Stanford University

Scalable Cross-Modal 3D Scene Understanding

Sayan Deb Sarkar

XR Munich Google Research Talk

November 4, 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

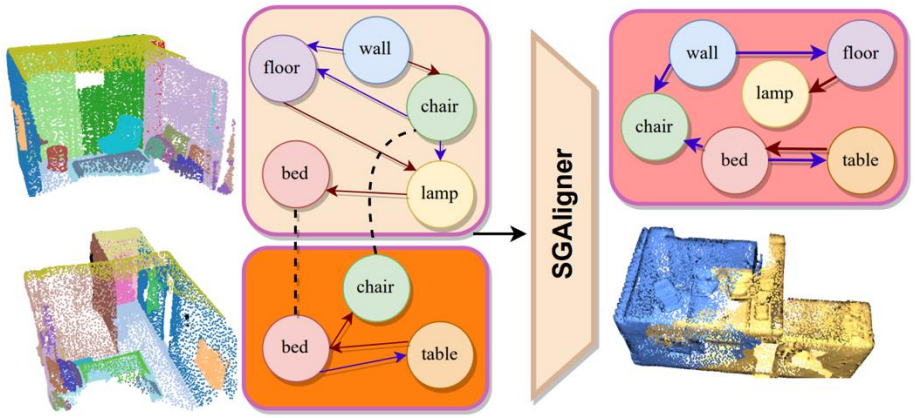
sayands.github.io

Microsoft

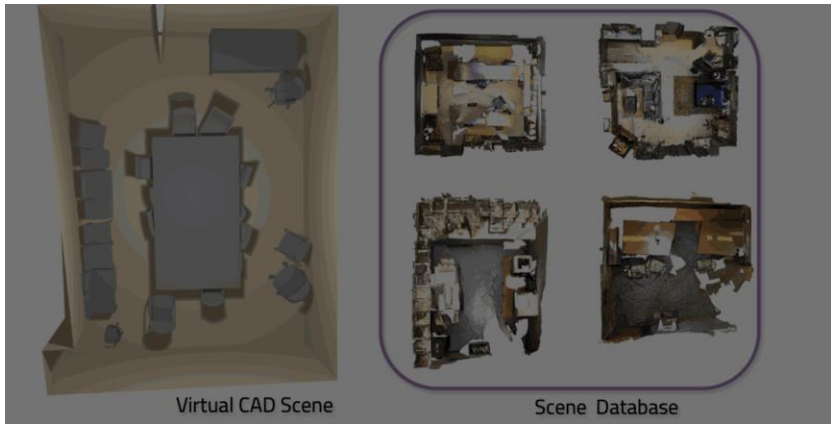
Qualcomm

TU Graz

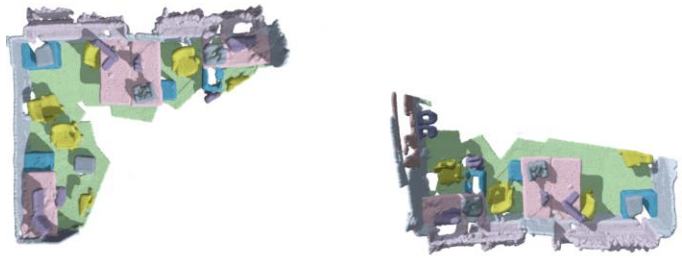
My Work Till Now



SGAligner [ICCV 2023]



CrossOver [CVPR 2025 Highlight]



SGAligner++ [under review]

GuideFlow3D [NeurIPS 2025]

