = \min\{|\hat{b} - b|, N_{\text{bin}} - |\hat{b} - b|\}$. FoRLoc identifies the facing anchor correctly for 86.4% of samples, and 52.1% have both a correct anchor and a bearing estimate within one bin of



(a) $R@5\,\mathrm{m}$ (b) $R@10\,\mathrm{m}$ (c) $R@15\,\mathrm{m}$

Figure 4: Localization recall (%) across EgoT2P condition subsets. Brackets indicate FoRLoc gains over VLM-Loc in percentage points; subsets are not mutually exclusive.

Table 5: One-stage and two-stage direct controls and FoRLoc intermediate-input ablations on EgoT2P.

$\hat{F}$	$\hat{A}$	$R@5m$	$R@10m$	$R@15m$
One-Stage Direct		24.52	55.06	76.03
Two-Stage Direct		23.08	54.94	75.85
×	✓	26.10	61.19	77.57
✓	×	29.53	64.58	80.72
✓	✓	**31.37**	**65.11**	**82.28**

Table 6: Effect of intermediate-state quality on localization. Pred., GT, and Shuf. denote predicted, ground-truth, and shuffled states.

F	A	$R@5m$	$R@10m$	$R@15m$
Pred.	Pred.	31.37	65.11	82.28
GT	Pred.	34.30	69.64	86.97
Pred.	GT	33.26	67.74	85.16
GT	GT	**36.42**	**72.60**	**88.07**
Shuf.	Shuf.	25.78	58.22	79.33

Table 7: Localization under different reference-state prediction qualities on EgoT2P. Samples are grouped by anchor correctness and bearing-bin error.

State	N (%)	$R@5/10/15\,\mathrm{m}\uparrow$	Err. (m)↓
Overall	5632(100.0)	31.37/65.11/82.28	9.38
Correct anchor, $e_b \leq 1$	2932(52.1)	39.43/77.32/91.98	7.30
Correct anchor, $e_b > 1$	1935(34.4)	23.67/53.59/72.66	11.40
Wrong anchor	765(13.6)	20.00/47.45/69.41	12.21

Table 8: Sensitivity of localization to the number of bearing bins on the EgoT2P test set.

N_{bin}	$R@5\,\mathrm{m}$	$R@10\,\mathrm{m}$	$R@15\,\mathrm{m}$
4	31.20	64.61	82.00
8	30.79	64.54	81.75
16	31.37	65.11	82.28
32	31.45	65.02	82.19

the ground truth. As shown in Table 7, among samples with a correct anchor, $R@10\,\mathrm{m}$ is 77.32% when $e_b \leq 1$ and 53.59% when $e_b > 1$. This association indicates that more accurate reference-state predictions coincide with better localization, although the groups may also differ in intrinsic localization difficulty and should not be interpreted causally. Table 8 examines the bearing-bin resolution. Across 4 to 32 bins, $R@10\,\mathrm{m}$ varies by only 0.57 percentage points, with similarly small changes at the other thresholds. Thus, localization performance remains stable across the evaluated resolutions.

5 Conclusion

We study Spatial Frame-of-Reference Mismatch in T2P fine localization, where egocentric directional relations are expressed in an implicit observer-centered frame and do not directly specify map-space constraints. We introduce EgoT2P, a controlled benchmark with paired map-frame and egocentric queries, and propose FoRLoc to investigate explicit target-centered reference-state modeling for egocentric localization. Experiments on EgoT2P show that the evaluated T2P methods achieve lower recall with paired egocentric queries than with their map-frame counterparts. The complete FoRLoc system improves localization performance over the strongest evaluated baseline, while ablation studies examine the roles of its intermediate predictions. Together, the paired-query analysis characterizes the performance gap under egocentric queries, while the component results provide empirical support for explicit target-centered reference-state modeling under the controlled EgoT2P protocol.

AI USE STATEMENT

We acknowledge the use of generative AI tools to assist in editing manuscript language, support certain parts of the experimental workflow, aid preliminary interpretation of empirical data, and offer structural recommendations for figures, tables, and page layout to improve clarity. We maintain complete scientific oversight at every stage. We critically review all AI-generated suggestions, and we manually inspect and validate all experimental implementations, empirical evaluations, and literature citations. We accept full accountability for the accuracy, integrity, and conclusions of the work presented in this paper.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pp. 422–440. Springer, 2020.

Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, and Yulan Guo. Spinnet: Learning a general surface descriptor for 3d point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11753–11762, 2021.

Sheng Ao, Yulan Guo, Qingyong Hu, Bo Yang, Andrew Markham, and Zengping Chen. You only train once: Learning general and distinctive 3d local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3949–3967, 2022.

Sheng Ao, Qingyong Hu, Hanyun Wang, Kai Xu, and Yulan Guo. Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1255–1264, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221. Springer, 2020.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

An-Chieh Cheng, Yang Fu, Yatai Ji, Ligeng Zhu, Guanqi Zhan, Zhuoyang Zhang, Zhaojing Yang, Song Han, Yao Lu, Pavlo Molchanov, Vidya Nariyambut Murali, Jan Kautz, Xiaolong Wang, Hongxu Yin, and Sifei Liu. Grounded 3d-aware spatial vision-language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16688–16700, June 2026.

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*, pp. 8296–8305, 2025.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15372–15383, 2023.

Edward J Hu. yelong shen, phillip wallis, zeyuan allen-zhu, yuanzhi li, shean wang, lu wang, and weizhu chen. 2022. lora: Low-rank adaptation of large language models. In *International conference on learning representations*, volume 4, 2022.

Suchae Jeong, Jaehwi Song, Haeone Lee, Hanna Kim, Jian Kim, Dongjun Lee, Dong Kyu Shin, Changyeon Kim, Dongyoon Hahm, Woogyeol Jin, et al. Learning multi-view spatial reasoning from cross-view relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2570–2581, 2026.

Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17307–17316, 2023.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

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*, pp. 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*, pp. 6687–6696, 2022.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9241–9251, 2025.

Jinpeng Li, Haiping Wang, Jiabin Chen, Yuan Liu, Zhiyang Dou, Yuexin Ma, Sibei Yang, Yuan Li, Wenping Wang, Zhen Dong, et al. Cityanchor: City-scale 3d visual grounding with multi-modality llms. In *ICLR*, 2025a.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3707–3717, 2025b.

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*, 45:3292–3310, 2021.

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*, volume 39, pp. 5397–5405, 2025a.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, and Xiangyang Xue. Reasongrounder: Lvlm-guided hierarchical feature splatting for open-vocabulary 3d visual grounding and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3718–3727, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yanghong Mei, Yirong Yang, Longteng Guo, Qunbo Wang, Ming-Ming Yu, Xingjian He, Wenjun Wu, and Jing Liu. Urbannav: Learning language-guided embodied urban navigation from web-scale human trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18505–18513, 2026.

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 *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Tanawan Premsri and Parisa Kordjamshidi. Forest: frame of reference evaluation in spatial reasoning tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34965–34991, 2025.

Tianyi Shang, Zhenyu Li, Pengjie Xu, and Jinwei 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. IEEE, 2025.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15768–15780. IEEE Computer Society, 2025.

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*, volume 37, pp. 2501–2509, 2023.

Jianshi Wu, Minghang Zhu, Dunqiang Liu, Wen Li, Sheng Ao, Siqi Shen, Chenglu Wen, and Cheng Wang. Leader: Learning reliable local-to-global correspondences for lidar relocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9932–9942, June 2026.

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*, pp. 19231–19242, 2023.

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*, pp. 14958–14967, 2024.

Yan Xia, Letian Shi, Yilin Di, Joao F Henriques, and Daniel Cremers. Text2loc++: Generalizing 3d point cloud localization from natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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*, pp. 6637–6647, 2025.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2051–2060, 2019.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Wanglu Wanglu. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11591–11609, 2025a.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025b.

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*, pp. 2928–2937, 2021.

Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, and Yang Liu. Cospace: benchmarking continuous space perception ability for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29569–29579, 2025.

A Overview

This appendix provides additional details on the EgoT2P benchmark, the FoRLoc implementation, and supplementary experiments. Section B describes the benchmark construction and paired map-frame queries. Section C specifies the structured map representation, prompts, training configuration, and evaluation protocols. Section D reports additional protocols and analyses, while Section E discusses the limitations of the present controlled setting.

B EgoT2P Benchmark Details

B.1 Local Maps and Target Sampling

Data Source. EgoT2P is constructed exclusively from KITTI-360. We adopt the CityLoc-K partition, using sequences *00*, *02*, *04*, *06*, and *07* for training, sequence *10* for validation, and sequences *03*, *05*, and *09* for testing. The sequence sets are disjoint, and all maps, target locations, and queries are generated only from data assigned to their respective split. When one target admits multiple valid reference states, all resulting query pairs remain in the split of that target.

Local Maps. Following CityLoc-K, each local map covers a $50\,\mathrm{m} \times 50\,\mathrm{m}$ region and is rasterized into a 224×224 BEV image. The aligned semantic-instance scene graph is represented in the same coordinate system. The pixel origin is at the upper-left corner, with x increasing rightward and y downward. A source map is retained in EgoT2P only if it contains at least one target that admits a valid reference state and a complete query.

Target Consolidation and Filtering. Duplicate target pixels within each local map are merged. Metric coordinates are retained for evaluation, while integer BEV pixel coordinates serve as localization labels. A target is retained only if it admits the reference state in Section B.2 and the complete six-mention query in Section B.3.

Target Sampling. We combine eligible CityLoc-K and KITTI360Pose targets with additional region and boundary samples (Table B1). Region candidates are sampled from interior pixels of area-like source categories, whereas boundary candidates are sampled from exterior pixels adjacent to wall and fence structures. All candidates are de-duplicated at their integer BEV pixels and processed under the same protocol for constructing queries.

Table B1: Composition of the EgoT2P candidate target pool.

Component	Included sources or categories
Original	Eligible CityLoc-K and KITTI360Pose targets
Region	Road, sidewalk, parking, terrain, building, garage, bridge, tunnel, and vegetation
Boundary	Wall and fence

B.2 Target-Centered Relation Regions

Observation Frame. EgoT2P defines directional relations in a target-centered observation frame. Given a target position $\mathbf{p}$ and a facing anchor centered at $\mathbf{o}_f$, the forward unit vector is

$$\mathbf{h} = \frac{\mathbf{o}_f - \mathbf{p}}{\|\mathbf{o}_f - \mathbf{p}\|_2}. \tag{8}$$

With $\mathbf{h} = (h_x, h_y)^\top$, we define the rightward perpendicular under the BEV raster convention as

$$\mathbf{h}_\perp = (-h_y, h_x)^\top. \tag{9}$$

Anchor Choice. Facing anchors are selected from *building*,*garage*, *wall*, *fence*, *pole*, *traffic sign*, and *traffic light*. These categories provide stable and recognizable scene cues for defining the observation heading. We require the target and anchor to be separated by at least 16 BEV pixels. A clear

