RAGE-I2P: RELIABILITY-AWARE GEOMETRIC EMBEDDINGS FOR IMAGE-TO-POINT CLOUD REGISTRATION IN CHANGING ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing image-to-point cloud registration methods have achieved strong performance in static scenes. However, real-world environments often contain moved, removed, or newly added objects. Such changes can preserve feature similarity while breaking physical correspondence, producing unreliable matches at both coarse and fine levels. We propose RAGE-I2P, a geometry-guided framework for reliable registration in changing environments. We learn coarse- and fine-level geometric embeddings from image and point-cloud normals. At the coarse level, a covariance-aware discrepancy between geometric embeddings provides compatibility weights for cross-modal attention. At the fine level, normal-derived embeddings enrich feature extraction with local surface information, while an additional embedding loss encourages correspondence-discriminative representations. We also propose ASIMLab, a seven-scene dataset captured under real scene changes, and construct Changing RGBDScenesV2 through controlled image edits. Experiments on both benchmarks show that RAGE-I2P achieves the best mean correspondence and registration performance. The source code and the datasets used in the paper will be released soon.

1 INTRODUCTION

Image-to-point cloud (I2P) registration is an important task in computer vision, with broad applications in world action models Huang et al. (2026), robotic navigation Yan et al. (2025), and 3D/4D reconstruction Liu et al. (2025). Existing methods improve correspondence quality through post-processing Peng et al. (2025), graph reasoning An et al. (2026); Bie et al. (2025), and specialized loss functions An et al. (2025a). However, they are mainly developed and evaluated on observations without substantial structural changes. In long-term operation, moved, removed, or newly added objects can produce feature-similar but physically invalid correspondences, which degrade pose estimation. Dynamic SLAM methods Jiang et al. (2024); Bescos et al. (2018) often address scene changes by detecting moving objects and discarding their measurements. This strategy may also remove useful static information and reduce scene completeness.

The main challenge of I2P registration in changing environments is to distinguish appearance-similar correspondences that differ in their geometric compatibility. Our key insight is that surface normals provide complementary cues for learning correspondence-discriminative geometric representations. Their spatial variation describes local surface structure and can enrich matching features when appearance alone is ambiguous. We therefore learn normal-derived geometric embeddings at both coarse and fine levels: coarse embeddings provide a compatibility signal for cross-modal interaction, while fine embeddings introduce local surface information into dense feature learning.

Based on this insight, we propose RAGE-I2P, which extends LDF-I2P Peng et al. (2025) to image-to-point cloud registration under scene changes. RAGE-I2P uses feature encoders to learn geometric embeddings from surface normals. Coarse embedding discrepancies modulate cross-modal attention, while fine embeddings and their supervision improve local correspondence discrimination. Our contributions are summarized as follows:

- We construct a dataset for image-to-point cloud registration under scene changes. The dataset contains seven scenes. Unlike existing image-to-point cloud registration datasets,

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.

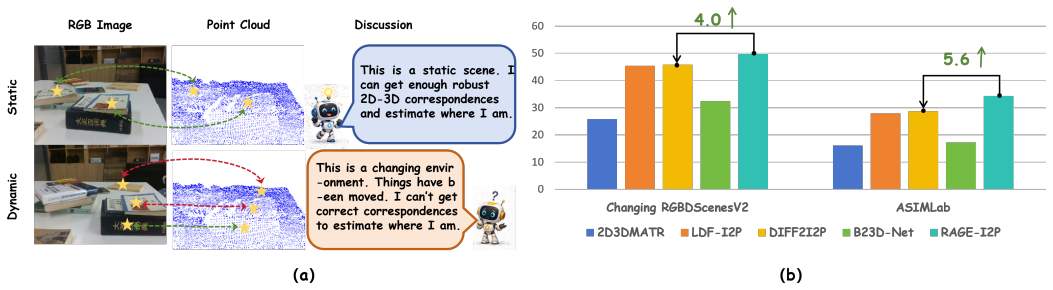


Figure 1: (a) Visualization of our motivation: In real-world scenes, objects may be moved, removed or added. This makes it harder for a robot to identify reliable correspondences to estimate its pose. (b) Registration recall on Changing RGBDScenesV2 and ASIMLab.

the images and point clouds in our dataset are captured from the same scene at different time points, **where objects are moved, removed, or added**. We believe this setting better reflects real-world image-to-point cloud registration scenarios.

- We propose RAGE-I2P, a framework that learns normal-derived geometric embeddings for coarse-to-fine image-to-point cloud registration under scene changes. At the coarse level, covariance-aware embedding discrepancies provide geometric compatibility weights for cross-modal attention. At the fine level, geometric embeddings and an additional embedding loss enhance correspondence discrimination.
- We conduct experiments on two datasets: a synthetically modified dataset and a newly collected dataset. The proposed method achieves improvements of up to 0.7 percentage points (pp) on IR and 5.6 pp on RR, outperforming previous state-of-the-art methods.

2 Related Work

2.1 Registration Methods

Intra-modality Registration. MAC Yang et al. (2024) utilizes compatibility graphs to mine local consensus via maximal cliques. It can generate more accurate pose hypotheses for robust registration. Based on MAC, MAC++ Zhang et al. (2025b) enhances graphs with voting-guided clique pools and hierarchical evaluation to handle extremely severe outliers and low-overlap challenges. HyperGCT Zhang et al. (2025a) proposes a dynamic hypergraph neural network that models 3D correspondences to capture high-order geometric consistency for robust registration. By implicitly constructing and optimizing the hypergraph structure, it effectively filters outliers and generates accurate pose hypotheses.

Inter-modality Registration. The main problem in inter-modality registration is how to align cross-modal features to obtain reliable correspondences. Early image-to-point cloud registration methods usually follow a detect-then-match paradigm, where 2D keypoints and 3D keypoints are first detected independently and then matched in a shared feature space Wang et al. (2021). Although these methods provide a feasible solution for sparse cross-modal correspondence estimation, they often process images and point clouds separately and ignore sufficient interaction between different modalities. To address this limitation, recent methods have shifted toward detection-free registration. Hg-I2P An et al. (2026) uses a heterogeneous graph to enhance cross-modal feature interaction and prune incorrect correspondences via geometric constraints. GraphI2P and its previous works Bie et al. (2025; 2024) use visual foundation models to transfer 2D-3D registration problem to 3D registration problem. After establishing initial correspondences, they use graph structure to filter outliers. Following this line of work, Angle-I2P Peng et al. (2026) constructs a global-to-local graph to reject outliers at different scales. R23Net Cheng et al. (2026) selects high-value matching regions and aligns cross-modal zone representations, achieving state-of-the-art image-to-point cloud registration with improved accuracy and robustness.