Theme: Multimodal Data Representations For Spatial Understanding

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

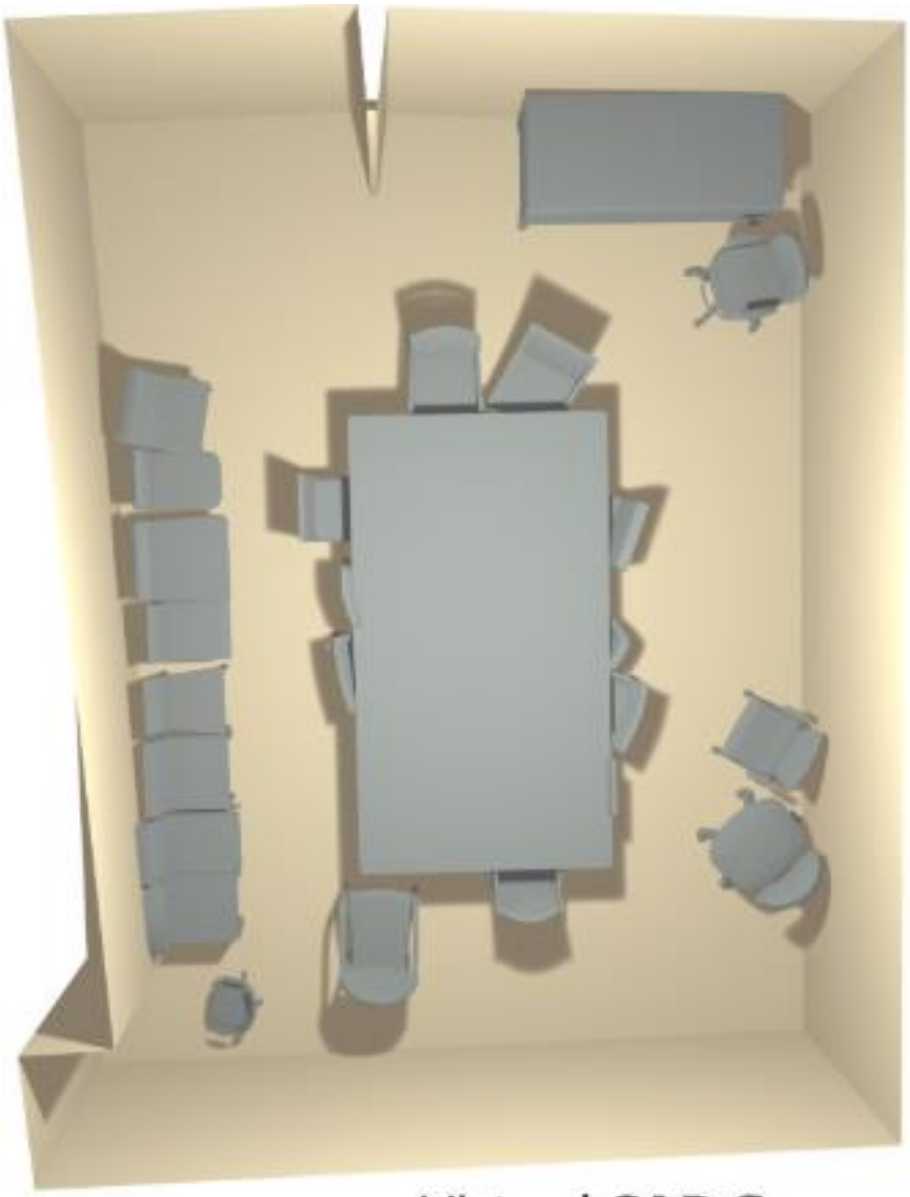
Computer Vision and Geometry Lab

Stanford University

Gradient Spaces

HUN REN SZTAKI

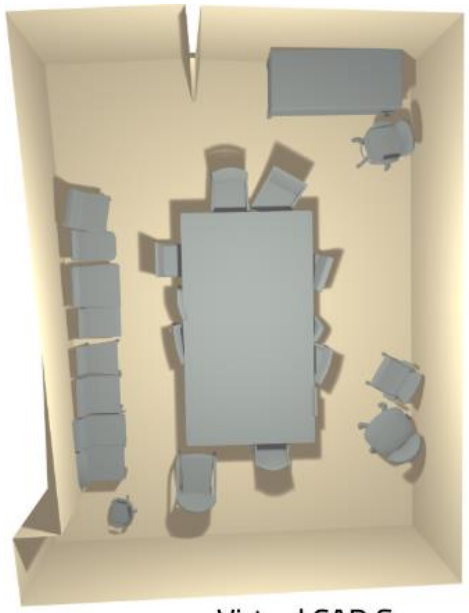
Application



Virtual CAD Scene

Scene Database

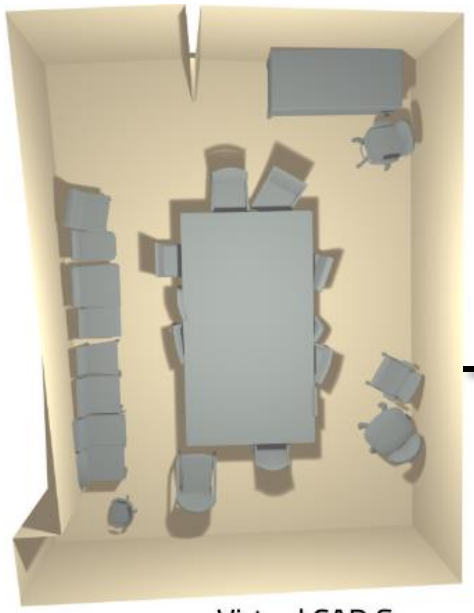
Application



Virtual CAD Scene

Scene Database

Application



Virtual CAD Scene

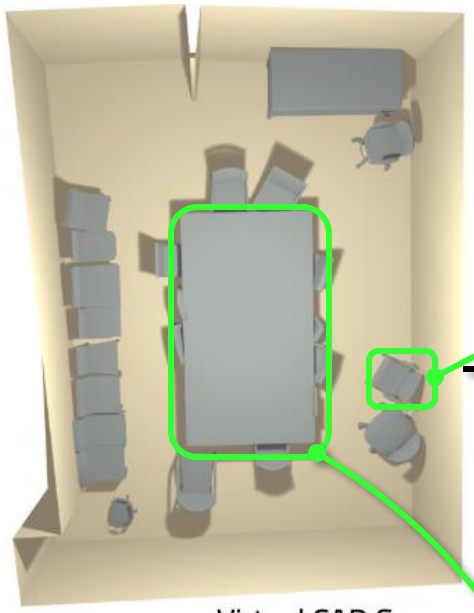
Scene Database

CrossOver

Retrieved

Application

Goal: Aligning digital 3D design with real-world reconstruction.

