

CrossOver: 3D Scene Cross-Modal Alignment

Supplementary Material

Abstract

In the supplementary material, we provide:

1. Impact of scaling up data (Sec. A)
2. Results on training with all pairwise modalities (Sec. B)
3. Results on same modality scene retrieval (Sec. C)
4. Results on scene retrieval with one modality input to the scene-level encoder (Sec. D)
5. Results on cross-modal coarse visual localization (Sec. E)
6. Additional qualitative results on scene retrieval (Sec. F)
7. Details on the camera view sampling algorithm (Sec. G)
8. Analysis of inference runtime (Sec. H)
9. Further details on the experimental setup (Sec. I)

A. Data Scale-up Improvements

We investigate the impact of scaling up training data by merging different datasets and its effect on CrossOver’s performance, particularly for instance- and scene-level matching recall. Figure 7 demonstrates the advantages of joint training on the ScanNet and 3RScan datasets compared to training on each dataset individually. Please note that 3RScan includes only the \mathcal{I} , \mathcal{P} , and \mathcal{R} modalities. Joint training significantly enhances scene-level recall performance and also improves instance-level recall. These results highlight CrossOver’s ability to effectively leverage diverse data sources, enabling better generalization across varying scenes and object arrangements, ultimately boosting overall performance.

B. All Pairwise Modality Training

As mentioned in Sec. 3.1 of the main paper, training with all pairwise modality combinations, as in prior work [18, 43], directly aligns all modality pairs in a shared embedding space. However, this approach underperforms compared to alignment with a single reference modality, as evidenced by the results in Tabs. B.1 and B.2. Note that ‘Ours’ results are copied from Fig. 4 of the main paper. The key limitation of aligning all modality pairs lies in its added complexity, which dilutes focus and leads to lower scene-level recall metrics. In contrast, intra-modal alignment enhances robustness, particularly in cases of missing modality inputs, by concentrating learning on specific modality relationships. This focused alignment not only improves performance but also facilitates *emergent modality* behavior. Similar insight is also noticed when training the unified encoders with the raw scene data using all pairwise modalities,

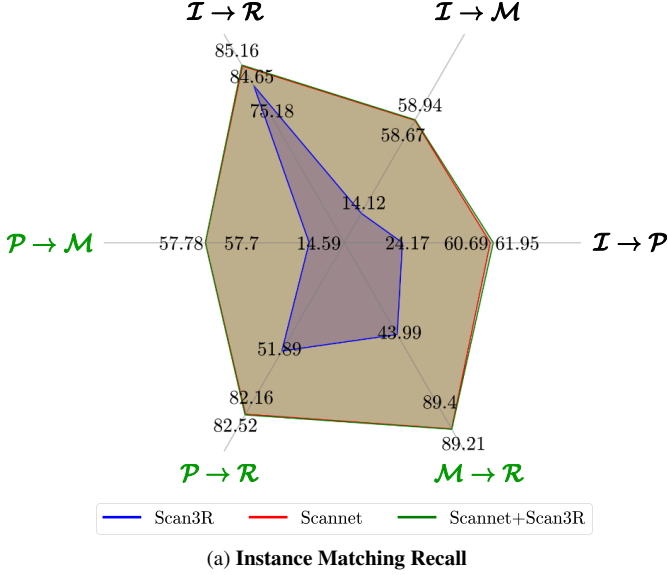
Scene-level Recall \uparrow			
	R@25%	R@50%	R@75%
$\mathcal{I} \rightarrow \mathcal{P}$			
All Pairs	97.12	75.00	15.06
Ours	98.08	76.92	23.40
$\mathcal{I} \rightarrow \mathcal{R}$			
All Pairs	100	98.08	75.95
Ours	99.66	98.28	76.29
$\mathcal{I} \rightarrow \mathcal{M}$			
All Pairs	87.82	63.14	33.97
Ours	86.54	63.46	34.29
$\mathcal{P} \rightarrow \mathcal{R}$			
All Pairs	99.66	97.25	75.26
Ours (<i>emergent</i>)	99.31	96.56	70.10
$\mathcal{P} \rightarrow \mathcal{M}$			
All Pairs	89.42	65.71	35.26
Ours (<i>emergent</i>)	87.50	61.54	30.77
$\mathcal{M} \rightarrow \mathcal{R}$			
All Pairs	100	98.08	83.52
Ours (<i>emergent</i>)	99.23	97.70	83.91

Table B.1. **Scene-level matching results on ScanNet.** ‘All Pairs’ refers to training our instance-level encoder with all pairwise modality combinations. As shown, training on all pairwise combinations does not provide drastically improved performance, as one would expect, even in the modality pairs that are not directly aligned in ‘Ours’ (*emergent*).

ties, namely \mathbf{F}_{1D} , \mathbf{F}_{2D} , \mathbf{F}_{3D} and \mathbf{F}_S . This is shown as ‘All Pairs’ in Tabs. D.1 and D.2.