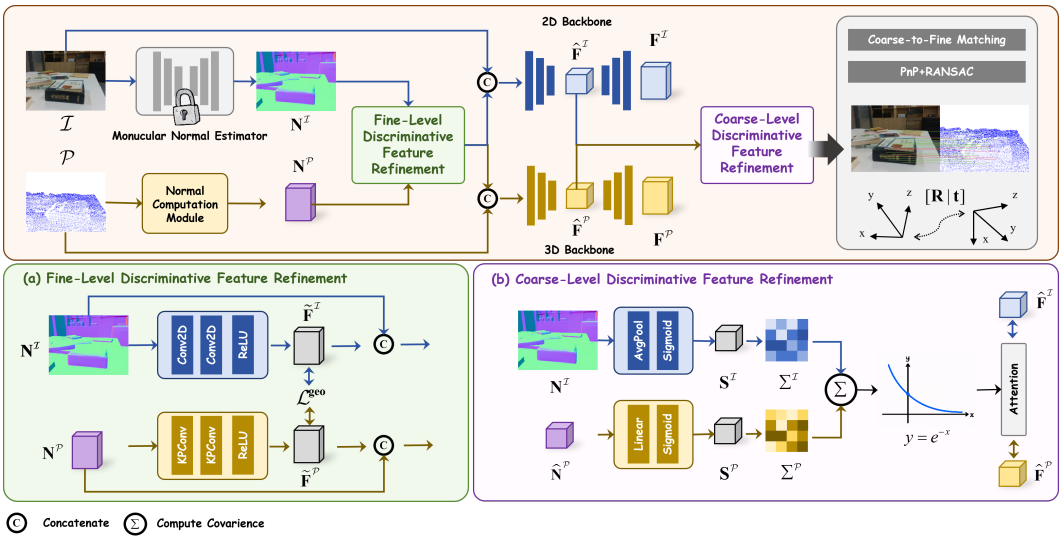


Figure 2: Overview of RAGE-I2P. Image and point-cloud normals are encoded into geometric embeddings at two resolutions. (a) Fine-level embeddings enrich the inputs to the feature extraction networks and are regularized for correspondence discrimination. (b) Coarse-level embedding discrepancies are normalized using covariance statistics and converted into compatibility weights for cross-modal attention.

2.2 Localization and Reconstruction in Dynamic Scenes

Most existing methods leverage semantic segmentation results as a prior and mitigate the impact of dynamic objects on localization by discarding or filtering out information from dynamically changing regions Bescos et al. (2018; 2021). These methods may lose information and reduce precision of the system. DROID-W Li et al. (2026) is a robust RGB SLAM system that handles dynamic scenes by leveraging differentiable uncertainty-aware bundle adjustment with per-pixel uncertainty estimated from multi-view visual feature inconsistency. DynamicVGGT He et al. (2026) unifies dynamic point maps, motion-aware temporal attention, future point prediction, and dynamic 3D Gaussian splatting to achieve 4D reconstruction without camera parameters or dense annotations. Unlike dynamic SLAM, which exploits temporal observations to detect or track moving objects, our setting considers asynchronously captured images and point-cloud maps after structural changes have occurred.

3 Methods

3.1 Overview

Given a set of image pixels $\{\mathbf{p}_i\}_{i=1}^{N}$ and a point cloud $\{\mathbf{q}_j\}_{j=1}^{M}$. $\mathbf{p}_i$ and $\mathbf{q}_j$ form an inlier correspondence if their reprojection error is below a threshold θ (Eq. 1):

$$< \mathbf{p}_i, \mathbf{q}_j > \Longleftrightarrow ||\pi(\mathbf{y}_j, \mathcal{K}, \mathbf{T}) - \mathbf{x}_i||_2 \leq \theta \tag{1}$$

where $< \mathbf{p}_i, \mathbf{q}_j >$ means $\mathbf{p}_i$ and $\mathbf{q}_j$ are a pair of inliers. $\pi(\cdot)$ is the projection method and $\mathcal{K}$, $\mathbf{T}$ are intrinsics and extrinsics of the camera. Recent `Detect-then-Match` methods like 2D3DMATR Li et al. (2023); Bie et al. (2025) use features of images and point cloud in latent space to get correspondences. Cross-modal features within a certain threshold are selected as inliers (Eq. 2):

$$< \mathbf{p}_i, \mathbf{q}_j > \Longleftrightarrow \Delta\mathbf{f} = ||\mathbf{f}_i^{\mathcal{I}} - \mathbf{f}_j^{\mathcal{P}}||_2 \leq \theta_f \tag{2}$$

In changing environments, appearance-similar image regions and point-cloud regions may cease to form valid correspondences under the camera-to-map transformation. Consequently, a small feature distance alone does not establish correspondence validity. We propose normal-derived geometric embeddings to provide complementary surface information for matching. The embeddings are learned through modality-specific encoders, and their discrepancies are used to guide cross-modal feature interaction and correspondence discrimination.

To address these challenges, RAGE-I2P learns geometric embeddings from image and point-cloud normals at two resolutions, as illustrated in Fig. 2. At the fine level, modality-specific geometric encoders provide local surface information for feature extraction, with an additional loss encouraging correspondence-discriminative embeddings (Sec. 3.2). At the coarse level, we compute covariance-aware discrepancies between geometric embeddings and use the resulting compatibility weights to modulate cross-modal attention (Sec. 3.3). The resulting features are used to get coarse-to-fine correspondences and estimate camera pose.

3.2 Fine-Level Discriminative Feature Learning

In dynamically changing scenes, objects may move, disappear, or appear, so simply using cross-modal features can easily lead to mismatches. After feature extraction, the differences between the extracted features no longer arise solely from modality discrepancies, but also include variations in geometric information. In such changing scenes, correct correspondences should share similar local geometric information. Based on this observation, we leverage robust geometric information to learn fine-level cross-modal features.

$$\mathbf{G}_{\mathrm{f}}^{\mathcal{I}} = \phi_{\mathrm{f}}^{\mathcal{I}}(\mathbf{N}^{\mathcal{I}}), \;\; \mathbf{G}_{\mathrm{f}}^{\mathcal{P}} = \phi_{\mathrm{f}}^{\mathcal{P}}(\mathbf{N}^{\mathcal{P}}) \tag{3}$$

Here, $\mathbf{N}^{\mathcal{I}}$ and $\mathbf{N}^{\mathcal{P}}$ denote the image and point-cloud normals. The modality-specific encoders $\phi_{\mathrm{f}}^{\mathcal{I}}$ and $\phi_{\mathrm{f}}^{\mathcal{P}}$ map normals into learned geometric embeddings. Cross-modal compatibility is learned in the embedding space, allowing each encoder to process normals expressed in its modality's coordinate frame. Their geometric embeddings are concatenated with the original inputs and initial features: $\mathbf{F}_{\mathrm{in}}^{\mathcal{I}} = [\mathcal{I}; \mathbf{N}^{\mathcal{I}}; \mathbf{G}_{\mathrm{f}}^{\mathcal{I}}] \in \mathbb{R}^{H\times W\times 5}$, $\mathbf{F}_{\mathrm{in}}^{\mathcal{P}} = [\mathbf{1}; \mathbf{N}^{\mathcal{P}}; \mathbf{G}_{\mathrm{f}}^{\mathcal{P}}] \in \mathbb{R}^{N\times 5}$, and feed them into the 2D (ResNet He et al. (2016)) and 3D (KPConv Thomas et al. (2019)) feature extraction networks. We additionally supervise the fine-level geometric embeddings with the loss in Eq. 4, which encourages smaller embedding distances for positive training pairs and larger distances for negative pairs.

$$\mathcal{L}_{\mathrm{g}} = \frac{1}{\gamma}\log[1 + \sum_{\mathbf{d}_j \in \mathcal{D}_i^{\mathcal{P}}} e^{\beta_p^{i,j}(d_i^j - \Delta_p)} \cdot \sum_{\mathbf{d}_k \in \mathcal{D}_i^{\mathcal{N}}} e^{\beta_n^{i,k}(\Delta_n - d_i^k)}] \tag{4}$$

where d_i^j is the pairwise feature distance between $\mathbf{G}_{\mathrm{f}}^{\mathcal{I}}$ and $\mathbf{G}_{\mathrm{f}}^{\mathcal{P}}$. In changing scenes, when matching is performed using fine-level features, such geometric characteristics can better assist the model in identifying correct correspondences, rather than relying solely on texture information. Further details of the loss are provided in Sec. 3.4.