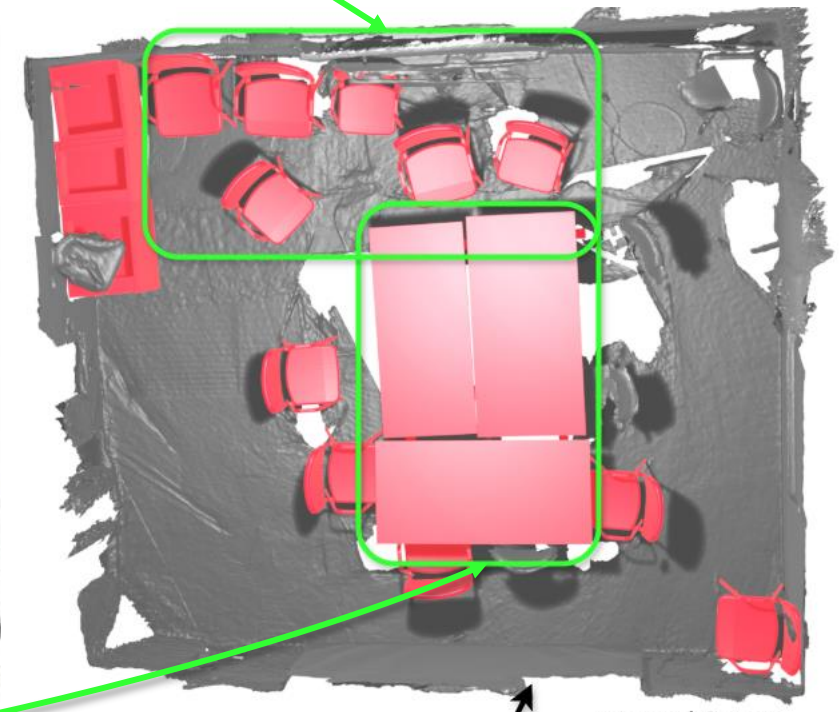


Virtual CAD Scene

CrossOver

Scene Database

Retrieved

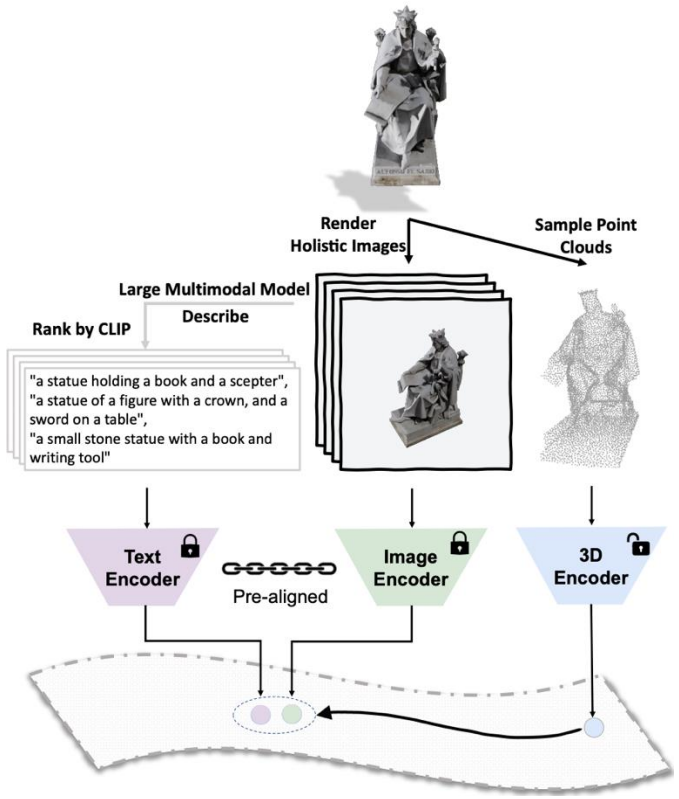


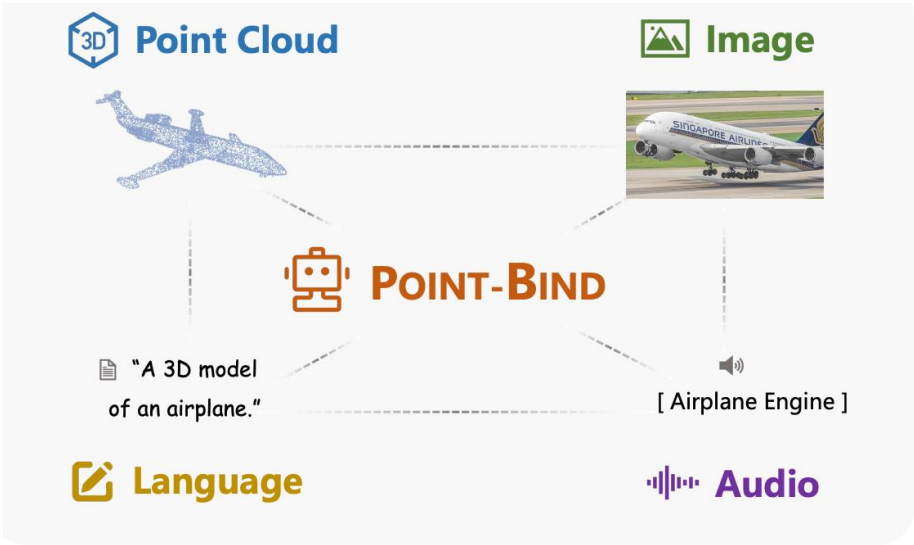
Merged Scene

Matched CAD Models

Motivation