C. Same-Modality Scene Retrieval

We present results for *same-modality scene retrieval* in Tabs. C.1 and C.2, evaluated on the ScanNet and 3RScan datasets. Metrics include scene category recall, temporal recall, and intra-category recall. Our method is compared to ULIP-2 [43], PointBind [18], and our instance baseline. The instance baseline is not evaluated on the floorplan modality \mathcal{F} due to the lack of floorplan representation at the instance level. Additionally, the scene-level encoder combines *all* instance modalities to generate the \mathcal{F}_S encoding, utilizing ground truth instance segmentation that is consistent across all modalities. This can serve as an upper bound of performance for our method. Results indicate that individual modalities in our method are closely aligned within the embedding space, despite the cross-modal training objective. Consistent with cross-modal results, our



Scene-level Recall ↑			
Trained on	R@25%	R@50%	R@75%
$\mathcal{P} \rightarrow \mathcal{M}$			
3RScan	22.44	8.01	2.24
Scannet	86.54	64.42	33.97
3RScan + Scannet	86.54	63.46	34.29
$\mathcal{P} \rightarrow \mathcal{R}$			
3RScan	84.54	48.80	24.74
Scannet	99.31	96.22	68.38
3RScan + Scannet	99.31	97.25	70.10
$\mathcal{M} \rightarrow \mathcal{R}$			
3RScan	68.97	48.28	22.22
Scannet	99.62	98.47	82.38
3RScan + Scannet	99.23	97.70	83.91

(b) Scene-Level Matching Recall

Figure 7. **Scaled-up training performance on ScanNet.** When training on both ScanNet and 3RScan datasets together, results improve from any individual dataset training. As expected, training on 3RScan and evaluating on ScanNet will have limited performance. Note that the 3RScan includes only the \mathcal{I} , \mathcal{P} , and \mathcal{R} modalities.

Scene-level Recall ↑			
	R@25%	R@50%	R@75%
$\mathcal{I} \rightarrow \mathcal{P}$			
All Pair loss	99.36	77.71	17.20
Ours	99.36	79.62	22.93
$\mathcal{I} \rightarrow \mathcal{R}$			
All Pair Loss	100	97.32	62.42
Ours	100	97.32	67.79
$\mathcal{P} \rightarrow \mathcal{R}$			
All Pair Loss	100	93.96	54.36
Ours (<i>emergent</i>)	100	89.26	50.34

Table B.2. **Scene-level matching results on 3RScan.** ‘All Pairs’ refers to training our instance-level encoder with all pairwise modality combinations. Similar to ScanNet, training on all pairwise combinations does not provide improved performance, as one would expect, even in the modality pairs that are not directly aligned in ‘Ours’ (*emergent*).

method performs better than the instance baseline in most cases, highlighting the importance of scene-level understanding. Moreover, it achieves significantly better or comparable performance to ULIP-2 and PointBind. Notably, our method achieves 100% accuracy on the intra-category recall metric in all modalities, consistently distinguishing the same, *e.g.*, *kitchen* among a database of *kitchens*, with ULIP-2 following closely. ULIP-2 and PointBind show decreased performance on the text referral \mathcal{R} modality, likely due to training on simple object descriptions (*e.g.*, “a point

cloud of a chair”) without scene context. Finally, while our scene-level encoder excels when all modalities are available, challenges arise with missing modalities, as discussed in Sec. D.

D. Uni-modal Scene-Level Encoder Inference

In Sec. 3.3 of the main paper, we highlighted two key advantages of unified dimensionality encoders over the scene-level encoder: (i) they eliminate the need for instance-level modalities or instance information, and (ii) the scene-level encoder struggles when provided with only a single modality (uni-modal) instead of all. To validate the latter, cross-modal scene retrieval results are presented in Tabs. D.1 and D.2. Our method significantly outperforms the uni-modal scene-level encoder in most cases, underscoring the effectiveness and value of the unified modality encoders.

E. Cross-Modal Coarse Visual Localization

We evaluate our method on the task of cross-modal coarse visual localization of a single image against a database of multi-modal reference maps, comparing it to SceneGraphLoc [29] and its baselines LipLoc [36] and Lidar-CLIP [19] on the 3RScan dataset. SceneGraphLoc uses 3D scene graphs during inference as the multi-modal reference maps, incorporating object instance point clouds, their attributes and relationships, and the scene’s structure (for a formal definition of these modalities we point the reader to [29, 34]). For a fair comparison, we use the 2D unified dimensionality encoder to process the input image into an

