DISCO: DISAMBIGUATING MULTI-MODAL FLOORPLAN LOCALIZATION VIA $SE(2)$-AWARE CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual **F**loorplan **Loc**alization (FLoc) aims to estimate camera poses by aligning visual observations with floorplans, yet it remains challenging due to structural aliasing in repetitive indoor environments. Such ambiguity causes distinct poses to exhibit similar visual-floorplan compatibility, resulting in multi-modal localization distributions with limited spatial separability and angular discriminability. Existing ray matching-based methods can generate geometry-consistent pose candidates, but lack effective mechanisms to disambiguate them. To address this issue, we propose **DisCo**, a visual-floorplan **Co**ntrastive **Dis**ambiguation framework for robust visual FLoc. First, we employ a depth-aware ray regression predictor to project monocular RGB observations into 2D ray primitives and generate multi-modal pose candidates through ray-floorplan matching. We further introduce a clustering-based hypotheses preparation strategy to extract representative FLoc hypotheses by exploiting the $SE(2)$ distribution of candidate poses. Second, we propose an $SE(2)$-aware contrastive disambiguation strategy to learn fine-grained visual-floorplan compatibility beyond geometric matching. Through structured pose perturbations, DisCo improves local pose smoothness, spatial separability, and angular discriminability by capturing implicit structural semantics in local floorplan layouts without relying on explicit semantic annotations. Experiments on three visual FLoc benchmarks show that DisCo improves pose recall over its candidate-generation baselines. On S3D (full), it achieves 75.0% R@1 m and 74.3% R@1 m, 30° without explicit semantic annotations.

1 INTRODUCTION

Camera localization is a fundamental task in computer vision, supporting various applications including robotics (Zhong et al., 2026; Chen et al., 2026a;b) and augmented reality (Chen et al., 2022). Indoor visual localization remains challenging due to complex environments and the absence of reliable satellite signals. Conventional approaches typically rely on pre-built 3D models (Liu et al., 2017; Sarlin et al., 2019; Sattler et al., 2016) or large-scale image retrieval databases (Balntas et al., 2018; Arandjelovic et al., 2017), which require substantial storage and continuous maintenance. In contrast, building floorplans provide persistent geometric references that are robust to appearance variations, making visual **F**loorplan **Loc**alization (FLoc) (Chen et al., 2024) a promising solution for scalable indoor localization.

However, visual FLoc faces two fundamental challenges. The first challenge is the cross-modal gap between RGB observations and minimalist floorplans, as shown in Fig. 1(a)-(b). Monocular images contain rich appearance information, such as furniture, illumination changes, and object-level clutter, whereas floorplans mainly preserve scene layouts. Existing methods bridge this gap by extracting compact geometric representations from observations (Min et al., 2022) or transforming images into ray primitives for ray-floorplan matching (Chen et al., 2025a;b; Grader & Averbuch-Elor, 2025). Although these approaches can generate geometry-consistent pose candidates, they mainly focus on candidate retrieval, while the ambiguity among candidates remains unresolved.

The second challenge is structural aliasing caused by repetitive indoor layouts, which severely limits *spatial separability* and *angular discriminability*. *Spatial separability* refers to distinguishing physically distant locations with similar local structures, while *angular discriminability* denotes identify-

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.

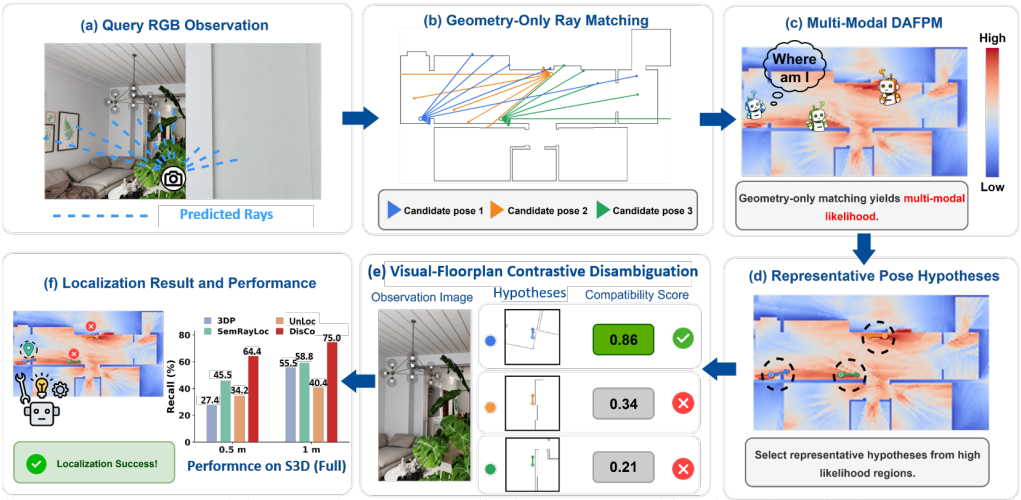

Figure 1: Due to modal gap and structural aliasing, ray matching-based methods generate multi-modal FLoc distributions (a)-(c). Our method first extracts representative FLoc hypotheses from DAFPM and then performs visual-floorplan contrastive disambiguation to compute compatibility scores (d)-(e), achieving superior visual FLoc without relying on explicit semantic annotations (f).

ing the correct orientation among symmetric observations. As shown in Fig. 1(c), structural aliasing leads to multiple plausible poses with similar matching scores, resulting in inherently multi-modal localization distributions. Existing approaches introduce additional semantic cues, such as object categories or room functions, to alleviate this ambiguity (Grader & Averbuch-Elor, 2025). However, explicit semantics are costly to obtain, and high-level semantic categories may still fail to capture fine-grained visual-floorplan correspondences required for precise $SE(2)$ pose disambiguation.

To address these challenges, we propose **DisCo**, a visual-floorplan **Co**ntrastive **Dis**ambiguation framework by learning $SE(2)$-aware compatibility between RGB observations and floorplan structures. First, we adopt a depth-aware **R**ay **R**egression **P**redictor (RRP) as a geometry-aware candidate generator. RRP projects monocular RGB observations into compact 2D ray primitives by suppressing irrelevant vertical visual clutter. Matching the predicted rays with floorplan layouts produces a **D**epth-**A**ware **F**Loc **P**robability **M**ap (DAFPM), which provides high-recall multi-modal pose candidates. However, directly disambiguating all candidates is inefficient since high-probability candidates usually concentrate around several local modes, as shown in Fig.1(c). Therefore, we further introduce a clustering-based hypotheses preparation strategy to extract representative FLoc hypotheses from pose candidates according to their $SE(2)$ spatial and angular distributions. This process removes redundant candidates while preserving distinct ambiguity modes, providing compact and informative hypotheses for subsequent disambiguation, as shown in Fig.1(d).

Second, we propose an $SE(2)$-aware contrastive disambiguation strategy to refine these FLoc hypotheses. Through contrastive learning, slack positives generated by local pose perturbations encourage local *pose smoothness*, while position-level and orientation-level negatives explicitly enhance *spatial separability* and *angular discriminability*. Instead of relying on explicit semantic annotations, DisCo learns a visual-floorplan compatibility function between RGB observations and local floorplan structures, as shown in Fig.1(e). Such compatibility captures implicit structural semantics encoded in local layouts, including spatial configurations and functional characteristics, enabling fine-grained correspondence between observations and FLoc hypotheses. Normalizing the visual-floorplan compatibility scores over the retained FLoc hypotheses yields a **D**isambiguation **P**robability **M**ap (DPM), which is fused with DAFPM to obtain the final visual FLoc result, as shown in Fig.3(c).

Experiments on Gibson(f/g), S3D (full), and ZInD show gains from DisCo's hypothesis reranking, with the strongest comparative results on Gibson(f/g) and at the 1 m and joint position–orientation thresholds on S3D (full). Moreover, DisCo is complementary to existing semantic-based candidate generators (Grader & Averbuch-Elor, 2025) and can serve as a general ambiguity resolution module. Our main contributions are summarized as follows:

(1) We propose DisCo, a visual-floorplan contrastive disambiguation framework for visual FLoc. Based on ray-based candidate generation, we introduce a clustering-based hypotheses preparation strategy to extract representative FLoc hypotheses and reduce redundant ambiguity modes. **(2)** We design an $SE(2)$-aware contrastive disambiguation strategy that learns fine-grained visual-floorplan compatibility without relying on explicit semantic annotations. Through pose-structured positive and negative sampling, the proposed method improves local pose smoothness, spatial separability, and angular discriminability. **(3)** Comparative and ablation studies on three visual FLoc benchmarks demonstrate the effectiveness of the proposed reranking method. The code and checkpoints will be made publicly available to facilitate research in related fields.

2 Related Work

Visual Localization. Visual localization is a fundamental problem in computer vision and has been extensively studied. Image retrieval-based methods (Balntas et al., 2018; Arandjelovic et al., 2017) estimate camera poses by searching similar images from large-scale databases. Structure-from-motion approaches (Panek et al., 2022; Sarlin et al., 2019) construct 3D scene models and establish 2D-3D correspondences through local feature matching and robust pose estimation techniques (Kukelova et al., 2008; Fischler & Bolles, 1981; Barath & Matas, 2021). Scene coordinate regression methods (Brachmann et al., 2017) learn pixel-wise 3D correspondences, while pose regression methods (Kendall & Cipolla, 2017) directly predict camera poses from images. However, these approaches typically require scene-specific 3D models or large-scale image databases, limiting scalability and long-term deployment. In contrast, visual FLoc leverages widely available floorplans as lightweight geometric references for indoor localization.