1. Current multi-modal models align isolated objects but ignore spatial and semantic relationships within scenes.





Point-LLM for 3D Q&A

Any-to-3D Generation

3D Embedding Arithmetic

3D Zero-shot Learning

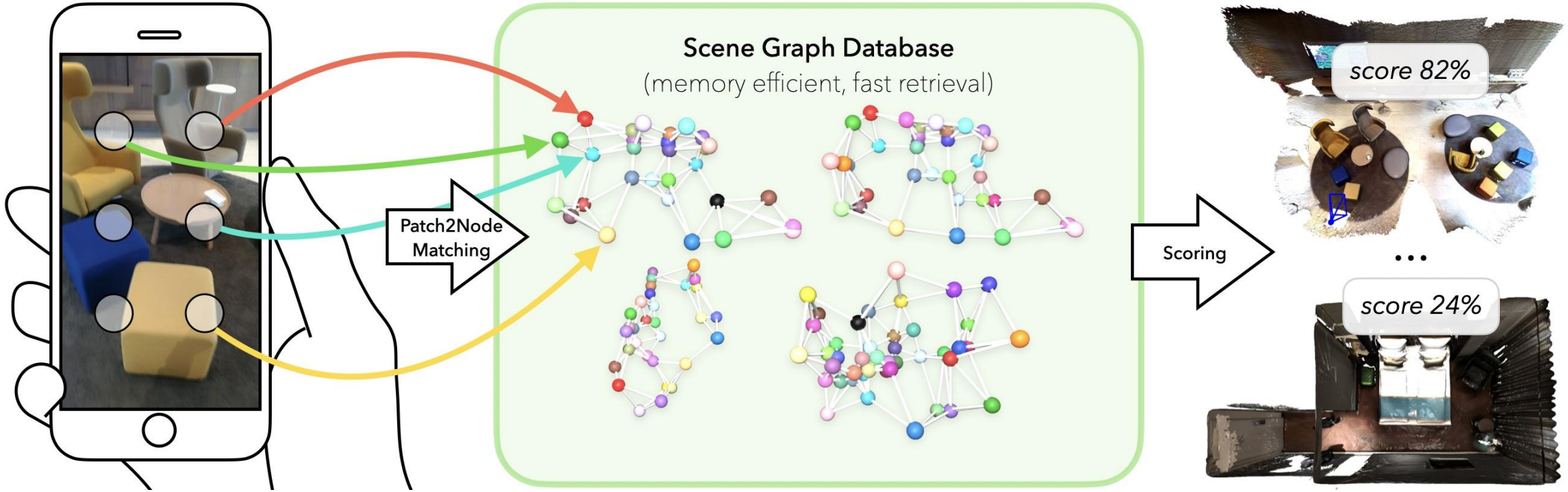
Object-level alignment ≠ scene-level understanding.