Method	Scene Category Recall \uparrow			Temporal Recall \uparrow			Intra-Category Recall \uparrow		
	top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-3	top-5
$\mathcal{I} \rightarrow \mathcal{I}$									
ULIP-2 [43]	35.9	44.23	56.73	1.00	2.00	30.00	89.75	96.91	96.91
PointBind [18]	93.59	96.79	98.08	22.00	59.00	99.00	90.21	100	100
Inst. Baseline (Ours)	89.74	95.19	97.12	22.00	58.00	99.00	80.22	98.84	99.87
Ours	91.67	97.76	98.08	11.00	59.00	98.00	100	100	100
$\mathcal{R} \rightarrow \mathcal{R}$									
ULIP-2 [43]	11.34	18.56	24.05	1.00	2.00	4.00	36.63	57.12	66.17
PointBind [18]	11.34	18.56	24.05	1.00	2.00	4.00	36.63	57.12	66.17
Inst. Baseline (Ours)	69.42	91.75	94.16	13.00	51.00	83.00	86.56	97.65	99.20
Ours	76.98	91.75	94.85	14.00	40.00	79.00	100	100	100
$\mathcal{P} \rightarrow \mathcal{P}$									
ULIP-2 [43]	13.14	13.14	23.72	1.00	2.00	3.00	21.52	42.12	57.25
PointBind [18]	17.63	58.33	71.47	7.00	23.00	45.00	59.54	90.36	96.46
Inst. Baseline (Ours)	38.14	75.00	85.38	14.00	42.00	73.00	86.31	97.14	99.81
Ours	86.54	95.51	96.79	19.00	57.00	96.00	100	100	100
$\mathcal{F} \rightarrow \mathcal{F}$									
ULIP-2 [43]	13.78	24.36	41.03	1.00	2.00	5.00	99.27	99.89	99.89
PointBind [18]	63.78	82.37	89.10	7.00	37.00	67.00	100	100	100
Ours	59.95	83.65	90.38	14.00	43.00	74.00	100	100	100
$\mathbf{F}_S \rightarrow \mathbf{F}_S$									
Ours	94.23	97.44	98.08	17.00	57.00	99.00	100	100	100

Table C.1. **Same-Modality Scene Retrieval on ScanNet.** Our method performs on par with or better than baselines in same-modality scene retrieval across most metrics, indicating that individual modalities in our method are closely aligned within the embedding space, despite the cross-modal training objective.

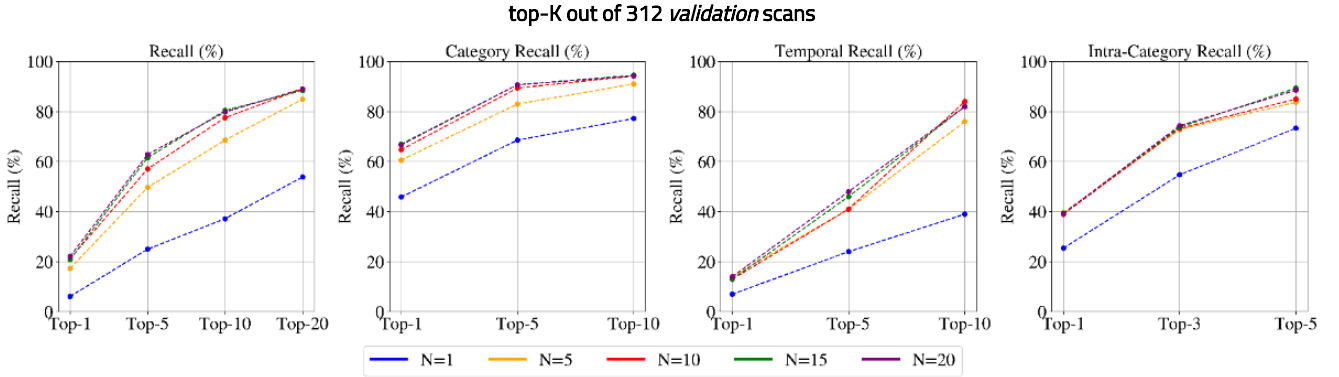


Figure 8. **Cross-Modal $\mathcal{I} \rightarrow \mathcal{P}$ Scene Retrieval on ScanNet.** Plots showcase scene matching recall (Recall), category recall, temporal recall, and intra-category recall for different number of camera views evaluated on several Top- k matches. Note that maximum k differs per recall since the amount of eligible matches depends on the criteria for each recall type: scene similarity, category, temporal changes.

\mathcal{F}_{2D} feature vector, which is then compared to the \mathcal{F}_S feature vectors of all scenes in the database, extracted by our scene-level encoder. As shown in Tab. E.1, despite encoding less information in our multi-modal maps, our method performs competitively with SceneGraphLoc.