Visual Floorplan Localization. Visual FLoc is closely related to LiDAR-based **M**onte **C**arlo **L**ocalization (MCL) (Dellaert et al., 1999; Chu et al., 2015; Mendez et al., 2018; Winterhalter et al., 2015), which estimates poses by matching sensor observations with geometric maps. However, LiDAR dependence limits its deployment on common visual platforms. Early visual FLoc methods (Boniardi et al., 2019; Chu et al., 2015; Howard-Jenkins & Prisacariu, 2022; Howard-Jenkins et al., 2021; Min et al., 2022) explore monocular or panoramic localization by matching visual cues with floorplan structures, using handcrafted priors (Boniardi et al., 2019), learned representations (Min et al., 2022), or synthetic views rendered from floorplans (Howard-Jenkins & Prisacariu, 2022; Howard-Jenkins et al., 2021). However, their reliance on panoramic inputs or known camera configurations limits general monocular applications.

Recent monocular visual FLoc methods (Karkus et al., 2018; Min et al., 2022) mainly formulate localization as matching visual observations with floorplan-derived representations. Ray matching-based approaches (Chen et al., 2024; 2025a;b; Wüest et al., 2025) further represent observations as compact geometric primitives and align predicted rays with floorplan geometry. Despite their effectiveness, these methods mainly focus on retrieving geometry-consistent candidates, while structural aliasing remains unresolved. Repetitive layouts often produce multiple candidates with similar compatibility scores, resulting in multi-modal localization distributions. To alleviate this ambiguity, semantic-enhanced methods (Min et al., 2022; Mendez et al., 2020; Grader & Averbuch-Elor, 2025) introduce cues such as room categories, doors, and windows, while (Chen et al., 2025b) explores unsupervised multi-modal alignment to reduce annotation dependency. However, these approaches mainly improve candidate generation and lack explicit modeling of fine-grained visual-floorplan compatibility for disambiguating poses within the same environment.

Different from existing approaches, DisCo is designed as a general ambiguity resolution framework. It accepts FLoc hypotheses derived from DAFPM and then learns an $SE(2)$-aware visual-floorplan compatibility function to refine these hypotheses. By exploiting structured contrastive constraints, DisCo improves spatial separability and angular discriminability without relying on explicit semantic annotations or additional 3D scene information.

3 Methodology

3.1 Problem Formulation

This work formulates visual FLoc as estimating a camera pose in the $SE(2)$ space by aligning a monocular RGB observation with a 2D floorplan. Given a floorplan F and an RGB image $\mathcal{I} \in \mathbb{R}^{h\times w\times 3}$, the objective is to estimate the absolute camera pose $S_{\mathcal{I},F} = (x, y, \theta)$, where (x, y) and θ denote the spatial position and viewing orientation, respectively. Instead of deterministic regression,

we model FLoc as estimating the posterior distribution:

$$\hat{S}_{\mathcal{I},F} = \arg\max_{S_i \in \mathcal{S}} p(S_i|\mathcal{I}, F), \tag{1}$$

where the continuous $SE(2)$ space is discretized into candidate poses $\mathcal{S} = \{S_i\}$, forming a probability volume $P \in \mathbb{R}^{\hat{\mathcal{H}} \times \hat{\mathcal{W}} \times O}$, where each element $P(S_i)$ represents the probability $p(S_i|\mathcal{I}, F)$ for a candidate pose S_i. Here, $\hat{\mathcal{H}}$ and $\hat{\mathcal{W}}$ denote the number of discretized cells in the x and y dimensions, respectively. O represents the number of orientation bins. In repetitive indoor environments, structural aliasing causes multiple spatially distant or orientation-inconsistent poses to exhibit similar visual-floorplan compatibility. A reliable FLoc method requires not only generating plausible pose candidates, but also effectively disambiguating competing candidates.

3.2 Stage 1: RRP-based Hypotheses Preparation

Given an RGB image, existing methods predict purely geometric rays (e.g., F^{3}Loc (Chen et al., 2024) and UnLoc (Wüest et al., 2025)) and/or additional semantic rays (e.g., SemRayLoc Grader & Averbuch-Elor (2025)), which are matched with floorplan-derived ground-truth rays. Despite their differences, these approaches share the same paradigm: generating multi-modal pose distribution through ray-based matching.

RRP-based Candidate Generator. As shown in Fig. 2, we adopt a depth-aware RRP (Wüest et al., 2025) as an example ray-based candidate generator. Specifically, the depth-aware visual features pass through a convolutional layer to generate keys and values. A vertically weighted summation is subsequently applied to form 1D queries. This operation explicitly suppresses vertical visual clutter and isolates horizontal layout structures, effectively minimizing the impact of 3D occlusions. For spatial awareness, the query's positional encodings are derived from 1D coordinates, while the positional encodings of key and value are mapped from 2D image coordinates. For each query, attention is applied across the image to predict the distance to the nearest wall. The depth prediction is formulated as a weighted sum of discrete bins:

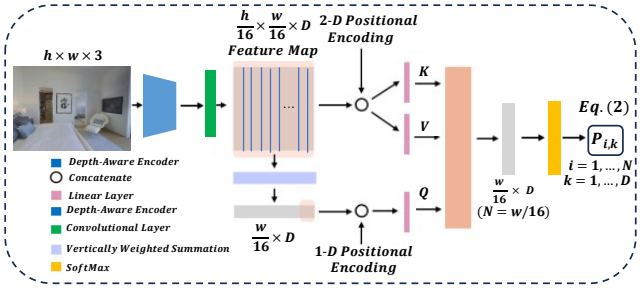

Figure 2: An illustration of the depth-aware RRP, which maps depth estimation expertise to probabilities across different bins in Eq. (2). The RRP employs a depth-aware encoder, such as DINO-V2 (Oquab et al., 2023) in Depth Anything V2 (Zhao, 2024), to extract visual features.

$$d_i = \sum_{k=1}^{D} P_{i,k}\, d_k, \quad d_k = \left(d_{\text{min}}^{\gamma} + \frac{k}{D}(d_{\text{max}}^{\gamma} - d_{\text{min}}^{\gamma}) \right)^{1/\gamma}, \tag{2}$$

where $P_i \in \mathbb{R}^D$ denotes the predicted probability distribution across D bins for the i-th ray, ensuring $\sum_{k=1}^{D} P_{i,k} = 1$ via a softmax function. The parameter N ($i = 1, \ldots, N$) denotes the total number of predicted rays. The terms d_{max} and d_{min} define the depth boundaries, while the power-law discretization parameter γ controls the resolution distribution. The RRP is optimized using a joint L_1 and cosine similarity-based shape loss:

$$\mathcal{L}_{FLoc} = ||\mathbf{d} - \mathbf{d}^*||_1 + 1 - \frac{\mathbf{d}^\top \mathbf{d}^*}{\max\{||\mathbf{d}||_2 ||\mathbf{d}^*||_2, \epsilon\}}, \tag{3}$$

where $\mathbf{d}$ and $\mathbf{d}^*$ represent the predicted and ground-truth 2D-ray depths respectively, and ϵ is a small constant to prevent division by zero.

During inference, the predicted rays are matched with floorplan-derived ground-truth rays (Chen et al., 2024), producing a probability volume $P_d \in [0, 1]^{\hat{\mathcal{H}} \times \hat{\mathcal{W}} \times O}$. Instead of a deterministic prediction, P_d provides a multi-modal candidate pose distribution, *i.e.*, a DAFPM, as shown in Fig. 1(c). In this work, we first extract Top-K pose candidates from DAFPM and then propose DisCo to disambiguate them.

Clustering-based Hypotheses Preparation. However, since high-likelihood pose candidates frequently cluster around a few local modes, directly disambiguating all Top-K candidates introduces

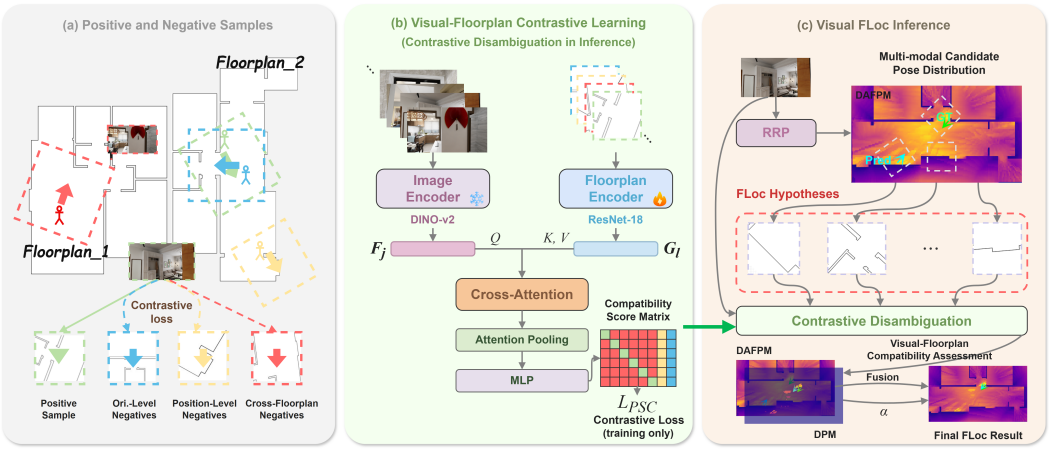

Figure 3: (a) shows the spatially perturbed sampling strategy, which constructs slack positives, position-level negatives, and orientation-level negatives according to the $SE(2)$ pose structure. (b) details the proposed visual-floorplan contrastive disambiguation. (c) DisCo performs a visual-floorplan compatibility assessment of the FLoc hypotheses derived from the DAFPM to obtain a DPM. Fusing DAFPM with DPM suppresses ambiguous hypotheses, and yields the final FLoc.