[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

Motivation

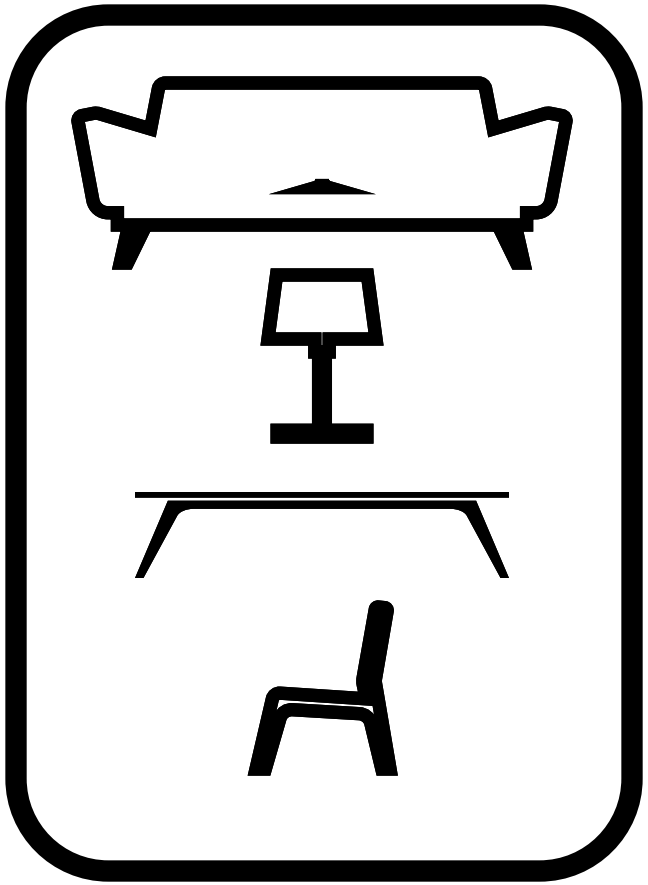
2. Existing scene understanding methods depend on explicit semantic annotations or 3D scene graphs.



How can we achieve cross-modal scene understanding **without complete and paired data?**