F. Qualitative Results

We present additional qualitative results in Figs. 11 and 12 for cross-modal scene retrieval of the pairwise modalities $\mathcal{F} \rightarrow \mathcal{P}$. Fig. 11 illustrates a success case for our method, where the correct scene is retrieved in the first match. In contrast, PointBind [18] and our instance baseline fail to

Method	Temporal Recall \uparrow		
	top-1	top-5	top-10
$\mathcal{I} \rightarrow \mathcal{I}$			
ULIP-2 [43]	2.13	8.51	29.79
PointBind [18]	10.64	51.06	93.62
Inst. Baseline (Ours)	4.26	65.96	100
Ours	17.02	61.70	100
$\mathcal{R} \rightarrow \mathcal{R}$			
ULIP-2 [43]	2.13	6.38	8.51
PointBind [18]	2.13	6.38	8.51
Inst. Baseline (Ours)	19.15	46.81	91.49
Ours	12.77	51.06	87.23
$\mathcal{P} \rightarrow \mathcal{P}$			
ULIP-2 [43]	0.04	4.26	6.38
PointBind [18]	2.13	17.02	36.17
Inst. Baseline (Ours)	6.38	29.79	3.83
Ours	19.15	65.96	97.87
$\mathbf{F}_S \rightarrow \mathbf{F}_S$			
Ours	17.02	59.57	97.87

Table C.2. **Same-Modality Scene Retrieval on 3RScan.** Our method performs on par with or better than baselines in same-modality scene retrieval across most metrics. The lower performance on this dataset is likely due to limited training data availability.

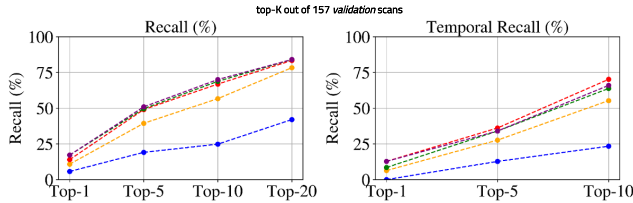


Figure 9. **Cross-Modal $\mathcal{I} \rightarrow \mathcal{P}$ Scene Retrieval on 3RScan.** Plots showcase scene matching recall (Recall) and temporal recall for different number of camera views.

retrieve the correct scene within the first four matches. Notably, our instance baseline does not retrieve any bedrooms. Fig. 12 illustrates a failure case of our method. Despite this, it successfully retrieves all office scenes with a layout similar to the query one. In comparison, the baselines also fail to retrieve the correct scene but instead find matches in bedrooms and meeting rooms. Fig. 13 shows success and failure cases on 3RScan dataset for cross-modal scene retrieval of the pairwise modalities $\mathcal{R} \rightarrow \mathcal{P}$.

G. Camera View Sampling

To sample camera views for the unified 2D encoder (Sec. 3.3 of the main paper), we represent each camera pose as a 7D grid, combining its 3D translation and quaternion-based rotation (4 quaternion + 3 translation components).

Our method selects N camera poses to maximize 3D spatial separation in rotation and translation. Starting with a random pose, we iteratively select the pose farthest from *all* previously chosen ones. This method ensures an even and diverse sampling of camera viewpoints across the scene. We analyze the impact of the number of selected cameras and present results for N values of 1, 5, 10, and 20) in Figs. 8 and 9. The results show that performance stabilizes after $N = 10$, with additional frames providing only slight improvements, indicating full scene coverage is not necessary for training CrossOver. Consequently, we use $N = 10$ for all reported results in our method.

H. Runtime Analysis

Our scene retrieval model consists of 1.5B-parameter. On an NVIDIA 4090 GPU, our model takes $1.01s \pm 0.26s$ for a single modality and 1.98s for all modalities in 1D, 2D and 3D. It can be adapted to lightweight encoders for faster inference in compute-limited scenarios, with potential performance trade-off.

I. Experimental Setup Details

Location Encoding & Instance Spatial Relationships. Given \mathcal{P}_i , we compose features $f_i^{\mathcal{P}}$ and the location l_i (ie, 3D position, length, width and height) to form instance tokens $\hat{f}_i^{\mathcal{P}}$ [48]. A similar process is followed for \mathcal{M}_i . Since we do not use scene graph representations, for instance modality \mathcal{P} , we embed the pairwise spatial relationships between objects in a spatial transformer [20, 48] to encode the scene layout and context. For any two objects \mathcal{O}_i and \mathcal{O}_j present in a scene, we define relationship $s_{ij} = [d_{ij}, \sin(\theta_h), \cos(\theta_h), \sin(\theta_v), \cos(\theta_v)]$, where d_{ij} is the Euclidean distance between the centroids of objects i and j , and θ_h and θ_v are the horizontal and vertical angles of the line connecting the centers of objects i and j . The pairwise spatial feature matrix $S = \{s_{ij}\}$ is used to update the attention weights in the self-attention layers of the transformer.

Evaluation Setup. Our results are reported on the *validation* sets of ScanNet [11] and 3RScan [38], as their corresponding *test* sets lack public annotations or is unavailable. For the experiments in Sec. E, we follow the dataset split provided by SceneGraphLoc [29] to ensure fairness.