severe computational redundancy. Therefore, we perform a $SE(2)$-aware candidate refinement to extract representative FLoc hypotheses before contrastive disambiguation. We first sort all pose candidates by their DAFPM likelihoods. Then, we iteratively retain the highest-scoring unassigned candidate and suppress redundant neighbors within its local $SE(2)$ basin:

$$d_{SE(2)}(S_i, S_j) = \sqrt{\frac{\|\mathbf{t}_i - \mathbf{t}_j\|_2^2}{\sigma_t^2} + \lambda_\theta \frac{\Delta(\theta_i, \theta_j)^2}{\sigma_\theta^2}}, \tag{4}$$

where $\mathbf{t}_i = (x_i, y_i)$ and $\Delta(\theta_i, \theta_j) = \text{wrap}(\theta_i - \theta_j)$. A pose candidate S_j is suppressed if $d_{SE(2)}(S_i, S_j) \leq \rho$. *We conduct parametric studies of σ_t, σ_θ, λ_θ, and ρ in the Appendix C.1.* The retained set $\mathcal{R}$ subsequently consists only of spatially and angularly distinct representative pose hypotheses. We experimentally find that the number of pose hypotheses typically remains within 10^2, which is manageable. By controlling the value of K or the number of FLoc hypotheses, we can strike a balance between preserving the correct pose mode and the computational cost.

3.3 Stage 2: $SE(2)$-Aware Contrastive Disambiguation

Visual-Floorplan Compatibility Modeling. For each candidate pose $S_l = (x_l, y_l, \theta_l)$, we crop the floorplan at (x_l, y_l) and rotate it by $-\theta_l$, aligning the candidate heading to a canonical direction. We score the query image against this view-aligned crop to disambiguate Stage 1 hypotheses without explicit semantic annotations.

Let $\mathbf{F}_j$ and $\mathbf{G}_l$ denote image and local floorplan tokens at candidate S_l. Cross-attention uses $\mathbf{F}_j$ as queries and $\mathbf{G}_l$ as keys/values; residual addition and layer normalization yield candidate-conditioned tokens $\mathbf{H}_{j,l}$. Learned attention pooling and an MLP produce the compatibility score:

$$c(\mathcal{I}_j, F, S_l) = s_\theta(\mathbf{F}_j, \mathbf{G}_l) = \text{MLP}(\text{AttnPool}(\mathbf{H}_{j,l})). \tag{5}$$

The shared network scores each pair independently (Appendix A.1). We construct contrastive samples according to the structures of the $SE(2)$ pose space rather than treating all poses uniformly, as shown in Fig.3(a). These pose-structured constraints guide the model to learn meaningful visual-floorplan compatibility, maintaining local pose smoothness while enhancing spatial separability and angular discriminability.

Slack Positives for Local Pose Smoothness. For a query image $\mathcal{I}_j$ with a ground-truth pose $S_j = (x_j, y_j, \theta_j)$, we construct a slack positive by applying a minor pose perturbation:

$$S_j^+ = S_j \oplus u, \quad u \sim \mathcal{U}(-B, B), \tag{6}$$

where B defines the perturbation boundary for translation and orientation, and $\oplus$ denotes pose composition in $SE(2)$. The local floorplan structure cropped at S_j^+ is encoded into tokens $\mathbf{G}_j^+$. This slack positive establishes a local pose neighborhood around the ground truth, encouraging consistent visual-floorplan compatibility under minor pose variations.

Position-Level Negatives for Spatial Separability. We construct position-level negatives $(\mathbf{F}_j, \mathbf{G}_m^{p-})$ by sampling poses with distant translations while maintaining comparable orientations. These samples often correspond to visually similar local structures at different physical positions. By explicitly distinguishing them from positives, the model learns to suppress spatially ambiguous matches and improves *spatial separability*.

Orientation-Level Negatives for Angular Discriminability. We construct orientation-level negatives $(\mathbf{F}_j, \mathbf{G}_m^{a-})$ by changing the viewing orientation at the correct physical position. These samples enforce the discrimination of FLoc hypotheses with correct positions but incorrect headings, thereby improving *angular discriminability* and mitigating directional ambiguity.

Cross-Floorplan Negatives. We retain standard cross-floorplan negatives $(\mathbf{F}_j, \mathbf{G}_m^{c-})$ by treating floorplan features from other batch samples as generic negatives. They provide coarse scene-level separation, while the proposed position-level and orientation-level negatives specifically address structural aliasing within the same floorplan.

We learn the visual-floorplan compatibility function by optimizing a single-positive **P**ose-**S**tructured **C**ontrastive (PSC) loss. For concise notation, we define the positive score and three negative partition terms as follows:

$$\begin{aligned} z_j^+ &= \exp(s_\theta(\mathbf{F}_j, \mathbf{G}_j^+)/\tau), \quad Z_j^c = \sum_{m \in \mathcal{N}_j^c} \exp(s_\theta(\mathbf{F}_j, \mathbf{G}_m^{c-})/\tau), \\ Z_j^p &= \sum_{m \in \mathcal{N}_j^p} \exp(s_\theta(\mathbf{F}_j, \mathbf{G}_m^{p-})/\tau), \quad Z_j^a = \sum_{m \in \mathcal{N}_j^a} \exp(s_\theta(\mathbf{F}_j, \mathbf{G}_m^{a-})/\tau). \end{aligned} \tag{7}$$

Here, $\mathcal{N}_j^c$, $\mathcal{N}_j^p$, and $\mathcal{N}_j^a$ index the three negative types. Image and floorplan tokens use a frozen DINOv2-S backbone (Oquab et al., 2023) and trainable ResNet-18 (He et al., 2016), respectively. With a learned temperature τ and batch $\mathcal{B}$, we optimize:

$$\mathcal{L}_{PSC} = -\frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \log \frac{z_j^+}{z_j^+ + Z_j^c + Z_j^p + Z_j^a}. \tag{8}$$

$SE(2)$-Aware DisCo. At inference, the same network scores each hypothesis: $c_m = s_\theta(\mathbf{F}_j, \mathbf{G}_m)$. The compatibility scores are normalized to obtain the DPM: $P_c(R_m) = \mathrm{softmax}_m(c_m)$. We fuse the DPM with the geometric likelihood to select the pose:

$$\hat{S}_{\mathcal{I}_j, F} = \arg\max_{R_m \in \mathcal{R}} \left(\log P_d(R_m | \mathcal{I}_j, F) + \alpha P_c(R_m) \right). \tag{9}$$

Here, α weights raw scores without the training temperature. Image tokens are cached and candidate scoring is batched.

4 Experiments

4.1 Experimental Setup

Datasets. Based on existing studies (Chen et al., 2025b; Grader & Averbuch-Elor, 2025), we conduct comparative and ablation studies on three visual FLoc benchmarks, including the Gibson (f/g) (Chen et al., 2024), S3D (full) (Zheng et al., 2020), and ZinD (Cruz et al., 2021) datasets. They provide complementary evaluation settings across synthetic and real-world scenes. Gibson(g) consists of general motions, including in-place steering motions, while Gibson(f) consists of only forward motions. Therefore, Gibson(g) is intuitively more challenging than Gibson(f). S3D (full) is a large-scale synthetic dataset containing 3,500 designer-created houses with photorealistic rendered observations. The heavy presence of furniture, 3D object occlusions, and repetitive indoor layouts makes this dataset particularly challenging. ZInD is a real-world dataset comprising 1,575 unfurnished residential homes and 59,361 panoramic images. Following standard practice (Grader & Averbuch-Elor, 2025), panoramas are converted into perspective views for visual FLoc.

Baselines. We compare our method against four categories of baselines. *The implementation details of our method can be found in Appendix A.1.*

(1) Early Methods. PF-net (Karkus et al., 2018), MCL (Dellaert et al., 1999), and LASER (Min et al., 2022). **(2) Strong Baselines under the F^{3}Loc Framework.** F^{3}Loc (Chen et al., 2024) is a representative probabilistic method that combines a ray-based observation module with histogram filtering. 3DP (Chen et al., 2025a) injects 3D geometric priors into the F^{3}Loc framework to improve localization accuracy without semantic labels. RSK (Chen et al., 2025b) employs unsuper-

Table 1: Comparative studies on the Gibson(f) and Gibson(g) datasets. * indicates the code is not open source and we reproduced the results by ourselves.

