SEERAYLOC: GENERALIZABLE FLOORPLAN LOCALIZATION WITH WHAT YOU SEE

Anonymous authors
Paper under double-blind review

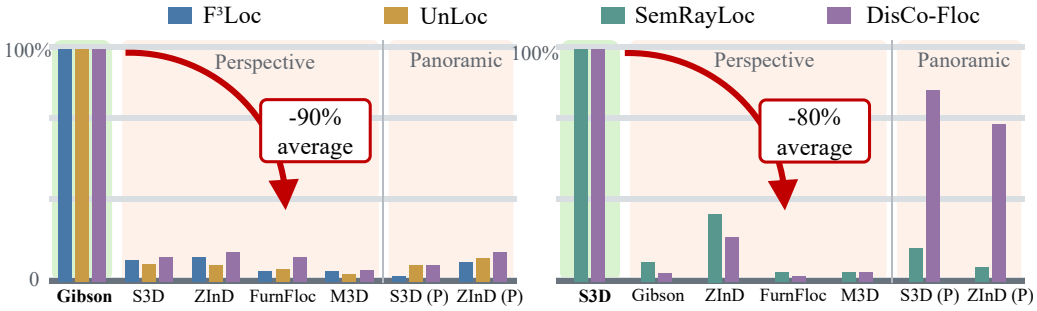
ABSTRACT

Ray-based floorplan localization estimates camera pose by matching image-derived depth rays against floorplan geometry. Existing methods directly predict floorplan-aligned wall distances, including geometry that is often not visible in the image. Such limitation on implicit scene completion can overfit to dataset-specific domain priors and harm cross-domain generalization. We introduce **SeeRayLoc**, a framework that decouples visible geometry estimation from localization-oriented ray construction. SeeRayLoc estimates visible scene depth and adaptively selects localization-relevant rays for floorplan matching. Floorplan supervision guides ray selection instead of hidden-geometry prediction, yielding generalizable floorplan localization. We evaluate SeeRayLoc through source-to-target transfer across six indoor localization datasets, including **FurnFloc**, our newly curated dataset for furnished scenes. The evaluation spans diverse scene distributions, camera configurations, fields of view, and both perspective and panoramic imagery. Extensive experiments show that SeeRayLoc has stronger transfer robustness than direct floorplan-depth prediction, achieving the highest R@1m30° recall in all seven Structured3D source transfer settings across perspective and panoramic imagery. It proves the effectiveness of visible geometry providing a more generalizable basis for ray-based floorplan localization.
• Project page and code: https://anonymous.4open.science/r/SeeRayLoc-9573.

1 INTRODUCTION

Floorplan-based indoor localization estimates the position and orientation of a camera with respect to a known floorplan. Ray-based methods provide an efficient geometric solution by predicting a compact depth profile from a query image and comparing it with rays rendered from candidate poses on the floorplan (Chen et al., 2024; Grader & Averbuch-Elor, 2025; Chen et al., 2025; 2026; Wueest et al., 2026; Ye et al., 2026). This setting avoids the need for a complete 3D reconstruction while maintaining a direct geometric connection between visual observations and the map.

A major difficulty is that images and floorplans describe different geometry. Floorplans primarily represent structural boundaries from the allocentric perspective, whereas image rays often terminate on foreground objects from the ego-centric perspective. Existing methods typically address



(a) Gibson (f) as the source domain. (b) Structured3D as the source domain.

Figure 1: **Performance drop** from source to target domain. Source-normalized (as 100%) and cross-dataset scores are measured by Recall@1m. Target domains involve perspective and panoramic imagery.

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.

this discrepancy by supervising the predicted rays with floorplan-derived wall distances (Chen et al., 2024; Grader & Averbuch-Elor, 2025; Chen et al., 2025; 2026; Wueest et al., 2026; Ye et al., 2026). When a wall is occluded, however, its distance is not directly observable from the ego image. The predictor might fail or must infer hidden structural geometry. This implicit amodal completion can work well within the training domain, but the hidden wall distance is underdetermined by visual evidence. A model may instead rely on source-specific correlations involving room dimensions, furniture placement, and object-wall relationships. These correlations can change considerably across indoor datasets, causing models that perform well in-domain to degrade when transferred to new scene distributions or camera configurations, as reflected by the substantial cross-dataset performance drops in Figure 1 (detail in Appendix A.1). Therefore, we raise an under-explored question: ***how to improve the cross-dataset generalization of floorplan localization methods?***

To address this problem, we introduce **SeeRayLoc**, which separates visible geometry estimation from localization-oriented structural reasoning. As shown in Figure 2, our approach estimates dense depth for observable scene surfaces and constructs a structural ray profile by adaptively selecting localization-relevant measurements from the visible geometry. The resulting horizontal ray layout remains compatible with conventional floorplan matching, while each image-side ray can select among observations according to structural reliability.

We evaluate SeeRayLoc under source-to-target transfer, where models trained on a source dataset are applied directly to unseen target datasets **without target-domain fine-tuning or calibration**. Our experiments comprise six indoor localization settings based on **Matterport3D** (Chang et al., 2017), **Structured3D** (Zheng et al., 2020), **ZInD** (Cruz et al., 2021), **Gibson(f)** (Chen et al., 2024), **Gibson(g)** (Chen et al., 2024), and **FurnFloc** (ours), a new dataset of furnished indoor environments. They cover synthetic and real environments, different camera intrinsics and fields of view, and both **perspective and panoramic observations**. The results demonstrate that constructing localization rays from visible geometry provides a more robust basis for transfer than directly predicting floorplan-aligned wall distances.

In summary, our contributions are as follows:

- **Egocentric Depth Adaptation.** We adapt a frozen Depth Anything V3 with a lightweight dense adapter to improve indoor dense depth estimation without floorplan-specific egocentric depth supervision.
- **Adaptive Ray Selection.** We introduce a learnable selector that adaptively extracts localization-relevant rays from dense depth predictions, replacing fixed ray sampling.
- **Cross-Dataset Generalization.** SeeRayLoc directly transfers to unseen datasets without fine-tuning, supporting different fields of view and both perspective and panoramic queries.
- **FurnFloc and Multi-Dataset Evaluation.** We curate FurnFloc from HM3D, comprising 132 validated floorplans and 402 k usable images, and establish a unified source-to-target evaluation across six indoor localization datasets.

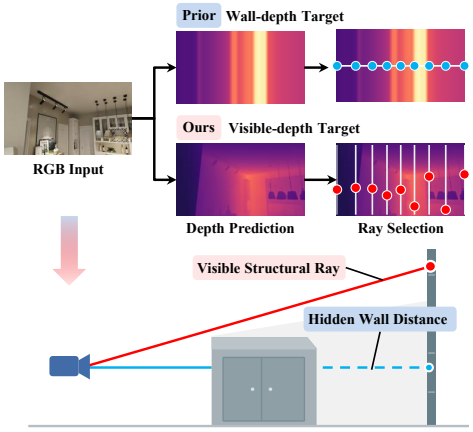


Figure 2: **Floorplan localization paradigm comparison**. Unlike using fixed floorplan-aligned wall distances despite foreground occlusions, our SeeRayLoc predicts visible depth and adaptively selects floorplan-compatible observations.