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

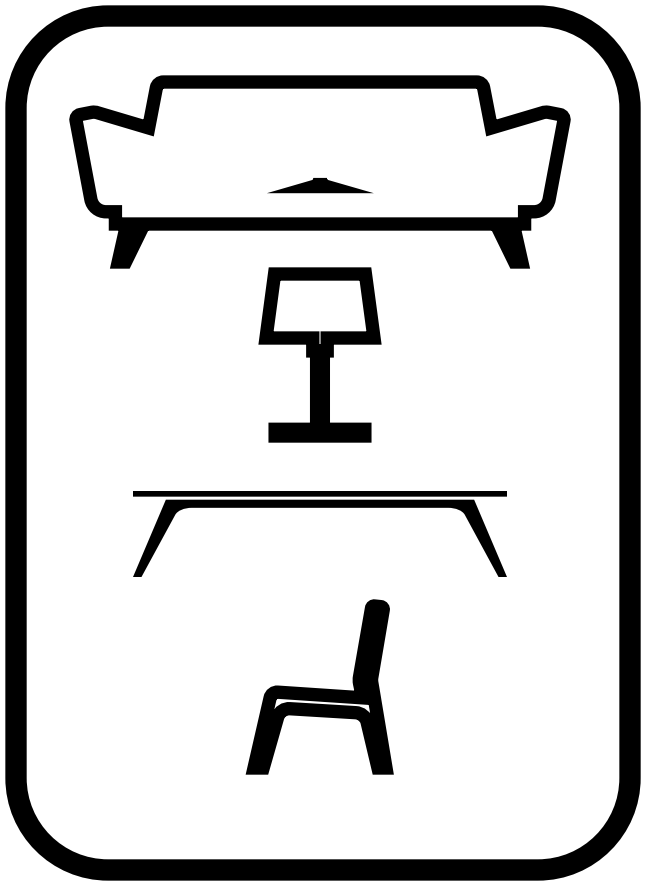
CrossOver

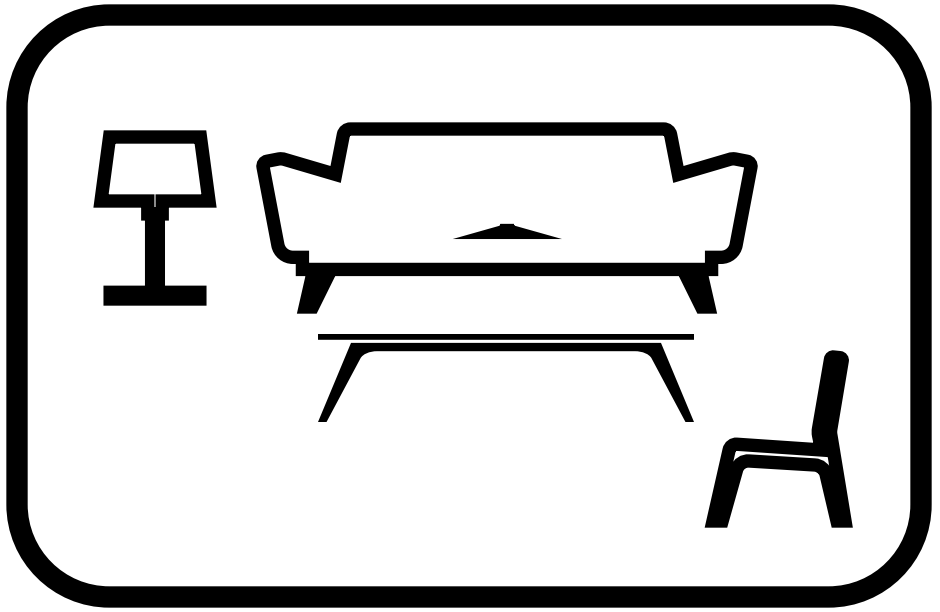


I. Instance-Level Alignment

to leverage large pre-trained models

CrossOver