Method (Venue)	Gibson(f) R@				Gibson(g) R@			
	0.1 m↑	0.5 m↑	1 m↑	1 m 30°↑	0.1 m↑	0.5 m↑	1 m↑	1 m 30°↑
PF-net(CoRL 2018)	0	2.0	6.9	1.2	1.0	1.9	5.6	1.9
MCL(ICRA 1999)	1.6	4.9	12.1	8.2	2.3	6.2	9.7	7.3
LASER(CVPR 2022)	0.4	6.7	13.0	10.4	0.7	7.0	11.8	9.5
F^3Loc(CVPR 2024)	4.7	28.6	36.6	35.1	4.3	26.7	33.7	32.3
3DP(ACM MM 2025)	5.3	33.2	39.8	38.4	9.4	37.4	43.1	41.5
RSK(AAAI 2026)	8.3	38.5	45.3	43.6	8.7	36.4	42.3	40.4
3DP & RSK	10.9	42.7	47.9	46.5	10.7	38.8	44.4	42.8
UnLoc* (ICLR 2026)	11.8	45.0	49.6	48.3	12.3	45.0	49.9	48.2
Ours	**13.1**	**50.9**	**56.7**	**55.4**	**12.4**	**48.0**	**55.6**	**54.0**

vised room-style knowledge learning to mitigate FLoc ambiguity. 3DP & RSK adaptively leverages both geometric priors and room-style knowledge (*see Appendix A.2 for details*). UnLoc (Wüest et al., 2025) employs a deep estimation encoder trained on large-scale data as the observation model and incorporates localization uncertainty. **(3) 3D Reconstruction-based Strong Baseline.** SceneAligner (Cho et al., 2026) projects 3D reconstructions onto dense maps and aligns them with floorplans as a reconstruction-based visual FLoc baseline. **(4) Strong Baselines under the SemRayLoc Framework.** SemRayLoc$_s$ (Grader & Averbuch-Elor, 2025) leverages sparse floorplan semantics to predict semantic rays to reduce geometric ambiguity. By adapting visual pre-trainings from 3DP and RSK to SemRayLoc$_s$, three enhanced variants are derived: + 3DP, + RSK, and + 3DP & RSK (*see Appendix A.3 for details*). SemRayLoc$_r$ further uses room-type labels; FoD (Ye et al., 2026) fuses depth and floorplan semantics.

Metrics. Following existing work (Chen et al., 2024), we report recall metrics computed at positional localization accuracies of 0.1 m, 0.5 m, and 1 m. To evaluate both positional and angular localization accuracy, we also report recall under a joint threshold of 1 m positional error and 30° orientation error. Recall is calculated as the percentage of predictions falling within these thresholds.

Table 2: Comparative studies on the S3D (full) dataset. Sem. indicates whether semantics are used.

Method (Venue)	S3D (full) R@				Sem.
	0.1 m↑	0.5 m↑	1 m↑	1 m 30°↑	
PF-net(CoRL 2018)	0.2	1.3	3.2	0.9	no
MCL(ICRA 1999)	1.3	5.2	7.8	6.4	
LASER(CVPR 2022)	0.7	6.4	10.4	8.7	
F^3Loc(CVPR 2024)	1.5	14.6	22.4	21.3	
3DP(ACM MM 2025)	5.6	27.4	55.5	24.0	
RSK(AAAI 2026)	6.4	28.6	56.9	25.2	
3DP & RSK	6.7	26.8	54.7	24.2	
UnLoc(ICLR 2026)	5.5	34.2	40.4	39.3	
SceneAligner(Arxiv 2026)	3.5	37.5	53.8	51.6	
SemRayLoc$_s$(ICCV 2025)	5.4	41.9	53.5	52.6	yes
+ 3DP(ACM MM 2025)	5.5	46.6	56.7	56.2	
+ RSK(AAAI 2026)	6.2	48.1	59.9	58.8	
+ 3DP & RSK	7.1	48.9	61.5	60.0	
SemRayLoc$_r$(ICCV 2025)	5.7	45.5	58.8	57.5	
FoD$_d$(CVPR 2026)	11.4	56.4	61.4	60.8	
FoD(CVPR 2026)	**12.0**	65.2	71.9	71.4	
Ours (InfoNCE)	8.9	52.6	60.8	59.9	no
Ours (DisCo)	10.0	64.4	75.0	74.3	
SemRayLoc+DisCo	9.0	**67.5**	**81.5**	**81.1**	yes

4.2 Comparisons with SOTA Methods

Results on Gibson(f) and Gibson(g). Tab. 1 presents comparisons with SOTA methods on the Gibson benchmarks. Existing ray matching-based methods improve visual FLoc by generating geometry-consistent pose candidates, but remain vulnerable to structural aliasing among these candidates. DisCo consistently achieves the best performance on both datasets. Specifically, DisCo improves R@1 m/R@1 m 30° to $56.7\%/55.4\%$ on Gibson(f) and $55.6\%/54.0\%$ on Gibson(g), outperforming the previous best semantic-free method UnLoc by large margins. This improvement demonstrates the effectiveness of our $SE(2)$-aware hypothesis preparation and contrastive disambiguation, which further distinguish ambiguous FLoc hypotheses beyond geometric matching.

Results on S3D (full). Tab. 2 reports results on the challenging S3D (full) dataset, which contains severe occlusions and repetitive indoor structures. Existing methods, including semantic-enhanced approaches, remain limited by ambiguous FLoc hypotheses generated from similar visual-floorplan patterns. DisCo achieves 75.0% R@1 m and 74.3% R@1 m30° without using any explicit semantic annotations, exceeding the listed baselines at these two thresholds. Compared with SceneAligner, DisCo improves R@1 m 30° by 22.7 percentage points, demonstrating the effectiveness of hypothesis preparation and $SE(2)$-aware contrastive disambiguation. Moreover, combining DisCo with

SemRayLoc further boosts performance to 81.5% R@1 m and 81.1% R@1 m 30°, validating that DisCo provides a general ambiguity resolution module complementary to different candidate generation strategies.

Results on ZInD. Tab. 3 presents results on the real-world ZInD dataset. Despite the large appearance variations and complex residential layouts, DisCo achieves 53.8% R@1 m and 48.0% R@1 m 30° without using semantic annotations, exceeding the semantic-free baselines listed in the table at these thresholds. This demonstrates that the learned visual-floorplan compatibility can effectively capture implicit structural cues beyond explicit semantics. Furthermore, integrating DisCo with SemRayLoc improves R@0.5 m from 23.4% to 42.7% and improves R@1 m 30° from 29.2% to 54.0%, validating its effectiveness as a general ambiguity resolution module for visual FLoc.

Table 3: Comparative experimental results on the ZInD dataset. Sem. indicates whether to predict semantic rays during inference.

Method(Venue)	**Sem.**	**0.1 m**↑	0.5 m↑	**1 m**↑	**1 m 30°**↑
LASER(CVPR 2022)	No	1.4	11.1	17.6	13.6
F^3Loc(CVPR 2024)	No	0.7	7.9	15.1	11.5
UnLoc*(ICLR 2026)	No	1.5	11.4	16.7	13.8
Ours	No	4.2	37.3	53.8	48.0
SemRayLoc(ICCV 2025)	Yes	3.3	23.4	31.9	29.2
SemRayLoc+DisCo	Yes	5.0	42.7	60.0	54.0

Table 4: Ablation studies on S3D (full). (a) evaluates inner-floorplan position-level negatives (I-Pos-N) and orientation-level negatives (Ori-N), while retaining cross-floorplan negatives. (b) evaluates slack positive sampling with positional perturbation (P-Pert) and angular perturbation (A-Pert), while retaining all negatives.

(a) Within-floorplan hard negatives

Ablations		**S3D (full) R@**			
I-Pos-N	Ori-N	**0.1 m**↑	**0.5 m**↑	**1 m**↑	**1 m 30°**↑
✗	✗	9.0	60.4	70.7	70.0
✓	✗	9.5	62.4	73.5	72.8
✗	✓	9.5	62.8	73.7	73.0
✓	✓	10.0	64.4	75.0	74.3

(b) Slack positives

Ablations		**S3D (full) R@**			
P-Pert	A-Pert	**0.1 m**↑	**0.5 m**↑	**1 m**↑	**1 m 30°**↑
✓	✗	8.6	54.5	63.4	62.7
✗	✓	8.9	57.8	67.0	66.1
✓	✓	10.0	64.4	75.0	74.3

4.3 Ablation Studies

Effectiveness of Pose-Structured Contrastive Disambiguation. Compared with standard InfoNCE, DisCo improves R@1m from 60.8% to 75.0% and R@1 m 30° from 59.9% to 74.3% under the same candidate generation pipeline (Tab. 2). The significant gain verifies that uniform contrastive learning cannot effectively handle $SE(2)$-structured ambiguities. By introducing pose-aware positives and negatives, DisCo learns finer visual-floorplan compatibility and improves spatial and angular disambiguation.

Effect of $SE(2)$-Aware Contrastive Constraints. Tab. 4(a) evaluates the impact of within-floorplan hard negatives. Removing both position-level and orientation-level negatives degrades R@1 m 30° from 74.3% to 70.0%, indicating that generic cross-floorplan negatives are insufficient for resolving structural aliasing. Introducing either position-level or orientation-level negatives consistently improves localization accuracy, while combining both achieves the best performance. This demonstrates that spatial and angular constraints provide complementary benefits for enhancing spatial separability and angular discriminability.

Effect of Slack Positive Sampling. Tab. 4(b) investigates the influence of slack positives with different $SE(2)$ perturbations. Removing either positional or angular perturbations leads to noticeable performance drops, especially in the joint metric. Combining both perturbations improves R@1 m 30° from $62.7\%/66.1\%$ to 74.3%, demonstrating that slack positives effectively establish local pose neighborhoods and encourage smooth visual-floorplan compatibility under minor pose variations.

Table 5: Experimental results for average hypothesis numbers, localization performance, and inference time when Top-K takes different values.

