Stanford University

Scalable Cross-Modal 3D Scene Understanding

Sayan Deb Sarkar

Microsoft Spatial AI Lab

July 18, 2025

Who Am I?

PhD student since 2024.09

- Advisor: Prof. Iro Armeni
- Gradient Spaces Research Group, part of Stanford Vision and Learning Lab (SVL)

Stanford University

Gradient Spaces

Computer Science MSc 2022.09 – 2024.08

- Advisor: Prof. Marc Pollefeys
- Computer Vision And Geometry Group (CVG)

CVG Computer Vision and Geometry Lab

ETH zürich

At Industry

- Internships at Microsoft Spatial AI Lab & Qualcomm XR
- Computer Vision Engineer at Mercedes Benz R & D

sayands.github.io

Microsoft

Qualcomm

TU Graz

CVPR Nashville JUNE 11-15, 2025

Highlight

CrossOver

3D Scene Cross-Modal Alignment

Sayan Deb Sarkar

Ondrej Miksik

Marc Pollefeys

Dániel Béla Baráth

Iro Armeni

ETH zürich

Microsoft

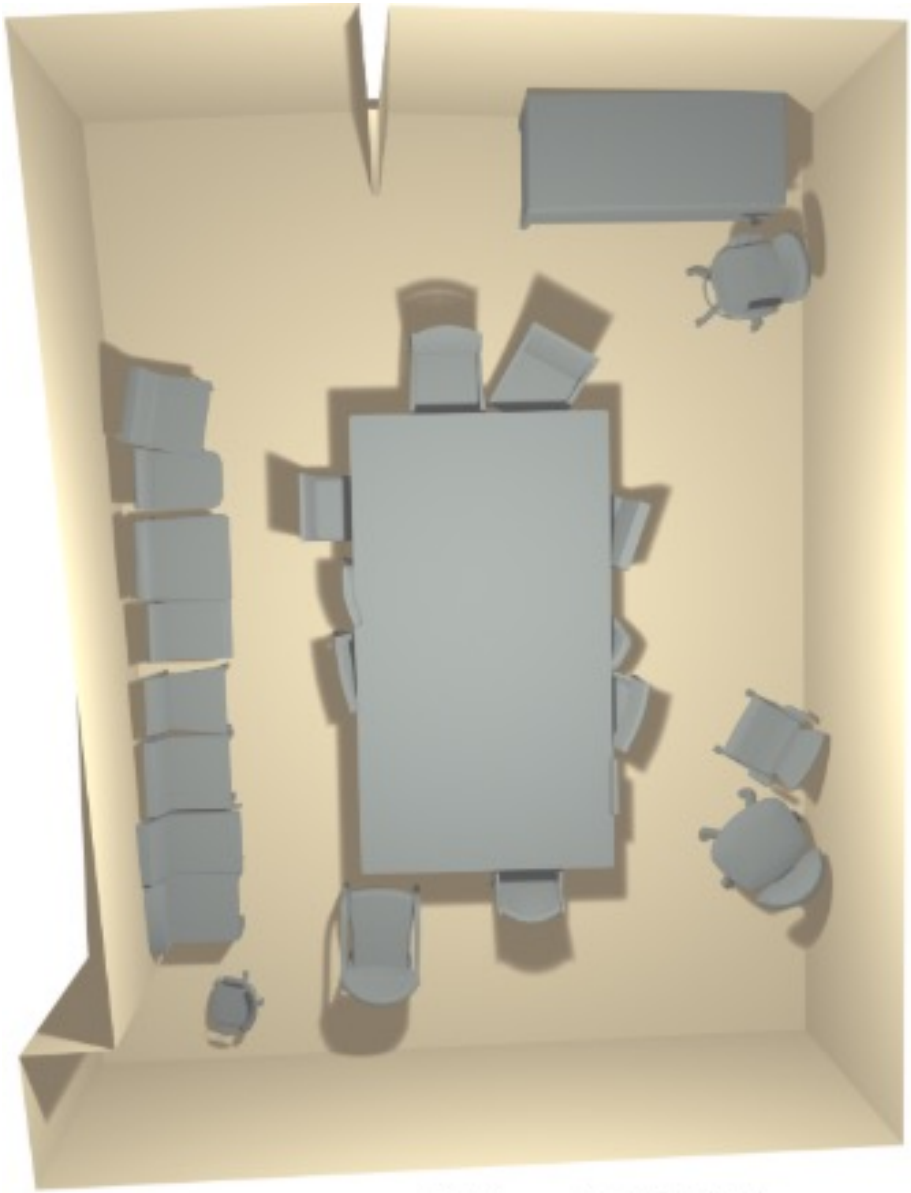
CVG Computer Vision and Geometry Lab

Stanford University

Gradient Spaces

HUN REN | SZTAKI

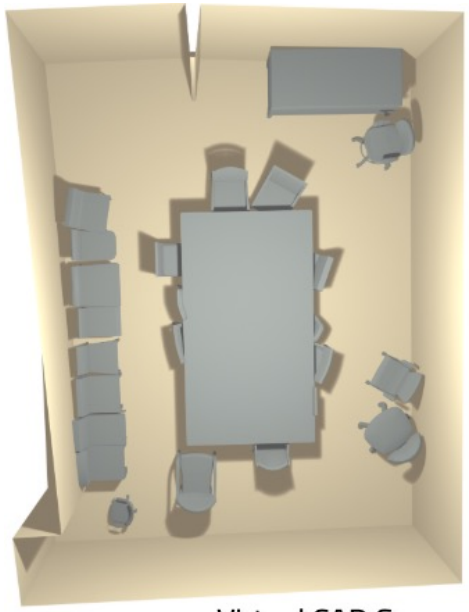
Application



Virtual CAD Scene

Scene Database

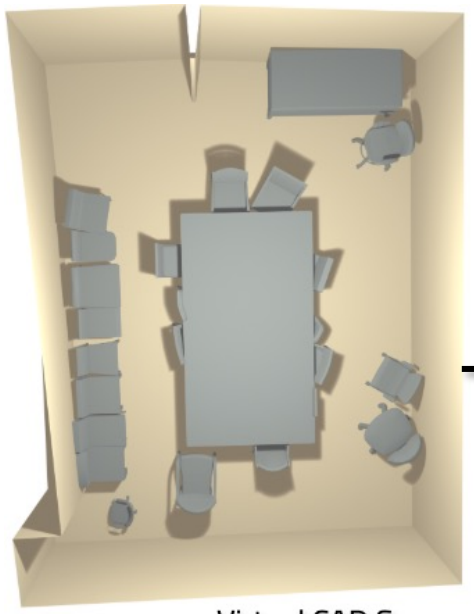
Application



Virtual CAD Scene

Scene Database

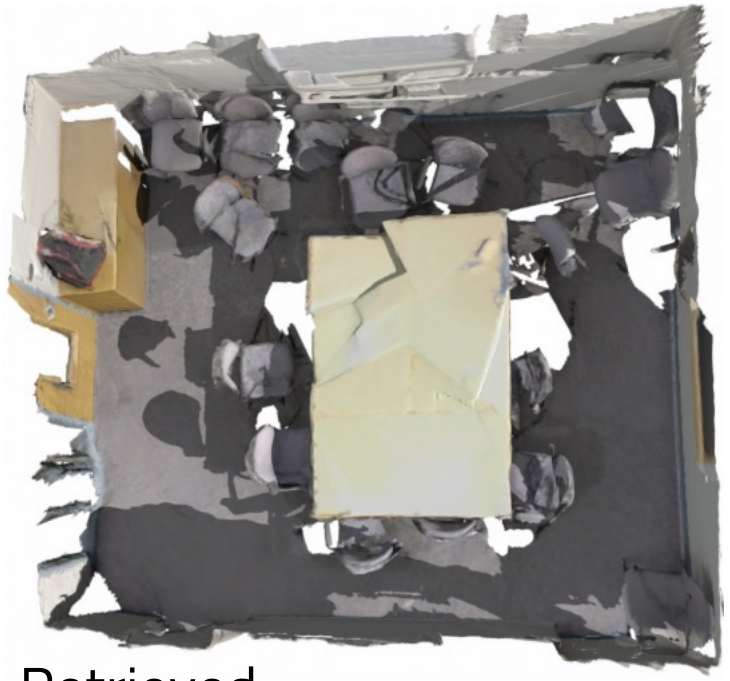
Application



Virtual CAD Scene

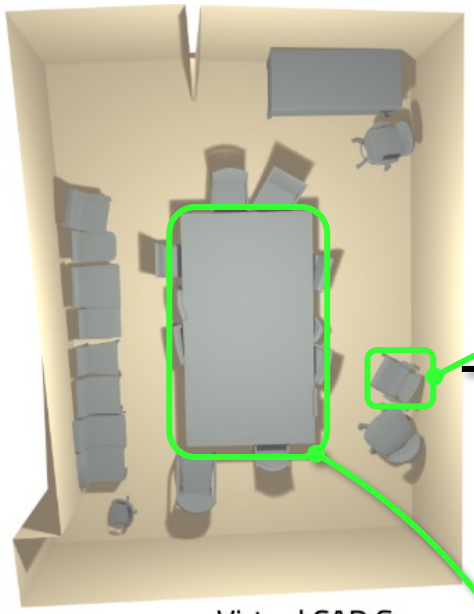
Scene Database

CrossOver



Retrieved

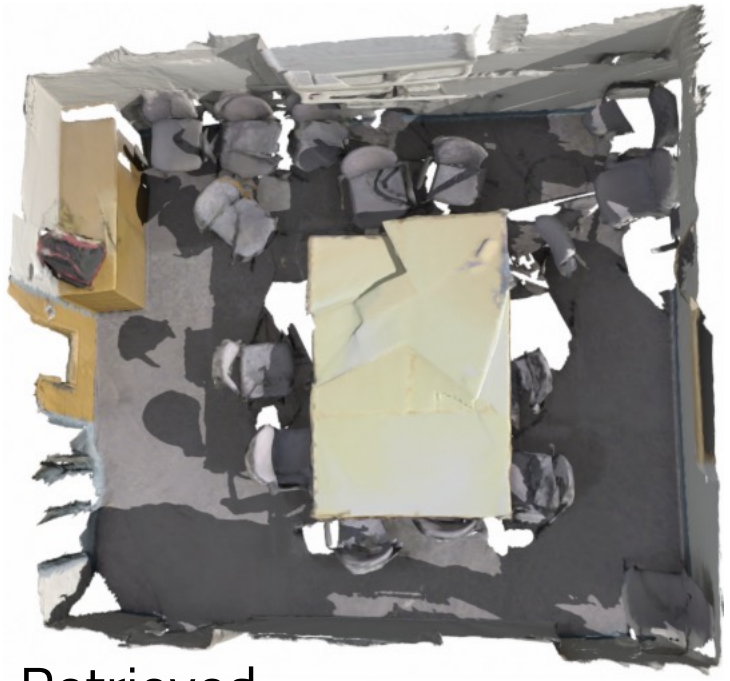
Application



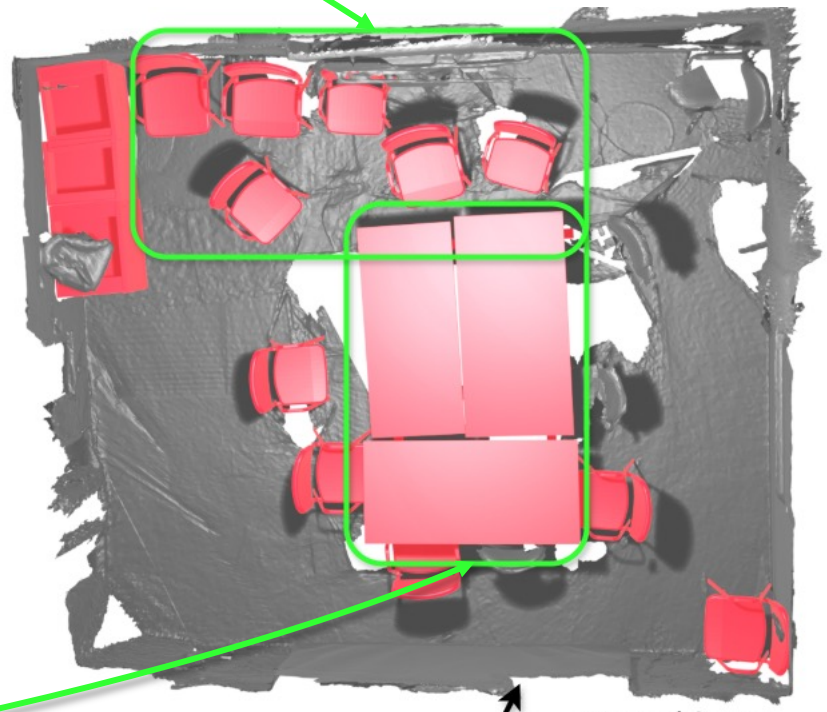
Virtual CAD Scene

CrossOver

Scene Database



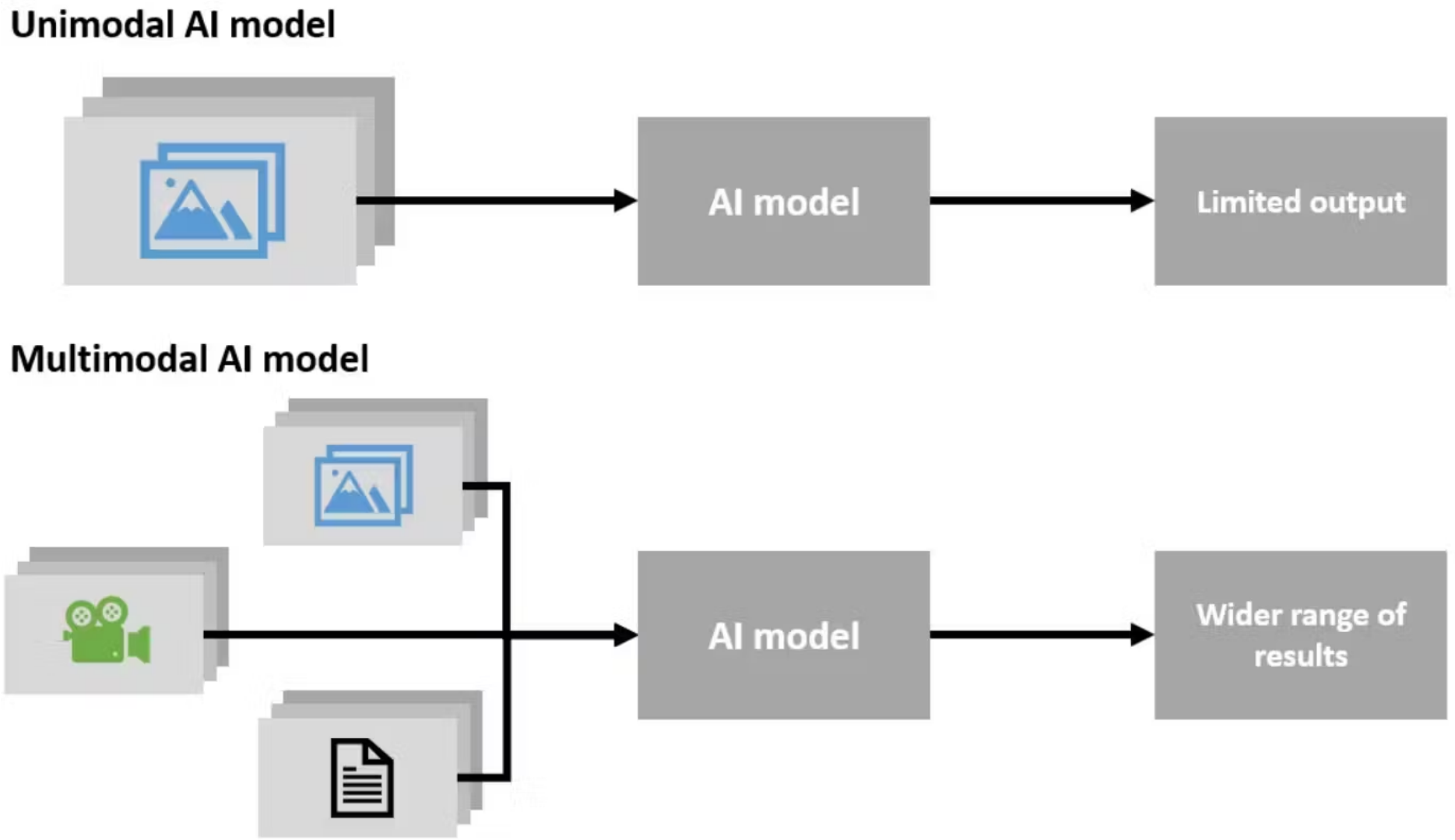
Retrieved



Merged Scene

Matched CAD Models

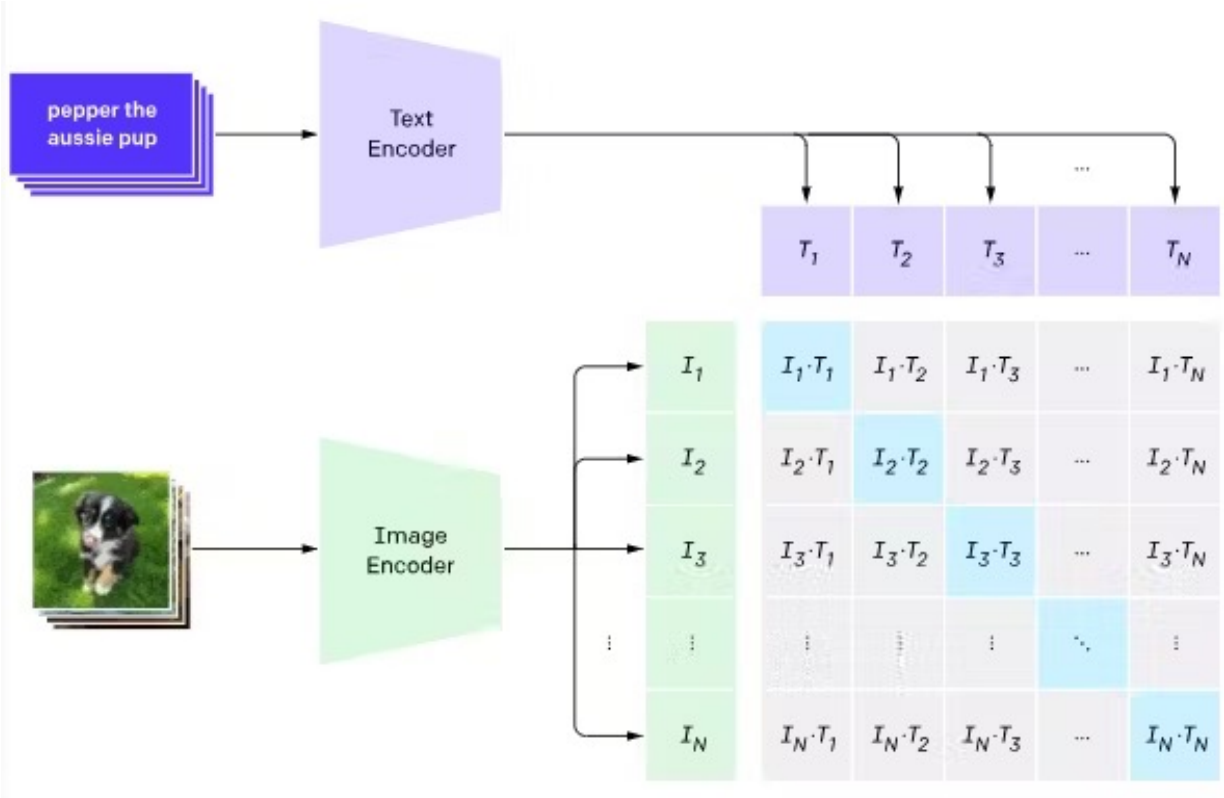
Uni-modal vs Multi-modal



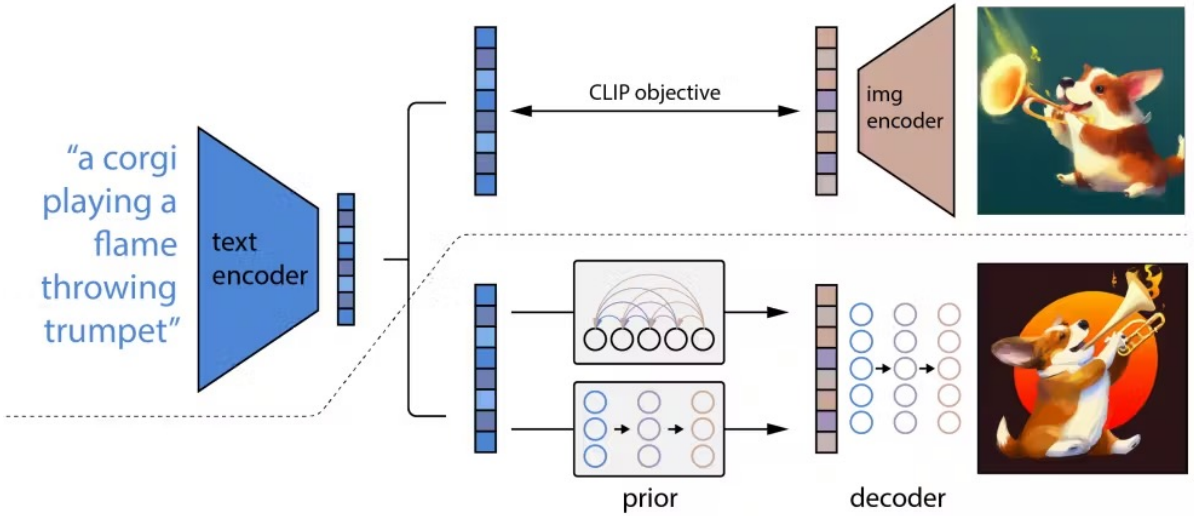
They enable knowledge and experience transfer across modalities.

Image: Link

Multimodal Models Today in 2D



CLIP