3.3 Coarse-Level Discriminative Feature Learning

Due to object motion and scene changes, image and point cloud features extracted from corresponding regions may no longer be strictly consistent in the feature space. In particular, patch-level features from dynamic objects or changed regions can still show high feature similarity, but they no longer represent valid correspondences under the same rigid camera transformation. Therefore, directly relying on feature similarity may introduce unstable coarse-level correspondences, which further affects subsequent fine-level matching.

To introduce geometric information into coarse-level matching, we learn normal-derived embeddings for image patches and point-cloud superpoints. Their cross-modal discrepancies provide a geometric compatibility signal that modulates feature interaction. The coarse geometric embeddings are computed as follows:

$$\mathbf{G}_{\mathrm{c}}^{\mathcal{I}} = \phi_{\mathrm{c}}^{\mathcal{I}}(\mathrm{AvgPool2D}(\mathbf{N}^{\mathcal{I}})), \;\; \mathbf{G}_{\mathrm{c}}^{\mathcal{P}} = \phi_{\mathrm{c}}^{\mathcal{P}}(\mathbf{N}^{\mathcal{P}}) \tag{5}$$

Let $\mathbf{g}_i^{\mathcal{I}}$ and $\mathbf{g}_j^{\mathcal{P}}$ denote the coarse geometric embeddings in $\mathbf{G}_{\mathrm{c}}^{\mathcal{I}}$ and $\mathbf{G}_{\mathrm{c}}^{\mathcal{P}}$, respectively. We compute the covariance of the embeddings within each modality to characterize feature-wise dispersion and inter-channel correlations. The sum of these covariance matrices defines the normalization metric for cross-modal embedding differences. Interpreting this sum as a residual covariance amounts to neglecting cross-modal covariance terms. We use the resulting covariance-aware distance as a measure of geometric embedding discrepancy, as defined in Eq. 6 and Eq. 7.

$$\Sigma^{\mathcal{I}} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{g}_i^{\mathcal{I}} - \mu^{\mathcal{I}})(\mathbf{g}_i^{\mathcal{I}} - \mu^{\mathcal{I}})^{\mathrm{T}}, \;\; \Sigma^{\mathcal{P}} = \frac{1}{N}\sum_{j=1}^{N}(\mathbf{g}_j^{\mathcal{P}} - \mu^{\mathcal{P}})(\mathbf{g}_j^{\mathcal{P}} - \mu^{\mathcal{P}})^{\mathrm{T}} \tag{6}$$

$$\mathbf{r}_{ij} = (\mathbf{g}_i^{\mathcal{I}} - \mathbf{g}_j^{\mathcal{P}})^{\mathrm{T}}(\Sigma^{\mathcal{I}} + \Sigma^{\mathcal{P}} + \epsilon\mathbf{I})^{-1}(\mathbf{g}_i^{\mathcal{I}} - \mathbf{g}_j^{\mathcal{P}}) \tag{7}$$

where $\mu^{\mathcal{I}}$, $\mu^{\mathcal{P}}$ are the mean values of $\mathbf{g}^{\mathcal{I}}$, $\mathbf{g}^{\mathcal{P}}$. Following geometry-modulated attention designs in PointDSC Bai et al. (2021) and CAST Huang et al. (2024), we convert the embedding discrepancies into geometric compatibility weights and apply them to cross-modal attention logits before softmax normalization.

$$\mathrm{Attention} = \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}} \odot \mathbf{W}^{\mathbf{c}}}{\sqrt{\mathrm{d}}})\mathbf{V} \tag{8}$$

Here, $\mathbf{R} = [\mathbf{r}_{ij}]$ contains the covariance-aware embedding discrepancies, and $\mathbf{W}^{\mathbf{c}} = \exp(-\mathbf{R}/\tau)$ converts them into compatibility weights. In cross-attention, $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are computed from features of the two modalities. Multiplication by $\mathbf{W}^{\mathbf{c}}$ makes the attention logits depend on both feature similarity and normal-derived geometric compatibility. We use correspondence reliability to refer to this relative compatibility weighting and its effect on matching and registration is evaluated empirically.

3.4 Loss Function

Following 2D3DMATR Li et al. (2023), we retain the coarse-to-fine matching losses, which pull positive image-point cloud pairs together and push negative pairs apart in the latent feature space. However, feature similarity alone becomes ambiguous under scene changes because visually similar regions may no longer preserve the same geometric relationship. Reliable registration should therefore maintain both cross-modal feature similarity and local geometric consistency. We propose the geometric loss $\mathcal{L}_{\mathrm{g}}$ in Eq. 4 as additional supervision. It pulls geometrically consistent pairs closer and separates inconsistent pairs, encouraging normal-guided features that preserve local geometry rather than relying on appearance alone.

The final training objective is formulated as Eq. 9:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{c}} + \lambda_{\mathrm{f}}\mathcal{L}_{\mathrm{f}} + \lambda_{\mathrm{g}}\mathcal{L}_{\mathrm{g}} \tag{9}$$

where $\mathcal{L}_{\mathrm{c}}$ and $\mathcal{L}_{\mathrm{f}}$ denote the patch-level and dense pixel-point matching losses inherited from 2D3DMATR. λ_{c}, λ_{f} and λ_{g} balance the coarse-level matching loss, fine-level matching loss, and geometric loss, respectively. Compared with the original 2D3DMATR objective, the additional $\mathcal{L}_{\mathrm{g}}$ explicitly injects geometric priors into feature learning. This makes the learned representation more sensitive to static geometric consistency and more robust to scene changes, thereby improving the reliability of the subsequent coarse-to-fine matching and pose estimation.

4 Experiments

To our knowledge, existing I2P benchmarks do not explicitly model scene changes between image and point-cloud observations. We therefore collect the ASIMLab dataset (Sec. 4.3) and construct Changing RGBDScenesV2 from the original RGBDScenesV2 dataset Lai et al. (2014) (Sec. 4.2). We also evaluate the effectiveness of each module in Sec. 4.4.

Table 1: Evaluation Results on Changing RGBDScenesV2 Dataset. **Boldfaced** numbers highlight the best and the second best are underlined. † indicates that we reproduce this method based on the published paper.