Top-K	**Avg. Hypo.**	**Oracle R@1 m**↑	**R@1 m**↑	**R@1 m 30°**↑	**Inference Time (ms)**↓
50	8.7	83.20	67.90	66.98	26.16 ± 4.49
100	12.2	87.81	70.43	69.54	27.00 ± 5.48
250	20.2	92.29	72.65	71.82	29.95 ± 5.77
500	31.8	95.19	73.74	72.91	35.74 ± 6.88
1000	52.9	97.17	75.00	74.30	45.61 ± 8.32

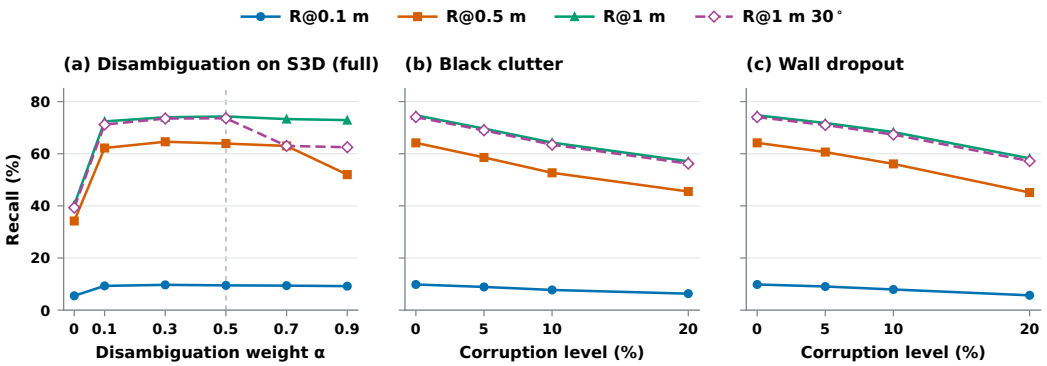

Figure 4: (a) An illustration of the parametric study of the disambiguation weight α on the S3D (full) dataset. (b) and (c) show that DisCo exhibits graceful degradation without crashing as the degree of corruption in the floorplan increases.

4.4 Parametric Studies

Impact of Top-K. All ablation studies are conducted on the S3D (full) dataset. Tab. 5 evaluates the impact of Top-K on candidate coverage and efficiency. Larger Top-K improves oracle recall from 83.20% to 97.17%, providing better coverage of the correct pose mode. After $SE(2)$-aware clustering, the retained FLoc hypotheses increase moderately, resulting in improved R@ 1m from 67.90% to 75.00%. Meanwhile, the inference time increases predictably from 26.16 ms to 45.61 ms, demonstrating that DisCo offers a controllable accuracy-efficiency trade-off and remains suitable for real-time localization. *More detailed parametric studies and analyses can be found in Appendix C.2.*

Impact of Disambiguation Weight α and Local Floorplan Size. As shown in Fig. 4(a), setting α to 0.3 or 0.5 helps achieve optimal visual FLoc performance. In this paper, we set α = 0.5. *For parametric studies and analyses of the local floorplan size, please see Appendix C.3.* In this paper, we set the local floorplan size to 5 m $\times$ 5 m.

Robustness to Imperfect Floorplans. We further evaluate the robustness of DisCo under imperfect floorplans by introducing two types of perturbations: black clutter, which adds spurious structures, and wall dropout, which removes partial wall structures. As shown in Tab. 4(b) and (c), DisCo exhibits graceful degradation as the corruption level increases. Even under severe 20% corruption, the performance remains stable without collapse, achieving over 57% R@1 m for both perturbation types. These results demonstrate that DisCo is robust to floorplan inaccuracies and environment-floorplan mismatches, benefiting from learned visual-floorplan compatibility rather than relying on exact geometric correspondence.

More computational efficiency analyses, comparative studies, visualizations, and discussion of failure cases can be found in Appendices D, E, F, and G, respectively.

5 Conclusion

In this work, we proposed DisCo, a general $SE(2)$-aware visual-floorplan contrastive disambiguation framework for visual FLoc. Different from existing methods that mainly focus on generating geometry-consistent pose candidates, DisCo explicitly addresses the structural ambiguity among candidate poses through clustering-based hypothesis preparation and pose-structured contrastive learning. The proposed framework improves spatial separability and angular discriminability without requiring explicit semantic annotations or scene-specific designs. Experiments across three benchmarks support the effectiveness of DisCo's pose-structured reranking, including substantial gains under fixed candidate-generation pipelines and strong 1 m and joint position–orientation recall on S3D (full). Moreover, DisCo exhibits strong robustness against imperfect floorplans, gracefully degrading rather than collapsing under substantial structural perturbations. We believe this work provides a general and reliable solution for visual FLoc and encourages future research on robust cross-modal scene understanding.

AI USE STATEMENT

We used generative AI tools to assist with building the experimental code framework, running experiments, and polishing the manuscript's language. The authors take responsibility for the experimental results, scientific claims, and final manuscript.

REFERENCES

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1–1, 2017.

Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 751–767, 2018.

Daniel Barath and Jiri Matas. Graph-cut ransac: Local optimization on spatially coherent structures. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4961–4974, 2021.

Federico Boniardi, Abhinav Valada, Rohit Mohan, Tim Caselitz, and Wolfram Burgard. Robot localization in floor plans using a room layout edge extraction network. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5291–5297. IEEE, 2019.

Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6684–6692, 2017.

Bolei Chen, Jiaxu Kang, Haonan Yang, Ping Zhong, and Jianxin Wang. Perspective from a higher dimension: Can 3d geometric priors help visual floorplan localization? In *Proceedings of the 33nd ACM International Conference on Multimedia*, 2025a.

Bolei Chen, Shengsheng Yan, Yongzheng Cui, Jiaxu Kang, Ping Zhong, and Jianxin Wang. Perspective from a broader context: Can room style knowledge help visual floorplan localization? *arXiv preprint arXiv:2508.01216*, 2025b.

Bolei Chen, Jiaxu Kang, Haonan Yang, Ping Zhong, Rui Fan, and Jianxin Wang. Treasure hunting: Embodied contrastive learning-enhanced coarse-to-fine object seeking with explorer and discriminator cooperation. *IEEE Transactions on Neural Networks and Learning Systems*, 37(3), 2026a.

Changan Chen, Rui Wang, Christoph Vogel, and Marc Pollefeys. F3loc: Fusion and filtering for floorplan localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18029–18038, 2024.

Haoran Chen, Lantian Wei, Haomin Liu, Boxin Shi, Guofeng Zhang, and Hongbin Zha. Mount: Learning 6dof motion prediction based on uncertainty estimation for delayed ar rendering. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3166–3179, 2022.

Kehan Chen, Yan Huang, Dong An, Jiawei He, Yifei Su, Jing Liu, Nianfeng Liu, and Liang Wang. Floorplan-vln: A new paradigm for floor plan guided vision-language navigation. 2026b.

Junhyeong Cho, Ruojin Cai, and Hadar Averbuch-Elor. Scenealigner: 3d-grounded floorplan localization in the wild. *arXiv preprint arXiv:2605.22581*, 2026.

Hang Chu, Dong Ki Kim, and Tsuhan Chen. You are here: Mimicking the human thinking process in reading floor-plans. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2210–2218, 2015.

Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2133–2143, 2021.

Frank Dellaert, Dieter Fox, Wolfram Burgard, and Sebastian Thrun. Monte carlo localization for mobile robots. In *Proceedings 1999 IEEE international conference on robotics and automation (Cat. No. 99CH36288C)*, volume 2, pp. 1322–1328. IEEE, 1999.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Yuval Grader and Hadar Averbuch-Elor. Supercharging floorplan localization with semantic rays. *arXiv preprint arXiv:2507.09291*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Henry Howard-Jenkins and Victor Adrian Prisacariu. Lalaloc++: Global floor plan comprehension for layout localisation in unvisited environments. In *European Conference on Computer Vision*, pp. 693–709. Springer, 2022.

Henry Howard-Jenkins, Jose-Raul Ruiz-Sarmiento, and Victor Adrian Prisacariu. Lalaloc: Latent layout localisation in dynamic, unvisited environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10107–10116, 2021.

Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter networks with application to visual localization. In *Conference on robot learning*, pp. 169–178. PMLR, 2018.

Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5974–5983, 2017.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Automatic generator of minimal problem solvers. In *European Conference on Computer Vision*, pp. 302–315. Springer, 2008.

Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2372–2381, 2017.

Oscar Mendez, Simon Hadfield, Nicolas Pugeault, and Richard Bowden. Sedar-semantic detection and ranging: Humans can localise without lidar, can robots? In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6053–6060. IEEE, 2018.

Oscar Mendez, Simon Hadfield, Nicolas Pugeault, and Richard Bowden. Sedar: reading floorplans like a human—using deep learning to enable human-inspired localisation. *International Journal of Computer Vision*, 128(5):1286–1310, 2020.

Zhixiang Min, Naji Khosravan, Zachary Bessinger, Manjunath Narayana, Sing Bing Kang, Enrique Dunn, and Ivaylo Boyadzhiev. Laser: Latent space rendering for 2d visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11122–11131, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Meshloc: Mesh-based visual localization. In *European Conference on Computer Vision*, pp. 589–609. Springer, 2022.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12716–12725, 2019.

Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wera Winterhalter, Freya Fleckenstein, Bastian Steder, Luciano Spinello, and Wolfram Burgard. Accurate indoor localization for rgb-d smartphones and tablets given 2d floor plans. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3138–3143. IEEE, 2015.

Matthias Wüest, Francis Engelmann, Ondrej Miksik, Marc Pollefeys, and Daniel Barath. Unloc: Leveraging depth uncertainties for floorplan localization. *arXiv preprint arXiv:2509.11301*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Kecheng Ye, Mao Chen, Xiangkai Zhang, and Xu Yang. Fusion of depth and semantics for probabilistic floorplan localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19413–19422, 2026.