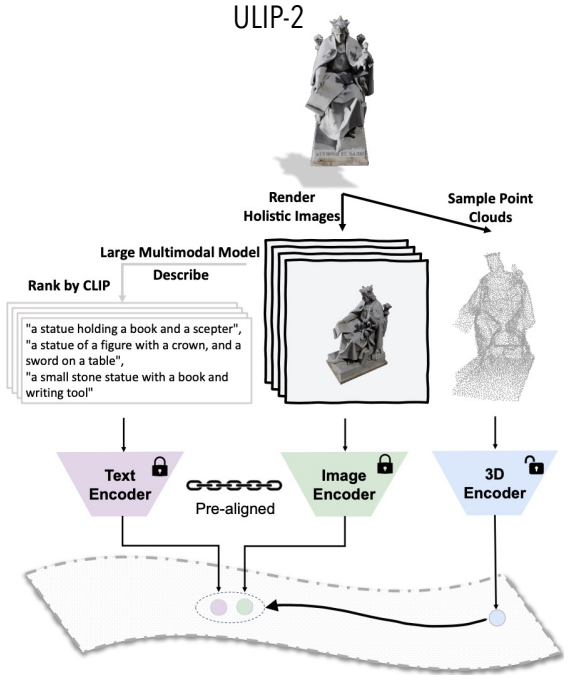


DALL-E

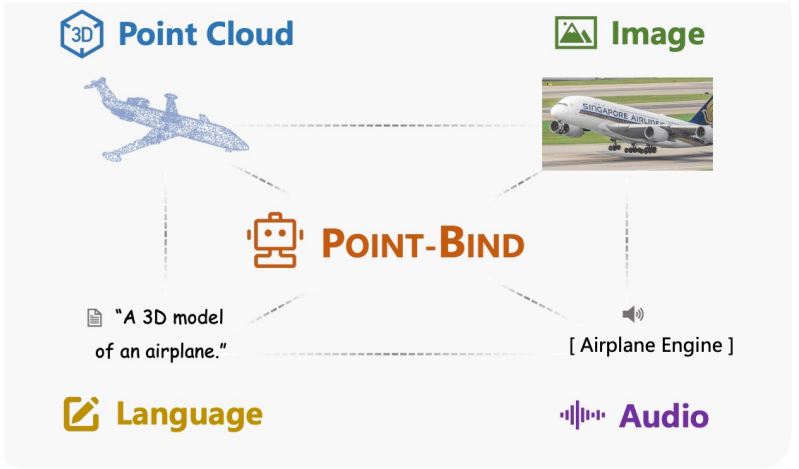
Main focus is on dual-modality, connecting images with text.

Motivation

ULIP-2



PointBind



Point-LLM for 3D Q&A

Any-to-3D Generation

3D Embedding Arithmetic

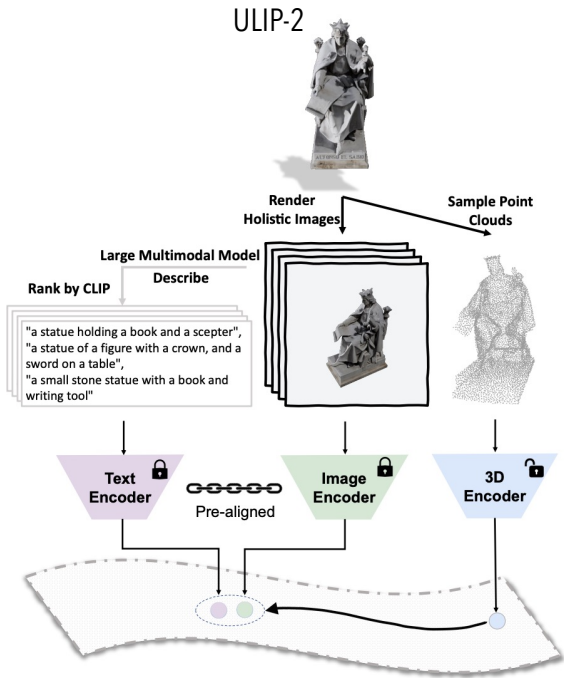
3D Zero-shot Learning

[1] Xue et al, *ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding*, CVPR 2024

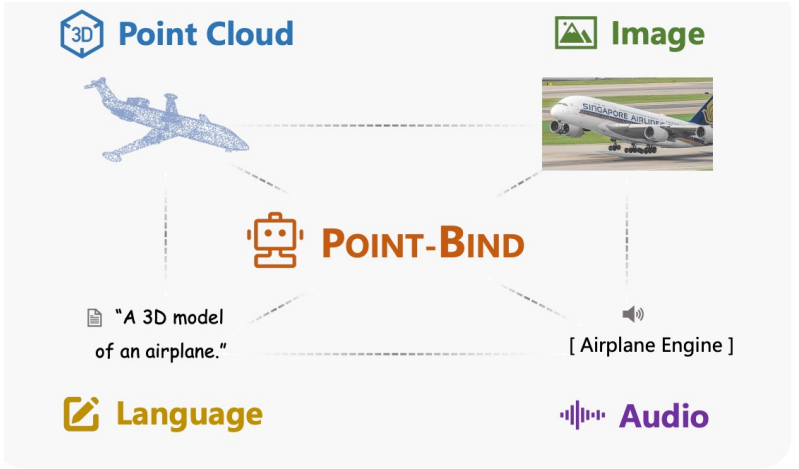
[2] Guo et al, *Point-Bind & Point-LLM: Aligning 3D with Multi-modality*, arXiv 2024

Motivation

ULIP-2



PointBind



Point-LLM for 3D Q&A

Any-to-3D Generation

3D Embedding Arithmetic

3D Zero-shot Learning

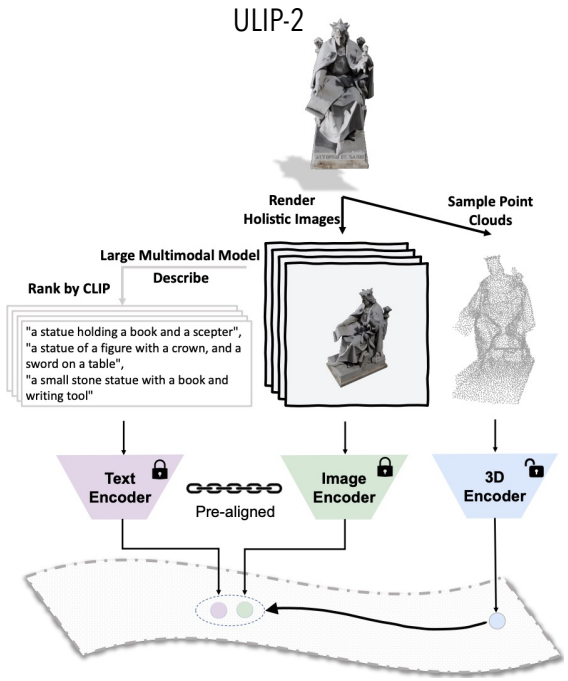
- Perfect Modality Alignments + Complete Object Data

[1] Xue et al, *ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding*, CVPR 2024

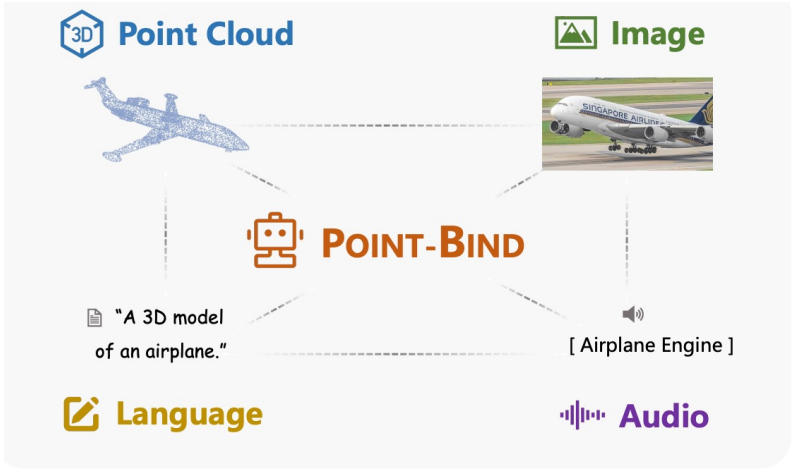
[2] Guo et al, *Point-Bind & Point-LLM: Aligning 3D with Multi-modality*, arXiv 2024

Motivation

ULIP-2



PointBind



Point-LLM for 3D Q&A

Any-to-3D Generation

3D Embedding Arithmetic

3D Zero-shot Learning

- Perfect Modality Alignments + Complete Object Data

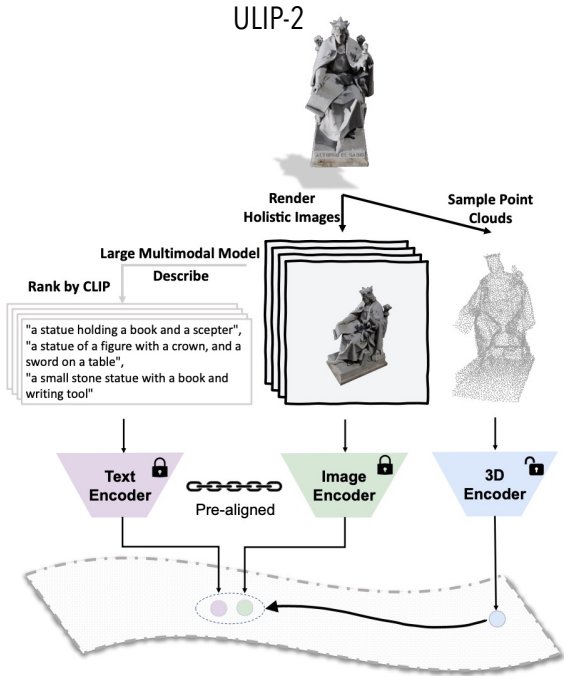
- No Context Within A Scene

[1] Xue et al, *ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding*, CVPR 2024

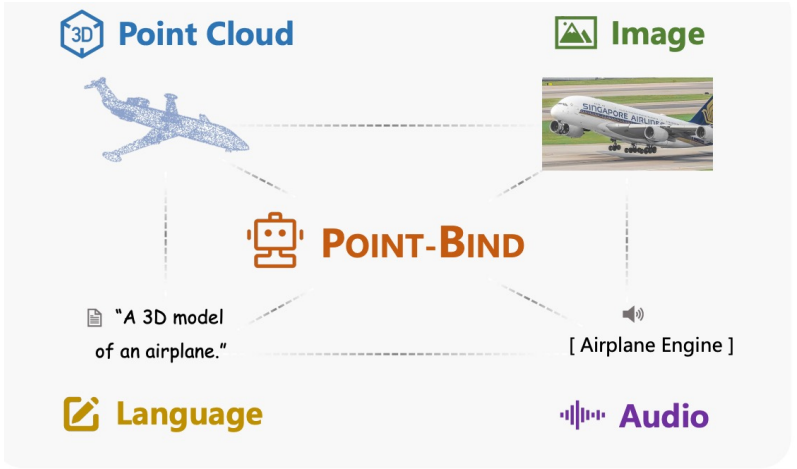
[2] Guo et al, *Point-Bind & Point-LLM: Aligning 3D with Multi-modality*, arXiv 2024

Motivation

ULIP-2



PointBind



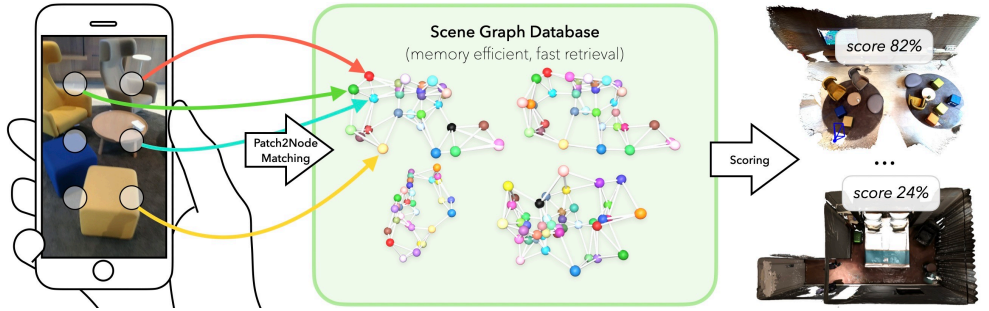
Point-LLM for 3D Q&A

Any-to-3D Generation

3D Embedding Arithmetic

3D Zero-shot Learning

- Perfect Modality Alignments + Complete Object Data
- No Context Within A Scene
- Need For Semantic Annotations/Explicit Scene Graphs for 3D Scene Understanding