Model	Scene-11	Scene-12	Scene-13	Scene-14	Mean
Mean Depth (m)	1.84	1.49	1.16	1.24	1.43
Edit Type	*Add*	*Add*	*Remove*	*Add+Remove*	-
Patch Inlier Ratio (PIR, @0.3)					
MATR Li et al. (2023)	18.5	9.6	4.5	8.3	10.2
LDF-I2P Peng et al. (2025)	26.8	10.1	8.2	11.3	14.1
Diff2I2P Mu et al. (2025)	27.2	**12.4**	9.6	10.9	15.0
B23D-Net† Cheng et al. (2025)	20.6	11.6	8.3	8.0	12.1
RAGE-I2P	**28.4**	11.7	**10.4**	**13.4**	**16.0**
Inlier Ratio (IR, @0.1m)					
MATR Li et al. (2023)	36.4	**12.6**	12.7	10.1	18.0
LDF-I2P Peng et al. (2025)	38.3	10.4	17.7	15.2	20.4
Diff2I2P Mu et al. (2025)	**40.8**	9.7	19.8	13.5	21.0
B23D-Net† Cheng et al. (2025)	30.2	9.9	18.3	8.5	16.7
RAGE-I2P	36.2	10.0	**23.6**	**16.0**	**21.4**
Registration Recall (RR, @5°, 0.15m)					
MATR Li et al. (2023)	23.6	41.2	12.4	26.1	25.8
LDF-I2P Peng et al. (2025)	**63.9**	45.1	34.0	38.5	45.4
Diff2I2P Mu et al. (2025)	57.7	52.4	35.6	37.4	45.8
B23D-Net† Cheng et al. (2025)	31.9	42.2	34.8	20.8	32.4
RAGE-I2P (ours)	56.9	**57.8**	**41.2**	**43.4**	**49.8**

4.1 Implementation Details

Network Architecture. To make the performance gains attributable to the proposed geometric refinement rather than to a larger backbone, we follow the detection-free coarse-to-fine design shared by recent I2P methods Li et al. (2023); Mu et al. (2025); An et al. (2025b). Specifically, the image and point-cloud branches use four-stage feature pyramids built on ResNet He et al. (2016) and KPConv Thomas et al. (2019), respectively. For the geometric branch, we use the pretrained DSINE model Bae & Davison (2024) to predict surface normals directly from the monocular RGB image, with its parameters frozen during registration training. Point-cloud normals are computed using local PCA. The two normal fields serve as inputs to the modality-specific geometric encoders. This task-oriented design avoids the potential error accumulation caused by monocular depth estimation, back-projection, and subsequent neighborhood-based normal fitting (Ablation studies are conducted in Sec. 4.4).

Baselines. We select four learning-based baselines that cover complementary I2P registration paradigms. 2D3D-MATR Li et al. (2023) is the primary detection-free coarse-to-fine baseline. LDF-I2P Peng et al. (2025) measures the benefit of extending discriminative feature learning from static to changing scenes. Diff2I2P Mu et al. (2025) represents the diffusion-prior and fully differentiable registration paradigm, whereas B23D-Net Cheng et al. (2025) represents uncertainty-aware hierarchical alignment with structural cross-modal reasoning.

Metrics. We adopt *Patch Inlier Ratio* (PIR) to assess the correctness of coarse-level feature matching, *Inlier Ratio* (IR) to evaluate fine-level feature matching, and *Registration Recall* (RR) to measure the accuracy of the estimated 6-DoF pose. The calculation of all metrics follows previous works, with details provided in Appendix A.4. We use an overlap threshold of 0.3 for PIR and a 3D distance threshold of 0.1 m for IR on both datasets. For RR, an image–point cloud pair is considered successfully registered only when both its relative rotation error and relative translation error are below the specified thresholds. We use thresholds of 5° and 0.15 m on Changing RGBDScenesV2, and 15° and 0.3 m on ASIMLab.

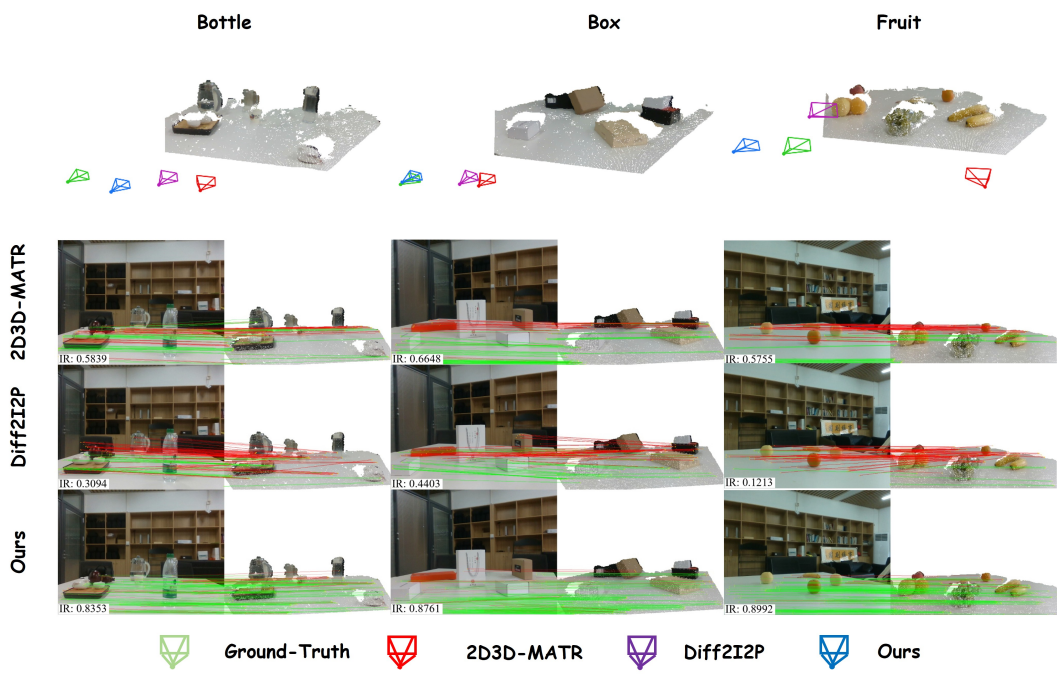


Figure 3: We present visual comparisons of camera pose estimation and matching results for different methods. The first column shows the estimated camera poses under various scenes. The right two columns display the matching visualizations obtained by different approaches, with red lines denoting outliers and green lines denoting inliers. Our method achieves more accurate camera poses and a higher inlier ratio.

4.2 Results on Changing RGBDScenesV2 Dataset

Datasets. Motivated by recent advances in image editing models, we applied ChatGPT to edit images from the RGBDScenesV2 dataset Lai et al. (2014) through operations such as object removal, object insertion, and color modification. After editing the images, we used a metric-scale monocular depth estimator MoGe-2 Wang et al. (2026) to estimate the depth. We modified only the images, while keeping the point clouds, camera extrinsics, and camera intrinsics unchanged. The mean depth of each test scene is also reported. This process resulted in `1748` training pairs, `236` validation pairs, and `497` testing pairs. More details about this process are shown in Appendix A.1.