Implementation. Inspired by CLIP [32], we adopt an embedding space of size 768, consistent across instance-level, scene-level, and unified training stages. Each model is trained for 300 epochs on an NVIDIA GeForce RTX 4090 Ti GPU using the AdamW optimizer [24] with a learning rate of $1e-3$, and a cosine annealing scheduler with warm restarts. To fine-tune the pre-trained encoders (BLIP [23], DinoV2 [12, 31], and I2PMAE [46]), we employ a 2-layer MLP projection head with dropout and Layer Normaliza-

Method	Scene Matching Recall \uparrow				Scene Category Recall \uparrow			Temporal Recall \uparrow			Intra-Category Recall \uparrow		
	top-1	top-5	top-10	top-20	top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-3	top-5
$\mathcal{I} \rightarrow \mathcal{P}$													
Uni-modal	16.67	51.92	66.67	85.26	36.22	73.72	85.26	14.00	36.00	67.00	49.05	85.15	91.91
All Pairs	16.35	54.17	75.32	91.35	65.71	86.54	93.91	11.00	42.00	77.00	41.51	71.38	84.85
Ours	21.15	57.05	77.56	89.10	64.74	89.42	94.23	13.00	41.00	84.00	38.98	73.28	85.00
$\mathcal{I} \rightarrow \mathcal{R}$													
Uni-modal	2.75	11.00	18.21	29.90	19.59	46.74	62.89	2.00	14.00	19.00	26.12	55.80	66.71
All Pairs	7.56	33.68	50.17	65.64	65.98	83.16	88.66	8.00	28.00	52.00	29.99	58.42	72.64
Ours	8.59	31.27	45.70	59.79	57.39	82.82	87.63	3.00	25.00	51.00	29.04	57.85	70.75
$\mathcal{P} \rightarrow \mathcal{R}$													
Uni-modal	2.06	5.15	12.03	21.31	11.68	39.86	57.04	3.00	6.00	11.00	25.82	53.52	68.08
All Pairs	6.87	24.05	37.46	58.42	56.70	74.57	82.82	3.00	22.00	41.00	31.94	56.12	70.22
Ours	7.22	27.49	44.33	57.73	57.73	79.04	85.57	5.00	20.00	46.00	26.79	56.57	68.63

Table D.1. **Uni-modal & All pair-wise modality training on Scene-Level Encoder Inference on Cross-Modal Scene Retrieval on ScanNet.** ‘All Pairs’ refers to training our unified encoder with all pairwise modality combinations. ‘Uni-modal’ refers to the scene-level encoder with single-modality input. As shown in the Table, our approach outperforms the scene-level encoder and ‘All Pairs’ in most cases. Unlike the unified dimensionality encoders, the scene-level encoder relies on instance-level data, even when operating on a single modality.

Method	Scene Matching Recall \uparrow				Temporal Recall \uparrow		
	top-1	top-5	top-10	top-20	top-1	top-5	top-10
$\mathcal{I} \rightarrow \mathcal{P}$							
Uni-modal	11.46	42.68	64.33	84.71	12.77	31.91	68.09
All Pairs	12.74	43.31	64.97	80.89	8.51	44.68	74.47
Ours	14.01	49.04	66.88	83.44	12.77	36.17	70.21
$\mathcal{I} \rightarrow \mathcal{R}$							
Uni-modal	3.36	14.77	28.86	51.01	8.51	21.28	42.55
All Pairs	8.05	30.20	46.98	60.40	8.51	31.91	59.57
Ours	6.04	26.85	42.28	62.42	2.13	34.04	63.83
$\mathcal{P} \rightarrow \mathcal{R}$							
Uni-modal	1.34	12.08	19.46	36.91	4.26	14.89	29.79
All Pairs	7.38	21.48	37.58	59.73	4.26	29.79	55.32
Ours	6.71	19.46	32.21	51.01	8.51	27.66	51.06

Table D.2. **Uni-modal & All pair-wise modality training on Scene-Level Encoder Inference on Cross-Modal Scene Retrieval on 3RScan.** ‘All Pairs’ refers to training our unified encoder with all pairwise modality combinations. ‘Uni-modal’ refers to the scene-level encoder with single-modality input. As shown in the Table, our approach outperforms the scene-level encoder in all but one case. Unlike the unified dimensionality encoders, the scene-level encoder relies on instance-level data, even when operating with a single modality.

tion [17, 29]. The τ parameter in the contrastive loss formulation is treated as a learnable parameter. Consistent with practices in [20], we pre-train object-level and scene-level encoders and freeze them during unified dimensionality encoder training.