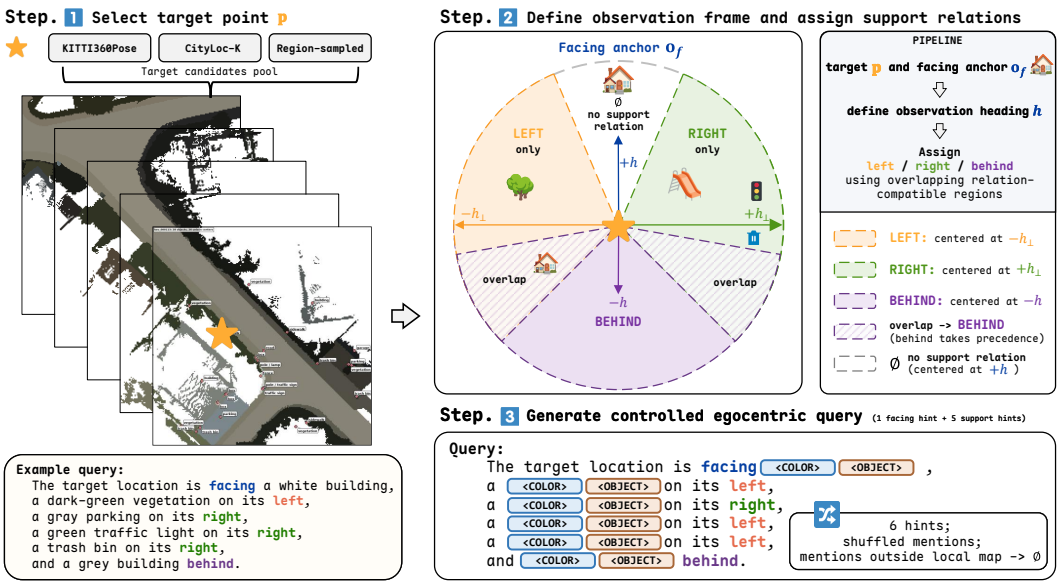


Figure B1: Construction of EgoT2P. A facing anchor defines the target-centered forward axis. Support objects are selected from continuous left, right, and behind regions, after which the facing and support hints are shuffled.

reference has one eligible anchor of the facing category in the local scene graph. In a same-category ambiguous reference, the surrounding supports provide additional cues for identifying the selected anchor among same-category candidates. When multiple valid anchor choices are available, the released constructions are fixed by an instance-specific deterministic ordering.

Relation Assignment. For a candidate support object with center $\mathbf{o}_j$, we compute its offset relative to the target-centered frame as

$$\alpha_j = (\mathbf{o}_j - \mathbf{p})^\top \mathbf{h}, \qquad \beta_j = (\mathbf{o}_j - \mathbf{p})^\top \mathbf{h}_\perp. \tag{10}$$

The three acceptance regions are

$$\begin{aligned} \textit{behind} &= \{j : \alpha_j < 0,\ |\beta_j| \le 2|\alpha_j|\}, \\ \textit{right} &= \{j : \beta_j > 0,\ |\alpha_j| \le 2|\beta_j|\}, \\ \textit{left} &= \{j : \beta_j < 0,\ |\alpha_j| \le 2|\beta_j|\}. \end{aligned} \tag{11}$$

Because these continuous regions may overlap, we apply the fixed priority *behind*, *right*, then *left*; each candidate therefore receives at most one relation. The non-strict dominance boundaries are included, whereas candidates outside all three regions are not used as grounded supports. A retained query must contain grounded *left* and *right* supports and exactly six mentions after constructing ungrounded mentions in Section B.3; otherwise, the candidate query is discarded.

B.3 Egocentric Query Generation

Query Composition. For each retained target pixel, we generate an egocentric query with exactly six object mentions: one facing anchor and five directional supports. Section B.2 defines the observation frame and directional regions; the following paragraphs specify the mention attributes and sentence realization.

Support Selection. The five supports are selected jointly from the eligible relation regions, without repeating a selected object. When more than one valid support set is available, its released realization is fixed by the same instance-specific ordering used for anchor choices.

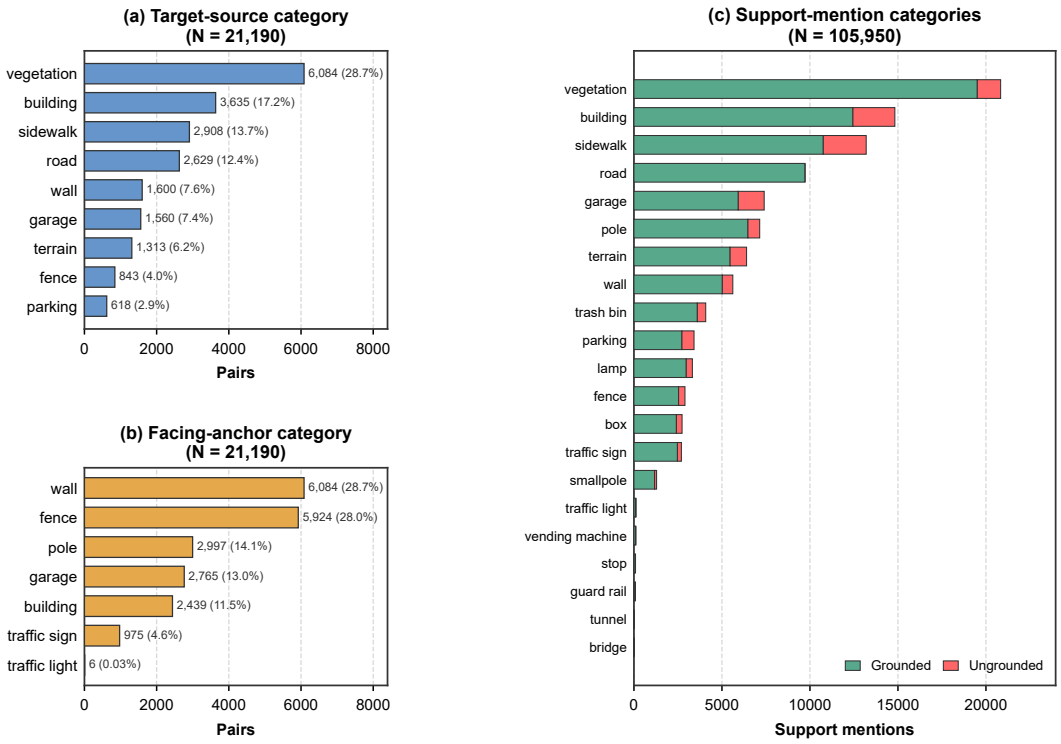


Figure B2: Target and mention composition of EgoT2P: (a) target-source categories, (b) facing-anchor categories, and (c) semantic categories of support mentions, separated into grounded and ungrounded mentions.