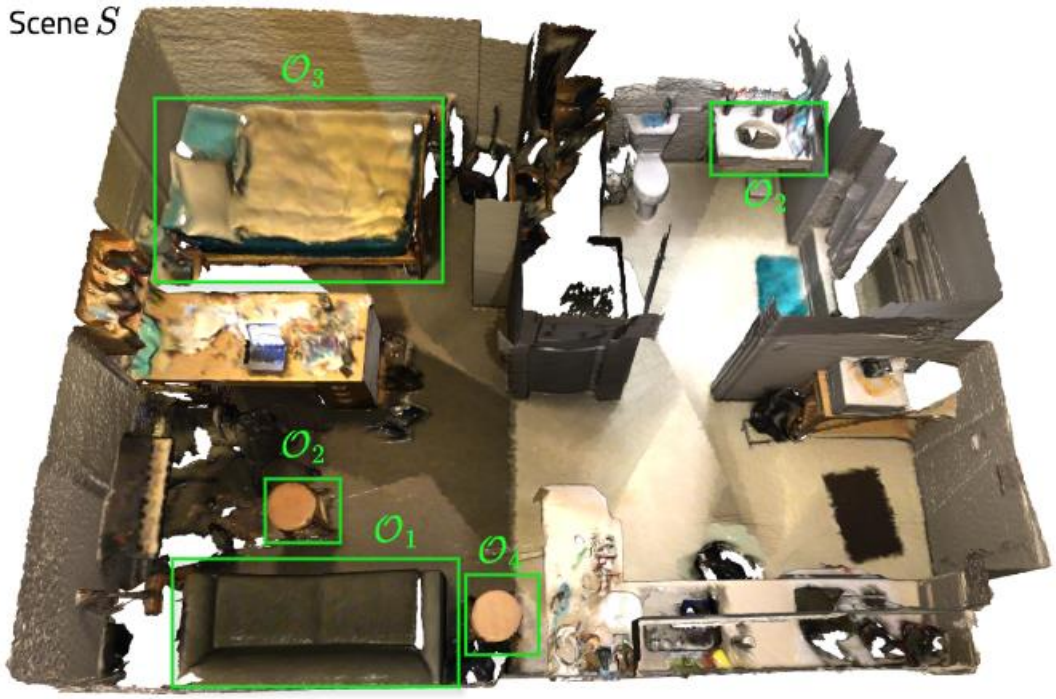
I. Instance-Level Alignment

to leverage large pre-trained models

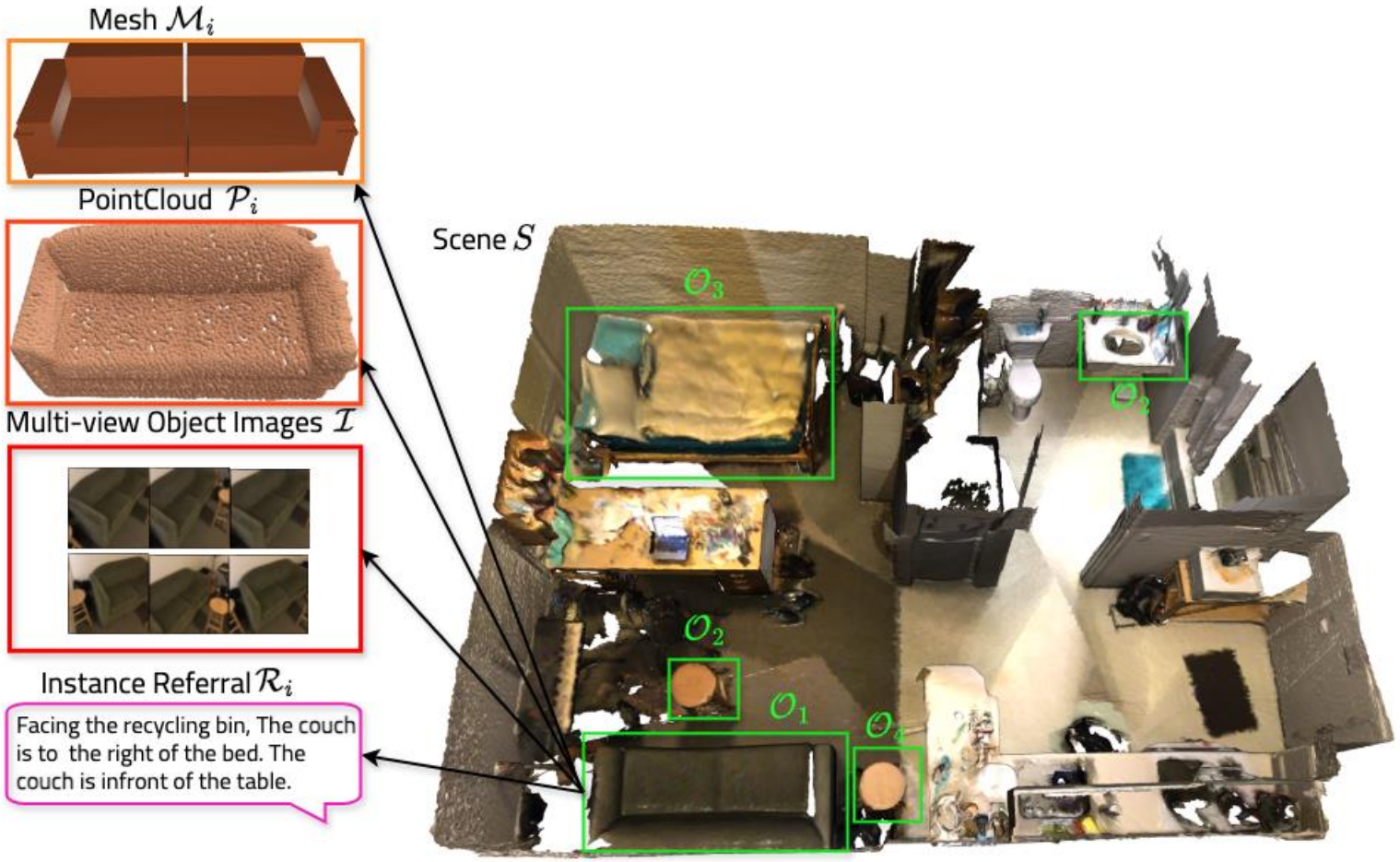
II. Scene-Level Alignment

From **rigid object-level alignment** to **flexible, modality-agnostic** scene understanding.