Method	Static Scenario					
	R out of 10 \uparrow			R out of 50 \uparrow		
	top-1	top-5	top-10	top-1	top-5	top-10
LidarCLIP [19]	16.30	41.40	60.60	4.70	11.00	16.30
LipLoc [36]	14.00	35.80	57.90	2.00	8.60	15.20
SceneGraphLoc [29]	53.60	81.90	92.80	30.20	50.20	61.20
Ours	46.00	77.97	90.58	18.69	39.16	51.62

Table E.1. **Cross-Modal Coarse Visual Localization on 3RScan.** Given a single image of a scene, the goal is to retrieve the corresponding scene from a database of multi-modal maps. We evaluate following the experimental setup in SceneGraphLoc [29] and compare our method to it and its baselines. Despite encoding less information in our multi-modal maps, our method performs competitively with SceneGraphLoc.

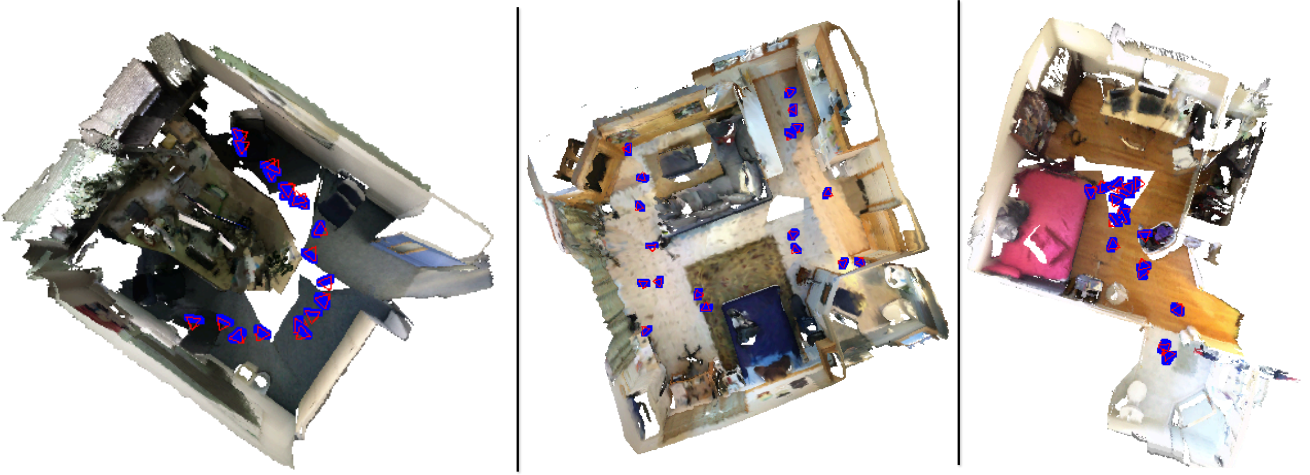


Figure 10. **Camera View Sampling Visualisation on ScaNnet dataset.** Here, we visualise the $N = 20$ selected views (in purple projected onto the ground truth scene mesh) using our camera sampling method. Although, the selected cameras may not cover the entire scene, they are spatially well-separated.