Color Attributes. Each object phrase combines a discrete color attribute with its semantic category. We inherit the precomputed CityLoc-K color label, obtained by mapping the representative RGB color to a fixed palette, for grounded and ungrounded mentions alike. Thus, an object phrase takes the form

$$\langle \text{obj} \rangle = \texttt{a} \;\; \langle \text{color} \rangle \, \langle \text{category} \rangle, \tag{12}$$

such as *a black wall*.

Sentence Template and Ordering. We generate each query as one complete natural-language sentence:

$$\textit{The target location is } \langle h_1 \rangle, \ldots, \langle h_5 \rangle, \textit{ and } \langle h_6 \rangle. \tag{13}$$

Here, each clause $\langle h_i \rangle$ is instantiated from its relation label and object phrase:

$$\langle h \rangle = \begin{cases} \texttt{facing } \langle \text{obj} \rangle, & \text{facing}, \\ \langle \text{obj} \rangle \texttt{ on its left}, & \text{left}, \\ \langle \text{obj} \rangle \texttt{ on its right}, & \text{right}, \\ \langle \text{obj} \rangle \texttt{ behind}, & \text{behind}. \end{cases} \tag{14}$$

After instantiation, we apply a deterministic instance-specific permutation to the six clauses.

Ungrounded Support Mentions. An ungrounded support retains its category, color, and egocentric relation, but has no usable corresponding node in the local scene graph. Its assignment is therefore recorded as $\varnothing$. This preserves partial contextual evidence while keeping the facing anchor grounded.

B.4 Benchmark Statistics

Dataset Scale. Table B2 reports the final EgoT2P composition. Across the nine sequences in the split, the benchmark contains 3,306 retained local maps, 20,472 unique target locations, and 21,190

query pairs. Target locations are unique within each local map, whereas multiple valid reference states may yield different query pairs for the same target. Each pair is released with an egocentric query and its paired map-frame query (Section B.5).

Table B2: EgoT2P statistics by split. The first block reports sequence, local-map, target-location, and query-pair counts. The final two rows count query pairs; percentages use the query-pair count of the corresponding split as denominator. The two conditions may overlap.

Statistic	Train	Val.	Test	Total
Sequences	5	1	3	9
Source local maps	2,767	300	1,027	4,094
Retained local maps	2,180	237	889	3,306
Unique target locations	13,916	1,131	5,425	20,472
Query pairs	14,377	1,181	5,632	21,190
Query pairs with ambiguous references	5,032 (35.0%)	692 (58.6%)	1,971 (35.0%)	7,695 (36.3%)
Query pairs with ≥ 1 ungrounded mention	6,690 (46.5%)	355 (30.1%)	2,521 (44.8%)	9,566 (45.1%)

Target and Mention Composition. Figure B2 reports the categories used for target sources, facing anchors, and directional support mentions. The support distribution is separated by grounded and ungrounded mentions.

Reference Frame and Relation Composition. Figure B3 summarizes the directions from targets to their facing anchors, support configurations, and ungrounded support counts. Panel (a) visualizes the distribution of target-to-anchor directions by grouping the continuous directions into 16 equal sectors in the fixed BEV frame. The underlying directions remain continuous during query construction. Panel (b) reports support configurations, where each configuration indicates the numbers of mentions in the target-centered *left*, *right*, and *behind* regions. Panel (c) reports the number of support mentions whose assignment is $\varnothing$.

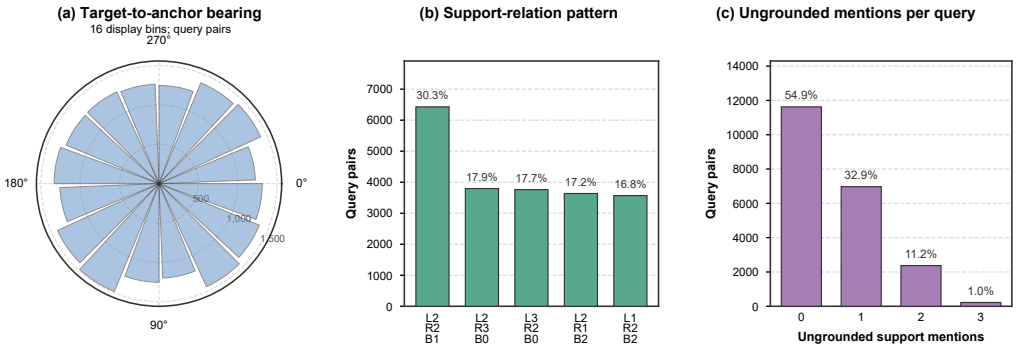


Figure B3: Reference frame and relation composition of EgoT2P. Counts are query pairs; percentages above the bars use all 21,190 query pairs as the denominator.

B.5 Construction of Paired Map-Frame Queries

Paired Query Construction. For each EgoT2P instance, we construct one map-frame counterpart for the controlled comparison in the main paper. Each pair retains the same BEV raster, scene graph, target, object phrases, assignment sequence, mention order, grounding status, and split membership. The non-directional query content therefore remains unchanged, while the relations are rewritten under the fixed BEV map axes.