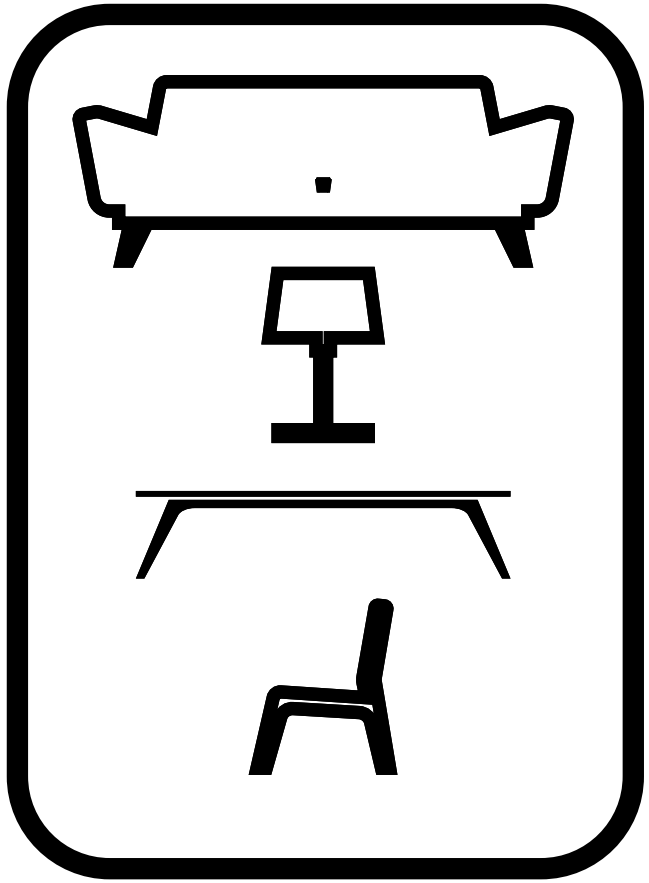


[1] Xue et al, *ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding*, CVPR 2024

[2] Guo et al, *Point-Bind & Point-LLM: Aligning 3D with Multi-modality*, arXiv 2024

[2] Miao et al, *SceneGraphLoc: Cross-Modal Coarse Visual Localization on 3D Scene Graphs*, ECCV 2024