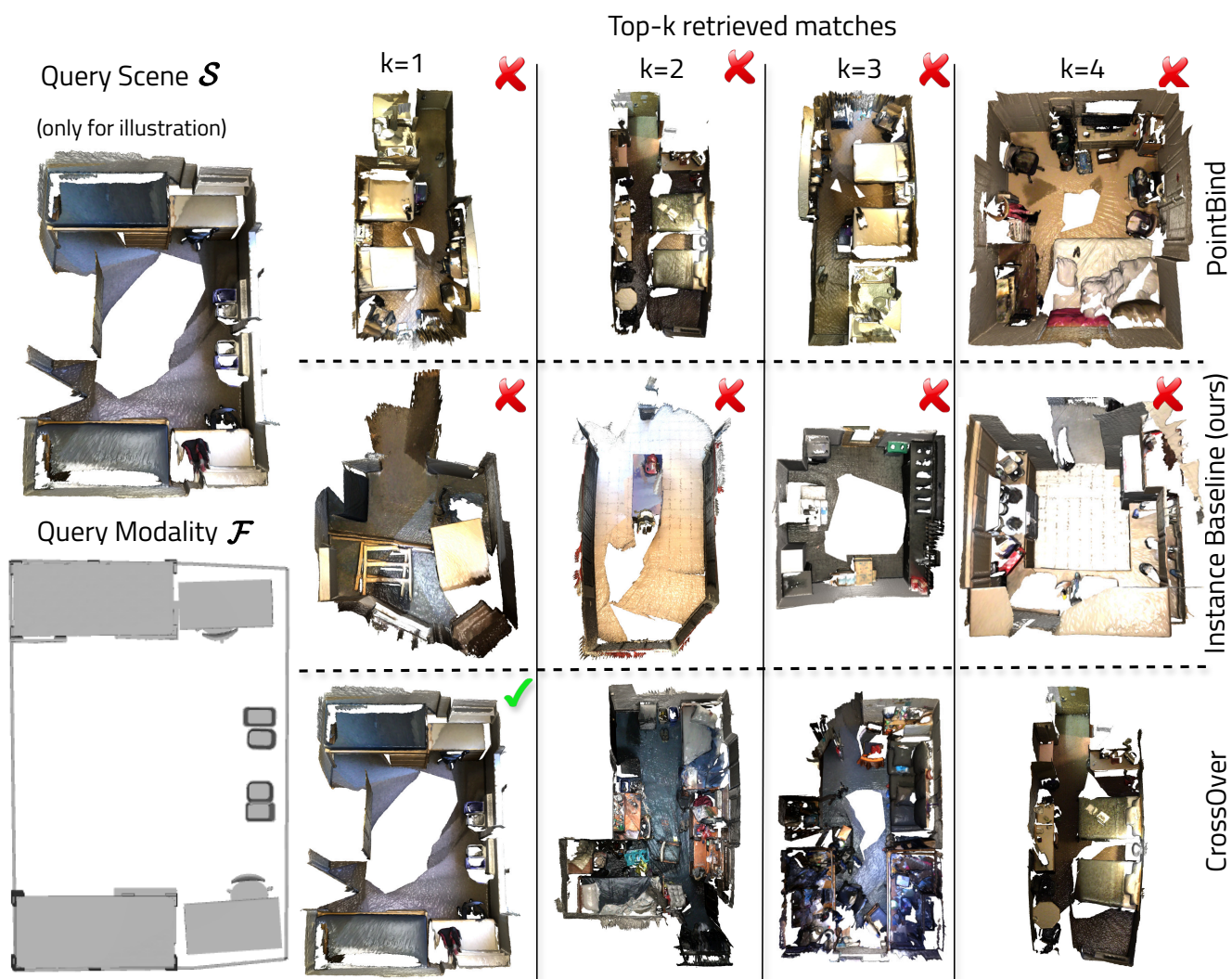


Figure 11. **Cross-Modal Scene Retrieval Success Qualitative Results on ScanNet.** Given a scene in query modality \mathcal{F} , we aim to retrieve the same scene in target modality \mathcal{P} . While PointBind and the Instance Baseline do not retrieve the correct scene within the top-4 matches, CrossOver identifies it as the top-1 match.