Map-Frame Conversion. The sentence structure and mention order follow Section B.3. Each clause is rewritten as

$$\begin{aligned} \langle h^{\mathrm{map}} \rangle &= \langle d \rangle \ \ \texttt{of} \ \ \langle \mathrm{obj} \rangle, \\ \langle d \rangle &\in \{\mathit{north}, \mathit{south}, \mathit{east}, \mathit{west}\}. \end{aligned} \tag{15}$$

These directions describe the target relative to the referenced object in the fixed BEV frame, whose origin is at the upper-left, with x increasing eastward and y increasing southward. For a grounded mention j, let $\mathbf{p}$ denote the target pixel and $\mathbf{o}_j$ the center of its assigned node. With

$$(\Delta x_j, \Delta y_j) = \mathbf{p} - \mathbf{o}_j, \tag{16}$$

the map-frame relation is

$$r_j^{\text{map}} = \begin{cases} \textit{east}, & |\Delta x_j| \geq |\Delta y_j| \wedge \Delta x_j \geq 0, \\ \textit{west}, & |\Delta x_j| \geq |\Delta y_j| \wedge \Delta x_j < 0, \\ \textit{south}, & |\Delta x_j| < |\Delta y_j| \wedge \Delta y_j \geq 0, \\ \textit{north}, & |\Delta x_j| < |\Delta y_j| \wedge \Delta y_j < 0. \end{cases} \tag{17}$$

Thus, horizontal directions are selected on equal-magnitude boundaries. For an ungrounded mention, the category, color, mention position, and null assignment are retained. Because no local node center is available, its map-frame wording is converted from the stored source relation using the query heading employed in the paired construction.

As a consistency check, we independently reconstruct the egocentric labels under Section B.2 and the map-frame labels under the conversion above using the released target, anchor, and scene-graph coordinates. Each reconstructed label agrees with its corresponding released label for all 30,583 grounded mentions in the 5,632 test pairs, including 3,111 pairs in which all six mentions are grounded.

C FoRLoc Implementation Details

This section complements the method description in the main paper by fixing the concrete representation, supervision, and model interfaces used by FoRLoc. We first describe how an annotated local map is materialized as an SP-SHG, then specify the two structured prediction targets and their prompts, and finally report the optimization and inference settings.

C.1 SP-SHG Materialization and Serialization

Nodes. We construct the spatial-primitive scene hypergraph (SP-SHG) directly from the semantic-instance scene graph of each local map. Each instance is encoded as $v_i = (\ell_i, \tau_i, \mathbf{u}_i)$, comprising its semantic label, primitive type, and centroid in the 224×224 BEV raster. The SP-SHG is constructed from the semantic-instance annotations available with each local map.

Spatial primitives. Each semantic category is mapped to a point-like landmark (P), region-like area (R), or boundary-like structure (B). Table C1 gives the mapping used throughout EgoT2P.

Table C1: Semantic-to-primitive mapping used to construct the SP-SHG.

Type	Semantic categories
P	pole, traffic light, smallpole, lamp, trash bin, vending machine, stop, box, traffic sign
R	tunnel, building, road, vegetation, terrain, sidewalk, garage, parking, bridge
B	wall, guard rail, fence

Hyperedges. We use two query-independent group families. A label group contains every node of a repeated semantic label, and a type group contains every node of one populated primitive type. The graph contains no target point, heading, query-derived relation, or ground-truth assignment, so it can be reused by all queries from the same local map.

Object-level graph control. The object-level graph used in the component analysis retains the same indexed object nodes, semantic labels, and centroid coordinates as the SP-SHG. It removes the primitive-type annotations and the semantic and primitive hyperedges. Neither representation contains the target point, target heading, query-derived relations, or ground-truth assignments.

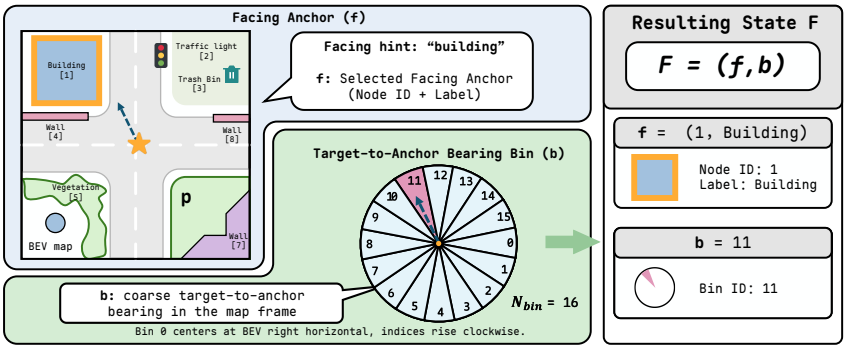


Figure C1: Target-centered reference state $F = (f, b)$, where facing anchor f and bearing bin b specify the corresponding target-to-anchor direction in the map frame.

(1) `type_codes`: Primitive code definitions listed in Table C1.

(2) `objects`: ID-sorted map instances. Each instance is represented as $\{\texttt{id}, \texttt{label}, \texttt{type}, \texttt{xy}\}$, where `id`, `label`, `type`, and `xy` denote the instance ID, semantic category, primitive code, and BEV centroid coordinate, respectively.

(3) `hypergraph`: ID-based group dictionaries, where `label_groups` and `type_groups` store object IDs grouped by semantic labels and primitive types, respectively.

```
{"type_codes": {"P": "point", "R": "region", "B": "boundary"},
 "objects": [
  {"id": 0, "label": "road", "type": "R", "xy": [112, 96]},
  {"id": 1, "label": "building", "type": "R", "xy": [53, 41]},
  {"id": 2, "label": "building", "type": "R", "xy": [181, 47]}],
 "hypergraph": {
  "label_groups": {"building": [1, 2]},
  "type_groups": {"R": [0, 1, 2]}}}
```

Listing 1: Illustrative SP-SHG serialization.

C.2 Reference-State and Assignment Representation

Figure C1 illustrates the target-centered reference state $F = (f, b)$. Facing-anchor grounding selects the map node f associated with the query's facing mention. In the EgoT2P construction, the observation direction is defined from the annotated target position toward this anchor (Section B.2). The ground-truth bin b is therefore quantized from a target-to-anchor bearing: it contains coarse target geometry and is not a separately observed observer-heading label. The bin does not specify the target-to-anchor distance or uniquely determine the final position. The target coordinate used to construct the training label is not an input to Stage 1 at inference.