2 Related Work

Visual Localization and Floorplan Matching. Conventional visual localization commonly relies on image retrieval or 2D-3D matching against a reference-image database or a reconstructed 3D map (Arandjelović et al., 2016; Sarlin et al., 2019; Taira et al., 2021). Floorplan localization instead relies on compact, appearance-independent 2D layouts. Earlier systems aligned RGB-D or LiDAR observations directly with architectural floorplans (Winterhalter et al., 2015; Boniardi et al., 2017), while learning-based methods enabled localization from RGB images. PF-Net (Karkus et al., 2018) embeds learned observation and motion models in a differentiable particle filter. LaLaLoc (Howard-

Jenkins et al., 2021) and LaLaLoc++ (Howard-Jenkins & Prisacariu, 2022) learn cross-modal representations between panoramic observations and floorplan geometry, whereas LASER (Min et al., 2022) matches query images with pose-conditioned latent representations rendered from a floorplan. Within this broader floorplan-localization setting, we focus on explicit geometric matching and its cross-dataset transferability.

Geometry- and Depth-based Floorplan Localization. Recent methods compare image-derived depth observations with geometric rays rendered from candidate floorplan poses. F^{3}Loc (Chen et al., 2024) predicts and fuses single- and multi-view horizontal depths and accumulates their evidence through probabilistic filtering. 3DP (Chen et al., 2025) introduces self-supervised 3D geometric priors through multi-view constraints and view scene alignment based on surface reconstruction, while UnLoc (Wueest et al., 2026) leverages pretrained monocular depth models (Depth Anything V2 (Yang et al., 2024)) and explicitly represents predicted floorplan depths as probability distributions to model uncertainty. Methods supervised with floorplan-aligned depths nevertheless face ambiguity under occlusion, because the target structure may lie behind the visible surface. We instead separate dense visible-depth estimation from localization-specific selection, adaptively selecting localization-relevant measurements from visible geometry for floorplan matching.

Semantic- and Context-aware Floorplan Localization. Semantic and contextual cues can disambiguate repetitive rooms and geometrically similar locations. SemRayLoc (Grader & Averbuch-Elor, 2025) jointly predicts depth and semantic rays and combines them in a structural-semantic probability volume. Subsequent work further studies the probabilistic fusion of geometric and semantic observations (Ye et al., 2026). Beyond explicit map semantics, Perspective from a broader context learns room-style representations from unlabeled images to provide scene-level localization cues (Chen et al., 2026). These approaches improve localization by enriching geometric observations with semantic or contextual information. SeeRayLoc instead learns localization-oriented geometric observations for cross-dataset transfer without target-domain fine-tuning.

3 SeeRayLoc Framework

3.1 Dense Visible-Depth Adapter

Figure 3 illustrates the SeeRayLoc pipeline. We freeze Depth Anything V3 (DA3) (Lin et al., 2025) and introduce a lightweight dense adapter that refines its metric depth prediction for indoor visible-surface geometry. Given an input image I, DA3 produces an initial metric depth map D_0 together with intermediate features $\{\mathbf{F}_l\}_{l\in\mathcal{L}}$. The adapter projects these features into a shared space, fuses them with the log-transformed depth prior, and predicts a dense residual Δ in log-depth space.

We fomulate depth prediction as a multiplicative refinement of the DA3 estimate:

$$\Delta = \gamma \tanh\left[A\left(\{\phi_l(\mathbf{F}_l)\}_{l\in\mathcal{L}}, \log D_0\right)\right], \qquad D_{\text{vis}} = D_0 \odot \exp(\Delta), \tag{1}$$

where ϕ_l denotes a layer-specific feature projection, A is the refinement network, and γ bounds the residual magnitude. Operating in log-depth space converts the residual into a multiplicative correction in depth space, allowing consistent relative adjustments across different depth ranges. The output layer is initialized to zero, yielding $\Delta = 0$ at initialization and therefore $D_{\text{vis}} = D_0$. This initialization allows the adapter to learn only the residual adaptation needed for indoor scens.

The adapter is trained with image-aligned dense depth supervision rather than floorplan-aligned structural depth. Accordingly, D_{vis} is optimized to capture observed-surface geometry, without explicit supervision for hidden structural boundaries. For instance, When a wall is partially occluded by foreground objects, the predicted depth is anchored to the observed surface, leaving the occluded wall geometry unconstrained.

3.2 Localization-Aware Structural Ray Selector

With the trained dense adapter frozen, the selector operates on its dense depth predictions D_{vis} and detached local features. Localization-specific supervision therefore affects only the selector and does not alter the visible-depth estimator. However, a gravity-horizontal measurement may terminate on furniture or other foreground objects rather than on a structural boundary represented in the floorplan. We therefore introduce a localization-aware selector that searches over vertically

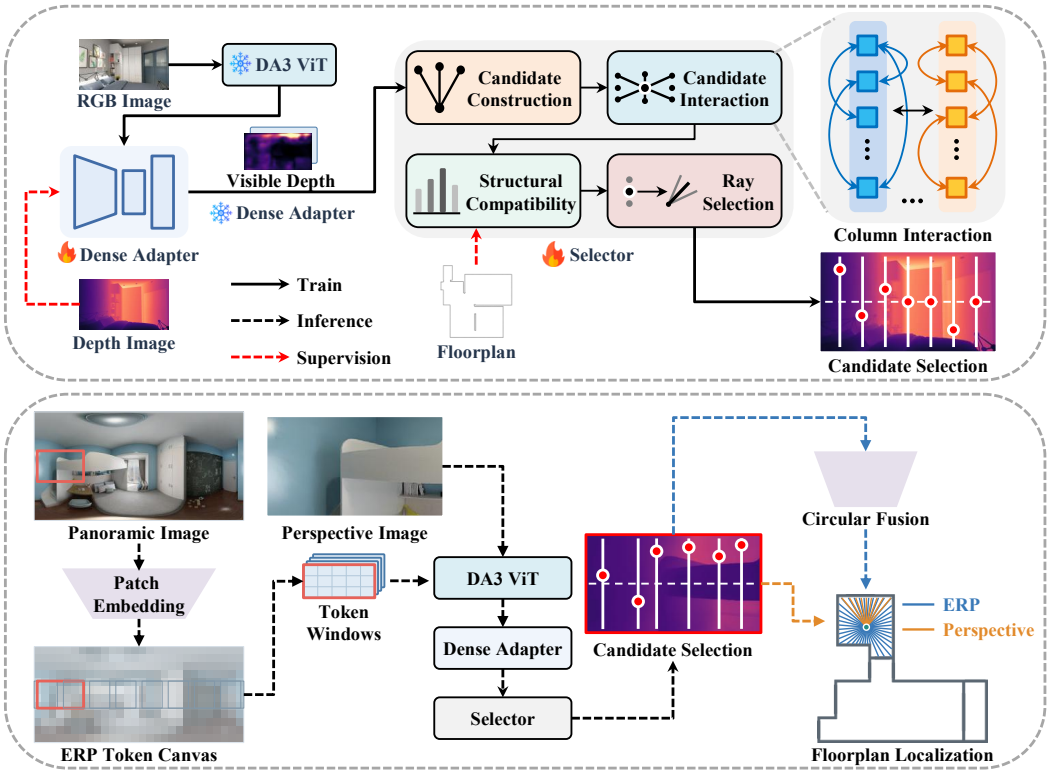


Figure 3: **Overview of SeeRayLoc framework**. **Top** (training): A frozen DA3 provides depth and features. Visible-depth supervision trains the dense adapter, while floorplan-depth supervision trains the selector to score vertically sampled candidates using within- and across-column context. **Bottom** (inference): A perspective image passes directly through DA3, the adapter, and the selector. A panoramic image is patch-embedded once, and circular token windows pass through the same modules. Their local ray predictions are fused before floorplan localization. The panoramic depth pathway uses no RGB cropping or perspective reprojection.

distributed candidates and identifies, for each horizontal direction, the visible measurement most suitable for floorplan matching.