Evaluation Results. To simulate changes in real-world scenes, all methods are trained on the original RGBDScenesV2 training split and evaluated directly on the edited test split without fine-tuning. By suppressing unreliable coarse correspondences and enforcing geometry-aware fine-level representations, our method achieves the highest mean PIR, IR, and RR. Diff2I2P learns cross-modal matching from static training scenes, so its performance may be affected when added, removed, or moved objects change the valid image-to-point-cloud correspondences. In contrast, RAGE-I2P explicitly models geometric information, providing an additional cue for identifying reliable matches.

4.3 Results on ASIMLab Dataset

Datasets. We use a well-calibrated Intel RealSense camera to capture the RGB-D images. We placed objects with different appearances in the scenes and changed their positions between the images and point clouds through operations such as moving, adding, and removing objects. In total, data from 7 scenes are collected, comprising color RGB images, depth images, global point clouds, and camera poses. The input point clouds are restricted to a tabletop region of interest, excluding surrounding room structures. The same preprocessing protocol is applied to all evaluated methods. For each scene, we acquired two configuration settings characterized by significant and minor vari-

Table 2: Evaluation Results on ASIMLab Dataset. **Boldfaced** numbers highlight the best and the second best are underlined. † indicates that we reproduce this method based on the published paper.

Model	Book→Bot.-L	Book→Bot.-S	Book→Fru.-L	Book→Fru.-S	Mean
Patch Inlier Ratio (PIR, @0.3)					
MATR Li et al. (2023)	4.3	3.8	4.5	4.1	4.2
LDF-I2P Peng et al. (2025)	3.8	3.3	4.7	4.6	4.1
Diff2I2P Mu et al. (2025)	3.8	3.7	4.5	4.2	4.0
B23D-Net† Cheng et al. (2025)	4.4	3.7	4.1	4.1	4.1
RAGE-I2P (Ours)	**5.5**	**5.1**	**6.1**	**5.2**	**5.5**
Inlier Ratio (IR, @0.1m)					
MATR Li et al. (2023)	17.5	16.1	**25.2**	**23.1**	20.5
LDF-I2P Peng et al. (2025)	18.7	15.6	21.3	17.4	18.3
Diff2I2P Mu et al. (2025)	14.0	12.9	13.4	12.5	13.2
B23D-Net† Cheng et al. (2025)	16.5	18.5	21.4	**22.6**	19.8
RAGE-I2P (Ours)	**19.5**	**18.8**	23.9	22.5	**21.2**
Registration Recall (RR, @15°, 0.3m)					
MATR Li et al. (2023)	17.2	12.3	13.2	21.8	16.1
LDF-I2P Peng et al. (2025)	26.5	26.4	**29.9**	28.7	27.9
Diff2I2P Mu et al. (2025)	30.8	28.3	24.1	31.6	28.7
B23D-Net† Cheng et al. (2025)	9.2	12.1	17.3	30.3	17.2
RAGE-I2P (Ours)	**35.4**	**42.4**	22.6	**36.8**	**34.3**

Table 3: Results of Ablation Studies on Changing RGBDScenesV2 Dataset.

Model	PIR	IR	RR	MRE	MTE
RAGE-I2P *full*	**16.0**	**21.4**	**49.8**	2.114	0.062
RAGE-I2P w/o *coarse-level feature refinement*	15.5	19.8	48.9	2.259	0.062
RAGE-I2P w/o *fine-level feature refinement*	15.6	20.2	48.6	2.122	**0.057**
RAGE-I2P w/o *geometric loss*	15.8	21.0	45.4	**2.095**	0.061
RAGE-I2P *full*	**16.0**	**21.4**	**49.8**	**2.114**	**0.062**
2D3DMATR w/ *DSINE normals*	15.7	20.6	44.2	2.157	**0.062**
RAGE-I2P w/ *MoGe-2 depth normals*	8.9	11.9	17.5	2.894	0.071

ations, denoted as "XXX-L" (large variation) and "XXX-S" (small variation), respectively. More details about ASIMLab dataset are shown in Appendix A.1.

Evaluation Results. On ASIMLab, we train all methods exclusively on the *Scene Book* and evaluate them on the remaining scenes without fine-tuning. This protocol is designed to assess cross-scene generalization under limited training data, reflecting practical scenarios where collecting sufficient training samples for every target environment is often difficult. By comparing small variation with large variation, we observed that substantial scene changes tend to degrade both the PIR and IR metrics, consequently impairing the accuracy of camera pose estimation. However, with the introduction of robust geometric priors, such as the LDF-I2P and RAGE-I2P methods, which enhance the discriminability of cross-modal features by leveraging geometric cues, we achieved consistent improvements in inlier ratios at both coarse and fine levels. Compared with Diff2I2P Mu et al. (2025), our approach yields gains of 8.0 percentage points (pp) in IR and 5.6 pp in RR. Visual comparisons are shown in Fig. 3.

4.4 Ablation Studies

To evaluate each module, we remove it and report the results in Tab. 3.

Effectiveness of coarse-level feature refinement. In changing scenes, object movement and variation make patch-to-superpoint matching based on feature distances more prone to false correspondences, which results in a low PIR value. To address this, we propose coarse-level feature residuals that assign larger weights to match pairs with small feature differences within the transformer, effectively preventing similar patches from being matched to the same superpoint. Both our method and the baseline approaches adopt a coarse-to-fine matching generation strategy. When the coarse-level feature refinement is removed, the PIR value drops, which in turn leads to a decline in the accuracy of IR and RR.

Effectiveness of fine-level feature refinement. In changing environments, the appearance of a local region may remain similar even when its underlying 3D structure has changed due to object movement, removal, or insertion. Consequently, fine-level correspondence estimation based solely on appearance-dominated cross-modal features is more likely to suffer from local matching ambiguity. To alleviate this issue, our fine-level refinement explicitly introduces normal-guided geometric representations into both image and point-cloud feature extraction. When this refinement module is removed, the model loses this complementary geometric cue and becomes more sensitive to appearance similarity in changed regions. As a result, the IR decreases from 21.4% to 20.2%, while the RR drops from 49.8% to 48.6%. These results demonstrate that geometry-aware fine-level refinement improves the discriminability of local cross-modal features and provides more reliable dense correspondences for pose estimation.

Effectiveness of geometric loss. Removing the geometric loss reduces IR from 21.4% to 21.0% and RR from 49.8% to 45.4%. This indicates that geometric supervision benefits pose recovery in a way that is not fully captured by the overall inlier ratio.

Directly predicted normals versus depth-derived normals. The last three rows of Table 3 compare two ways of obtaining image-side geometry. The full model predicts normals directly from the monocular RGB image with DSINE Bae & Davison (2024), whereas the other variants first estimate monocular depth with MoGe-2 Wang et al. (2026), back-project the depth map using the camera intrinsics, and then fit normals to local 3D neighborhoods. The 2D3D-MATR variant with DSINE normals reaches 44.2% RR, compared with 25.8% for the original baseline. The full RAGE-I2P further increases RR to 49.8%. This suggests that both the normal prior and the proposed refinement contribute to registration performance. Within the RAGE-I2P architecture, replacing directly predicted normals with normals derived from MoGe-2 depth reduces RR to 17.5%. This comparison concerns the complete normal acquisition pipelines and does not isolate the errors of depth estimation and normal fitting.