Stage 1 predicts the facing anchor and bearing bin, together with auxiliary mention assignments A; Stage 2 predicts the final target coordinate using these predicted intermediates and the query and map evidence. In the Stage 1 serialization, `anchor_mention_index` identifies the facing mention, `anchor_node_id` identifies its matched map node, and `heading_bin_16` stores the predicted bearing bin $\hat{b}$. The `assignments` field associates mentions with map nodes in query order, using `null` when no local-map match is identified. Thus, this protocol does not separately predict or evaluate an observer heading independent of the target-to-anchor bearing. Evaluating a separate observer-heading estimate would require a heading defined independently of the target–anchor pair. Ground-truth F, including its target-derived bearing, is used for Stage 1 supervision and explicitly labeled oracle diagnostics, not as a standard Stage 2 input.

C.3 System Prompts for the Two Stages

The two prompts use the same coordinate and serialization conventions but assign different responsibilities to the predictors. Stage 1 estimates the target-centered reference state and mention assignments, whereas Stage 2 conditions localization on these intermediate predictions. Tables C2 and C3 summarize the task instructions for each stage, coordinate conventions, and structured output formats.

Table C2: Stage 1 prompt specification for reference-state and assignment prediction.

ROLE. You are a spatial-reasoning assistant for target-centered localization in a bird's-eye-view map.

INPUT. You receive (1) a 224×224 BEV image, (2) an SP-SHG containing indexed nodes and semantic/primitive hyperedges, and (3) an ordered query with exactly six object mentions. Each query relation describes where an object lies relative to an unknown target in the target's egocentric frame.

COORDINATE SYSTEM. Image coordinates start at the upper-left; x points right and y points down. Heading is measured clockwise from the positive x-axis and represented by one of 16 bins indexed from 0 to 15.

GOAL. Identify which query mention is the facing anchor, ground it to an SP-SHG node, estimate the target-to-anchor heading bin, and assign every query mention to a map node when supported by the evidence.

RULES. Preserve query order. Return exactly six assignments. Use only node IDs present in the input graph; use `null` for an ungrounded mention. Do not predict a target coordinate. Resolve repeated semantic labels using the BEV geometry, appearance phrases, and the joint relational evidence.

OUTPUT FORMAT. Return one JSON object and no additional text. The object must contain `frame` with `anchor_mention_index`, `anchor_node_id`, and `heading_bin_16`, followed by six ordered `assignments`; each assignment contains `object_label`, `grounded`, and `matched_node`.

Table C3: Stage 2 prompt specification for reference-guided target localization.

ROLE. You are a reference-guided localizer in a bird's-eye-view map.

INPUT. You receive the same BEV image, SP-SHG, and six-mention ordered query as Stage 1, together with Stage 1's predicted reference state and initial assignments.

COORDINATE SYSTEM. Image coordinates start at the upper-left; x points right and y points down. The supplied heading bin follows the 16-bin clockwise convention defined for Stage 1.

GOAL. Interpret the egocentric relations using the supplied facing anchor and heading, reconcile the grounded objects with the BEV geometry, and estimate the target pixel.

RULES. Treat Stage 1 outputs as intermediate predictions, not immutable facts. Correct an assignment when the image, graph, and joint relations provide stronger evidence. Preserve query order, return exactly six assignments, use only valid node IDs or `null`, and place the point inside the 224×224 raster.

OUTPUT FORMAT. Return one JSON object and no additional text. The object must contain six ordered `assignments` and a two-element integer array `point_2d`.

C.4 Training and Inference Details

Optimization. We use Qwen3-VL-8B-Instruct as a shared frozen backbone and optimize an independent LoRA adapter for each stage. Table C4 reports the settings shared by both stages unless stated otherwise.

Table C4: FoRLoc training configuration.

Setting	Value
Backbone	Qwen3-VL-8B-Instruct
LoRA rank / scaling	8 / 16
LoRA target modules	all linear layers
Training epochs	1 per stage
Effective batch size	8
Optimizer	AdamW
Learning rate / warmup ratio	1×10^{-4} / 0.05
Numeric precision	BF16

Sequential training. Stage 1 and Stage 2 are trained to generate the respective structured outputs specified by their system prompts. After fitting Stage 1, we run its checkpoint offline on the training split and use the resulting predictions, rather than ground-truth intermediate states, to construct the Stage 2 training inputs. Validation and test intermediates are generated separately from their corresponding inputs. Ground-truth F or A is supplied to Stage 2 only in experiments explicitly labeled as oracle ablations. All ground-truth substitutions in Table 6 are applied only at test time to the same Stage 2 checkpoint trained on predicted intermediate inputs; Stage 2 is not retrained for these interventions.

Inference. At test time, the Stage 1 prediction is appended to the Stage 2 input. No ground-truth assignment or reference-state field is used in the standard two-stage pipeline.

Shuffled-Cue Protocol. The shuffled-cue diagnostic replaces each sample's predicted (F, A) with the jointly transferred Stage 1 output of another sample under a fixed cyclic shift. The donor cue retains the same JSON format but may not agree with the recipient query or local map. Stage 2 is evaluated without retraining; this test-time corruption assesses its reliance on query-specific intermediate cues.

C.5 Baseline Protocol

All baselines are adapted from their released implementations on the EgoT2P training split and evaluated under the same fine-localization protocol with the ground-truth local map. We retain each method's original architecture and map representation. For the VLM-based methods, VLM-Loc and FoRLoc use the same Qwen3-VL-8B-Instruct backbone and the training settings in Table C4. No method receives ground-truth target coordinates, reference states, or assignments at inference.