Projection-aware candidate construction. For each of the H horizontal profile locations, we construct J vertically distributed candidates $\{\beta_{i,j}\}_{j=1}^{J}$ over the visible elevation interval, including the gravity-horizontal candidate $\beta_{i,j_0} = 0$.

Because depth representations differ across camera models, each sampled measurement is mapped to a projection-normalized candidate depth:

$$z_{i,j} = \Pi_{\mathcal{C}} \left(D_{\text{vis}}(\mathbf{u}_{i,j}), \mathbf{u}_{i,j} \right), \tag{2}$$

where $\mathcal{C}$ denotes the input camera model, $\mathbf{u}_{i,j}$ is the image location of the candidate, and $\Pi_{\mathcal{C}}$ accounts for the corresponding projection and gravity geometry. Perspective candidates are constructed from camera intrinsics and gravity alignment, whereas panoramic candidates follow spherical ray geometry. Local panoramic context is extracted using circular feature windows to preserve horizontal continuity.

Contextual candidate interaction. Each candidate is represented using its depth, center-relative depth, elevation, and local dense-adapter features:

$$\mathbf{f}_{i,j} = \left[\log z_{i,j}, \log \frac{z_{i,j}}{z_{i,j_0}}, \sin \beta_{i,j}, \cos \beta_{i,j}, \boldsymbol{\mu}_{i,j}, \boldsymbol{\rho}_{i,j} \right]. \tag{3}$$

Here, $\boldsymbol{\mu}_{i,j}$ and $\boldsymbol{\rho}_{i,j}$ denote grouped mean and root-mean-square statistics of the local dense features. Candidate interactions are modeled first within each vertical set and then across horizontal directions, allowing the selector to assess each observation using both local evidence and the surrounding

geometric profile. The resulting contextual representation produces a compatibility logit $s_{i,j}$ for each candidate.

Structural compatibility learning. During training, the floorplan-aligned structural depth z_i^F at the ground-truth pose defines a soft compatibility target:

$$e_{i,j} = \frac{|z_{i,j} - z_i^F|}{\tau_i}, \qquad \tau_i = \max\left(\delta_{\text{abs}}, \delta_{\text{rel}} z_i^F\right), \qquad q_{i,j} = \exp(-e_{i,j}). \tag{4}$$

Here, δ_{abs} and δ_{rel} define absolute and depth-proportional tolerances. We supervise $p_{i,j} = \sigma(s_{i,j})$ against $q_{i,j}$ and further model the improvement of each candidate over the center:

$$g_{i,j}^* = e_{i,j_0} - e_{i,j}, \qquad \hat{g}_{i,j} = \log p_{i,j} - \log p_{i,j_0}. \tag{5}$$

A Smooth-L_1 loss aligns the predicted and target gains, while a center-ordering objective encourages structurally better candidates to outrank the center. The floorplan-aligned targets are used only for training and are not provided to the selector at inference time.

Center-preserving selection. Let $j_i^+ = \arg\max_{j \neq j_0} s_{i,j}$ denote the best non-center candidate. We replace the center candidate only when j_i^+ receives a higher compatibility score:

$$j_i^* = \begin{cases} j_i^+, & s_{i,j_i^+} > s_{i,j_0}, \\ j_0, & \text{otherwise.} \end{cases} \qquad \tilde{z}_i = z_{i,j_i^*}, \quad c_i = \sigma(s_{i,j_i^*}). \tag{6}$$

The selector outputs $\{(\tilde{z}_i, c_i)\}_{i=1}^H$. Importantly, it selects among visible measurements rather than predicting the distance to an unobserved wall.

3.3 Confidence-Weighted Floorplan Matching

We jointly resample the selected depths and confidences at the M horizontal directions used for localization:

$$\{(\bar{d}_k, \bar{c}_k, \alpha_k)\}_{k=1}^M = \mathcal{R}_{\mathcal{C}}\left(\{(\tilde{z}_i, c_i)\}_{i=1}^H\right), \tag{7}$$

where $\mathcal{R}_{\mathcal{C}}$ performs projection-aware angular resampling and converts the selected profile into horizontal ranges $\bar{d}_k$ in the same metric space as the floorplan-rendered rays.

We normalize the confidences to preserve the scale of conventional unweighted matching:

$$w_k = \frac{M\bar{c}_k}{\sum_{l=1}^M \bar{c}_l}. \tag{8}$$

For a candidate pose $\mathbf{p}$, let $d_k^F(\mathbf{p})$ denote the corresponding floorplan-rendered horizontal range. The confidence-weighted discrepancy and depth likelihood are

$$E(\mathbf{p}) = \sum_{k=1}^M w_k \left|\bar{d}_k - d_k^F(\mathbf{p})\right|, \qquad P(I \mid \mathbf{p}, F) = \exp\left(-\frac{E(\mathbf{p})}{\tau}\right). \tag{9}$$

The final pose is

$$\hat{\mathbf{p}} = \arg\max_{\mathbf{p} \in \mathcal{P}} P(I \mid \mathbf{p}, F). \tag{10}$$

4 Experiments

4.1 FurnFloc Dataset Construction

FurnFloc Dataset. To broaden the evaluation to furnished environments with diverse camera configurations, we construct **FurnFloc**, a floorplan-localization benchmark derived from HM3D (Ramakrishnan et al., 2021) and FloVerse-1.6K (Huang et al., 2026). Since HM3D does not provide native floorplans, the maps are derived from reconstructed meshes by filtering vertices according to height and surface-normal orientation to suppress non-structural objects, followed by wall extraction and top-down post-processing. Because automatically generated floorplans may contain

Table 1: **Floorplan localization dataset comparison**. **FoV** denotes filed of view. **Sem.** denotes using semantic information. **Furn.** indicates whether furniture included in the ego-centric image.

Dataset	Year	Floorplan	Image	FoV	Sem.	Furn.
360°-FoV panoramic domain						
Structured3D (Zheng et al., 2020)	ECCV'20	3,475	21,653	360°	✓	✓
ZInD (Cruz et al., 2021)	CVPR'21	2,453	59,361	360°	✓	✗
Narrow-FoV perspective domain						
Matterport3D (Chang et al., 2017)	3DV'17	183	57,948	62°	✓	✓
Structured3D (Zheng et al., 2020)	ECCV'20	3,296	78,399	80°	✓	✓
ZInD (Cruz et al., 2021)	CVPR'21	2,453	59,361	80°	✓	✗
Gibson(f) (Chen et al., 2024)	CVPR'24	118	99,116	108°	✗	✓
Gibson(g) (Chen et al., 2024)	CVPR'24	118	198,232	108°	✗	✓
FurnFloc (ours)	2026	132	402,906	90°	✗	✓

missing, spurious, or disconnected structural boundaries, we manually inspect the resulting maps and retain 132 valid floorplans from 101 HM3D buildings. We further subsample the rendered trajectories by retaining every fifth frame. This produces 403,302 RGB-depth pairs, of which 396 rendering failures are removed, resulting in 402,906 usable images.