Hengshuang Zhao. Depth anything v2. 2024.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 519–535. Springer, 2020.

Ping Zhong, Liangbai Liu, Bolei Chen, Tao Wu, Jiazhi Xia, Chaoxu Mu, and Jianxin Wang. Spatially generalizable mobile manipulation via adaptive experience selection and dynamic imagination. *arXiv preprint arXiv:2601.14649*, 2026.

Appendix Contents

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Methodology **3**

3.1 Problem Formulation 3

3.2 Stage 1: RRP-based Hypotheses Preparation 4

3.3 Stage 2: $SE(2)$-Aware Contrastive Disambiguation 5

4 Experiments **6**

4.1 Experimental Setup 6

4.2 Comparisons with SOTA Methods 7

4.3 Ablation Studies 8

4.4 Parametric Studies 9

5 Conclusion **9**

A More Experimental Details **14**

A.1 Implementation Details of Our Method 14

A.2 Implementation of 3DP & RSK 15

A.3 Variants of SemRayLoc 16

B Additional Controlled Studies **16**

C Additional Parametric Studies **16**

C.1 Study of the Parameters in Eq. 4 16

C.2 Study on Top-K and the Maximum Number of Hypotheses 17

C.3 Study on the Disambiguation Weight and Local FloorPlan Size 18

D Analysis of Computational Efficiency **18**

E Additional Evaluation for Long-Sequence Trajectory Tracking **18**

F More Visualizations **19**

G Failure Cases **19**

A More Experimental Details

A.1 Implementation Details of Our Method

To optimize the dense-to-ray geometric projection (RRP), we employ the Adam optimizer (Kingma, 2014) with a constant learning rate of 10^{-4} and a batch size of 64. The depth-aware visual encoder (DINOv2-S from Depth Anything V2 (Yang et al., 2024)) is frozen during training to preserve its depth estimation expertise. The remaining components are trained for 50 epochs on an NVIDIA RTX 3090 GPU. Our RRP matches 40 predicted rays to the floorplan to generate the initial geometry-aware pose likelihood map. For ray-depth discretization and candidate ray matching, we follow the same settings as F^{3}Loc (Chen et al., 2024), including the values of D, d_{min}, d_{max}, γ, and the $SE(2)$ pose discretization.

Scoring and fusion details. In Eq. 5, the candidate-conditioned tokens and pooling weights are

$$\mathbf{H}_{j,l} = \mathrm{LN}(\mathbf{F}_j + \mathrm{CrossAttn}(\mathbf{F}_j, \mathbf{G}_l, \mathbf{G}_l)),$$
$$a_{j,l,t} = \mathrm{softmax}_t(\mathbf{w}^\top \mathbf{H}_{j,l,t} + b), \qquad \mathrm{AttnPool}(\mathbf{H}_{j,l}) = \sum_t a_{j,l,t}\mathbf{H}_{j,l,t}.$$

Here, t indexes image tokens and θ denotes the trainable compatibility parameters. The MLP outputs an unnormalized scalar, with higher scores indicating better matches. During training, $\tau^{-1} = \exp(\beta)$ is learned with β initialized to $\log(1/0.07)$. At inference, Eq. 9 is equivalent to maximizing $P_d(R_m|\mathcal{I}_j, F)\exp(\alpha P_c(R_m))$. Cached image tokens are reused, while cross-attention, pooling, and MLP scoring are candidate-specific.

Compatibility network. The frozen DINOv2-S backbone is loaded from the Depth Anything V2 checkpoint (Yang et al., 2024). Its patch features are resized to a 6×40 grid and projected to 128 channels. Learned 2D positional encodings, layer normalization, and one four-head self-attention Transformer encoder layer produce the image tokens. The floorplan encoder retains the ResNet-18 spatial feature map, projects it to 128 channels, and adds learned 2D positional encodings and layer normalization. Four-head cross-attention uses image queries and floorplan keys/values, followed by a residual connection and layer normalization. A learned linear token scorer and a softmax over image tokens provide the pooling weights in Eq. 5. The pooled vector passes through a scoring head consisting of layer normalization, a $128 \to 128$ linear layer, GELU, and a $128 \to 1$ linear layer. The score has no sigmoid or cosine normalization. The backbone is frozen; token projections, positional encodings, attention layers, pooling, and the scoring head are trainable. Training evaluates all image–candidate pairs in chunks of 16 candidate maps and applies row-wise cross-entropy to the temperature-scaled score matrix.

On S3D, the contrastive disambiguation module is trained using AdamW with a batch size of 32 for 30 epochs on an NVIDIA RTX 3090 GPU. We use a cosine learning-rate schedule with an initial learning rate of 10^{-4} and a minimum learning rate of 10^{-5}. We select the checkpoint with the lowest loss on the validation split. Unless specified otherwise, the crop size for local floorplan structures is 5 m $\times$ 5 m. Each crop is centered at the sampled or proposed position and rotated by the inverse of its heading before floorplan encoding, aligning all candidate headings to a canonical direction. The same pose-conditioned crop construction is used for training samples and inference hypotheses. Cross-floorplan negatives are taken from other mini-batch samples and retained in all ablations. To encourage *pose smoothness* via slack positive sampling, we apply positional and orientational perturbations of 0.5 m and ± 0.26 radians to the GT pose, respectively. To improve *spatial separability*, inner-floorplan position-level negative samples are randomly drawn within a distance range of 1.5 m to 3.0 m from the GT pose. To improve *angular discriminability*, orientation-level negative samples are generated by applying a rotation sampled from $[30°, 180°]$ to the GT orientation. During inference, we first extract the top $K = 1000$ candidate poses from the DAFPM to ensure high candidate coverage. We then apply greedy $SE(2)$-aware candidate refinement with $\sigma_t = 0.6$ m, $\sigma_\theta = 30°$, $\lambda_\theta = 1.0$, and $\rho = 1.0$ before DisCo reranking. We set the compatibility fusion weight in Eq. 9 to $\alpha = 0.5$. All reported inference times are measured on an NVIDIA RTX 5880 Ada Generation GPU.

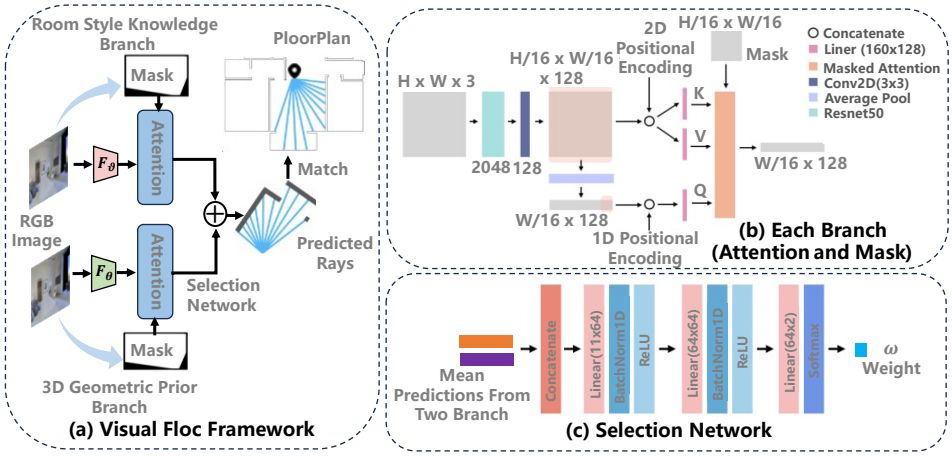

Figure 5: Illustrations of 3DP & RSK. (a) The pre-trained visual encoders, F_θ in 3DP and F_ϑ in RSK, are transferred to the visual FLoc framework for fine-tuning to further fit this task. The visual FLoc framework is a dual-branch model consisting of a 3D geometric prior branch and a RSK branch. (b) and (c) detail the masked attention mechanism and the selection network, respectively.

A.2 Implementation of 3DP & RSK

3DP & RSK is achieved by integrating the visual encoders pre-trained in 3DP (Chen et al., 2025a) and RSK (Chen et al., 2025b) into a unified FLoc framework. As shown in Fig. 5(a), the fully pre-trained encoders, F_θ in 3DP and F_ϑ in RSK, are transferred to the visual FLoc framework for fine-tuning to fit the task, which localizes by finding the pose in the floorplan that has the most similar 2D rays as the prediction. Our visual FLoc framework is implemented as a dual-branch model consisting of a 3D geometric prior branch and a RSK branch. In each branch, the image is first aligned with the gravity direction, as done in (Chen et al., 2024). Then, F_θ/F_ϑ and an attention (Vaswani et al., 2017)-based network are used to learn the probability distribution of planar depth over a range of depth hypotheses. Pixels that become unobservable due to gravity alignment are masked in the attention, as shown in Fig. 5(b). To adaptively leverage 3D geometric priors and RSK based on the current view, a selection network, shown in Fig. 5(c), implemented as a multilayer perceptron is adopted to learn a weight $0 \leq \omega \leq 1$ from the two predictions for adaptive selection:

$$\mathbf{P}_{\text{Fusion}} = \omega \mathbf{P}_{\text{3DP}} + (1-\omega)\mathbf{P}_{\text{RSK}}. \tag{10}$$

$\mathbf{P}_{\text{3DP}}$ and $\mathbf{P}_{\text{RSK}}$ denote the probability distributions of planar depth from the 3D geometric prior branch and the RSK branch, respectively. The expectation of $\mathbf{P}_{\text{Fusion}}$ provides the final prediction of 2D rays. ω is manually specified as 1 and 0, implying that only 3D geometric priors and RSK are used, respectively. For the training of FLoc models, we optimize an L1 loss and a cosine-similarity-based shape loss:

$$\mathcal{L}_{\text{FLoc}} = \|\mathbf{d} - \mathbf{d}^*\|_1 + 1 - \frac{\mathbf{d}^\top \mathbf{d}^*}{\max\{\|\mathbf{d}\|_2 \|\mathbf{d}^*\|_2, \epsilon\}}, \tag{11}$$

where $\mathbf{d}$ and $\mathbf{d}^*$ are predicted and GT 2D-ray depths, respectively. ϵ is a small constant to prevent division by zero.