5 Conclusion

RAGE-I2P shows that robust image-to-point-cloud registration in changing environments depends on more than feature similarity. A reliable match should also preserve a stable geometric relationship across modalities. Scene changes can preserve visual similarity while changing the underlying 3D structure. Such matches may appear plausible but cannot support accurate pose estimation. Geometry provides complementary evidence for evaluating correspondence reliability when appearance alone becomes ambiguous. Its role should therefore extend beyond final outlier removal. Geometry should guide both feature learning and correspondence evaluation. More broadly, future registration systems should focus on the reliability of correspondences rather than their quantity. They should jointly reason about appearance and geometric consistency and should also account for correspondence uncertainty. This direction may provide a more reliable foundation for localization in continually changing environments.

AI Use Statement

ChatGPT was used to improve the presentation of this paper, including correcting grammatical errors and revising the text to follow academic writing conventions. It was also used to assist with dataset processing so that the data met our required format and quality standards. However, ChatGPT was not used to design or optimize the proposed method, nor was it used to write the experimental code. All experimental code was based on publicly available repositories and was manually adapted and implemented by the authors. All experiments were conducted by the authors, who also independently collected and recorded the results.

Ethics Statement

All authors have read and agree to abide by the ICLR Code of Ethics. This work and all related conference activities, including submission, reviewing, and discussion, are conducted in accordance with its principles and requirements.

REFERENCES

Pei An, Jiaqi Yang, Muyao Peng, You Yang, Qiong Liu, Xiaolin Wu, and Liangliang Nan. MinCD-PnP: Learning 2D-3D Correspondences with Approximate Blind PnP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26519–26528, 2025a.

Pei An, You Yang, Jiaqi Yang, Muyao Peng, Qiong Liu, and Liangliang Nan. Enhance Image-to-Point-Cloud Registration with Beltrami Flow. *International Journal of Computer Vision*, 133(12): 8589–8616, 2025b.

Pei An, Junfeng Ding, Jiaqi Yang, Yulong Wang, Jie Ma, and Liangliang Nan. HG-I2P: Bridging Modalities for Generalizable Image-to-Point-Cloud Registration via Heterogeneous Graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Gwangbin Bae and Andrew J. Davison. Rethinking Inductive Biases for Surface Normal Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9535–9545, 2024.

Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. PointDSC: Robust Point Cloud Registration Using Deep Spatial Consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15854–15864. IEEE, 2021.

Berta Bescos, José M Fácil, Javier Civera, and José Neira. DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes. *IEEE Robotics and Automation Letters*, 3(4):4076–4083, 2018.

Berta Bescos, Carlos Campos, Juan D Tardós, and José Neira. DynaSLAM II: Tightly-coupled multi-object tracking and SLAM. *IEEE Robotics and Automation Letters*, 6(3):5191–5198, 2021.

Lin Bie, Shouan Pan, Kai Cheng, and Li Han. Build a Cross-Modality Bridge for Image-to-Point Cloud Registration. In *Proceedings of International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024.

Lin Bie, Shouan Pan, Siqi Li, Yining Zhao, and Yue Gao. GraphI2P: Image-to-Point Cloud Registration with Exploring Pattern of Correspondence via Graph Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22161–22171, 2025.

Zhixin Cheng, Jiacheng Deng, Xinjun Li, Baoqun Yin, and Tianzhu Zhang. Bridge 2D-3D: Uncertainty-Aware Hierarchical Registration Network with Domain Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 2491–2499, 2025.

Zhixin Cheng, Bohao Liao, Jiacheng Deng, Xiaotian Yin, Xinjun Li, Yujia Chen, Baoqun Yin, and Tianzhu Zhang. Rethinking 2D-3D Registration: A Novel Network for High-Value Zone Selection and Representation Consistency Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39052–39063, 2026.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012.

Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time RGB-D Camera Relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 173–179, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Zhuolin He, Jing Li, Guanghao Li, Xiaolei Chen, Jiacheng Tang, Siyang Zhang, Zhounan Jin, Feipeng Cai, Bin Li, Jian Pu, et al. DynamicVGGT: Learning Dynamic Point Maps for 4D Scene Reconstruction in Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35670–35679, 2026.

Renlang Huang, Yufan Tang, Jiming Chen, and Liang Li. A Consistency-Aware Spot-Guided Transformer for Versatile and Hierarchical Point Cloud Registration. *Advances in Neural Information Processing Systems*, 37:70230–70258, 2024.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. PointWorld: Scaling 3D World Models for In-The-Wild Robotic Manipulation. *arXiv preprint arXiv:2601.03782*, 2026.

Haochen Jiang, Yueming Xu, Kejie Li, Jianfeng Feng, and Li Zhang. Rodyn-Slam: Robust Dynamic Dense RGB-D Slam With Neural Radiance Fields. *IEEE Robotics and Automation Letters*, 9(9): 7509–7516, 2024.

Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised Feature Learning for 3D Scene Labeling. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3050–3057, 2014.

Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2D3D-MATR: 2D-3D Matching Transformer for Detection-Free Registration between Images and Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14128–14138, 2023.

Moyang Li, Zihan Zhu, Marc Pollefeys, and Daniel Barath. DROID-SLAM in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36498–36508, 2026.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time Dense Scene Reconstruction from Monocular RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16651–16662, 2025.

Juncheng Mu, Chengwei Ren, Weixiang Zhang, Liang Pan, Xiao-Ping Zhang, and Yue Gao. Diff2I2P: Differentiable Image-to-Point Cloud Registration with Diffusion Prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25777–25787, 2025.

Muyao Peng, Pei An, You Yang, and Qiong Liu. LDF-I2P: Learning Discriminative Cross-Modality Features for Image-to-Point Cloud Registration. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Muyao Peng, Shun Zou, Pei An, You Yang, and Qiong Liu. Angle-I2P: Angle-Consistent-Aware Hierarchical Attention for Cross-Modality Outlier Rejection. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. KPConv: Flexible and Deformable Convolution for Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6411–6420, 2019.

Julien Valentin, Angela Dai, Matthias Nießner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to Navigate the Energy Landscape. In *Proceedings of International Conference on 3D Vision (3DV)*, pp. 323–332, 2016.

Bing Wang, Changhao Chen, Zhaopeng Cui, Jie Qin, Chris Xiaoxuan Lu, Zhengdi Yu, Peijun Zhao, Zhen Dong, Fan Zhu, Niki Trigoni, et al. P2-net: Joint Description and Detection of Local Features for Pixel and Point Matching. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15984–15993, 2021.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate Monocular Geometry with Metric Scale and Sharp Details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026.

Zhijie Yan, Shufei Li, Zuoxu Wang, Lixiu Wu, Han Wang, Jun Zhu, Lijiang Chen, and Jihong Liu. Dynamic Open-Vocabulary 3D Scene Graphs for Long-Term Language-Guided Mobile Manipulation. *IEEE Robotics and Automation Letters*, 2025.