With our FunrFloc, we evaluate on six indoor floorplan-localization datasets: Structured3D, Matterport3D, ZInD, Gibson(f), Gibson(g), and FurnFloc. Dataset statistics are summarized in Table 1. Details and preprocessing for the five existing datasets, together with train/validation/test splits and semantic annotation availability for all six datasets, are provided in Appendix A.2.

4.2 Cross-Dataset Transfer Setting

We select source domains based on the availability of depth supervision, furnished indoor scenes, and reliable image to floorplan correspondence. Structured3D (S3D) satisfies these requirements and provides dense ground-truth depth. We additionally use Gibson(f) to broaden source diversity, using pseudo-depth labels generated by MoGe-3 ViT-L (Kong et al., 2026). We exclude Matterport3D due to unreliable depth and image-floorplan alignment in some samples, ZInD due to the lack of depth supervision and limited furniture occlusion, FurnFloc due to low-quality images and less standardized floorplans, and Gibson(g) because it shares scenes with Gibson(f).

For each source setting, SeeRayLoc is trained only on the source dataset and directly evaluated on unseen targets without fine-tuning. The same model is used for both perspective and panoramic queries. We compare with SemRayLoc (Grader & Averbuch-Elor, 2025), UnLoc (Wueest et al., 2026), and DA3 (Lin et al., 2025), retraining SemRayLoc and UnLoc under the same source-domain protocol. For datasets with semantic annotations, semantic cues are provided by the corresponding SemRayLoc semantic models, trained separately on S3D and ZInD. F^3Loc (Chen et al., 2024) is not reported separately because it corresponds to the depth-only formulation of SemRayLoc. We omit other methods when their implementations are unavailable or when their training pipelines cannot be adapted to the same source-only setting without introducing additional methodological changes.

4.3 Experiment Result

From S3D to target domains. For perspective queries (Table 2a-f), SeeRayLoc achieves the highest R@0.5m on S3D, Gibson(f), and Gibson(g), exceeding the strongest baseline by 19.2, 6.3, and 6.4 percentage points (pp), respectively. On ZInD, UnLoc leads at the position-only thresholds, while SeeRayLoc narrowly leads at R@1m30° (27.9% versus 27.8%). On FurnFloc, SeeRayLoc improves R@1m and R@1m30° over DA3 by 3.3 and 3.1 pp, although DA3 performs better at 0.1m. Matterport3D remains challenging for all methods, with only small differences between SeeRayLoc and DA3. This difficulty may partly stem from its larger-scale environments, which increase geometric ambiguity; floorplan-size statistics are provided in Appendix 7.

For panoramic queries (Table 2g and h), UnLoc outperforms DA3 on S3D-Panorama, but SeeRayLoc exceeds UnLoc by 16.0 pp at R@0.5m and 17.5 pp at R@1m30°. On ZInD-Panorama, SeeRayLoc exceeds the strongest baseline, DA3, by 8.1 and 7.6 pp on the same metrics. These results indicate that structural-ray selection learned from perspective images transfers effectively to

(a) Source : Structured3D

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	4.0	31.3	46.4	43.0	✓
UnLoc	5.3	34.0	45.7	42.6	✓
DA3	3.7	34.2	49.5	47.1	✓
SeeRayLoc	**10.5**	**53.4**	**62.2**	**60.5**	✓

(b) Target : Gibson(f)

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.1	3.4	8.7	6.0	✗
UnLoc	0.1	0.8	3.2	1.2	✗
DA3	6.7	38.0	46.0	44.6	✗
SeeRayLoc	**10.0**	**44.3**	**49.6**	**48.0**	✗

(c) Target : Gibson(g)

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.2	2.8	7.9	5.0	✗
UnLoc	0.0	1.2	3.6	1.8	✗
DA3	6.6	34.3	42.3	40.5	✗
SeeRayLoc	**9.2**	**40.7**	**46.4**	**44.7**	✗

(d) Target : ZInD

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.4	5.5	12.5	8.6	✓
UnLoc	**2.6**	**21.0**	**31.0**	27.8	✓
DA3	1.4	16.7	27.9	24.5	✓
SeeRayLoc	1.8	18.5	30.7	**27.9**	✓

(e) Target : FurnFloc

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.1	1.4	4.6	1.8	✗
UnLoc	0.0	0.6	2.3	0.5	✗
DA3	**2.2**	13.2	18.9	16.6	✗
SeeRayLoc	1.9	**13.5**	**22.2**	**19.7**	✗

(f) Target : Matterport3D

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.0	0.8	2.3	0.7	✗
UnLoc	0.2	2.2	4.8	3.5	✗
DA3	0.3	**3.7**	6.0	4.9	✗
SeeRayLoc	**0.3**	3.6	**6.3**	**5.1**	✗

(g) Target : Structured3D (Panorama)

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.7	6.0	10.3	7.3	✗
UnLoc	7.8	33.3	43.3	31.3	✗
DA3	5.0	28.7	37.2	28.7	✗
SeeRayLoc	**12.3**	**49.3**	**54.5**	**48.8**	✗

(h) Target : ZInD (Panorama)

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.1	2.0	5.5	3.3	✗
UnLoc	17.7	55.6	60.4	57.4	✗
DA3	25.8	64.9	68.6	67.3	✗
SeeRayLoc	**28.1**	**73.0**	**76.0**	**74.9**	✗

Table 2: Results from **S3D** to other target datasets. **Sem.** denotes semantic cues included.

panoramic queries. Panoramic images are processed directly in equirectangular (ERP) coordinates, with implementation details provided in Appendix A.3.

From Gibson(f) to target domains. For perspective queries(Table 3a-f), SeeRayLoc achieves the highest R@0.5m on Gibson(f), Gibson(g), and FurnFloc, exceeding the strongest baseline by 6.7, 7.9, and 3.9 pp, respectively. Gibson(g) shares scenes with Gibson(f), so this result does not measure generalization to unseen scenes. On S3D, ZInD, and Matterport3D, DA3 leads at R@0.5m by 1.0, 1.7, and 0.6 pp, respectively.

For panoramic queries (Table 3g and h), SeeRayLoc exceeds DA3 on S3D-Panorama by 6.2 pp at R@0.5m, whereas DA3 leads on ZInD-Panorama by 2.5 pp. Compared with Structured3D-source training (Table 2), Gibson(f)-source training yields lower recall on several targets, especially S3D and the panoramic datasets, but higher recall on FurnFloc. These results show that pseudo-depth can support training on an alternative source, although the gains vary across targets.

Evaluation under the original wall-distance task. The preceding source-transfer experiments evaluate SeeRayLoc with visible-depth supervision. To connect with the original floorplan-localization task, we also fine-tune DA3 to predict floorplan-aligned wall distance (ours). As shown in Table 4, our method ranks first at R@0.1m on both Gibson variants and second at R@0.5m on Gibson(f), Gibson(g), and Structured3D, reaching 49.3%, 47.0%, and 66.7%, respectively. Its R@0.5m on ZInD is 26.1%, below FoD and CF^{2}Loc. For ZInD, we follow SemRayLoc (Grader & Averbuch-Elor, 2025) preprocessing. CF^{2}Loc (Meng et al., 2026) follows its original protocol, while FoD (Ye et al., 2026) does not specify its preprocessing. Note that our SeeRayLoc does not consistently achieve the state-of-the-art performance while train and test on the source domain.