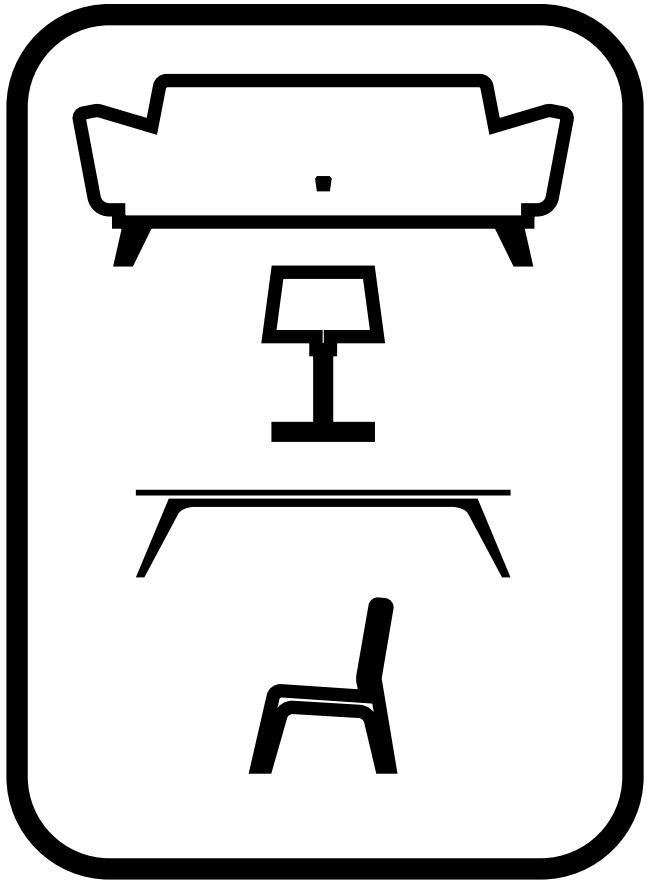
CrossOver



I. Instance-Level Alignment

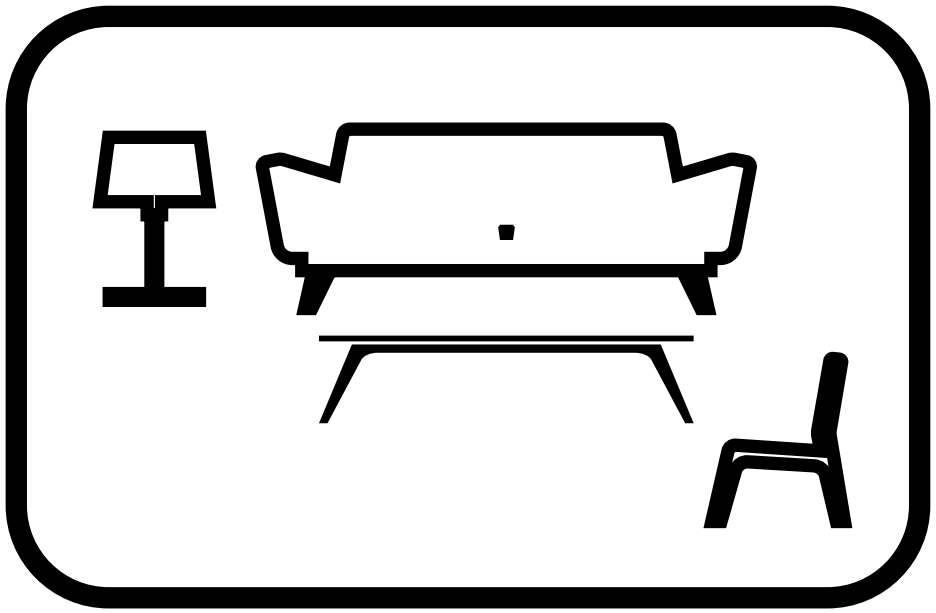
to leverage large pre-trained models

CrossOver



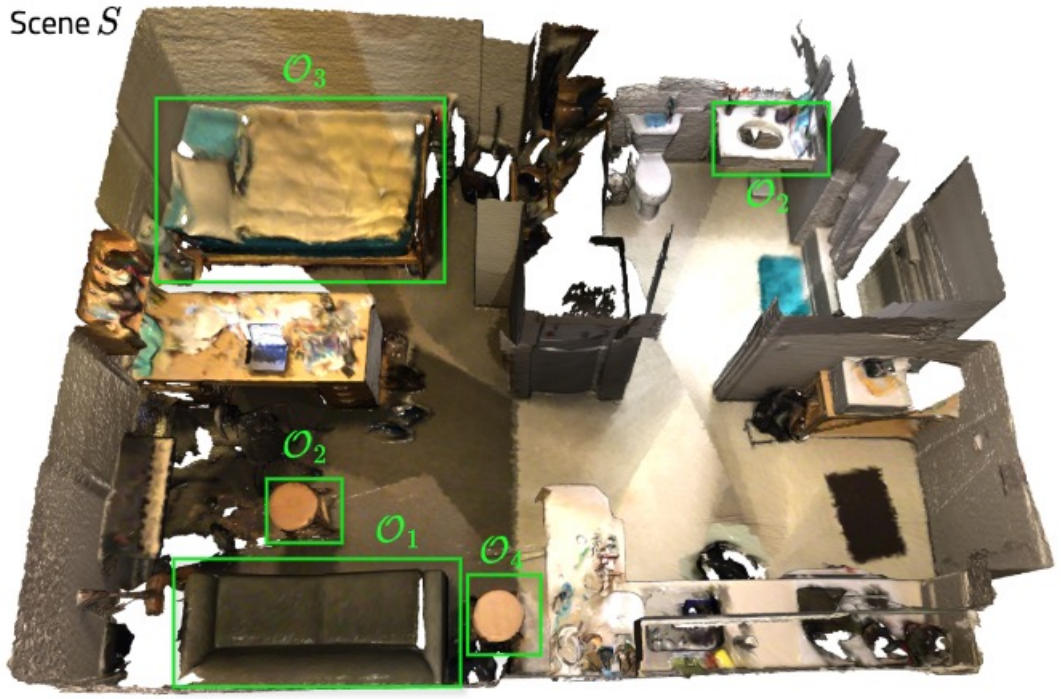
I. Instance-Level Alignment

to leverage large pre-trained models

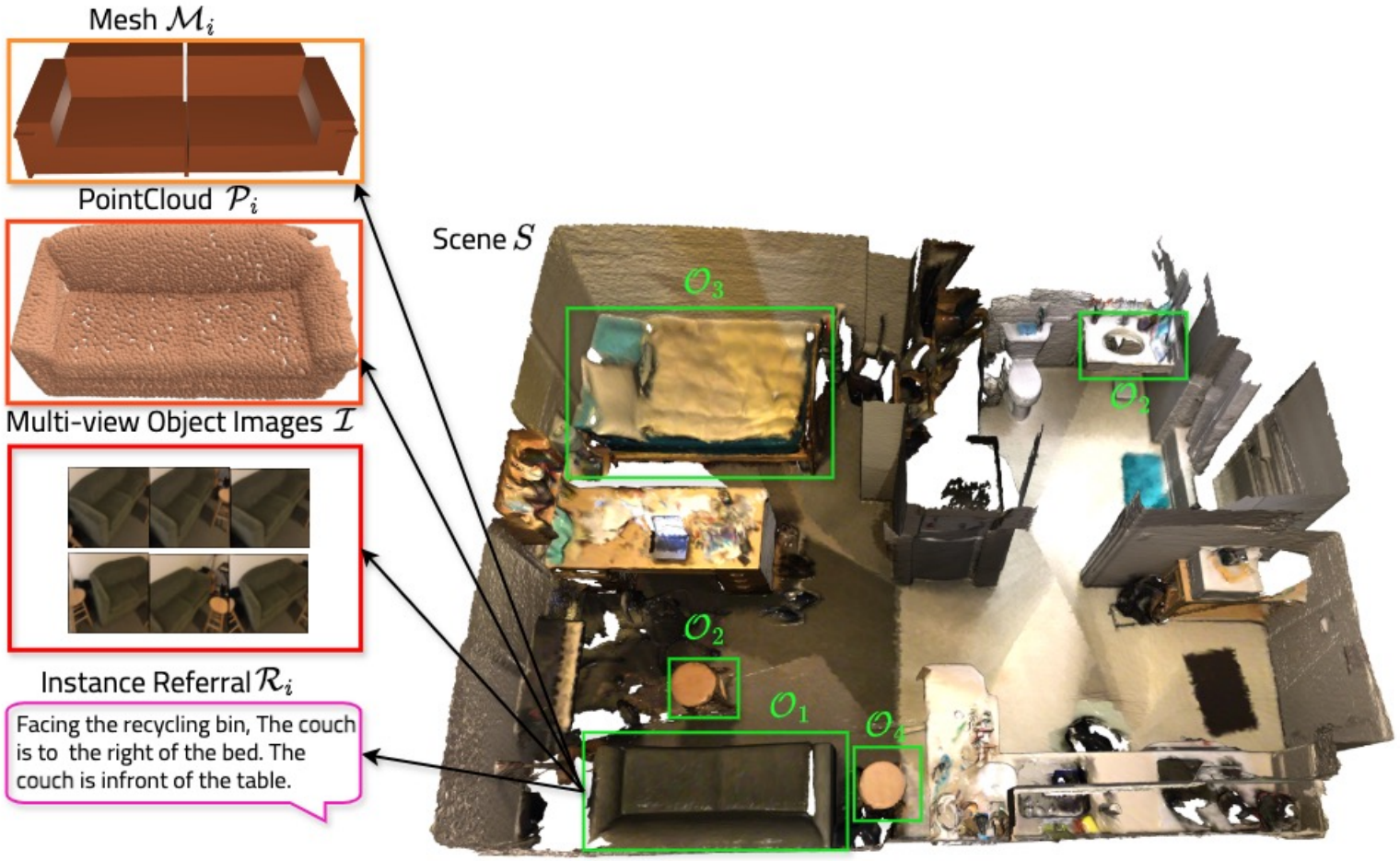


II. Scene-Level Alignment

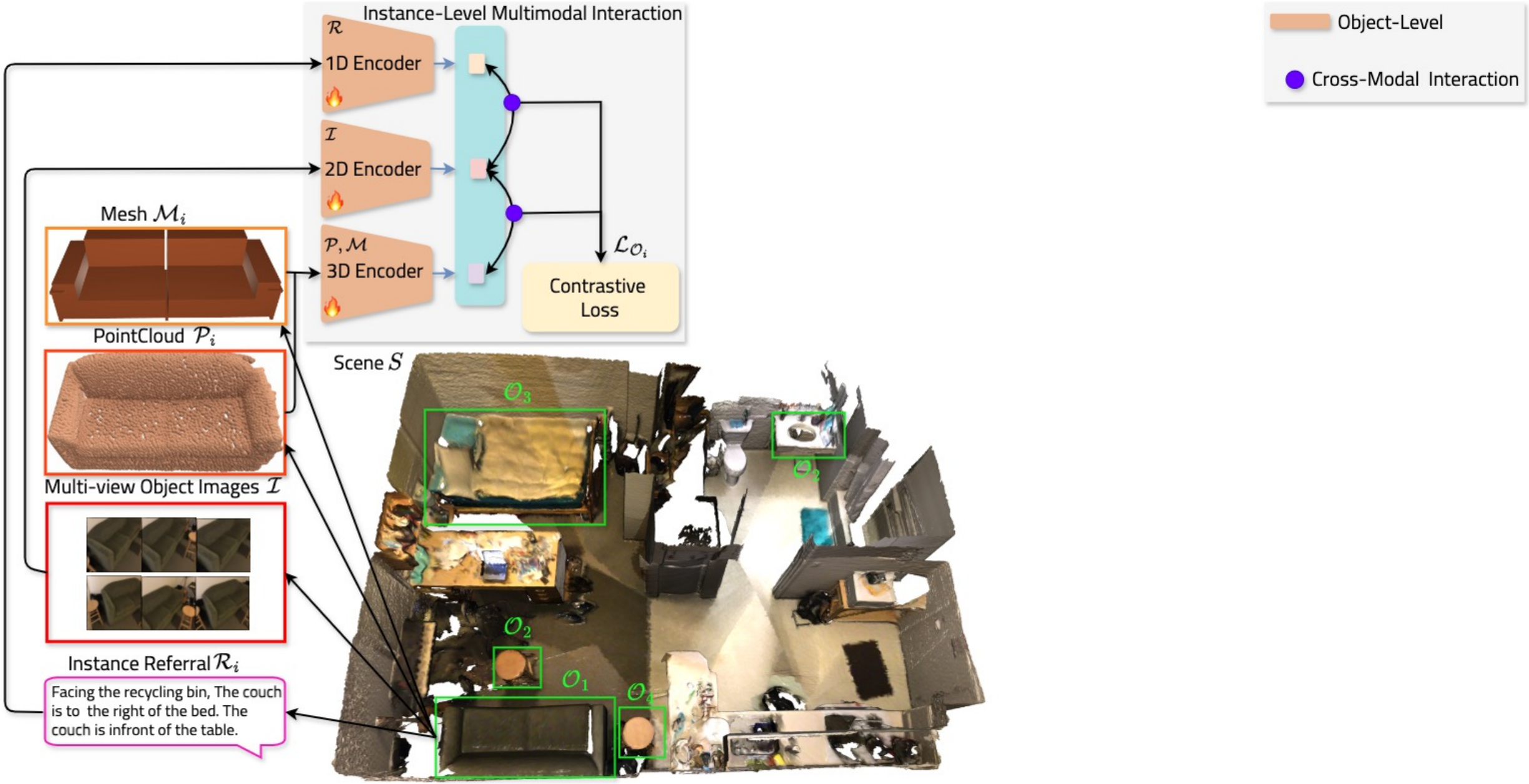
Overview of CrossOver