Jiaqi Yang, Xiyu Zhang, Peng Wang, Yulan Guo, Kun Sun, Qiao Wu, Shikun Zhang, and Yanning Zhang. MAC: Maximal Cliques for 3D Registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10645–10662, 2024.

Xiyu Zhang, Jiayi Ma, Jianwei Guo, Wei Hu, Zhaoshuai Qi, Fei Hui, Jiaqi Yang, and Yanning Zhang. HyperGCT: A Dynamic Hyper-GNN-Learned Geometric Constraint for 3D Registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24750–24759, 2025a.

Xiyu Zhang, Yanning Zhang, and Jiaqi Yang. MAC++: Going Further with Maximal Cliques for 3D Registration. In *Proceedings of International Conference on 3D Vision (3DV)*, pp. 261–275, 2025b.

A Appendix

A.1 More details about two datasets

ASIMLab Dataset. As shown in Table 4, existing I2P registration datasets are generally collected from the same scene at the same time, where the object categories and relative positions remain unchanged between the image and the point cloud. However, in real-world scenarios (e.g. a robot navigating and searching for objects with an onboard RGB camera based on a prior 3D point cloud map), the objects observed in the point cloud and the camera view often differ. To effectively simulate this situation and provide training data for neural networks, we collect the ASIMLab dataset in laboratory scenes. Examples of point clouds and images from several scenes are shown in Fig. 4.

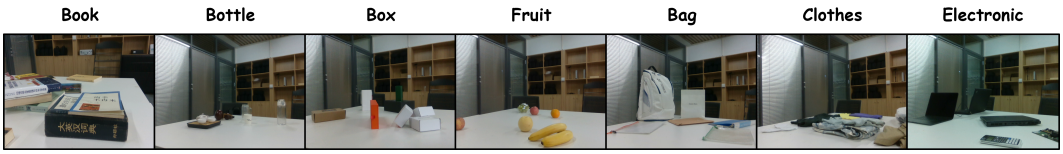


Figure 4: Details of the ASIMLab dataset.

The ASIMLab dataset contains 7 scenes, each with different types of objects, such as books and fruits. For each scene, we collect two levels of changes: large and small. Specifically, we simulate real-world scene changes by moving, removing, and adding objects.

Table 4: Parameters of commonly used I2P registration datasets

Name	Year	Type	Changing
7Scenes Glocker et al. (2013)	ISMAR 2013	Indoor	✗
RGBDScenesV2 Lai et al. (2014)	ICRA 2014	Indoor	✗
12Scenes Valentin et al. (2016)	3DV 2016	Indoor	✗
KITTI Geiger et al. (2012)	CVPR 2012	Outdoor	✗
ScanNet Dai et al. (2017)	CVPR 2017	Indoor	✗
ASIMLab	-	Indoor	✓

RGBDScenesV2 Dataset. We edited the original images using ChatGPT. Specifically, edited images were generated by using prompts such as "*Add ... on the table*", "*Remove ... from the table*", and "*Change the colors of ... and ... to ...*". For the generated images, we estimated their depth maps using MoGe-2 Wang et al. (2026). Details of the processed RGBDScenesV2 dataset are illustrated in Fig. 5.

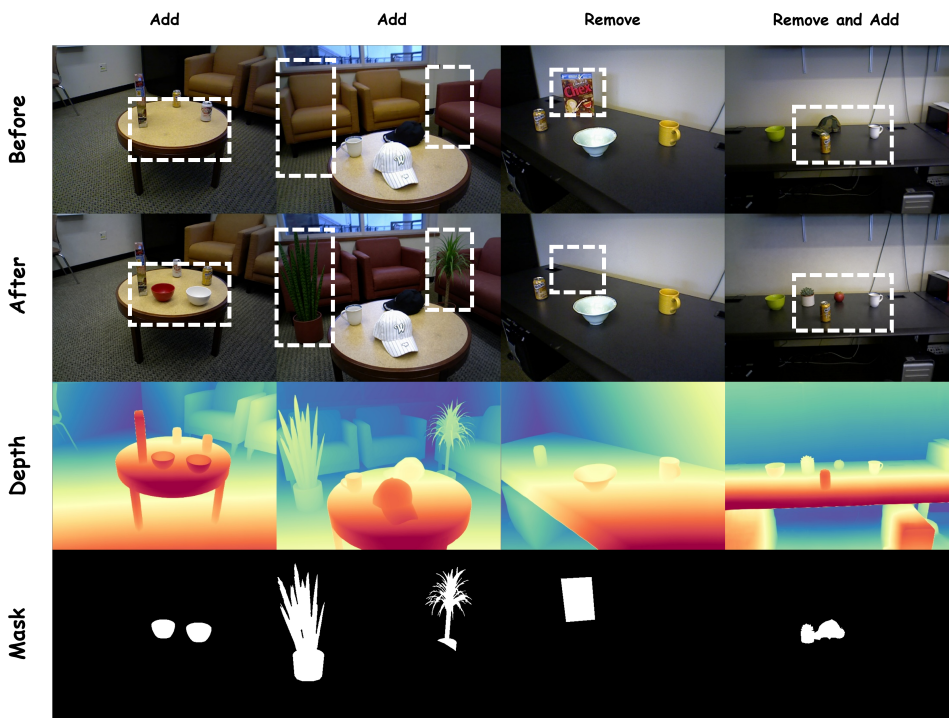


Figure 5: Details of the processed Changing RGBDScenesV2 dataset.

A.2 More implementation details

Training setup. All models are implemented in PyTorch 2.4.1 and trained on one NVIDIA GeForce RTX 4090 GPU. RAGE-I2P is trained for 20 epochs with a batch size of 1. We use the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}. The checkpoint with the highest validation registration recall is used for evaluation.

Training and evaluation protocol. For Changing RGBDScenesV2, all methods are trained on the original RGBDScenesV2 training split and evaluated directly on the edited test split without fine-tuning. The edited training and validation splits are provided for future research but are not used in the experiments reported in the main paper. For ASIMLab, the models are trained on *Scene Book* and tested on other scenes. No test image or point cloud is used during training or model selection.

Normal extraction and geometric embedding. We use the pretrained DSINE Bae & Davison (2024) model to estimate image normals. Its parameters are frozen during registration training. Point-cloud normals are computed using PCA over 50 nearest neighbors.

At the fine level, the image-side geometric embedding function $\phi_{\mathrm{f}}^{\mathcal{I}}$ consists of two 2D convolutional layers followed by a ReLU activation, while the point-cloud-side function $\phi_{\mathrm{f}}^{\mathcal{P}}$ comprises two KPConv layers followed by a ReLU activation. These two branches project the image and point-cloud normals into their respective fine-level geometric feature spaces. At the coarse level, $\phi_{\mathrm{c}}^{\mathcal{I}}$ applies average pooling followed by a sigmoid activation to aggregate the image normals at the patch level, whereas $\phi_{\mathrm{c}}^{\mathcal{P}}$ employs a linear projection followed by a sigmoid activation to encode the point-cloud normals at the superpoint level. To ensure stable matrix inversion, we add $\epsilon\mathbf{I}$ to the summed covariance matrix before inversion, where $\epsilon = 1 \times 10^{-8}$.