4.4 Ablation Study

Table 5 isolates the contribution of each component of SeeRayLoc on S3D. Adding the dense adapter to DA3 raises R@0.1m from 3.7% to 6.9% and R@0.5m from 34.2% to 39.5%. The selector further raises R@0.5m to 43.9%, while the full model reaches 53.4% with confidence and improves R@1m30° from 51.5% to 60.5% over the selector variant.

(a) Source : Gibson(f)

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	7.2	38.3	46.8	45.1	✗
UnLoc	8.6	39.7	46.3	45.0	✗
DA3	6.7	38.0	46.0	44.6	✗
SeeRayLoc	**11.9**	**46.4**	**50.0**	**49.0**	✗

(b) Target : Gibson(g)

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	7.0	35.5	43.4	42.1	✗
UnLoc	7.9	36.8	42.8	41.4	✗
DA3	6.6	34.3	42.3	40.5	✗
SeeRayLoc	**13.1**	**44.7**	**49.1**	**47.5**	✗

(c) Target : Structured3D

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.2	3.7	11.3	6.5	✓
UnLoc	0.1	3.2	12.1	5.4	✓
DA3	3.7	**34.2**	**49.5**	**47.1**	✓
SeeRayLoc	**3.8**	33.2	47.9	45.4	✓

(d) Target : ZInD

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.1	1.8	5.2	1.4	✓
UnLoc	0.0	1.8	5.8	2.5	✓
DA3	1.4	**16.7**	**27.9**	**24.5**	✓
SeeRayLoc	**1.8**	15.0	23.4	20.1	✓

(e) Target : FurnFloc

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.0	0.7	2.8	0.7	✗
UnLoc	0.3	4.6	11.8	9.0	✗
DA3	2.2	13.2	18.9	16.6	✗
SeeRayLoc	**2.7**	**17.1**	**25.4**	**23.1**	✗

(f) Target : Matterport3D

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.0	0.6	2.0	0.4	✗
UnLoc	0.0	0.6	1.9	0.5	✗
DA3	**0.3**	**3.7**	**6.0**	**4.9**	✗
SeeRayLoc	0.3	3.1	5.6	4.2	✗

(g) Target : Structured3D (Panorama)

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.6	3.3	7.5	1.9	✗
UnLoc	0.0	0.9	3.6	0.9	✗
DA3	5.0	28.7	37.2	28.7	✗
SeeRayLoc	**6.4**	**34.9**	**42.7**	**32.7**	✗

(h) Target : ZInD (Panorama)

Method	0.1m	0.5m	1m	1m30°	Sem.
SemRayLoc	0.1	1.7	3.8	1.0	✗
UnLoc	0.1	1.4	4.6	1.2	✗
DA3	**25.8**	**64.9**	**68.6**	**67.3**	✗
SeeRayLoc	23.6	62.4	65.2	63.6	✗

Table 3: Results from **Gibson(f)** to other target datasets. **Sem.** denotes semantic cues included.

(a) Train and test on Gibson(f)

Method	0.1m	0.5m	1m	1m30°	Sem.
LASER	0.4	6.7	13.0	10.4	✗
F^3Loc	4.7	28.6	36.6	35.1	✗
3DP	5.3	33.2	39.8	38.4	✗
RSK	8.3	38.5	45.3	43.6	✗
DisCo-FLoc†	13.1	**50.9**	**56.7**	**55.4**	✗
Ours	**16.8**	49.3	53.5	51.8	✗

(b) Train and test on Gibson(g)

Method	0.1m	0.5m	1m	1m30°	Sem.
LASER	0.7	7.0	11.8	9.5	✗
F^3Loc	4.3	26.7	33.7	32.3	✗
3DP	9.4	37.4	43.1	41.5	✗
RSK	10.7	38.4	44.3	42.4	✗
DisCo-FLoc†	12.4	**48.0**	**55.6**	**54.0**	✗
Ours	**15.6**	47.0	50.7	49.1	✗

(c) Train and test on Structured3D

Method	0.1m	0.5m	1m	1m30°	Sem.
F^3Loc	1.5	14.6	22.4	21.3	✗
SemRayLoc	5.7	45.5	58.8	57.5	✓
FoD	12.0	65.2	71.9	71.4	✓
DisCo-FLoc†	9.5	63.9	74.3	73.6	✗
CF^2Loc†	**15.0**	**71.0**	**79.1**	**78.4**	✓
Ours	12.8	66.7	72.7	72.0	✓

(d) Train and test on ZInD

Method	0.1m	0.5m	1m	1m30°	Sem.
LASER	1.4	11.1	17.6	13.6	✗
F^3Loc	0.7	7.9	15.1	11.5	✗
SemRayLoc*	1.6	17.7	26.3	23.3	✓
FoD	8.1	43.6	53.5	50.2	✓
CF^2Loc†	**11.5**	**51.7**	**63.2**	**54.7**	✓
Ours	3.8	26.1	33.5	30.8	✓

Table 4: Results on source datasets. † denotes two-stage, while * denotes our re-implemented results.

Figure 4 makes the changes to the 40-ray depth profiles visible. In (a)-(c), the adapter and selector progressively bring many measurements closer to the floorplan reference. In (d), the wardrobe occludes the wall: the adapter estimates visible geometry, while the selector uses exposed regions above the wardrobe for several rays. Panel (e) shows local adapter corrections even without a prominent furniture occlusion. Additional examples are provided in Appendix A.4.

Table 5: Component ablation on S3D. Semantic cues are shared by all configurations.

Components	0.1m	0.5m	1m	1m30°
DA3	3.7	34.2	49.5	47.1
+ Dense Adapter	6.9	39.5	51.5	48.7
+ Selector	8.3	43.9	53.8	51.5
+ Confidence (ours)	**10.5**	**53.4**	**62.2**	**60.5**

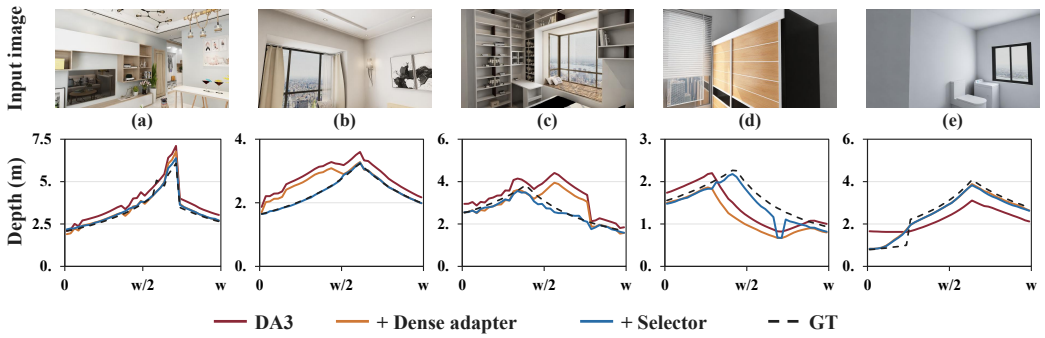


Figure 4: Depth-profile ablation on S3D examples. **Top**: input images. **Bottom**: 40-ray depth profiles from DA3, with the dense adapter, and with both the dense adapter and selector. Components are added cumulatively. Dashed lines denote floorplan depths at the ground-truth poses. The horizontal axis represents sampling position across the aligned image width. Depth is in meters. Our SeeRayLoc (blue line) aligns GT better.