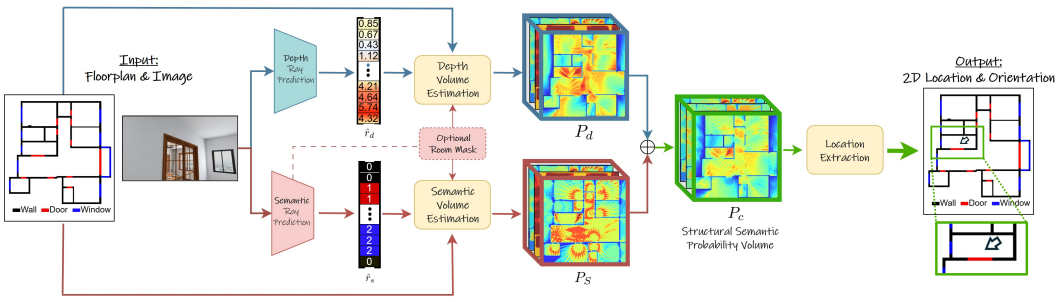

Figure 6: The pipeline of SemRayLoc from (Grader & Averbuch-Elor, 2025).

A.3 VARIANTS OF SEMRAYLOC

Similar to the implementation of 3DP & RSK in Subsection A.2, **SemRayLoc$_s$ + 3DP** is implemented by replacing the depth and semantic ray encoders in Fig. 6 with the fully pre-trained visual encoder F_θ from 3DP. Similarly, **SemRayLoc$_s$ + RSK** is implemented by replacing the depth and semantic ray encoders in Fig. 6 with the fully pre-trained visual encoder F_ϑ from RSK. **SemRayLoc$_s$ + 3DP & RSK** is implemented by replacing the depth and semantic ray encoders with F_θ and F_ϑ, respectively.

As shown in Tab. 10, increasing α from 0.0 substantially improves localization accuracy, confirming the importance of the learned visual-floorplan compatibility. The best overall performance is achieved at $\alpha = 0.5$, while overly large weights degrade the joint positional-directional metric, suggesting that excessive reliance on the learned compatibility may weaken the stabilizing effect of the geometry-based likelihood.

Table 6: A study on the effectiveness of DisCo under different datasets, candidate generation methods, and visual backbones.

Dataset	Candidate	Backbone	DisCo	R@0.1 m↑	R@0.5 m↑	R@1 m↑	R@1 m 30°↑
S3D (full)	RRP	DINOv2	✗	5.5	34.2	40.4	39.3
S3D (full)	RRP	DINOv2	✓	**10.0**	**64.4**	**75.0**	**74.3**
S3D (full)	F^{3}Loc-style	ResNet-50	✗	2.56	22.26	31.79	30.49
S3D (full)	F^{3}Loc-style	ResNet-50	✓	**6.62**	**52.46**	**63.87**	**62.83**
Gibson(f)	F^{3}Loc-style	ResNet-50	✗	4.63	28.55	36.62	35.03
Gibson(f)	F^{3}Loc-style	ResNet-50	✓	**7.17**	**40.06**	**48.38**	**46.94**

B ADDITIONAL CONTROLLED STUDIES

To investigate whether the effectiveness of DisCo mainly comes from the backbone or the candidate generation module, we conduct a controlled study by integrating DisCo with different candidate generators and visual backbones. As shown in Tab. 6, DisCo consistently improves localization performance across different settings, demonstrating its generality as a pose disambiguation module.

First, we compare RRP with and without DisCo under exactly the same candidate generator, image encoder, checkpoint, and candidate poses. The only difference is whether DisCo is applied for pose refinement. On S3D, DisCo improves R@1m from 40.4% to 75.0% and R@1m,30° from 39.3% to 74.3%, indicating that the performance gain mainly originates from the proposed SE(2)-aware disambiguation mechanism rather than a stronger visual backbone.

Furthermore, we integrate DisCo with an F^{3}Loc-style candidate generator using a ResNet-50 backbone. Without changing the upstream localization pipeline, DisCo improves R@1m from 31.79% to 63.87% on S3D and from 36.62% to 48.38% on Gibson(f). These results demonstrate that DisCo is not tied to a specific backbone or proposal strategy, but can consistently enhance different localization pipelines by resolving ambiguous pose hypotheses.

Overall, these experiments verify that candidate generation and pose disambiguation are complementary components. While the candidate generator provides high-recall pose hypotheses, DisCo effectively improves their spatial and angular separability through SE(2)-aware contrastive learning, serving as a general refinement module for visual floorplan localization.

C ADDITIONAL PARAMETRIC STUDIES

C.1 STUDY OF THE PARAMETERS IN EQ. 4

We investigate the sensitivity of the key parameters in Eq. 4 in the main paper, including σ_t, σ_θ, λ_θ, and ρ. As shown in Tab. 7, DisCo maintains stable performance under different parameter settings, demonstrating the robustness of the proposed SE(2)-aware refinement.

For σ_t and ρ, smaller values retain more hypotheses and slightly improve localization accuracy, while larger values reduce the hypothesis number with only marginal performance degradation. For

Table 7: Experimental results for the average hypothesis numbers and localization performance when the parameters in Eq. 4 are changed.

Setting	Avg. Hypotheses	R@1 m↑	R@1 m 30°↑
$\sigma_t = 0.3$ m	121.3	75.22	74.55
$\sigma_t = 0.6$ m	52.9	75.00	74.30
$\sigma_t = 1.0$ m	33.3	74.32	73.49
$\sigma_\theta = 15°$	62.7	74.91	74.22
$\sigma_\theta = 30°$	52.9	75.00	74.30
$\sigma_\theta = 45°$	47.8	74.75	74.00
$\lambda_\theta = 0.5$	48.0	74.71	73.96
$\lambda_\theta = 1.0$	52.9	75.00	74.30
$\lambda_\theta = 2.0$	57.3	74.71	73.99
$\rho = 0.5$	132.8	75.11	74.52
$\rho = 1.0$	52.9	75.00	74.30
$\rho = 1.5$	31.6	74.40	73.52

example, changing ρ from 0.5 to 1.5 reduces the average hypotheses from 132.8 to 31.6, while R@1m only decreases from 75.11% to 74.40%. Similarly, varying σ_θ and λ_θ results in negligible performance differences, indicating that DisCo is insensitive to the exact angular refinement settings.

Overall, these results demonstrate that DisCo provides a flexible accuracy-efficiency trade-off by adjusting the retained hypothesis number, while maintaining robust localization performance across different refinement configurations.

Table 8: Experimental results for average hypothesis numbers, localization performance, and inference time when Top-K takes different values.

Top-K	Avg. Hypo.	Oracle R@1 m↑	R@1 m↑	R@1 m 30°↑	Inference Time (ms)↓
50	8.7	83.20	67.90	66.98	26.16 ± 4.49
100	12.2	87.81	70.43	69.54	27.00 ± 5.48
250	20.2	92.29	72.65	71.82	29.95 ± 5.77
500	31.8	95.19	73.74	72.91	35.74 ± 6.88
1000	52.9	**97.17**	**75.00**	**74.30**	45.61 ± 8.32

Table 9: Experimental results on the average hypothesis numbers, localization performance, and inference time when the maximum hypothesis number is fixed.

Max. Hypo.	Avg. Hypo.	R@0.1 m↑	R@0.5 m	R@1 m↑	R@1 m 30°↑	Inference Time (ms)↓
16	16.0	9.51	61.05	71.18	70.34	31.35 ± 3.98
32	32.0	9.73	63.19	73.52	72.77	37.22 ± 4.05
48	45.4	9.80	64.03	74.52	73.72	42.40 ± 4.95
64	51.1	9.84	64.18	75.00	74.30	44.77 ± 6.03
Uncapped	52.9	**10.00**	**64.40**	**75.00**	**74.30**	46.62 ± 10.86

C.2 Study on Top-K and the Maximum Number of Hypotheses

We further investigate the influence of the hypothesis budget on localization performance and efficiency. As shown in Tab. 8, increasing the source Top-K mainly improves the proposal recall. Specifically, increasing Top-K from 50 to 1000 improves Oracle R@1m from 83.20% to 97.17% and final R@1m from 67.90% to 75.00%, at the cost of increased inference time. This demonstrates that the proposed framework provides a flexible accuracy-efficiency trade-off by adjusting the candidate budget.

We further analyze the maximum number of refined hypotheses in Tab. 9. The computational cost is mainly determined by the number of representative hypotheses after SE(2)-aware selection rather than the initial candidate pool. Reducing the hypothesis budget from 64 to 16 decreases the inference time from 44.77 ms to 31.35 ms, while maintaining competitive performance (R@1m: 71.18% vs. 75.00%).