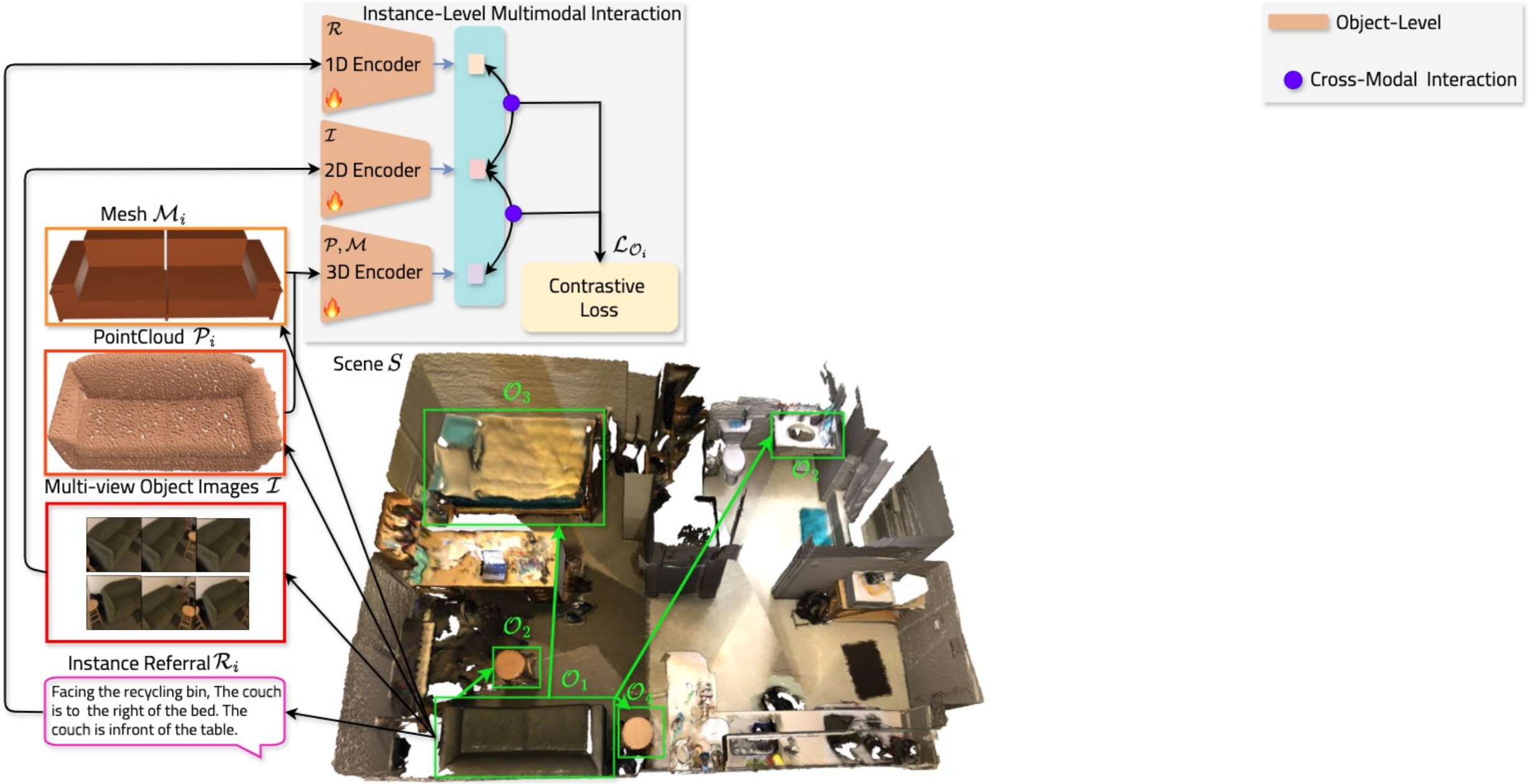
Overview of CrossOver



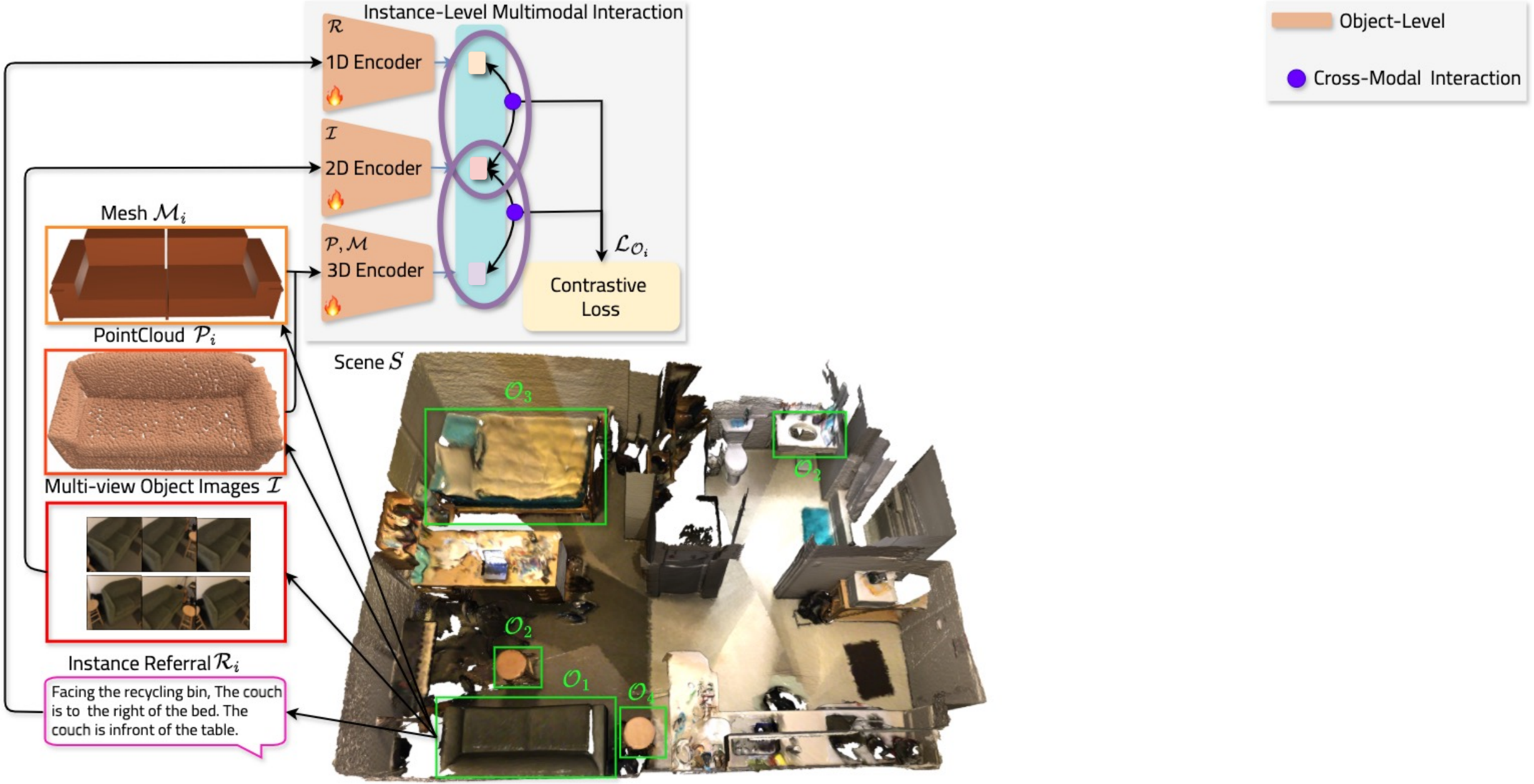
Overview of CrossOver



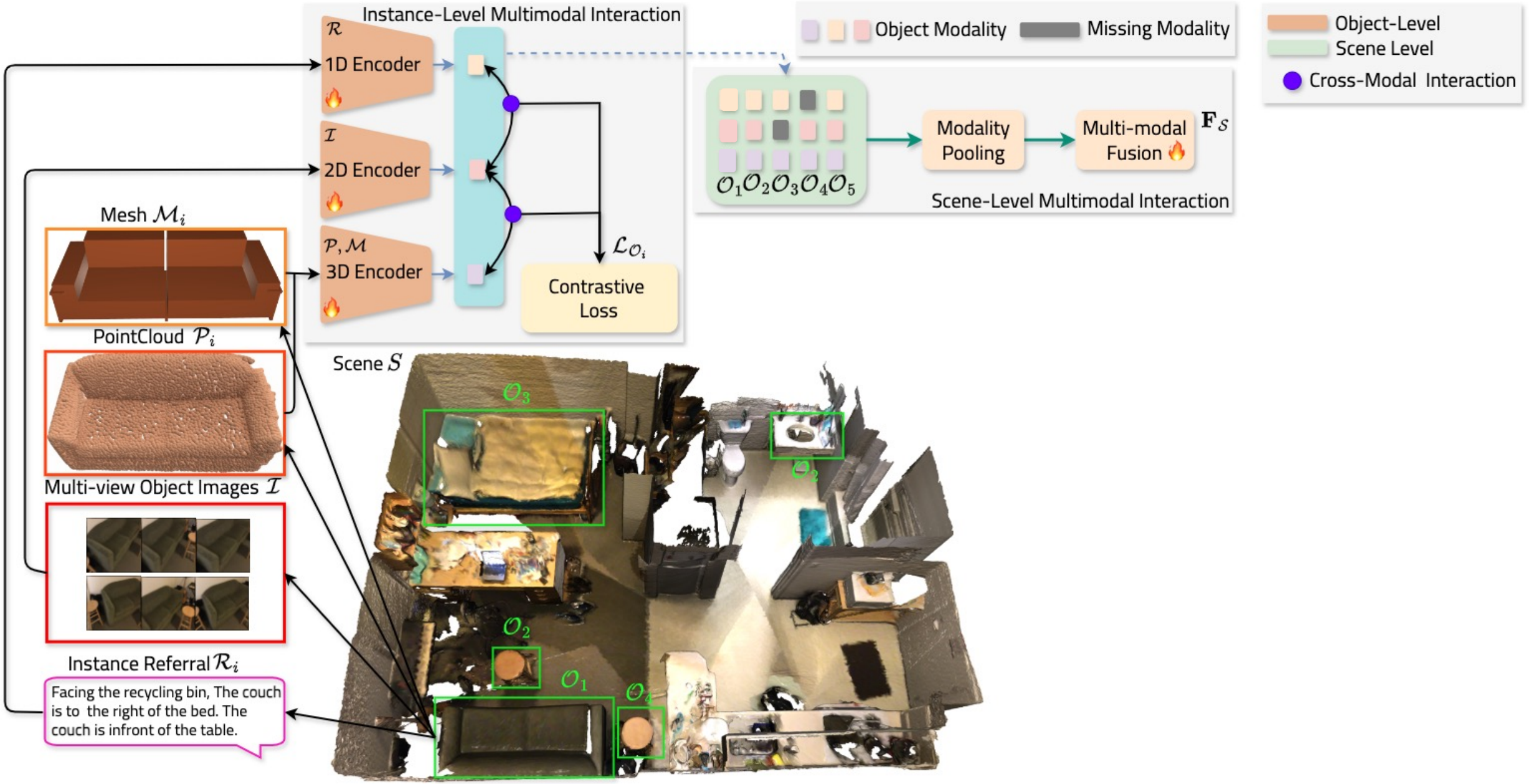
Overview of CrossOver



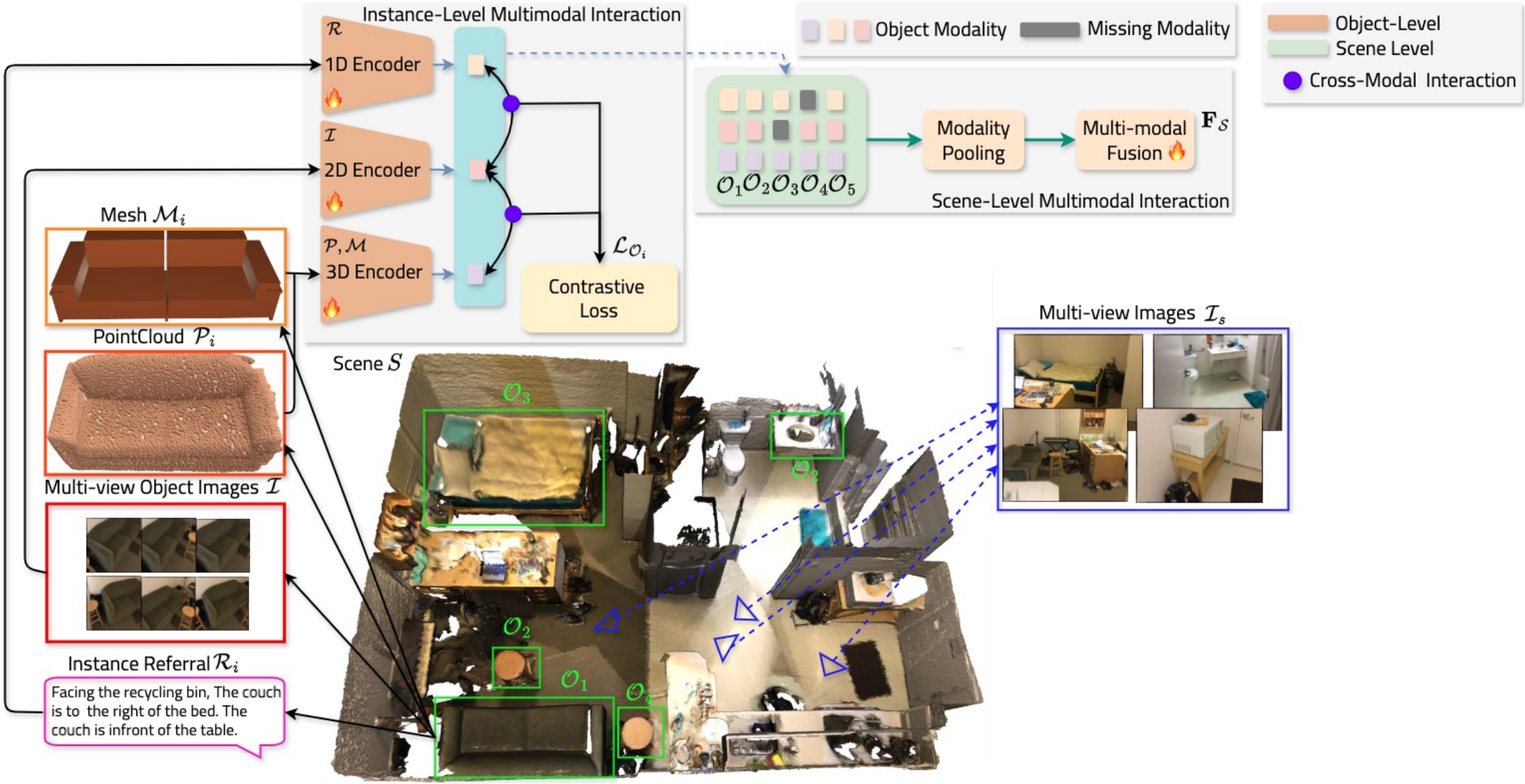
Overview of CrossOver



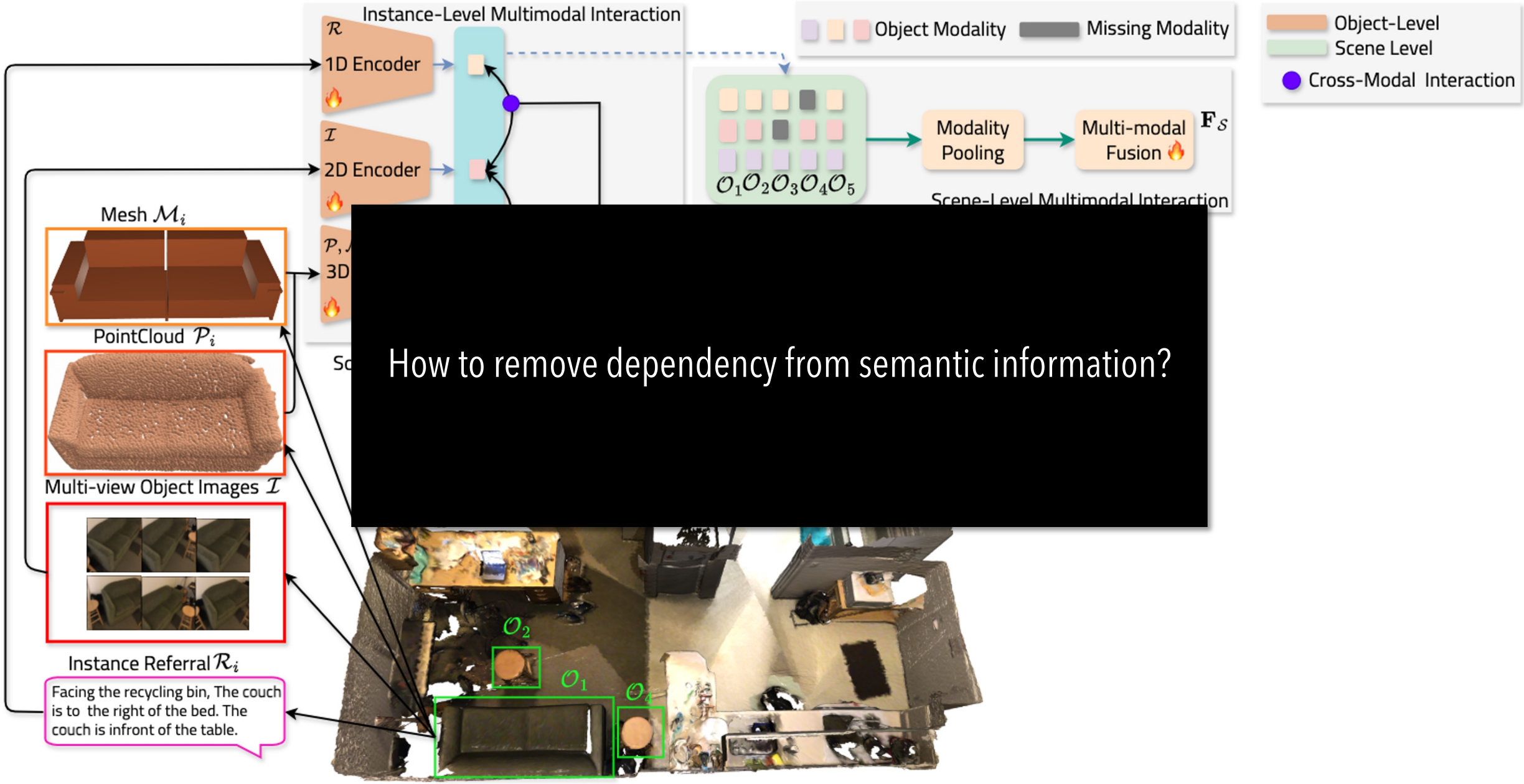
Overview of CrossOver



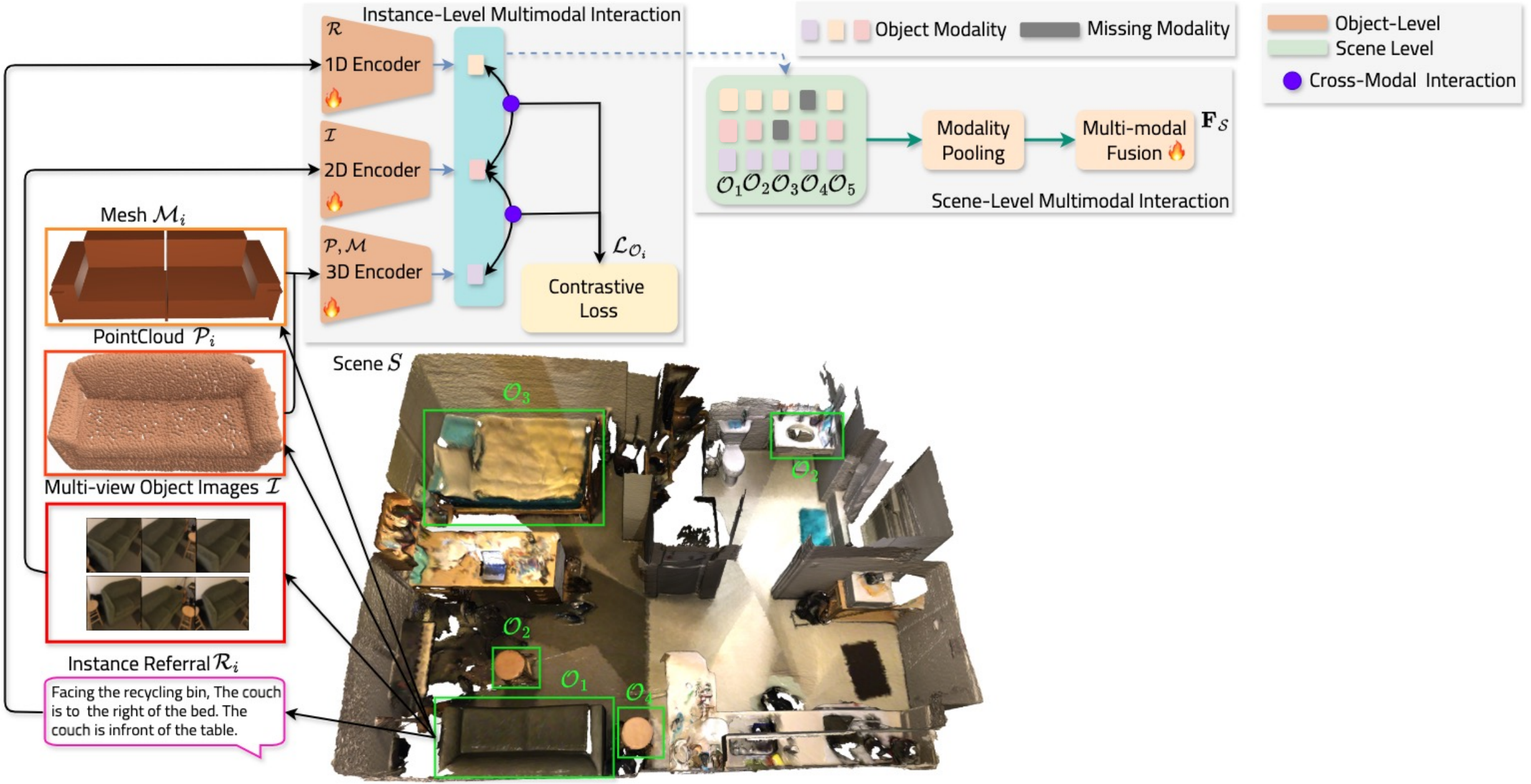
Overview of CrossOver