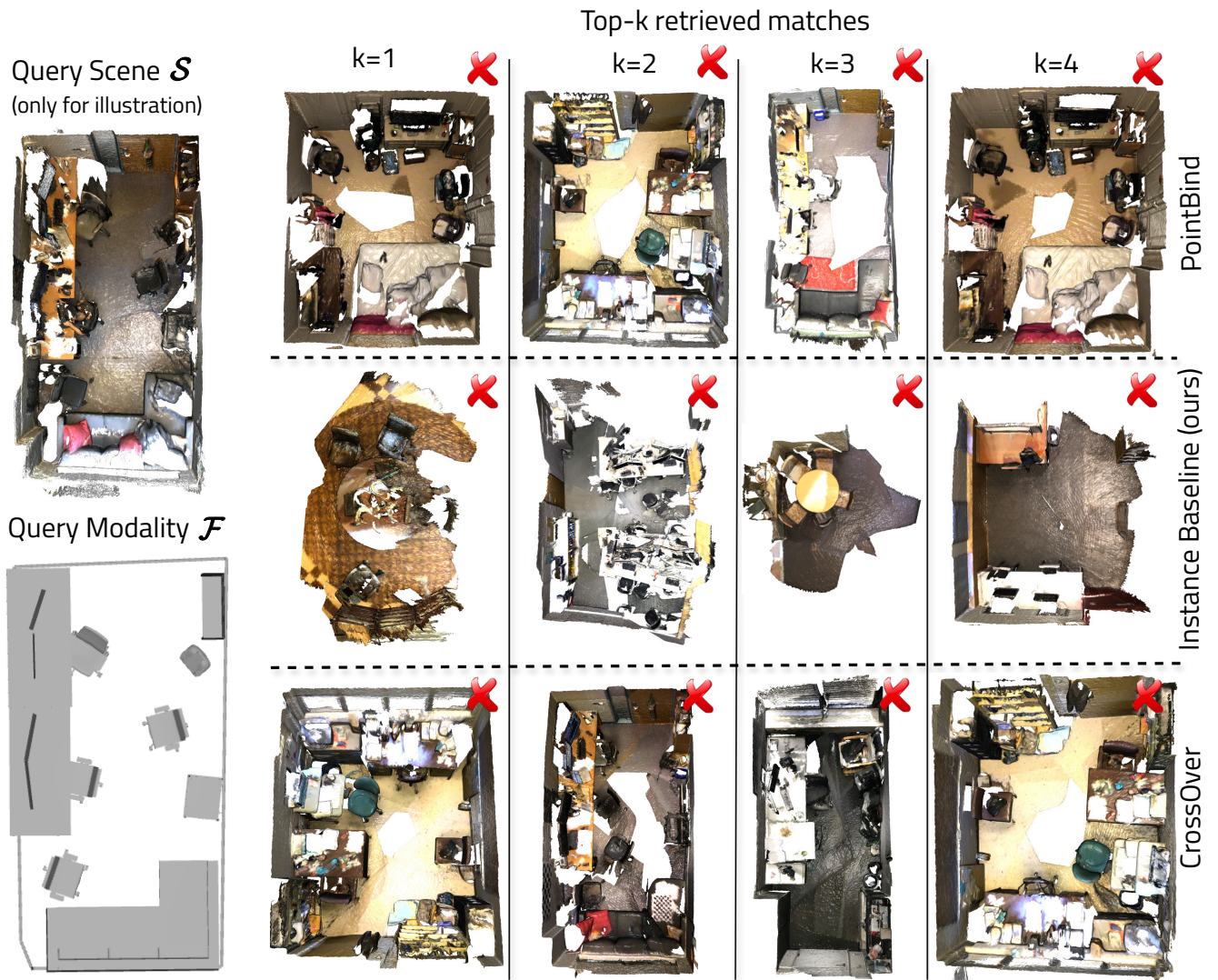


Figure 12. **Cross-Modal Scene Retrieval Failure Qualitative Results on ScanNet.** Given a scene in query modality \mathcal{F} , we aim to retrieve the same scene in target modality \mathcal{P} . While the baselines also fail to retrieve the same scene, CrossOver ($k = 2$) and PointBind ($k = 3$) retrieve a temporal scan as match.

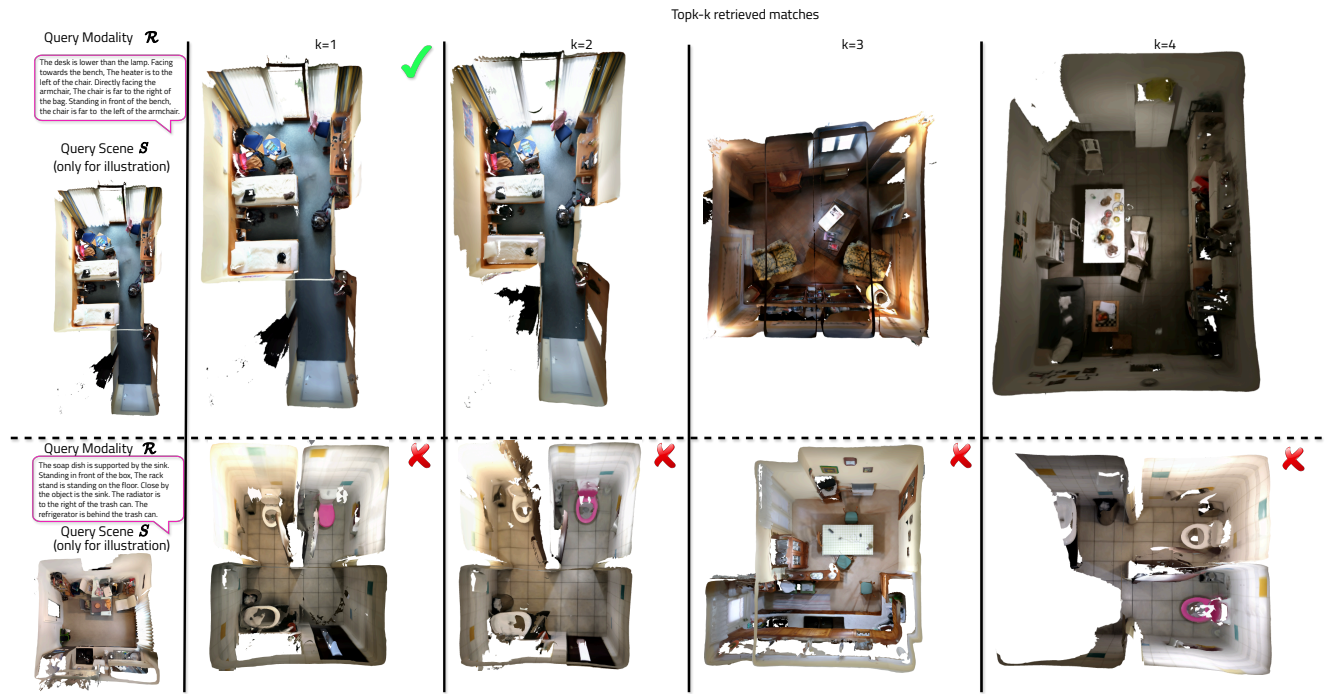


Figure 13. **Cross-Modal Scene Retrieval Qualitative Results on 3RScan. Top row - Success, Bottom row - Failure.** Given a scene in query modality \mathcal{R} , we aim to retrieve the same scene in target modality \mathcal{P} . Temporal scenes cluster naturally in the embedding space. However, query referrals may retrieve scans with similar objects across different scenes, especially when not discriminative enough (bottom).