Effect of confidence on localization. Figure 5 compares the localization score maps and predicted poses for two queries as the selector and confidence are added, complementing the depth-profile comparison in Figure 4. In Case 1, position and orientation errors decrease from 4.97 m and 158.2° without the selector to 1.27 m and 11.8° with the selector, then to 0.28 m and 1.8° with confidence. In Case 2, they decrease from 2.96 m and 108.8° to 2.05 m and 91.2°, then to 0.26 m and 1.2°, respectively. Confidence leaves the 40-ray profile unchanged; its additional gains come from resampling and weighting the rays used for floorplan matching.



Figure 5: Localization ablation on two S3D queries with furniture.

5 Conclusion

We propose **SeeRayLoc**, which revisits how visual geometry is connected to floorplan structure for localization. By separating visible depth estimation from structural observation selection, the framework grounds geometric prediction in observed surfaces while using floorplan supervision to guide their relevance to localization. This formulation connects pretrained depth representations with geometric floorplan matching and accommodates both perspective and panoramic observations. Furthermore, we construct **FurnFloc**, a furnished indoor localization dataset. Experiments across diverse indoor settings support the value of this separation for cross-dataset transfer, suggesting that learning how to use visible geometry is an effective direction for generalizable floorplan localization.

Limitation and future work. SeeRayLoc deliberately avoids predicting unobserved wall distances and instead selects localization-relevant measurements from visible geometry. This design also defines its main limitation: under severe furniture occlusion, the true structural boundary may be completely invisible, leaving no valid candidate for the selector to recover. Future work will investigate visibility-aware uncertainty modeling and multi-view evidence aggregation to handle such cases without reverting to direct hidden-wall regression.

AI USE STATEMENT

In this work, we used generative AI tools, including OpenAI ChatGPT, for brainstorming research concepts, supporting software development and debugging, refining the clarity and readability of the manuscript text, and formatting LaTeX code and tables. We did not use generative AI to generate synthetic datasets, formulate mathematical claims, or produce mathematical proofs. All research design, experimental methodologies, and result interpretations were independently conceived and executed by the authors. All AI-assisted code was inspected, tested, and verified through our own experimental pipeline. We take full responsibility for the final content, claims, and code.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive personal data. The proposed methods and experiments are conducted using existing public or research datasets and are intended for advancing floorplan localization research.

REPRODUCIBILITY STATEMENT

To facilitate reproduction of SeeRayLoc, we describe the model architecture, training objectives, and inference procedure in Section 3. Section 4.1 and Appendix A.2 provide details on the datasets, data splits, and preprocessing settings. Code at: https://anonymous.4open.science/r/SeeRayLoc-9573.

REFERENCES

R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Federico Boniardi, Tim Caselitz, Rainer Kümmerle, and Wolfram Burgard. Robust lidar-based localization in architectural floor plans. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3318–3324, 2017. doi: 10.1109/IROS.2017.8206168.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 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 33rd ACM International Conference on Multimedia*, MM '25, pp. 7181–7190, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754504.

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? In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i4.37268.

Changan Chen, Rui Wang, Christoph Vogel, and Marc Pollefeys. F3loc: Fusion and filtering for floorplan localization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18029–18038, 2024. doi: 10.1109/CVPR52733.2024.01707.

Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360º panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2133–2143, June 2021.

Yuval Grader and Hadar Averbuch-Elor. Supercharging floorplan localization with semantic rays, 2025.

Henry Howard-Jenkins and Victor Adrian Prisacariu. Lalaloc++: Global floor plan comprehension for layout localisation in unvisited environments. 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.

Weiqi Huang, Shuangyi Dong, Jiaxin Li, Yifei Guo, Zan Wang, and Wei Liang. Floverse: Floor plan-guided multi-modal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter networks with application to visual localization. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto (eds.), *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pp. 169–178. PMLR, 29–31 Oct 2018.

Lingyu Kong, Ruicheng Li, Ruicheng Wang, Sicheng Xu, Chengtang Yao, Jianfeng Xiang, and Jiaolong Yang. Fine-detail monocular geometry estimation with self-guided sparse volumetric refinement, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025.

Shiyong Meng, Bolei Chen, Ping Zhong, Yang Wan, Rongzhi Wang, Jiazhi Xia, and Jianxin Wang. From uncertainty to determinism: Coarse-to-fine visual floorplan localization without ray matching, 2026.

Zhixiang Min, Naji Khosravan, Zachary Bessinger, Manjunath Narayana, Sing Bing Kang, Enrique Dunn, and Ivaylo Boyadzhiev. Laser: Latent space rendering for 2d visual localization, 2022.

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12708–12717, 2019. doi: 10.1109/CVPR.2019.01300.

Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1293–1307, 2021. doi: 10.1109/TPAMI.2019.2952114.

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, 2015. doi: 10.1109/IROS.2015.7353811.

Matthias Wueest, Francis Engelmann, Ondrej Miksik, Marc Pollefeys, and Daniel Barath. Unloc: Leveraging depth uncertainties for floorplan localization. In *Proc. ICLR*, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 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 (CVPR)*, 2026.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020.

Ping Zhong, Shiyong Meng, Bolei Chen, Tao Zou, Chaoxu Mu, and Jianxin Wang. Disco-floc: Semantic-free floorplan localization via $se(2)$-aware contrastive disambiguation, 2026.

A Appendix

A.1 Cross-dataset Localization Performance

To examine whether existing indoor localization methods generalize beyond their training domains, we directly evaluate source-trained models on unseen datasets without target-domain adaptation. Fig. 6 shows the source-normalized Recall@1m, with source-domain performance set to 100%. Most source-target pairs exhibit substantial performance degradation, indicating limited cross-dataset transferability. This observation motivates a systematic cross-dataset evaluation under our visible-depth and ray selection formulation, with models trained on selected source domains and directly transferred to multiple unseen target datasets.

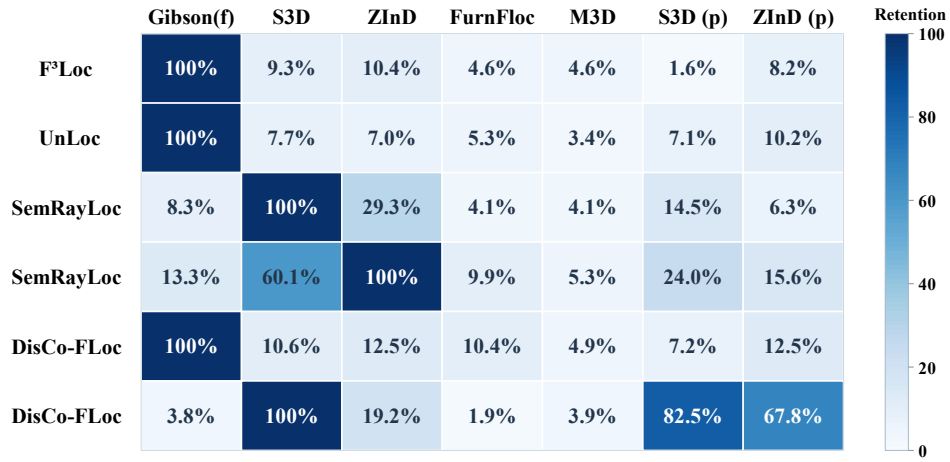


Figure 6: Source-normalized cross-dataset localization performance measured by Recall@1m. Each row represents a model trained on the dataset marked by its 100% cell. The remaining cells report target-domain Recall@1m as a percentage of the corresponding source-domain performance. Repeated method names denote models trained on different source datasets, and (p) denotes panoramic inputs. Detailed results are provided in Table 6.