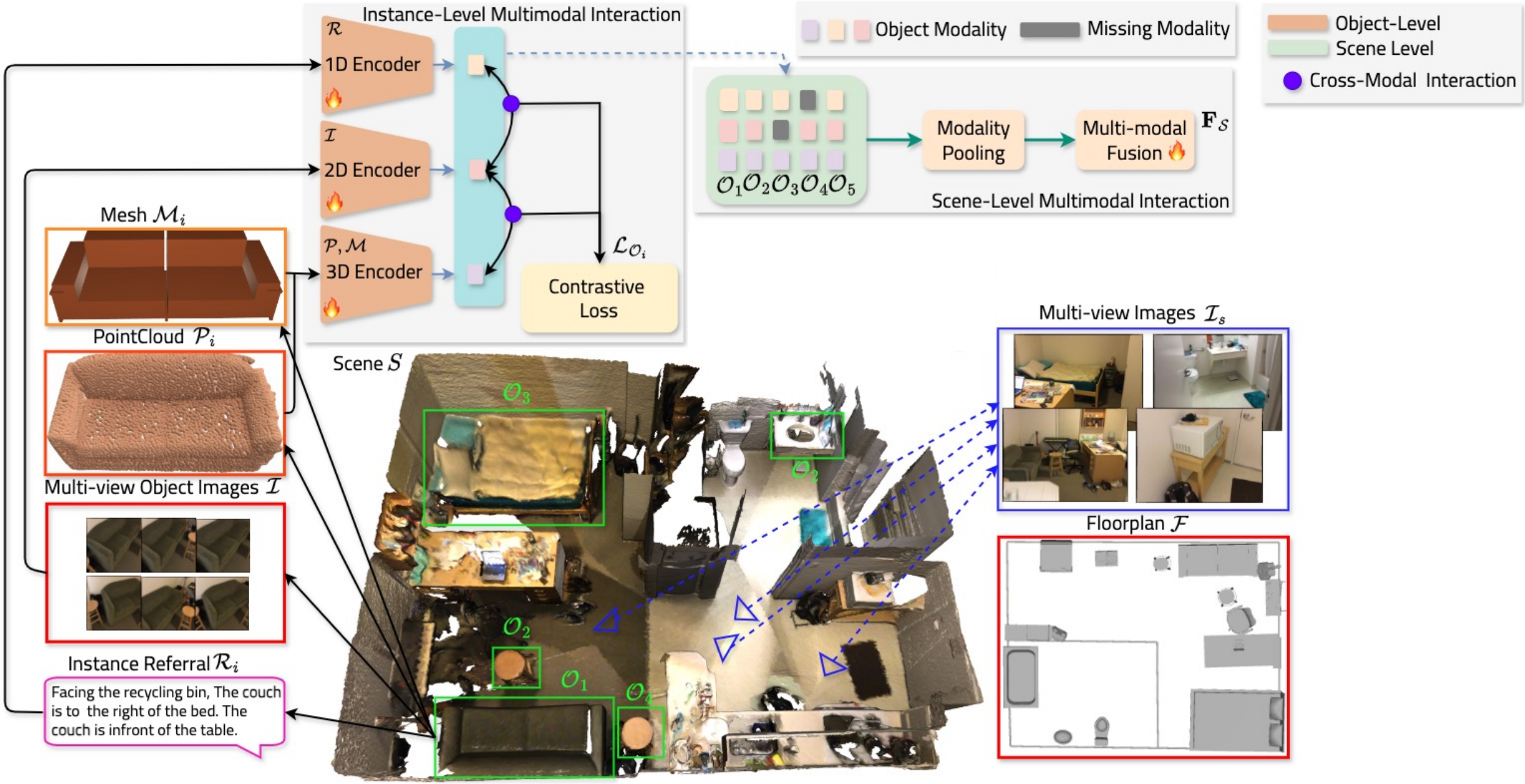
Overview of CrossOver



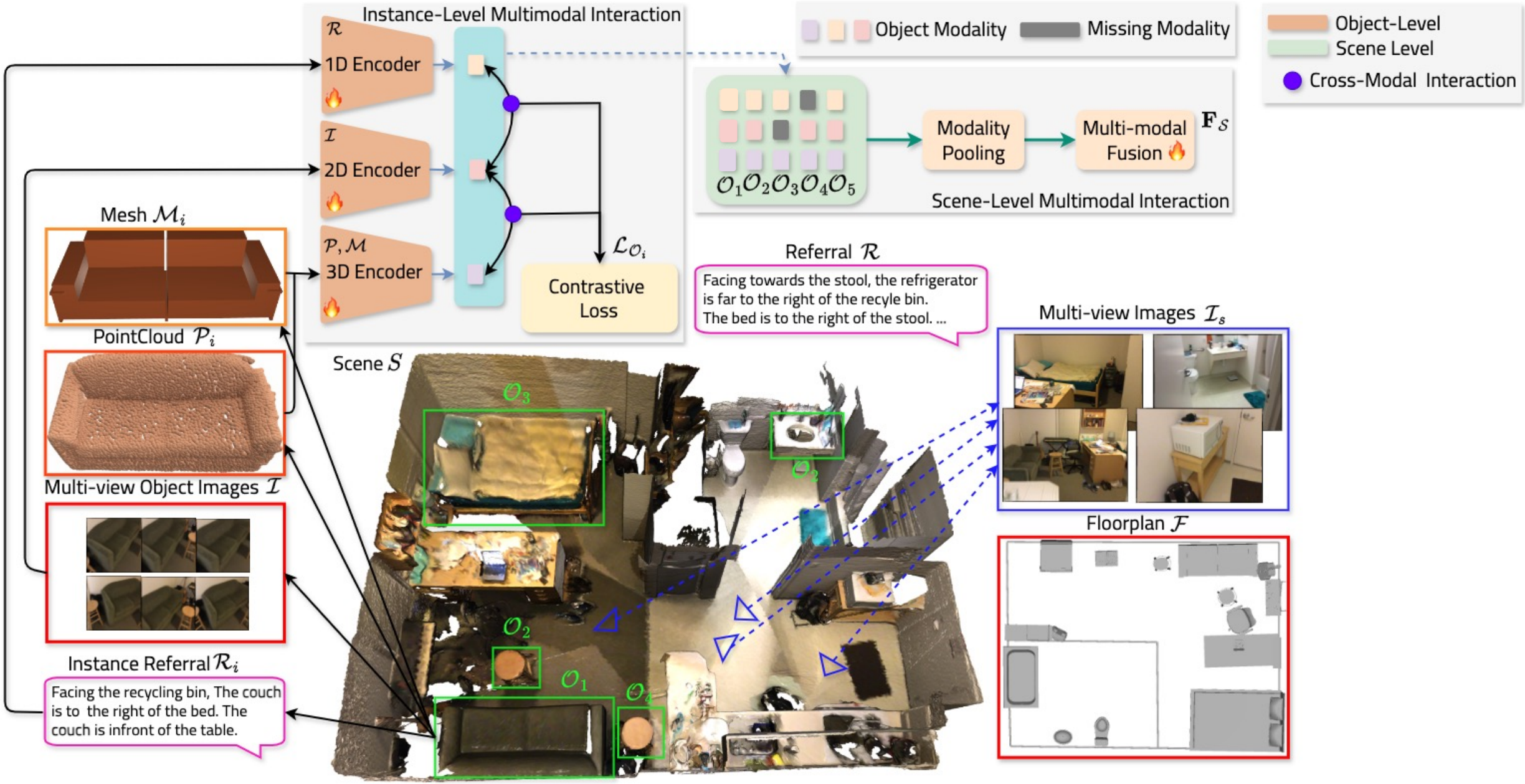
Overview of CrossOver



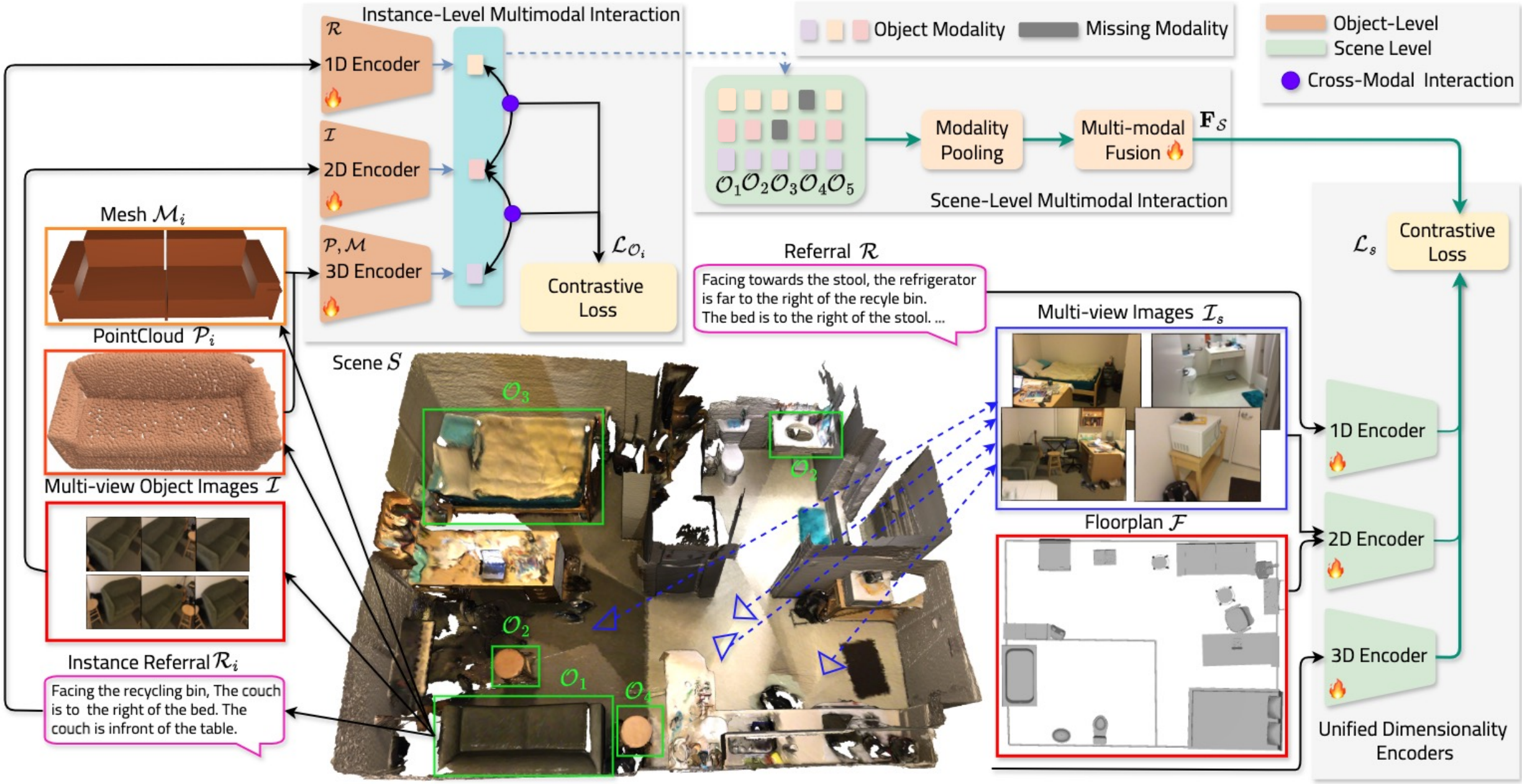
Overview of CrossOver



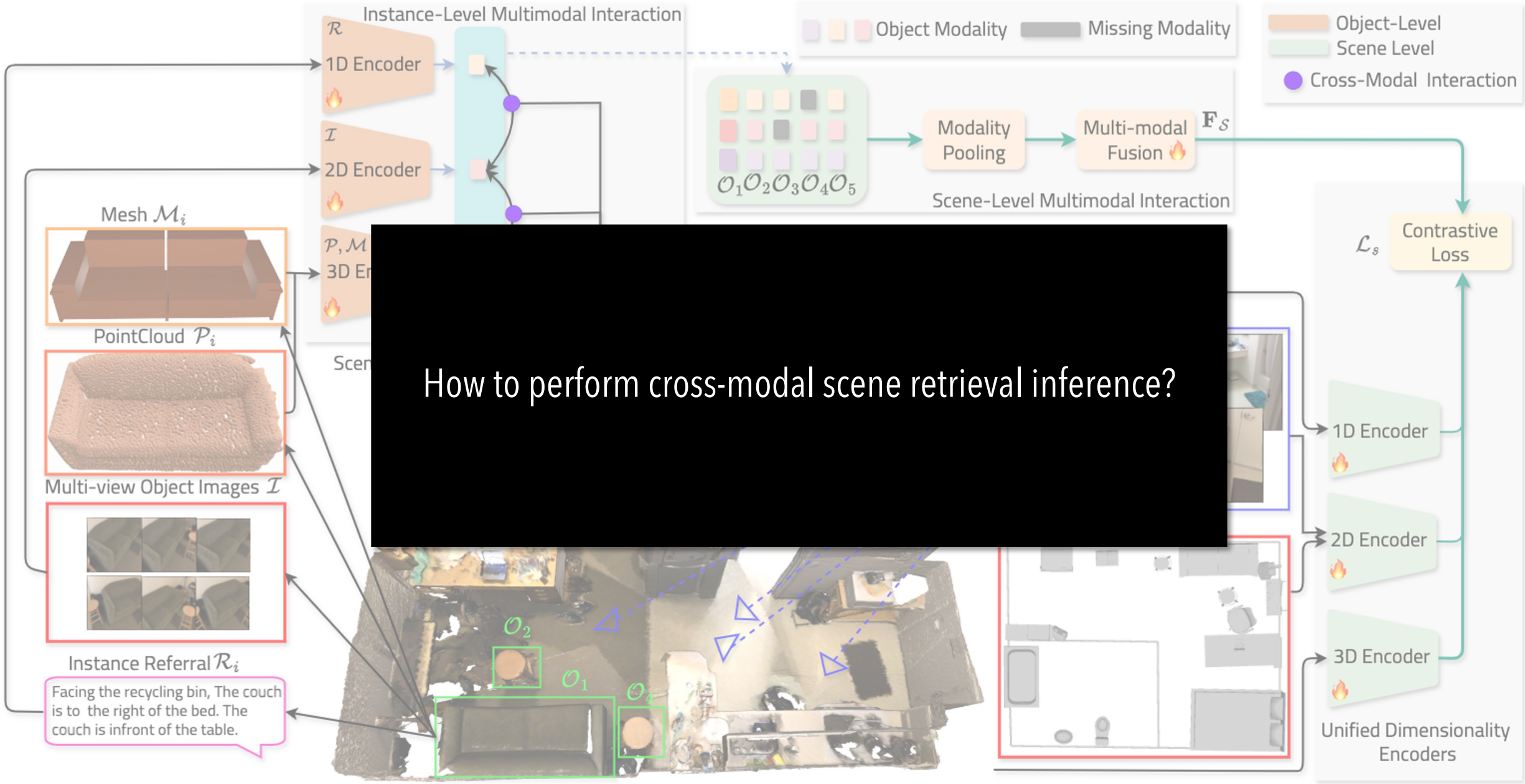
Overview of CrossOver



Overview of CrossOver

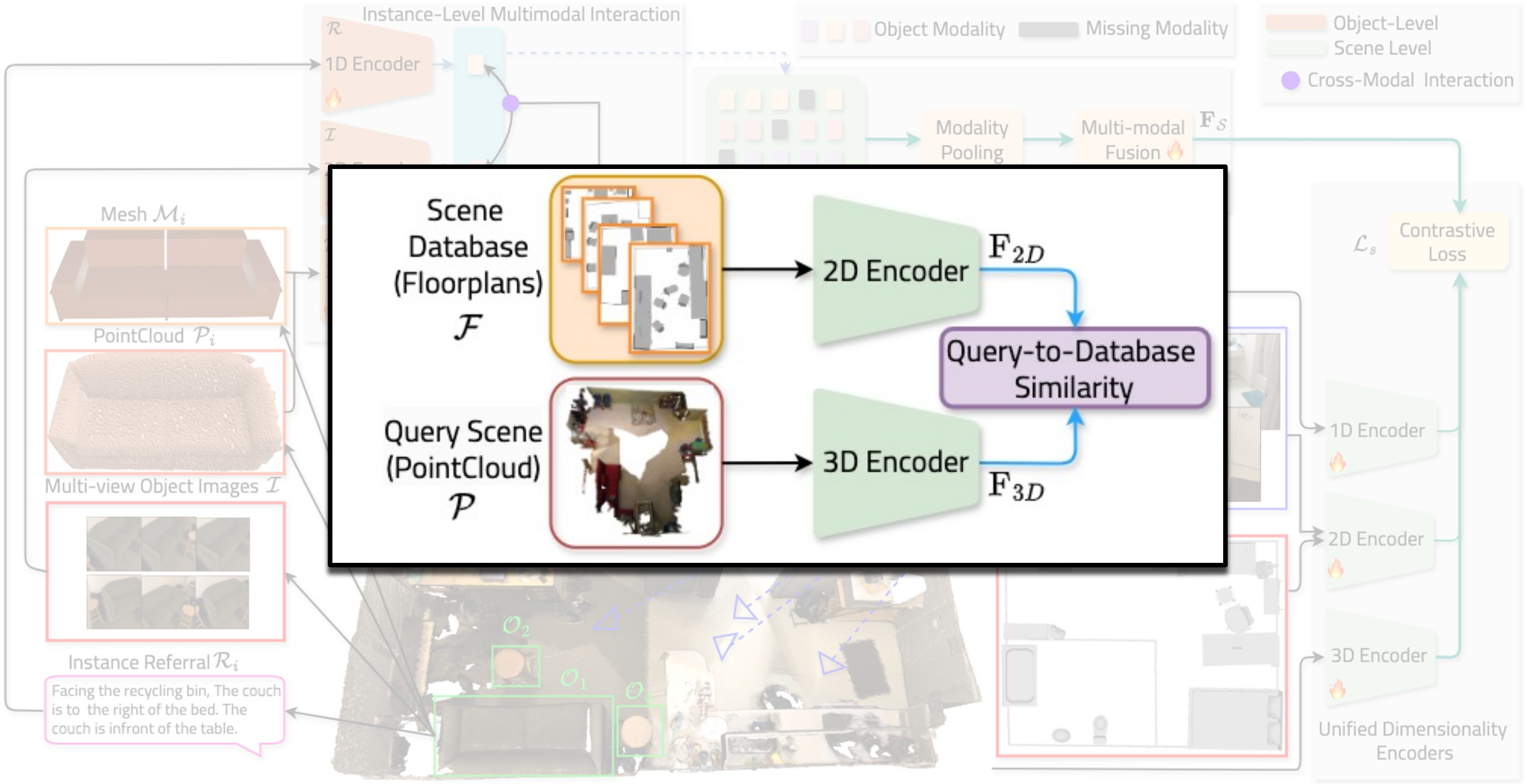


Overview of CrossOver



How to perform cross-modal scene retrieval inference?