Overall, these results demonstrate that DisCo achieves controllable scalability: the candidate budget improves robustness against missing pose modes, while the refined hypothesis number provides a direct trade-off between localization accuracy and inference efficiency.

Table 10: Additional parametric study on the disambiguation weight α on the S3D (full) dataset.

Disam. Weight	R@0.1 m↑	R@0.5 m↑	R@1 m↑	R@1 m 30°↑
0.0	5.5	34.2	40.4	39.3
0.1	9.3	62.2	72.4	71.2
0.3	9.7	64.6	74.0	73.5
0.5	10.0	64.4	75.0	74.3
0.7	9.4	63.0	73.3	63.0
0.9	9.2	52.0	72.9	62.5

Table 11: Additional parameter study on the crop size of floorplan structures on the S3D (full) dataset.

Crop Size	R@0.1 m↑	R@0.5 m↑	R@1 m↑	R@1 m 30°↑
3 m × 3 m	8.7	55.6	65.7	64.5
5 m × 5 m	10.0	64.4	75.0	74.3
7 m × 7 m	9.7	62.5	72.5	71.8

C.3 Study on the Disambiguation Weight and Local FloorPlan Size

In this section, we provide additional parametric studies on the disambiguation weight α and the crop size of local floorplan structures. As discussed in the main paper, we evaluate these parameters on the S3D (full) dataset. As shown in Fig. 10, setting α to 0.3 or 0.5 helps achieve optimal visual FLoc performance. Tab. 11 shows that a 5 m × 5 m crop achieves the best overall performance. Smaller crops may miss sufficient surrounding layout context, whereas larger crops can introduce irrelevant structures from neighboring regions, making visual-floorplan matching less discriminative.

Table 12: Average single-query inference time (ms) on an NVIDIA RTX 5880 Ada Generation GPU. The default DisCo result is taken from the Top-$K = 1000$ setting in Tab. 5.

Method	Time (ms) ↓
F^3Loc	6.4
SemRayLoc	169.4
RRP	11.56
Ours	45.61 ± 8.32

D Analysis of Computational Efficiency

We evaluate single-query inference time on an NVIDIA RTX 5880 Ada Generation GPU (Tab. 12). F^3Loc and RRP require 6.4 ms and 11.56 ms per query, respectively. With the default Top-$K = 1000$ setting, DisCo retains an average of 52.9 hypotheses and takes 45.61 ± 8.32 ms per query, corresponding to approximately 21.9 queries per second. SemRayLoc takes approximately 3.7 times as long per query (169.4 ms). This configuration achieves 75.0% R@1 m and 74.3% R@1 m 30° on S3D (full). The hypothesis-budget study in Tab. 9 further shows that capping the candidate set at 48 hypotheses reduces mean latency to 42.40 ms, with 74.52% R@1 m, providing a controllable accuracy-efficiency trade-off.

E Additional Evaluation for Long-Sequence Trajectory Tracking

In this section, we compare our method with existing SOTA methods on the long-sequence trajectory tracking task using the Gibson(t) dataset collected by F^3Loc (Chen et al., 2024). Gibson(t) consists of 118 long image sequences, each containing 280 ∼ 5152 image frames. The **R**oot-**M**ean-**S**quare **E**rror (RMSE) over the last 10 frames is employed to measure the accuracy of sequential trajectory tracking when localization is successful (RMSE(S)) and in all cases (RMSE(A)). Technically, we

Table 13: Comparative studies of long-sequence trajectory tracking methods on the Gibson(t) dataset.

Method (Venue)	**Gibson(t) R@**			
	R@0.2 m↑	**R@1 m↑**	**RMSE(S)↓**	**RMSE(A)↓**
LASER(CVPR 2022)	-	59.5	0.39	1.96
F^3Loc(CVPR 2024)	35.1	89.2	0.18	0.88
F^3Loc fusion(CVPR 2024)	62.2	94.6	0.12	0.51
3DP(ACM MM 2025)	70.3	97.3	0.12	0.34
RSK(AAAI 2026)	59.5	94.6	0.13	0.51
3DP & RSK	64.9	94.6	0.12	0.51
Ours	73.0	94.6	0.11	0.49
Oracle	-	100.0	0.07	0.07

combine the histogram filter proposed by F^3Loc with our method and perform visual FLoc with 100 historical frames.

As shown in Tab. 13, our method improves recall by 2.7% compared to 3DP at the localization accuracy of 0.2 m. The reduction in the RMSE(S) metric reflects the robustness of our method in sequential trajectory tracking. Additionally, our method achieves competitive results in terms of 1 m localization accuracy and RMSE(A). It is worth noting that F^3Loc fusion performs visual FLoc by adaptively utilizing single-frame and multi-frame images, yet it is only competitive on the RMSE(S) metric. Oracle achieved 100% Recall at 1 m accuracy by combining GT rays with a histogram filter, demonstrating the potential of visual FLoc.

F More Visualizations

Fig. 7 shows qualitative comparisons between our method and SemRayLoc (Grader & Averbuch-Elor, 2025). Due to the repetitive geometric structures in the floorplan, our RRP also fails when SemRayLoc encounters errors. However, our visual-floorplan contrastive disambiguation effectively addresses localization ambiguities and yields highly accurate final FLoc results.

G Failure Cases

Although DisCo effectively resolves structural ambiguities in semantic-free visual FLoc, it may still fail in several challenging scenarios. For example, when different regions exhibit highly similar geometric layouts but distinct functional semantics (e.g., room categories or object identities), semantic-based methods such as SemRayLoc can leverage additional scene context and achieve better discrimination. DisCo may also struggle when the floorplan geometry provides insufficient discriminative information, such as extremely repetitive structures or significant appearance–geometry mismatches.

These cases highlight the complementary nature of semantic information and our proposed pose disambiguation mechanism. DisCo focuses on improving spatial and angular separability without requiring costly semantic annotations, making it applicable to semantic-free settings. Integrating semantic cues or high-level scene priors into the candidate generation or refinement stages provides a promising direction for further improving robustness in such challenging scenarios.

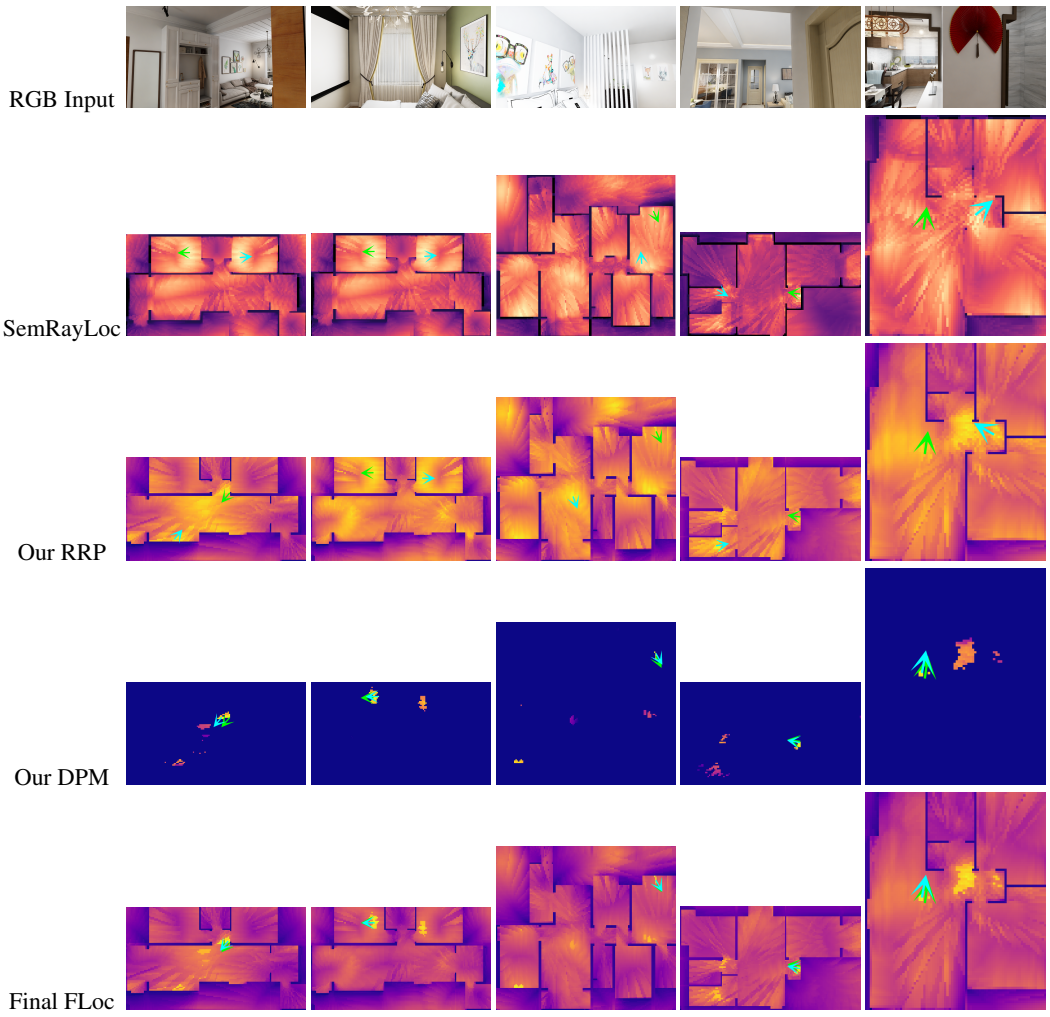

Figure 7: Qualitative comparison on 5 scenes.