A.2 Dataset details and preprocessing.

Matterport3D (M3D) (Chang et al., 2017) is a real-world RGB-D dataset containing 10,800 panoramic capture locations and 194,400 RGB-D images from 90 building-scale environments. Each capture location contains six azimuthal views at each of three elevation levels. We retain only the six level (middle-elevation) views, resulting in 64,800 candidate perspective images, and assign each camera to its corresponding floor level. Some capture locations lie on balconies, in courtyards, or outside the building, producing images dominated by outdoor content. We manually inspect and exclude 6,852 such outdoor or semi-outdoor views, leaving 57,948 images across 183 floorplans.

Structured3D (S3D) (Zheng et al., 2020) is a synthetic, photorealistically rendered dataset containing 3.5K house designs with detailed 3D structural and semantic annotations. It provides perspective and equirectangular panoramic renderings in both empty and fully furnished variants. We use the fully furnished RGB renderings from both the perspective and panorama modalities. During preprocessing, incomplete scenes are discarded, with panoramic data validated and filtered at the scene level.

ZInD (Cruz et al., 2021) is a real-world indoor dataset containing 67,448 equirectangular panoramas captured from 1,575 unfurnished homes. After excluding floorplans without a valid metric scale, we retain 2,453 floor-plan units. ZInD provides camera poses, room layouts, and semantic annotations for windows, doors, openings, and room types. We evaluate both the original ERP panoramas and a front-view variant constructed following SemRayLoc (Grader & Averbuch-Elor, 2025), in which each panorama is converted into an 80° perspective image.

(a) F^{3}Loc (Chen et al., 2024)

Dataset	0.1m	0.5m	1m	1m30°	Sem.
Source					
Gibson(f)	4.7	28.6	36.6	35.1	✗
Target					
S3D	0.0	0.8	3.4	1.5	✗
ZInD	0.0	0.9	3.8	0.9	✗
FurnFloc	0.0	0.4	1.7	0.4	✗
M3D	0.0	0.5	1.7	0.3	✗
S3D (p)	0.0	0.0	0.6	0.0	✗
ZInD (p)	0.0	0.2	3.0	0.4	✗

(b) UnLoc (Wueest et al., 2026)

Dataset	0.1m	0.5m	1m	1m30°	Sem.
Source					
Gibson(f)	19.7	61.1	64.7	63.8	✗
Target					
S3D	0.1	1.3	5.0	1.1	✗
ZInD	0.0	1.6	4.5	1.4	✗
FurnFloc	0.0	0.9	3.4	1.2	✗
M3D	0.0	0.7	2.2	0.7	✗
S3D (p)	0.0	1.6	4.6	0.9	✗
ZInD (p)	0.6	3.3	6.6	3.1	✗

(c) SemRayLoc (Grader & Averbuch-Elor, 2025)

Dataset	0.1m	0.5m	1m	1m30°	Sem.
Source					
S3D	5.7	45.5	58.8	57.5	✓
Target					
Gibson(f)	0.1	1.1	4.9	2.9	✗
ZInD	0.8	9.0	17.2	14.1	✓
FurnFloc	0.0	0.5	2.4	0.5	✗
M3D	0.0	0.7	2.4	1.0	✗
S3D (p)	0.3	4.2	8.5	6.2	✗
ZInD (p)	0.0	1.3	3.7	2.2	✗

(d) SemRayLoc (Grader & Averbuch-Elor, 2025)

Dataset	0.1m	0.5m	1m	1m30°	Sem.
Source					
ZInD	1.6	17.7	26.3	23.3	✓
Target					
Gibson(f)	0.0	0.7	3.5	0.9	✗
S3D	0.5	6.2	15.8	11.6	✓
FurnFloc	0.0	0.7	2.6	0.6	✗
M3D	0.0	0.4	1.4	0.4	✗
S3D (p)	0.1	2.2	6.3	3.2	✗
ZInD (p)	0.0	1.6	4.1	2.7	✗

(e) DisCo-FLoc (Zhong et al., 2026)

Dataset	0.1m	0.5m	1m	1m30°	Sem.
Source					
Gibson(f)	13.1	50.9	56.7	55.4	✗
Target					
S3D	0.0	1.3	6.0	1.0	✗
ZInD	0.1	2.4	7.1	1.8	✗
FurnFloc	0.0	1.8	5.9	3.7	✗
M3D	0.0	0.8	2.8	0.6	✗
S3D (p)	0.0	1.3	4.1	1.1	✗
ZInD (p)	0.5	3.4	7.1	3.1	✗

(f) DisCo-FLoc (Zhong et al., 2026)

Dataset	0.1m	0.5m	1m	1m30°	Sem.
Source					
S3D	9.5	63.9	74.3	73.6	✗
Target					
Gibson(f)	0.0	0.8	2.8	0.9	✗
ZInD	0.9	7.9	14.3	11.0	✗
FurnFloc	0.0	0.3	1.4	0.3	✗
M3D	0.0	0.9	2.9	1.4	✗
S3D (p)	15.4	57.2	61.3	59.7	✗
ZInD (p)	11.0	45.1	50.4	49.0	✗

Table 6: Cross-dataset localization recall of different methods using their respective pretrained weights. The pretrained weights of UnLoc are trained on Gibson(f) and evaluated on Gibson(t).

Gibson introduced by F^{3}Loc (Chen et al., 2024), contains images rendered from reconstructed real-world Gibson environments. Gibson(f) contains 24,779 four-view sequences generated under forward motion, whereas Gibson(g) contains 49,558 four-view sequences generated under general motion. The associated maps are geometric occupancy floorplans and do not contain door, window, or room-type semantics.

Dataset Splits and Semantic Annotations. For Structured3D, Matterport3D, ZInD, and Gibson, we follow their official train/validation/test splits. Multi-storey floors from the same building or home are always kept in the same partition. FurnFloc uses a custom leakage-free building-level split: 75 buildings are used for training, 10 for validation, and 16 for testing, corresponding to 98, 12, and 22 floor-plan units, respectively. The test buildings are drawn exclusively from the original HM3D validation partition, while the training and validation buildings are obtained from the HM3D training partition using a deterministic split. Structured3D, ZInD and Matterport3D provide layout and structural semantics, including doors, windows, and room categories. Gibson and FurnFloc contain geometry-only floorplans.

We additionally report floorplan-size statistics in Table 7, including the number of floorplans and the minimum, median, and maximum spatial extent of each dataset, to characterize differences in environment scale.

Table 7: Floorplan statistics across datasets. Area is reported in m^2, and dimensions are reported as length × width in m.