A.3 More experiments

Computational efficiency analysis All evaluated methods in Table 5 share 2D3D-MATR as their base architecture and introduce different modules to improve robust correspondence learning. This common backbone makes the comparison reflect the additional cost of each design. RAGE-I2P

achieves the highest RR of 49.8%, improving 2D3D-MATR Li et al. (2023) by 24.0 pp. Meanwhile, its computation increases only from 388.09 to 391.48 GFLOPs (+3.39 GFLOPs), its parameter count from 28.10M to 28.13M (+0.03 M), and its FPS from 7.74 to 7.59 (-0.15 FPS). Compared with Diff2I2P Mu et al. (2025), the strongest competing method, ours further improves RR by 4.0 pp while using 2.50 fewer GFLOPs and 0.03 M fewer parameters. These results show that the performance gain comes from targeted reliability and geometric modeling rather than indiscriminate model expansion, leading to the best accuracy-efficiency trade-off.

The computational statistics in Tab. 5 are reported for the registration network only. Image normal maps are precomputed using the frozen DSINE model and are therefore excluded from the reported parameters, GFLOPs.

Table 5: Computational cost and registration recall.

Model	GFLOPs	Parameters (M)	FPS	RR
MATR Li et al. (2023)	388.09	28.10	7.74	25.8
LDF-I2P Peng et al. (2025)	463.08	29.56	7.32	45.4
Diff2I2P Mu et al. (2025)	394.76	28.16	7.54	45.8
RAGE-I2P (Ours)	391.48	28.13	7.59	**49.8**

More experimental results on the ASIMLab dataset. We report more results on ASIMLab dataset, which are shown in Tab. 6

Table 6: Evaluation Results on ASIMLab Dataset. **Boldfaced** numbers highlight the best

Model	Bag.-L	Bag.-S	Mix.-L	Mix.-S	Mean
Registration Recall (RR, @15°, 0.3m)					
MATR Li et al. (2023)	6.7	10.1	2.9	7.7	6.9
B23D-Net† Cheng et al. (2025)	7.6	11.2	3.5	7.7	7.5
LDF-I2P Peng et al. (2025)	13.8	**18.4**	11.6	27.3	17.8
Diff2I2P Mu et al. (2025)	13.3	3.6	13.9	28.2	14.8
RAGE-I2P (Ours)	**17.1**	18.3	**17.9**	**31.7**	**21.2**

A.4 More details about the metrics

We follow the evaluation protocol of 2D3D-MATR Li et al. (2023) and report all ratios as percentages. Patch Inlier Ratio evaluates coarse-level patch matching, Inlier Ratio evaluates fine-level pixel-point matching, and Registration Recall measures the accuracy of the recovered camera pose.

Patch Inlier Ratio (PIR). Let $\mathcal{C}_{\mathrm{c}}$ denote the set of predicted patch-superpoint correspondences. For each correspondence, we compute its image-side and point-cloud-side overlap ratios under the ground-truth transformation. Following 2D3D-MATR, a pixel-point pair is regarded as overlapping when its 3D distance is below 3.75 cm and its 2D reprojection distance is below 8 pixels. The smaller of the two directional overlap ratios is used as the final overlap score o_{ij} for patch pair $< \hat{\mathbf{p}}_i, \hat{\mathbf{q}}_j >$. PIR is then defined as

$$\mathrm{PIR} = \frac{1}{|\mathcal{C}_{\mathrm{c}}|} \sum_{<\hat{\mathbf{p}}_i, \hat{\mathbf{q}}_j> \in \mathcal{C}_{\mathrm{c}}} \mathbf{1}[o_{ij} > \tau_{\mathrm{PIR}}] \tag{10}$$

where $\mathbf{1}[\cdot]$ is the indicator function. A higher PIR indicates that a larger fraction of the predicted coarse correspondences have sufficient geometric overlap.

Inlier Ratio (IR). Let $\mathcal{C}_{\mathrm{f}}$ denote the set of putative fine-level pixel-point correspondences. For a correspondence $< \mathbf{p}_i, \mathbf{q}_j >$, the point $\mathbf{q}_j$ is transformed into the camera coordinate system using the ground-truth transformation $\mathbf{T}^*_{\mathcal{P}\to\mathcal{I}}$. The pixel $\mathbf{p}_i$ is unprojected to 3D using its depth D_i and the camera intrinsics $\mathcal{K}$. IR measures the fraction of correspondences whose 3D distance is below τ_{IR}:

$$\mathrm{IR} = \frac{1}{|\mathcal{C}_{\mathrm{f}}|} \sum_{<\mathbf{p}_i, \mathbf{q}_j> \in \mathcal{C}_{\mathrm{f}}} \mathbf{1}\left[\left\|\mathbf{T}^*_{\mathcal{P}\to\mathcal{I}}(\mathbf{q}_j) - \pi^{-1}(\mathbf{p}_i, D_i, \mathcal{K})\right\|_2 < \tau_{\mathrm{IR}}\right] \tag{11}$$

where $\pi^{-1}(\cdot)$ denotes depth-based unprojection. A higher IR reflects more accurate fine-level correspondence estimation.

Registration Recall (RR). Let $\hat{\mathbf{T}}_k = [\hat{\mathbf{R}}_k \mid \hat{\mathbf{t}}_k]$ and $\mathbf{T}^*_k = [\mathbf{R}^*_k \mid \mathbf{t}^*_k]$ denote the estimated and ground-truth transformations for the k-th image–point cloud pair, respectively. We first compute the isotropic relative rotation error (RRE) as

$$\mathrm{RRE}_k = \frac{180}{\pi} \arccos\left(\mathrm{clip}\left(\frac{\mathrm{tr}(\hat{\mathbf{R}}_k^{\mathsf{T}}\mathbf{R}^*_k) - 1}{2}, -1, 1\right)\right), \tag{12}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace. The clipping operation avoids numerical errors outside the valid domain of $\arccos$. The relative translation error (RTE) is the Euclidean distance between the estimated and ground-truth translations:

$$\mathrm{RTE}_k = \left\|\hat{\mathbf{t}}_k - \mathbf{t}^*_k\right\|_2. \tag{13}$$

The k-th pair is considered successfully registered only when $\mathrm{RRE}_k < \tau_{\mathrm{RRE}}$ and $\mathrm{RTE}_k < \tau_{\mathrm{RTE}}$. For a test set containing M image–point cloud pairs, RR is defined as

$$\mathrm{RR} = \frac{1}{M} \sum_{k=1}^{M} \mathbf{1}[\mathrm{RRE}_k < \tau_{\mathrm{RRE}} \ \wedge \ \mathrm{RTE}_k < \tau_{\mathrm{RTE}}]. \tag{14}$$

We set $(\tau_{\mathrm{RRE}}, \tau_{\mathrm{RTE}}) = (5^\circ, 0.15\,\mathrm{m})$ for Changing RGBDScenesV2 and $(15^\circ, 0.3\,\mathrm{m})$ for ASIMLab. A higher RR indicates that accurate camera poses are recovered for a larger proportion of the test pairs.