D Additional Experiments

D.1 Native KITTI360Pose Protocol Details

Native map setting. We additionally evaluate all methods on the original KITTI360Pose native submaps, rather than on the enlarged local maps used by EgoT2P. We retain the released $30\,\mathrm{m} \times 30\,\mathrm{m}$ cells without re-tiling and render each cell as a 224×224 BEV map with its native scene graph.

Query conversion. We preserve the six released mentions and reuse the sentence template in Section B.3. Let d_i and a_i denote the source direction and node assignment of mention i, respectively. The facing anchor is selected in source order as

$$i_f = \min\{i : a_i \neq \varnothing,\ d_i \in \{\textit{north}, \textit{south}, \textit{east}, \textit{west}\}\}. \tag{18}$$

Poses for which i_f is undefined are discarded. The anchor defines the observation frame, and the remaining relations are assigned using Section B.2. We retain the pose when

$$a_i \neq \varnothing, \qquad r_i \in \{\textit{left}, \textit{right}, \textit{behind}\}, \quad \forall i \neq i_f, \tag{19}$$

and discard it otherwise. This retains the original six mentions without introducing replacement supports. The resulting split contains 28,807 training, 3,187 validation, and 11,505 test queries.

Native-submap fine localization. For the experiment reported in the main paper, every method receives the ground-truth native cell containing the target. All methods are retrained on the same native egocentric split, and predictions are transformed to world coordinates for evaluation with Recall@$\{5, 10, 15\}$ m. This protocol holds the coarse input fixed and evaluates fine localization within the original $30\,\mathrm{m} \times 30\,\mathrm{m}$ context.

Diagnostic with fixed retrieved submaps. The protocol above holds the native cell fixed to isolate fine localization. As a complementary diagnostic, we replace this cell with one frozen Top-1 native-cell prediction from CMMLoc for each validation and test query. The resulting retrieval manifest is fixed before fine-localization evaluation and shared by all compared methods. All methods are trained on the same native training split and receive the same retrieved cell for each query.

We transform predictions to world coordinates and report Recall@$\{5, 10, 15\}$ m over all 11,505 test queries. The selected Top-1 cell contains the target location for 4,735 of 11,505 queries (41.16%); the remaining queries are retained and scored. This diagnostic holds the coarse-stage input fixed rather than comparing jointly optimized end-to-end retrieval and localization systems. It measures fine-localization behavior under the same retrieved local context.

Table D1: Localization performance on native KITTI360Pose with CMMLoc-retrieved submaps under a fixed retrieval input. Bold and underline denote the best and second-best result in each column, respectively.

Method	R@5 m↑	R@10 m↑	R@15 m↑
Text2Pos	17.00	34.00	**41.00**
Text2Loc	<u>19.63</u>	33.86	40.56
MNCL	**19.67**	<u>34.53</u>	39.23
CMMLoc	17.71	34.02	40.36
VLM-Loc	18.66	34.02	40.26
FoRLoc (Ours)	17.45	**34.65**	<u>40.76</u>

Under this fixed retrieved-submap input, FoRLoc has the highest R@10 m but does not lead at R@5 m or R@15 m. With only 41.16% of selected submaps containing the target, this diagnostic does not show a uniform advantage under retrieval errors.

D.2 Query Rewrite Examples

We draw a scene-stratified random subset of 200 queries from the EgoT2P test split and manually rewrite each selected egocentric query. Each rewrite varies wording and sentence structure while preserving the referenced objects, their attributes, and spatial relations; the local map and target position remain unchanged. We manually verify each rewrite against its original query for semantic consistency. The paired localization recall in Table 3 uses all 200 original–rewrite pairs. For illustration, Table D2 presents ten of these pairs from three test scenes.

Table D2: Ten illustrative original–rewrite query pairs from three EgoT2P test scenes, selected from the 200 evaluated pairs.

Original query	Rewritten query
The target location is facing a black building, a dark-green vegetation on its left, a black fence on its right, a black garage on its right, a black pole behind, and a black terrain behind.	At the target, you face a black building. A black pole is behind you, while a black fence is on your right. Dark-green vegetation is on your left; black terrain is also behind you, and a black garage is on your right.

Continued on next page

Table D2: Original and rewritten egocentric queries (continued).

Original query	Rewritten query
The target location is a gray-green sidewalk on its left, a gray road on its right, a dark-green wall on its left, facing a green pole, a dark-green terrain on its right, and a black vegetation on its right.	The gray-green sidewalk is on your left at the target, where you face a green pole. A gray road is on your right; the dark-green wall is on your left. Black vegetation is also to your right, as is dark-green terrain.
The target location is a gray road on its left, a green vegetation on its right, a beige sidewalk on its right, a black pole behind, a green trash bin behind, and facing a gray building.	A black pole and a green trash bin are behind you at the target. You face a gray building; the gray road is on your left, and green vegetation is on your right. The beige sidewalk is also on your right.
The target location is a gray road on its left, facing a black wall, a dark-green building on its right, a gray sidewalk on its right, a gray lamp behind, and a gray pole behind.	Behind you are a gray lamp and a gray pole. The target is where you face a black wall; a dark-green building is on your right, while the gray road is on your left. The gray sidewalk is also on your right.
The target location is a gray building on its left, a gray-green road on its right, a green sidewalk on its right, facing a black fence, a gray-green vegetation on its left, and a green wall on its left.	A gray-green road is on your right at the target, where you face a black fence. The gray building is on your left; a green sidewalk is on your right. A green wall and gray-green vegetation are each on your left.
The target location is a dark-green terrain on its left, facing a black fence, a green building on its right, a dark-green traffic sign on its left, a black vegetation on its right, and a gray sidewalk behind.	The gray sidewalk is behind you at the target. You face a black fence; a green building is on your right, while dark-green terrain is on your left. Black vegetation is also on your right, and the dark-green traffic sign is on your left.
The target location is facing a dark-green wall, a green vegetation on its left, a green building on its right, a gray-green traffic sign on its right, a green pole behind, and a green terrain behind.	Green vegetation and a green building are respectively on your left and right. You face a dark-green wall at the target; behind you are a green pole and green terrain. A gray-green traffic sign is also on your right.
The target location is a gray sidewalk on its left, a gray-green lamp on its right, a dark-green vegetation on its left, facing a green fence, a green building behind, and a dark-green garage behind.	A green building is behind you, and a gray sidewalk is on your left. At the target you face a green fence; a gray-green lamp is on your right. Dark-green vegetation is also on your left, while the dark-green garage is behind you.
The target location is a gray lamp on its left, a green fence on its right, a gray pole on its left, facing a green wall, a green building on its right, and a green garage behind.	A green garage is behind you, while a green fence is on your right. The target is where you face a green wall; a gray pole is on your left, and a green building is also on your right. A gray lamp is on your left.
The target location is a green box on its left, a beige garage on its right, a green sidewalk on its left, facing a dark-green wall, a green building behind, and a dark-green garage behind.	A green building and dark-green garage are behind you; the beige garage is on your right. You face a dark-green wall at the target, and the green box is on your left. A green sidewalk is also on your left.