Dataset	Maps	Min. Area / Dimensions	Median Area / Dimensions	Max. Area / Dimensions
Panorama				
Structured3D	3,475	5.49 (3.20 × 1.80)	78.54 (9.57 × 10.99)	1185.45 (46.15 × 30.82)
ZInD	2,453	5.56 (2.13 × 4.11)	97.07 (16.09 × 7.51)	452.11 (23.85 × 27.93)
Perspective				
Matterport3D	183	1.54 (1.29 × 1.28)	164.16 (10.59 × 19.44)	1979.87 (81.06 × 59.87)
Structured3D	3,296	5.49 (3.20 × 1.80)	78.55 (9.74 × 10.76)	799.22 (35.68 × 25.92)
ZInD	2,453	5.56 (2.13 × 4.11)	97.07 (16.09 × 7.51)	452.11 (23.85 × 27.93)
Gibson(f)	118	18.12 (5.30 × 4.40)	72.94 (7.90 × 11.80)	271.73 (16.90 × 20.60)
Gibson(g)	118	18.12 (5.30 × 4.40)	72.94 (7.90 × 11.80)	271.73 (16.90 × 20.60)
FurnFloc	132	23.01 (4.20 × 5.80)	112.41 (12.20 × 10.20)	407.90 (27.20 × 22.20)

A.3 Direct ERP Processing and Circular Fusion

For panoramic queries, we process the full equirectangular projection (ERP) directly without perspective tiling or tangent-plane reprojection. The ERP image is patch-embedded once to form a global token canvas, from which overlapping local token windows are extracted in ERP coordinates and processed using the same frozen backbone and learned modules. Local predictions are then mapped to their corresponding global longitudes and circularly fused into a unified 360° structural profile, which is finally sampled into the 40 rays used for floorplan localization.

Let ψ_i denote the center longitude of the i-th local window and θ a target global ray direction. We define the wrapped angular offset as

$$\delta_i(\theta) = \mathrm{wrap}(\theta - \psi_i), \tag{11}$$

where $\mathrm{wrap}(\cdot)$ maps angular differences to the principal circular interval. A window contributes to direction θ when $|\delta_i(\theta)| < \alpha$, where α denotes the angular fusion support.

For each contributing window, we assign a center-weighted raised-cosine weight

$$\tilde{w}_i(\theta) = \cos^2\left(\frac{\pi}{2}\frac{|\delta_i(\theta)|}{\alpha}\right), \tag{12}$$

which is normalized over all valid windows:

$$w_i(\theta) = \frac{\tilde{w}_i(\theta)}{\sum_{j\in\mathcal{V}(\theta)} \tilde{w}_j(\theta)}, \tag{13}$$

where $\mathcal{V}(\theta)$ denotes the set of windows covering θ. Predictions near a window center therefore receive greater weight, while the contribution decreases smoothly toward its boundary.

Before fusion, local depth predictions are converted into horizontal ray ranges. Let $r_i(\theta)$ denote the range predicted by window i at direction θ. We aggregate overlapping predictions in log-range space:

$$r(\theta) = \exp\left(\sum_{i\in\mathcal{V}(\theta)} w_i(\theta)\log r_i(\theta)\right). \tag{14}$$

This corresponds to a weighted geometric mean and reduces the influence of scale-dependent discrepancies between overlapping predictions.

When confidence estimates $c_i(\theta)$ are available, they are fused linearly using the same normalized weights:

$$c(\theta) = \sum_{i\in\mathcal{V}(\theta)} w_i(\theta)c_i(\theta). \tag{15}$$

The resulting circular range and confidence profiles are sampled at the predefined global structural-ray directions, yielding the 40-ray 360° representation used by the floorplan-matching stage.

A.4 Additional Examples

Figures 7 and 8 show additional qualitative examples of the S3D-trained model directly transferred to unseen perspective-image and panorama-image datasets, respectively.

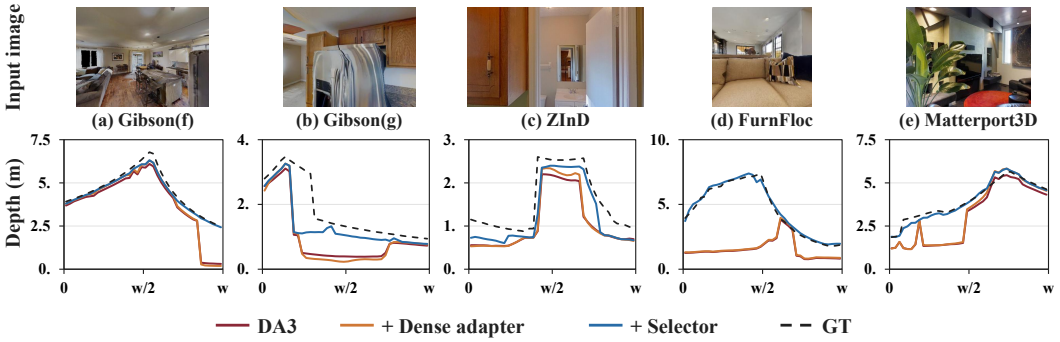


Figure 7: Additional perspective examples for depth-profile ablation. **Top**: input images. **Bottom**: 40-ray depth profiles from DA3, with the dense adapter, and with both the dense adapter and selector. Components are added cumulatively. Dashed lines denote floorplan depths at the ground-truth poses. The horizontal axis represents sampling position across the aligned image width. Depth is in meters. Our SeeRayLoc (blue line) aligns GT better.

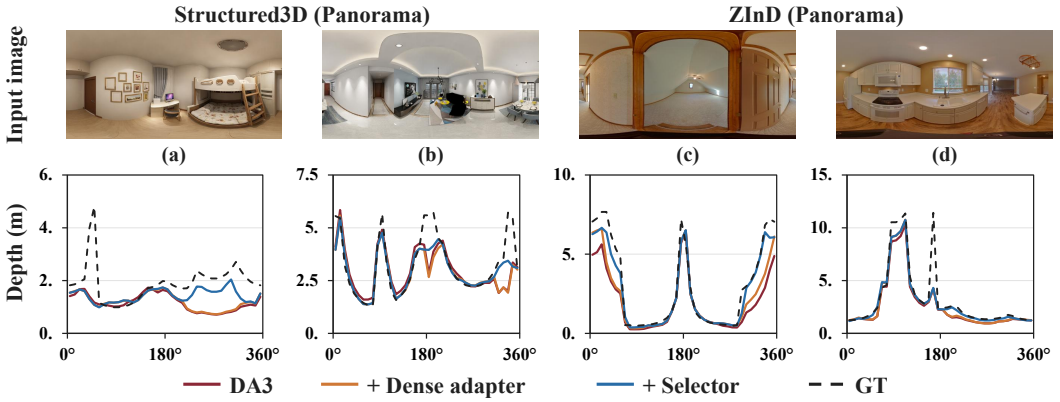


Figure 8: Additional panorama examples for depth-profile ablation. **Top**: input images. **Bottom**: 40-ray depth profiles from DA3, with the dense adapter, and with both the dense adapter and selector. Components are added cumulatively. Dashed lines denote floorplan depths at the ground-truth poses. The horizontal axis represents sampling position across the aligned image width. Depth is in meters. Our SeeRayLoc (blue line) aligns GT better.