Scene Retrieval Inference Pipeline



Experimental Results

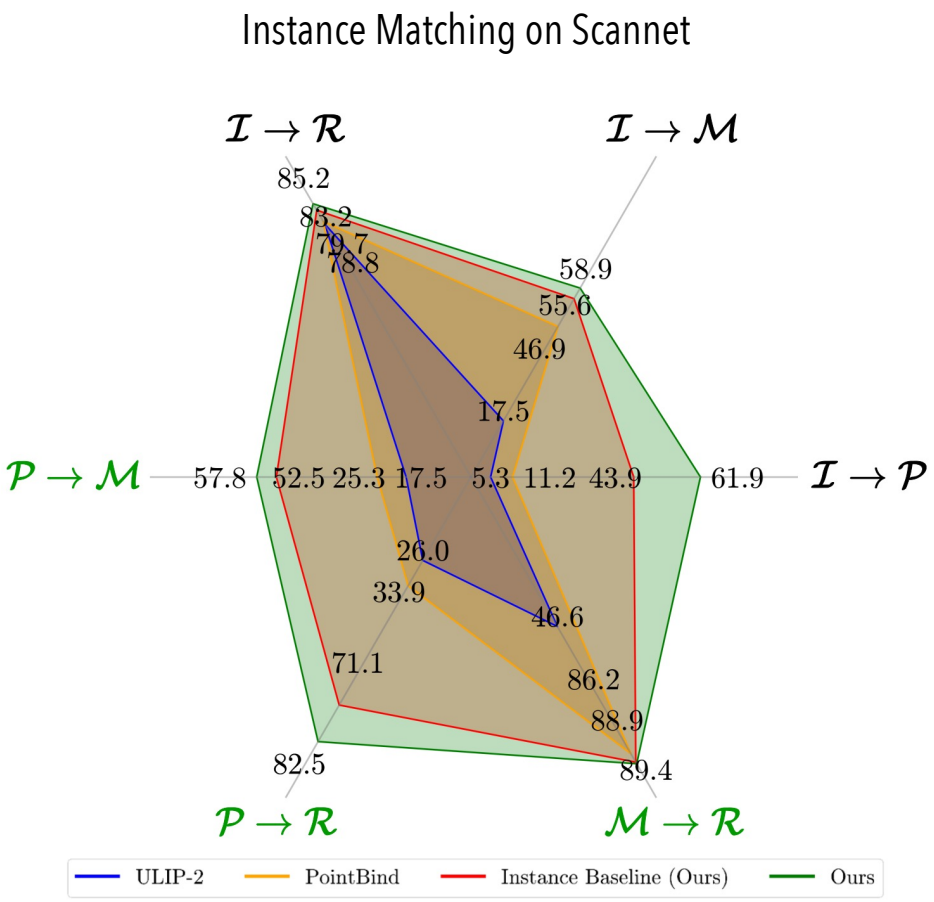
Cross-Modal Instance Matching

Experimental Results

Cross-Modal Instance Matching

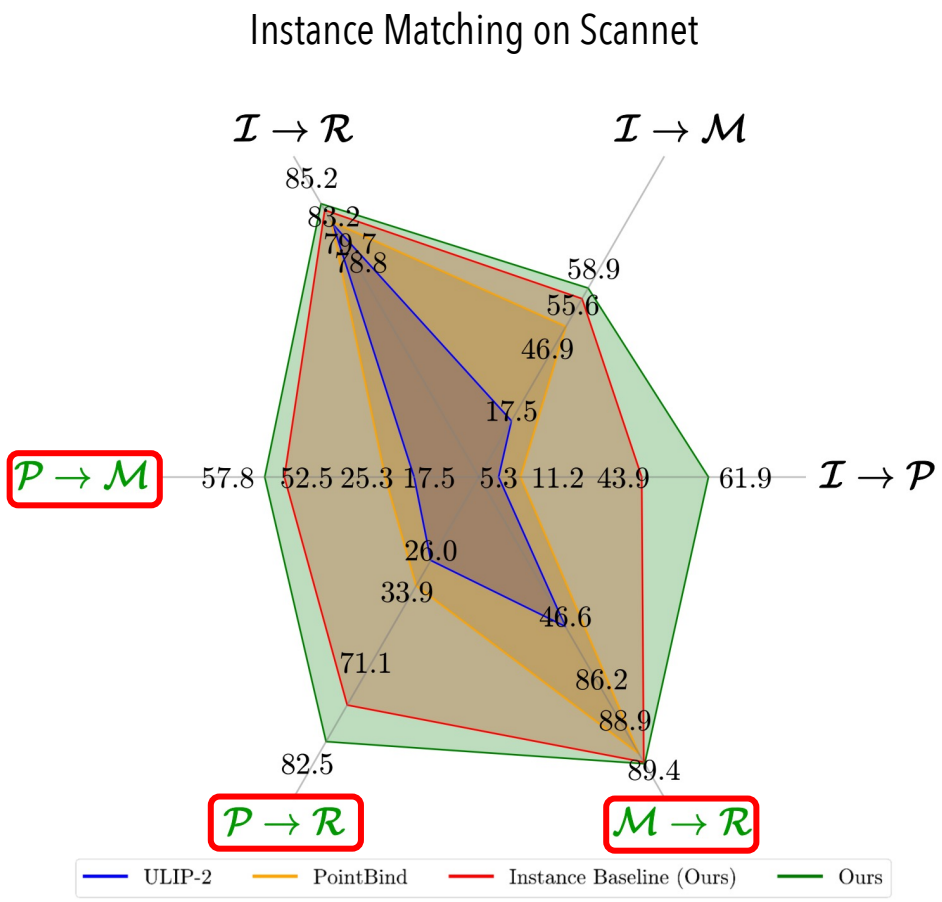
Experimental Results

Cross-Modal Instance Matching



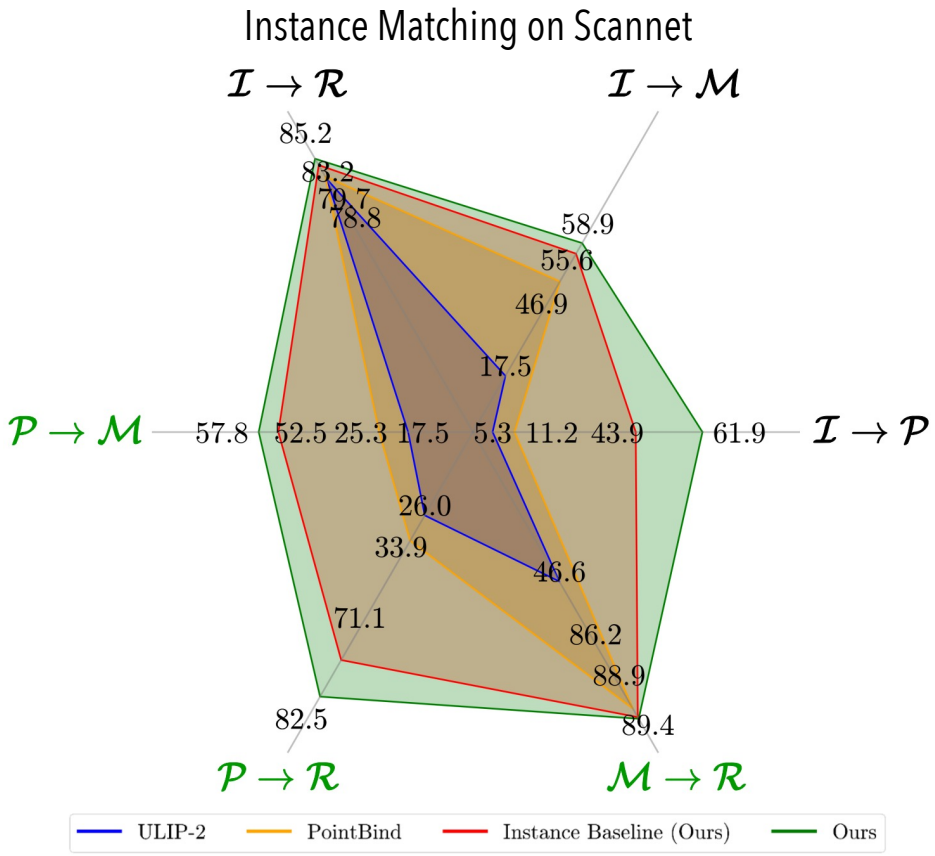
Experimental Results

Cross-Modal Instance Matching



Experimental Results

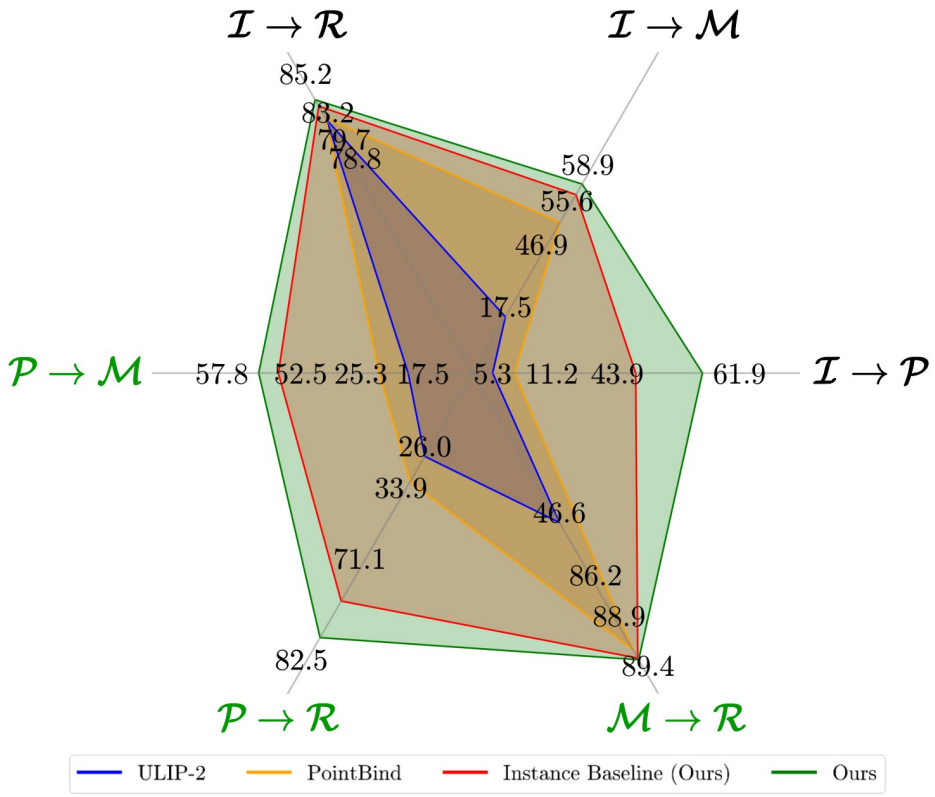
Cross-Modal Instance Matching



Experimental Results

Cross-Modal Instance Matching

Instance Matching on Scannet

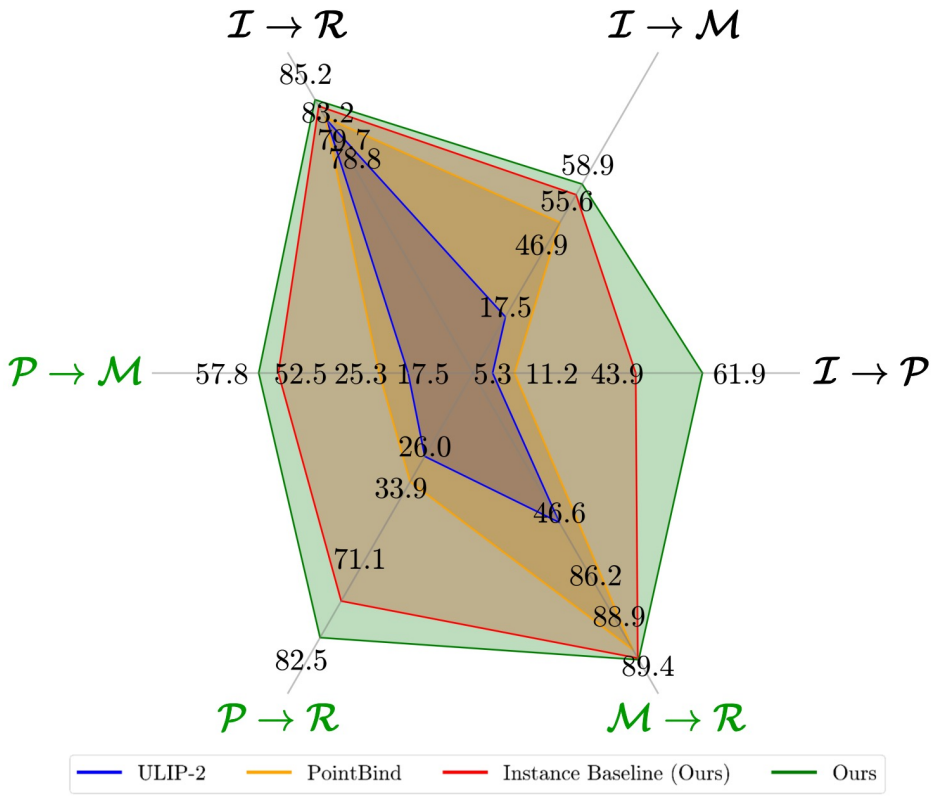


Scene Level Instance Matching on Scannet

Experimental Results

Cross-Modal Instance Matching

Instance Matching on Scannet



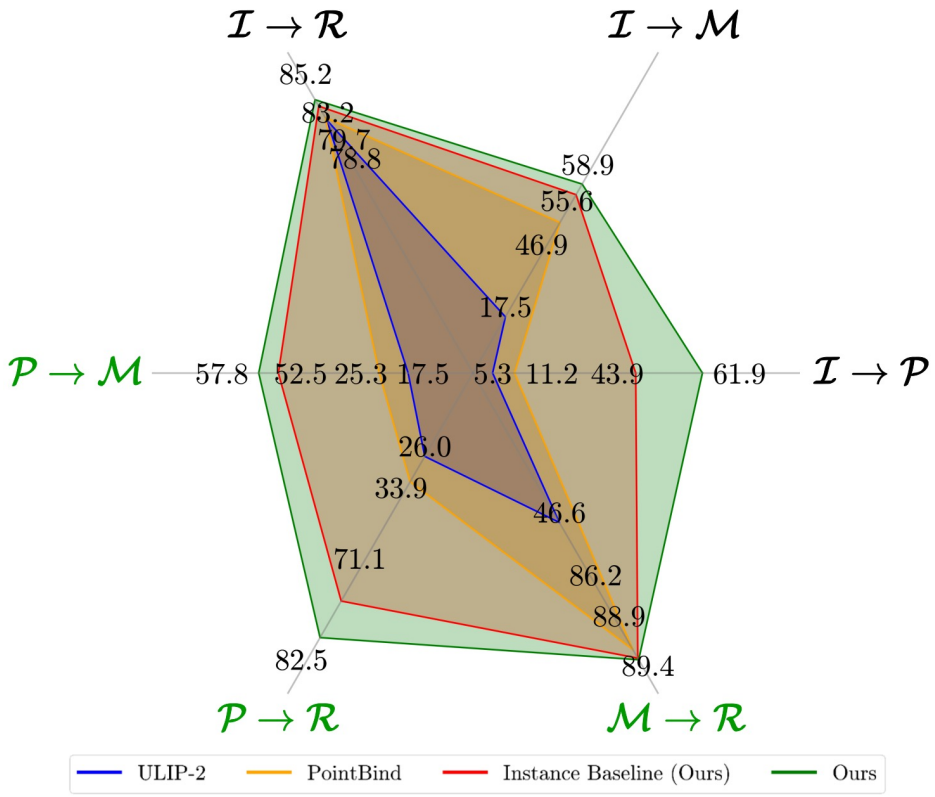
Scene Level Instance Matching on Scannet

	Scannet [11]			3RScan [38]		
Scene-level Recall ↑	**R@25%**	**R@50%**	**R@75%**	**R@25%**	**R@50%**	**R@75%**
$\mathcal{I} \rightarrow \mathcal{P}$						
ULIP-2 [43]	1.28	0.64	0.24	1.91	0.40	0.28
PointBind [18]	6.73	0.96	0.32	3.18	0.64	0.01
Inst. Baseline (Ours)	88.46	37.82	1.92	93.63	35.03	3.82
Ours	**98.08**	**76.92**	**23.40**	**99.36**	**79.62**	**22.93**
$\mathcal{I} \rightarrow \mathcal{R}$						
ULIP-2 [43]	98.12	96.21	60.34	98.66	85.91	36.91
PointBind [18]	98.22	95.17	62.07	**100**	87.25	41.61
Inst. Baseline (Ours)	99.31	97.59	71.13	**100**	92.62	55.03
Ours	**99.66**	**98.28**	**76.29**	**100**	**97.32**	**67.79**
$\mathcal{P} \rightarrow \mathcal{R}$						
ULIP-2 [43]	37.24	16.90	8.62	16.78	6.04	1.34
PointBind [18]	54.83	27.93	11.72	21.48	6.04	2.01
Inst. Baseline (Ours)	98.63	83.85	46.74	92.62	60.40	20.81
Ours	**99.31**	**96.56**	**70.10**	**100**	**89.26**	**50.34**

Experimental Results

Cross-Modal Instance Matching

Instance Matching on Scannet



Scene Level Instance Matching on Scannet

	Scannet [11]			3RScan [38]		
Scene-level Recall ↑	**R@25%**	**R@50%**	**R@75%**	**R@25%**	**R@50%**	**R@75%**
$\mathcal{I} \rightarrow \mathcal{P}$						
ULIP-2 [43]	1.28	0.64	0.24	1.91	0.40	0.28
PointBind [18]	6.73	0.96	0.32	3.18	0.64	0.01
Inst. Baseline (Ours)	88.46	37.82	1.92	93.63	35.03	3.82
Ours	**98.08**	**76.92**	**23.40**	**99.36**	**79.62**	**22.93**
$\mathcal{I} \rightarrow \mathcal{R}$						
ULIP-2 [43]	98.12	96.21	60.34	98.66	85.91	36.91
PointBind [18]	98.22	95.17	62.07	**100**	87.25	41.61
Inst. Baseline (Ours)	99.31	97.59	71.13	**100**	92.62	55.03
Ours	**99.66**	**98.28**	**76.29**	**100**	**97.32**	**67.79**
$\mathcal{P} \rightarrow \mathcal{R}$						
ULIP-2 [43]	37.24	16.90	8.62	16.78	6.04	1.34
PointBind [18]	54.83	27.93	11.72	21.48	6.04	2.01
Inst. Baseline (Ours)	98.63	83.85	46.74	92.62	60.40	20.81
Ours	**99.31**	**96.56**	**70.10**	**100**	**89.26**	**50.34**

- Outperforms all baselines in all datasets
- Emergent behaviour within the embedding space, highlighting robustness of learning cross-modal interactions

Experimental Results

Temporal Instance Matching

Experimental Results

Temporal Instance Matching

Method	Scene-level Recall ↑ R@25%	R@50%	R@75%
same-modal ($\mathcal{P} \rightarrow \mathcal{P}$)			
MendNet [15]	80.68	64.77	37.50
VN-DGCNN$_{\text{cls}}$ [13]	72.32	53.41	29.55
VN-ONet$_{\text{recon}}$ [13]	86.36	71.59	44.32
LivingScenes [47]	87.50	78.41	50.00
Ours	**92.31**	**84.62**	**57.69**
cross-modal (ours)			
$\mathcal{I} \rightarrow \mathcal{P}$	89.74	73.08	42.31
$\mathcal{I} \rightarrow \mathcal{R}$	62.33	38.96	18.18
$\mathcal{P} \rightarrow \mathcal{R}$	68.83	40.26	22.08

Experimental Results

Temporal Instance Matching

Scene-level Recall ↑			
Method	**R@25%**	**R@50%**	**R@75%**
same-modal ($\mathcal{P} \rightarrow \mathcal{P}$)			
MendNet [15]	80.68	64.77	37.50
VN-DGCNN$_{\text{cls}}$ [13]	72.32	53.41	29.55
VN-ONet$_{\text{recon}}$ [13]	86.36	71.59	44.32
LivingScenes [47]	87.50	78.41	50.00
Ours	**92.31**	**84.62**	**57.69**
cross-modal (ours)			
$\mathcal{I} \rightarrow \mathcal{P}$	89.74	73.08	42.31
$\mathcal{I} \rightarrow \mathcal{R}$	62.33	38.96	18.18
$\mathcal{P} \rightarrow \mathcal{R}$	68.83	40.26	22.08

- Better performance in the same-modal task compared to baselines, despite not being specifically trained on temporal data!

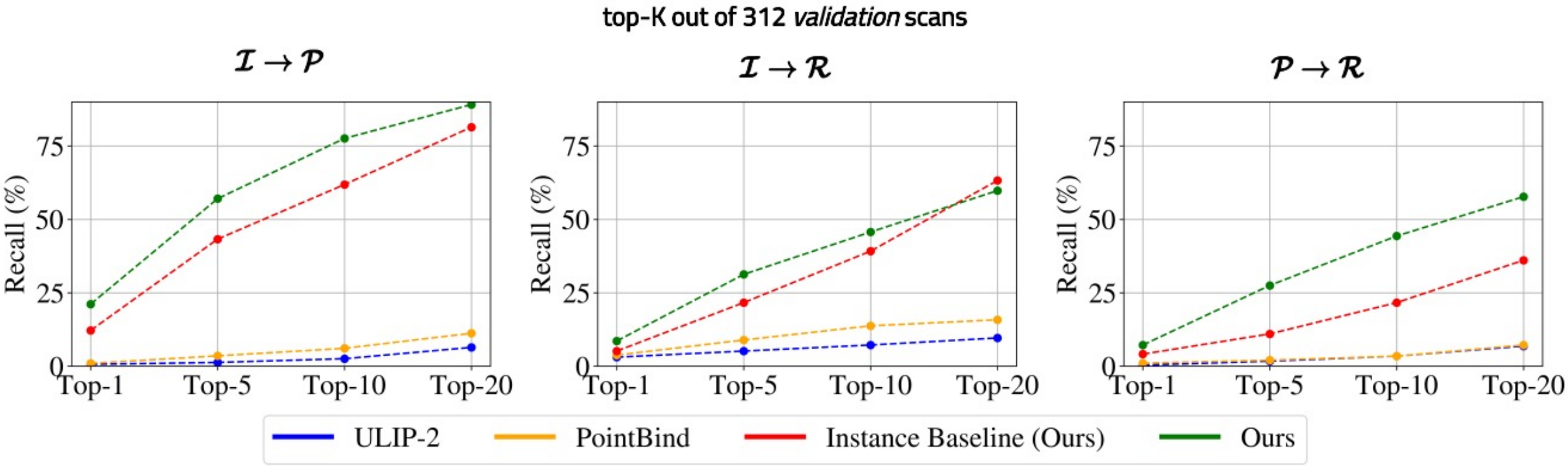
Experimental Results

Cross-Modal Scene Retrieval

Experimental Results

Cross-Modal Scene Retrieval

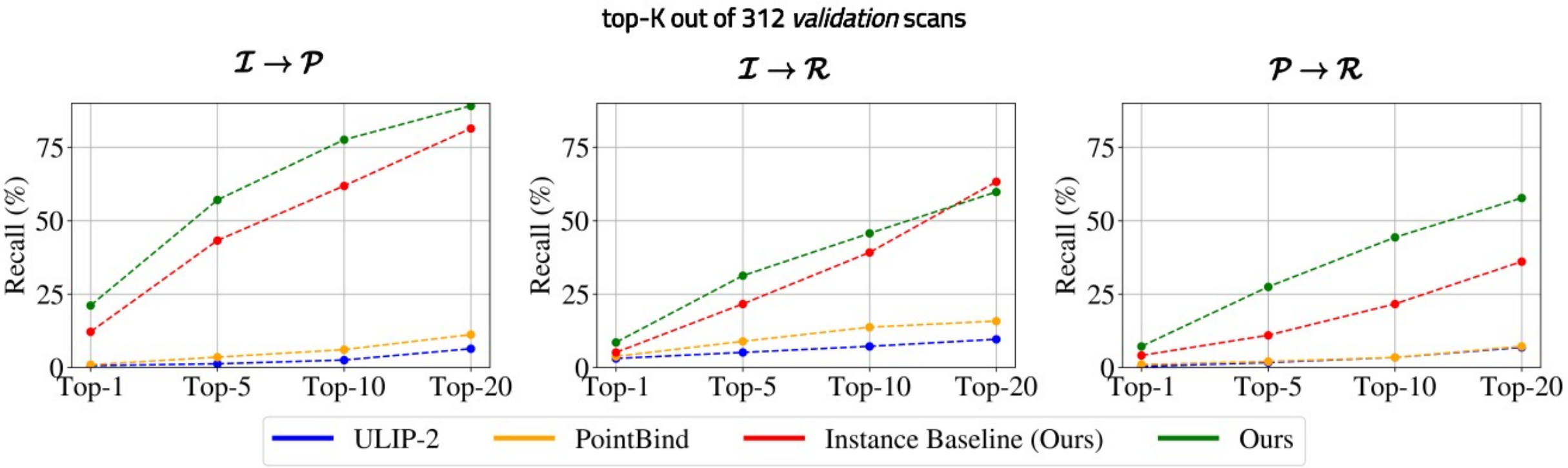
Scene matching recall of different methods on three modality pairs



Experimental Results

Cross-Modal Scene Retrieval

Scene matching recall of different methods on three modality pairs



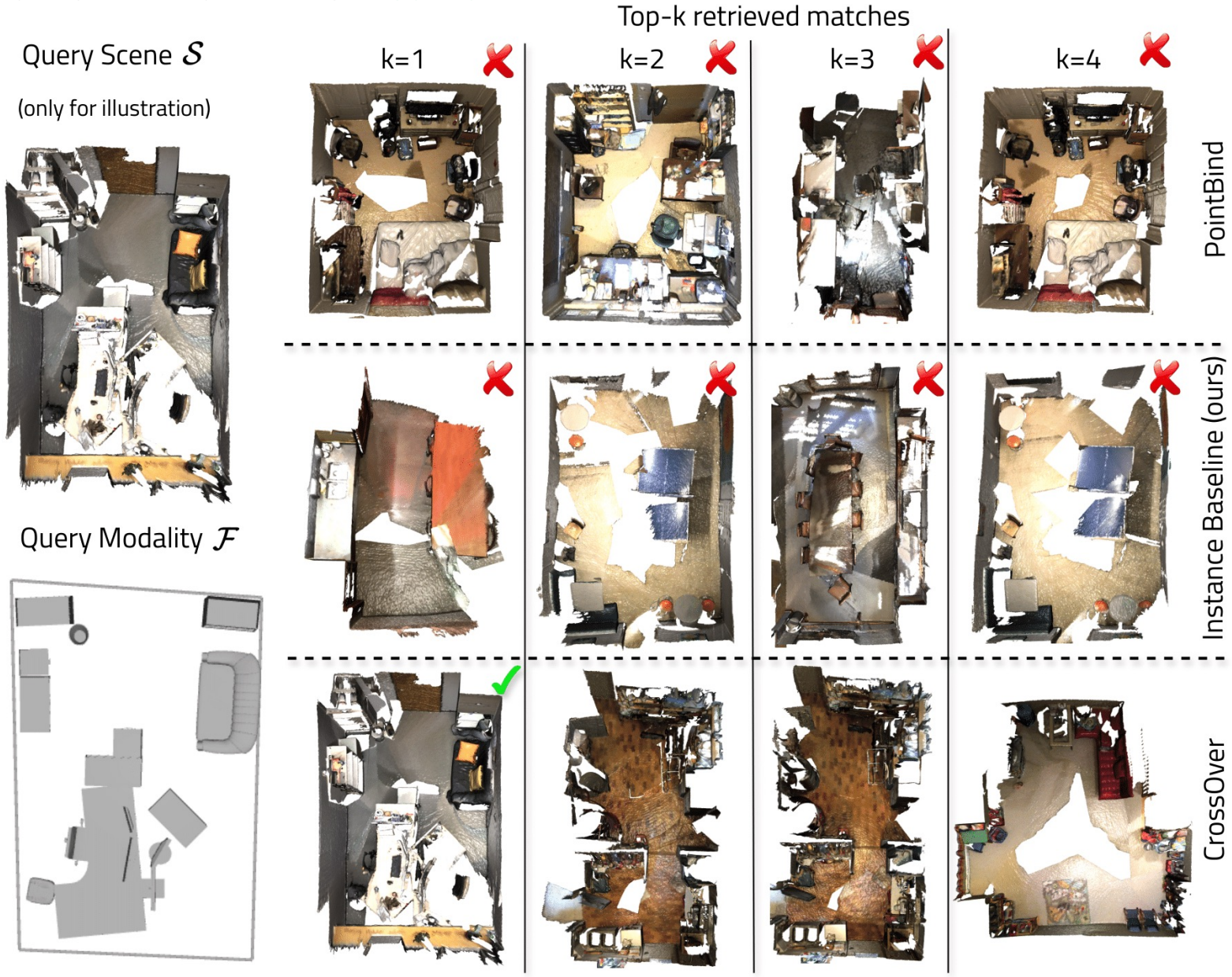
- Unified dimensionality encoders: do not rely on semantics, consistently outperform prior methods in all pairwise modalities