D.3 Detailed Egocentric Condition Subsets

We form three overlapping diagnostic subsets directly from the fixed EgoT2P test metadata. For query q, let n_q^f denote the number of same-category facing-anchor candidates, u_q the number of null assignments, and n_q^r the number of grounded support mentions with relation r. We define

$$\begin{aligned} \mathcal{A} &= \{q : n_q^f > 1\}, \\ \mathcal{P} &= \{q : u_q > 0\}, \\ \mathcal{W} &= \{q : \min_{r \in \{\text{left},\text{right},\text{behind}\}} n_q^r = 0\}. \end{aligned} \tag{20}$$

Table D3: Map-evidence controls for FoRLoc on the EgoT2P test set. The object-level graph replaces SP-SHG; the no-BEV variant omits the BEV image.

Map input	$R@5\,\mathrm{m}\uparrow$	$R@10\,\mathrm{m}\uparrow$	$R@15\,\mathrm{m}\uparrow$
SP-SHG + BEV (FoRLoc)	31.37	65.11	82.28
Object-level graph + BEV	30.98	64.04	80.91
SP-SHG without BEV	28.25	62.59	80.30

These sets correspond to Ambiguous Reference (1,971 queries), Partial Context (2,521), and Weak Directional Support (2,423), respectively. They are not mutually exclusive; each query retains its original input, target, and prediction in every subset to which it belongs.

D.4 Two-Stage Direct Control

The component analysis in Table 5 includes a two-stage direct-prediction control. Its first stage takes the same BEV, SP-SHG, and egocentric query as FoRLoc, but predicts mention assignments and an initial target pixel $(\hat{A}, \hat{\mathbf{u}}_{p,1})$ without estimating the reference state. The second stage receives these predictions and estimates a refined target pixel $\hat{\mathbf{u}}_{p,2}$, which is converted to a map-frame position for evaluation. Both stages use independently optimized LoRA adapters with the same backbone and training schedule as FoRLoc. Their measured inference costs are reported in Table D4.

D.5 Map Evidence Controls

We compare the map inputs used by FoRLoc while retaining its two-stage reference-state and assignment predictions. The object-level graph control replaces SP-SHG with the representation described in Section C.1, while the no-BEV variant omits the BEV image. Table D3 reports their localization results on the EgoT2P test set. Replacing SP-SHG with an object-level graph reduces $R@10\,\mathrm{m}$ by 1.07 percentage points; removing BEV reduces it by 2.52 percentage points.

D.6 Inference Cost

Table D4 reports localization recall and inference cost for the evaluated systems on the EgoT2P test set. Time is the total duration of batched evaluation divided by the number of queries; for Two-Stage Direct, we sum the separately measured times of its two stages. FoRLoc achieves 31.37% at $R@5\,\mathrm{m}$, compared with 26.01% for VLM-Loc, while requiring 2.14 versus 0.85 seconds and generating 239.07 versus 145.32 output tokens per query. These results characterize the accuracy–cost trade-off of the complete FoRLoc system under the evaluated protocol. Two-Stage Direct requires 0.89 seconds per query and attains 23.08% recall; its lower runtime makes it a two-stage alternative rather than a matched-compute control.

Table D4: Localization accuracy and amortized inference cost on the EgoT2P test set. Two-Stage Direct time sums both stages; queries/s is effective throughput.

Method	$R@5\,\mathrm{m}\uparrow$	Time (s/query)↓	Output tokens/query↓	Queries/s↑
VLM-Loc	26.01	0.85	145.32	1.18
Two-Stage Direct	23.08	0.89	130.37	1.12
FoRLoc	31.37	2.14	239.07	0.47

D.7 Qualitative Results

Figure D1 presents additional qualitative comparisons between FoRLoc and the baseline methods on EgoT2P. The examples cover different scene layouts and query configurations, with the final row showing a challenging case in which all compared methods remain outside the $5\,\mathrm{m}$ threshold. These results complement the qualitative examples in the main paper by illustrating both accurate localization and a remaining failure case.



Figure D1: Additional qualitative localization results on EgoT2P. Each row shows an egocentric query and the predictions of six compared methods on the corresponding BEV map. Black and red circles denote the ground-truth and predicted positions, respectively. The localization error is reported below each prediction; green and red borders indicate errors below and above $5\,\mathrm{m}$, respectively. The query is displayed as six relation statements are separated for readability; the model receives the original single-sentence query.

E LIMITATIONS

This work focuses on T2P fine localization within a controlled local-map setting. Paired EgoT2P queries enable a controlled comparison of map-frame and egocentric directional formulations, while FoRLoc's ablations examine its intermediate predictions within this setting. The benchmark uses template-generated descriptions and semantic-instance map structure from KITTI-360. Performance with fully naturally authored queries, other environments, and maps built from predicted semantic instances remains to be evaluated.