Experimental Results

Cross-Modal Scene Retrieval Visualization: Success

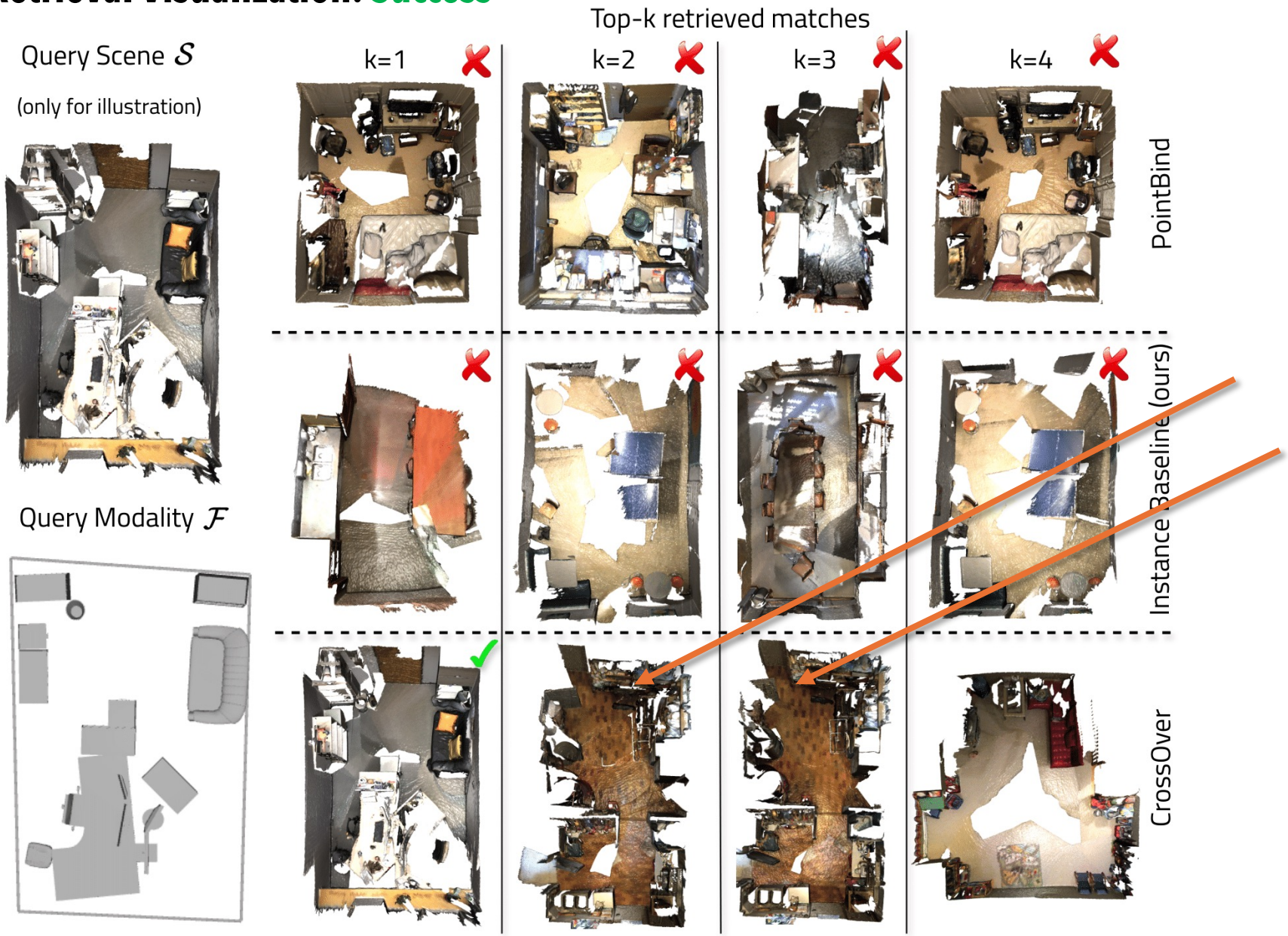
Experimental Results

Cross-Modal Scene Retrieval Visualization: Success



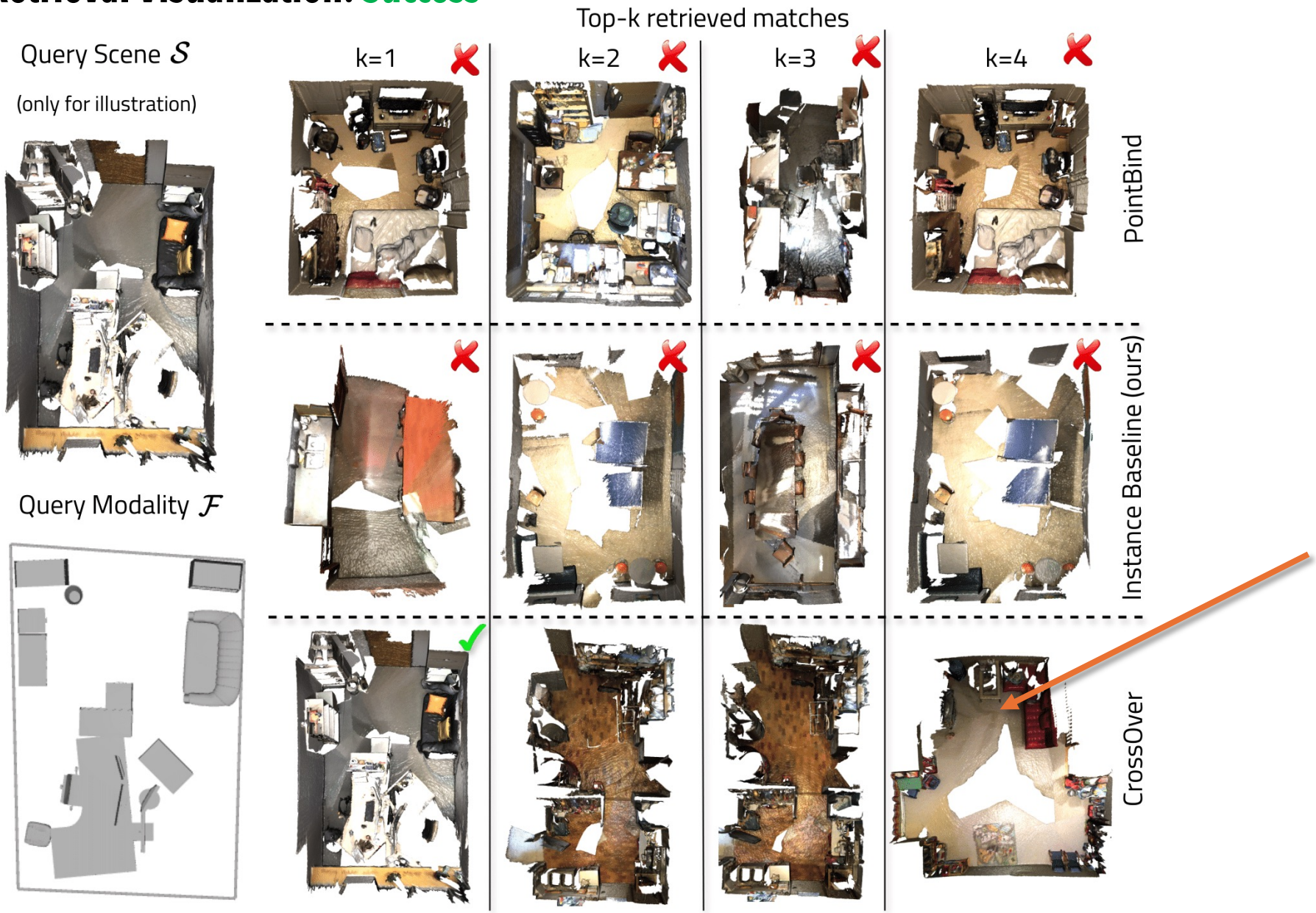
Experimental Results

Cross-Modal Scene Retrieval Visualization: Success



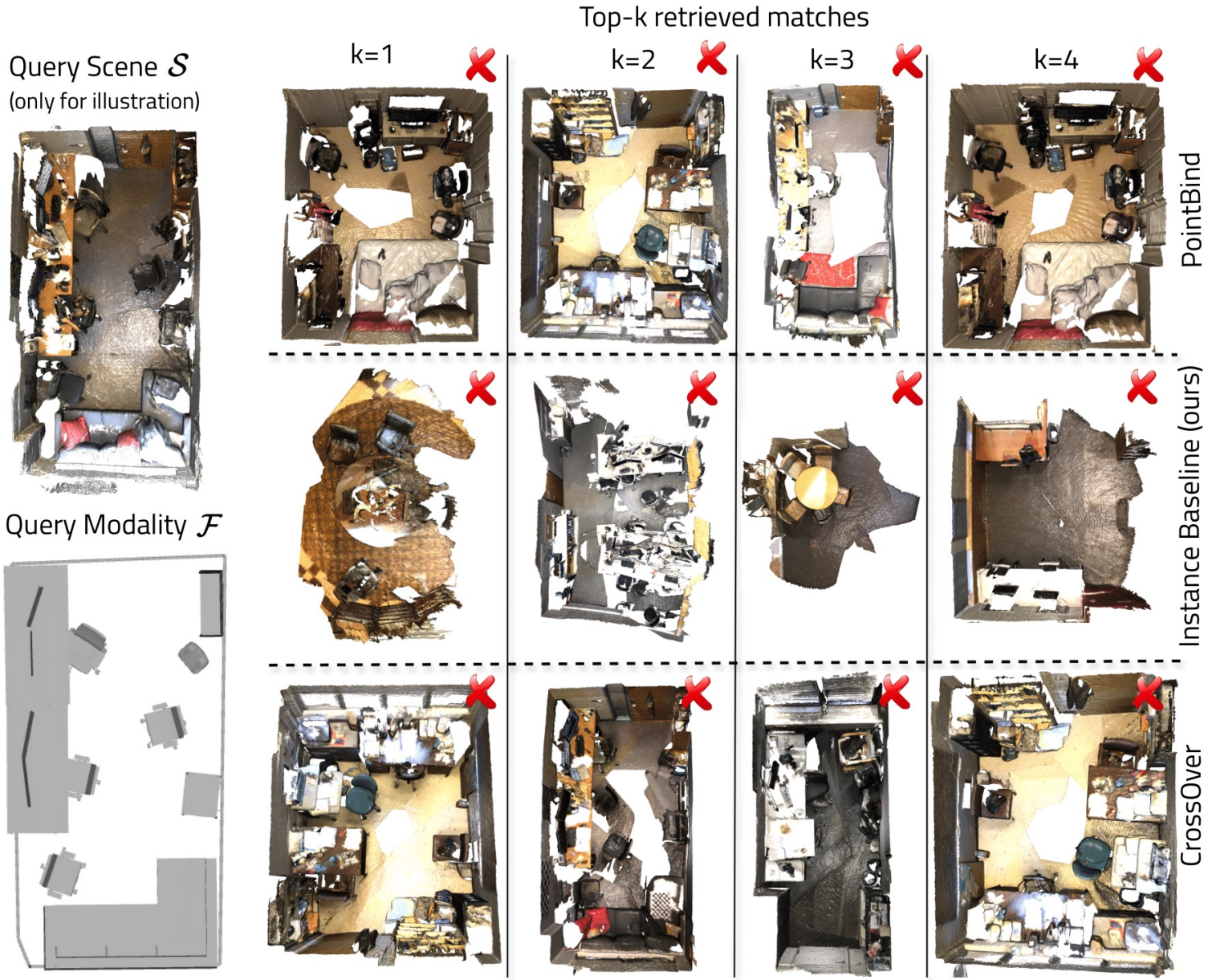
Experimental Results

Cross-Modal Scene Retrieval Visualization: Success



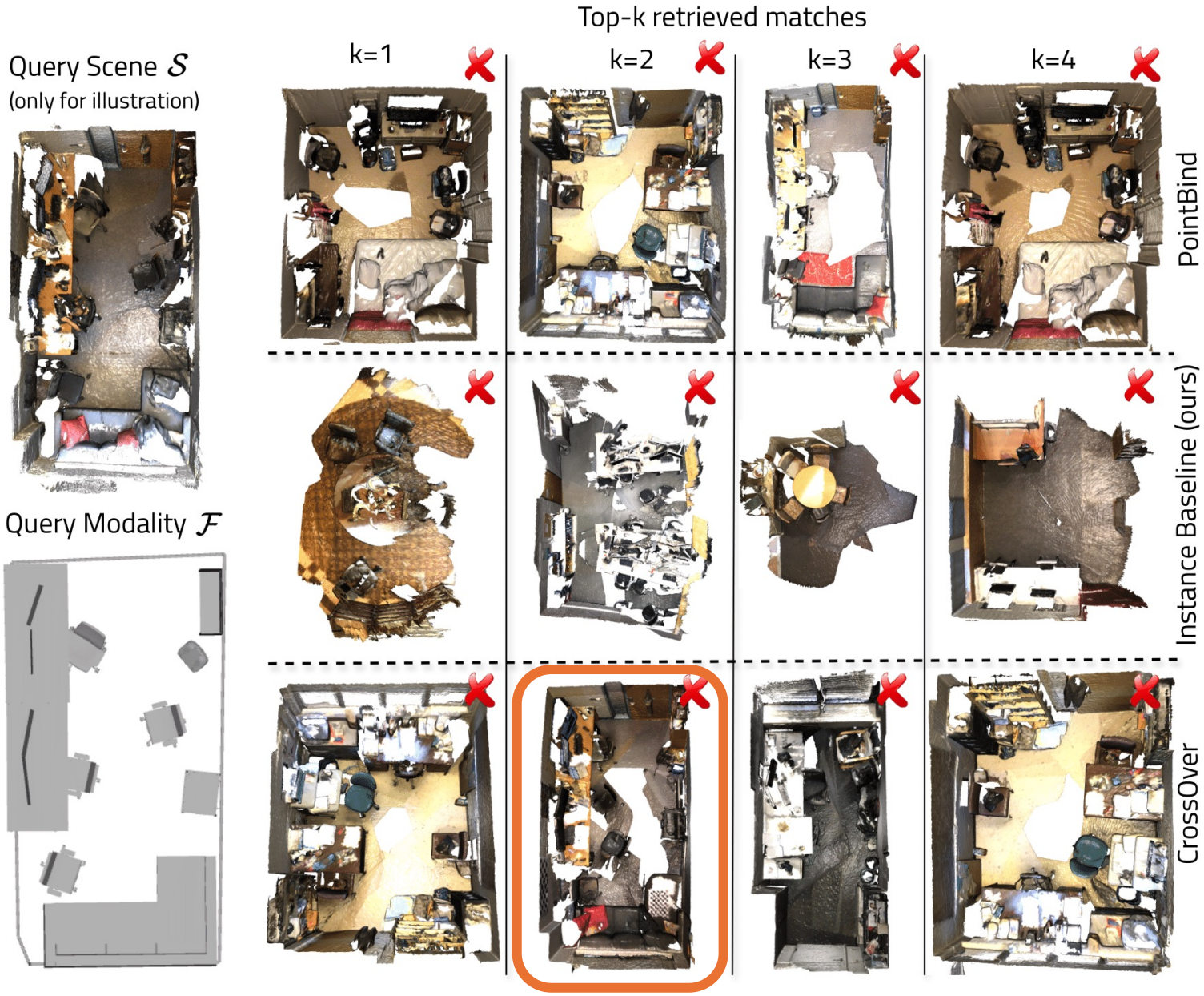
Experimental Results

Cross-Modal Scene Retrieval Visualization: Failure



Experimental Results

Cross-Modal Scene Retrieval Visualization: Failure



Experimental Results

Missing Modalities

Experimental Results

Missing Modalities

Ablation on instance matching on ScanNet with non-overlapping data per modality pair

Available Data		Instance Matching Recall ↑			
$\mathcal{I} \rightarrow \mathcal{P}$ (%)	$\mathcal{I} \rightarrow \mathcal{M}$ (%)	same	diff	top-1	top-3
25	75	86.32	**73.38**	55.46	79.73
50	50	**87.46**	70.02	57.49	79.94
75	25	87.35	67.65	54.99	79.45
100	100	87.44	72.46	**59.88**	**80.81**

- Although partial data availability decreases recall, our matching only decreases by 3% even when using 25% data
- Common in real-world applications, where certain modalities might be scarce!

Future Work

Key Takeaways:

Future Work

Key Takeaways:

- **Cross-modal alignment method** for 3D scenes that learns a unified, modality-agnostic embedding space, enabling a range of scene understanding tasks.

- CrossOver leverages a **unified embedding space** centered on image features, allowing it to generalize across **unpaired modalities** and **outperform existing methods** on real-world datasets

Future Work

Key Takeaways:

- **Cross-modal alignment method** for 3D scenes that learns a unified, modality-agnostic embedding space, enabling a range of scene understanding tasks.

- CrossOver leverages a **unified embedding space** centered on image features, allowing it to generalize across **unpaired modalities** and **outperform existing methods** on real-world datasets

Current Directions:

Future Work

Key Takeaways:

- **Cross-modal alignment method** for 3D scenes that learns a unified, modality-agnostic embedding space, enabling a range of scene understanding tasks.

- CrossOver leverages a **unified embedding space** centered on image features, allowing it to generalize across **unpaired modalities** and **outperform existing methods** on real-world datasets

Current Directions:

- How can we scale up cross-modal approaches on a scene level? Multiple challenges: noisy data, bad annotations, etc

- How can we **enable traditional 3D scene understanding** with CrossOver, eg, segmentation, 3D question answering & visual grounding?

- How our approach can be applied to **dynamic scene reconstruction and real-time navigation**, leading to interactive and immersive mixed-reality experiences?

Future Work

Key Takeaways:

- **Cross-modal alignment method** for 3D scenes that learns a unified, modality-agnostic embedding space, enabling a range of scene understanding tasks.

- CrossOver leverages a **unified embedding space** centered on image features, allowing it to generalize across **unpaired modalities** and **outperform existing methods** on real-world datasets

Current Directions:

- How can we scale up cross-modal approaches on a scene level? Multiple challenges: noisy data, bad annotations, etc

- How can we **enable traditional 3D scene understanding** with CrossOver, eg, segmentation, 3D question answering & visual grounding?

- How our approach can be applied to **dynamic scene reconstruction and real-time navigation**, leading to interactive and immersive mixed-reality experiences?

🔥 Stay Tuned!

- New models trained on much bigger datasets with zero-shot capabilities, releasing soon!

Thank You!