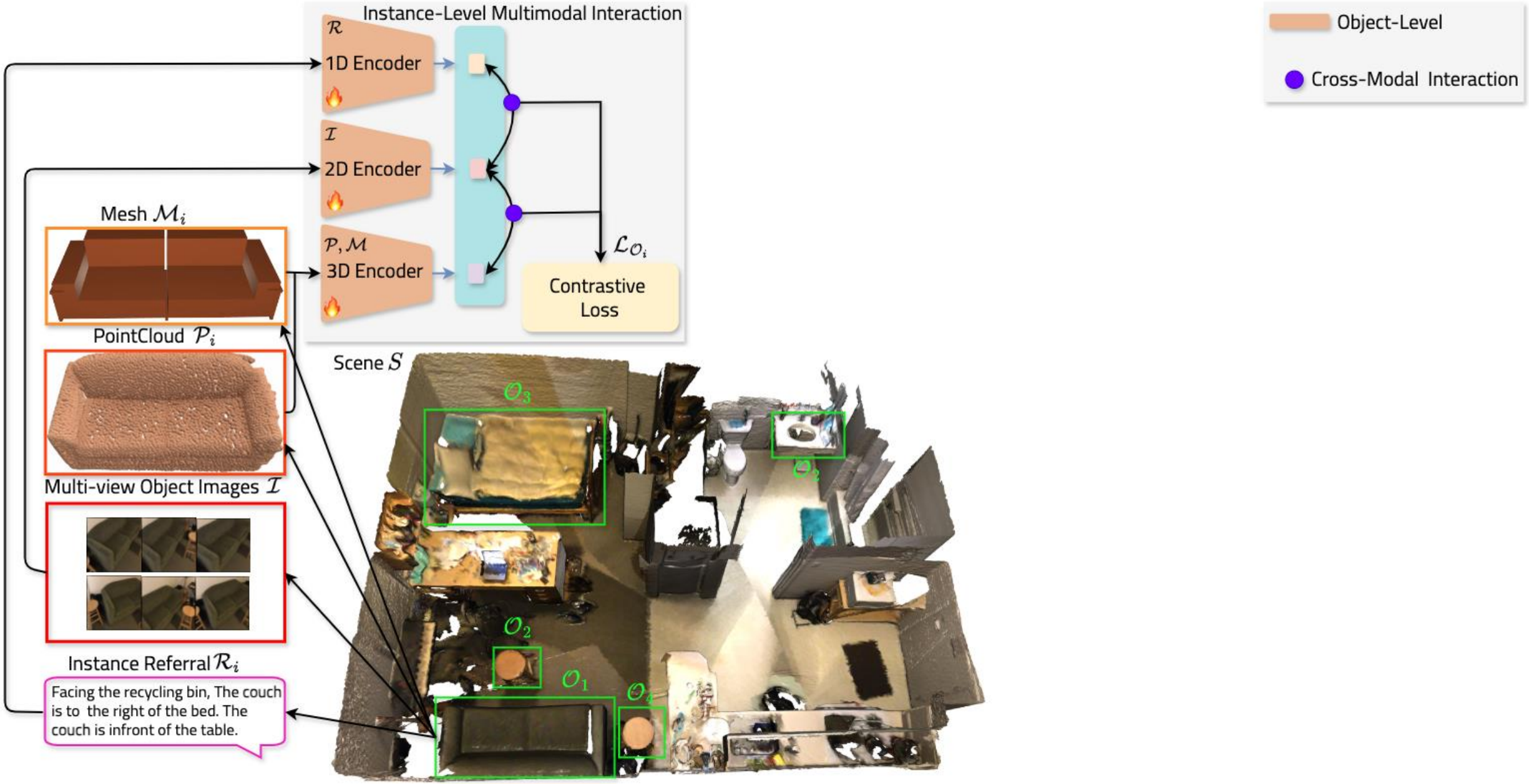
Overview of CrossOver



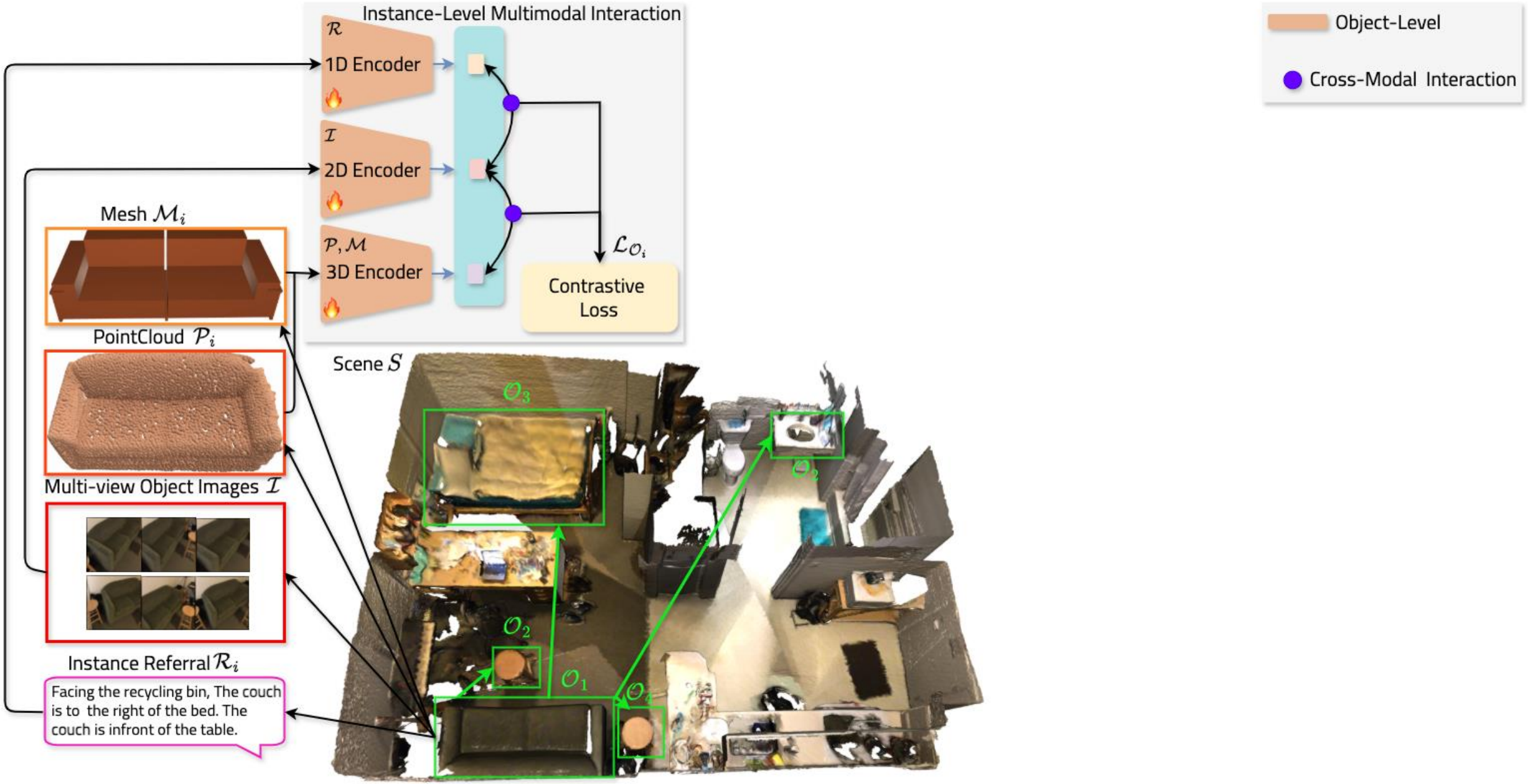
Overview of CrossOver



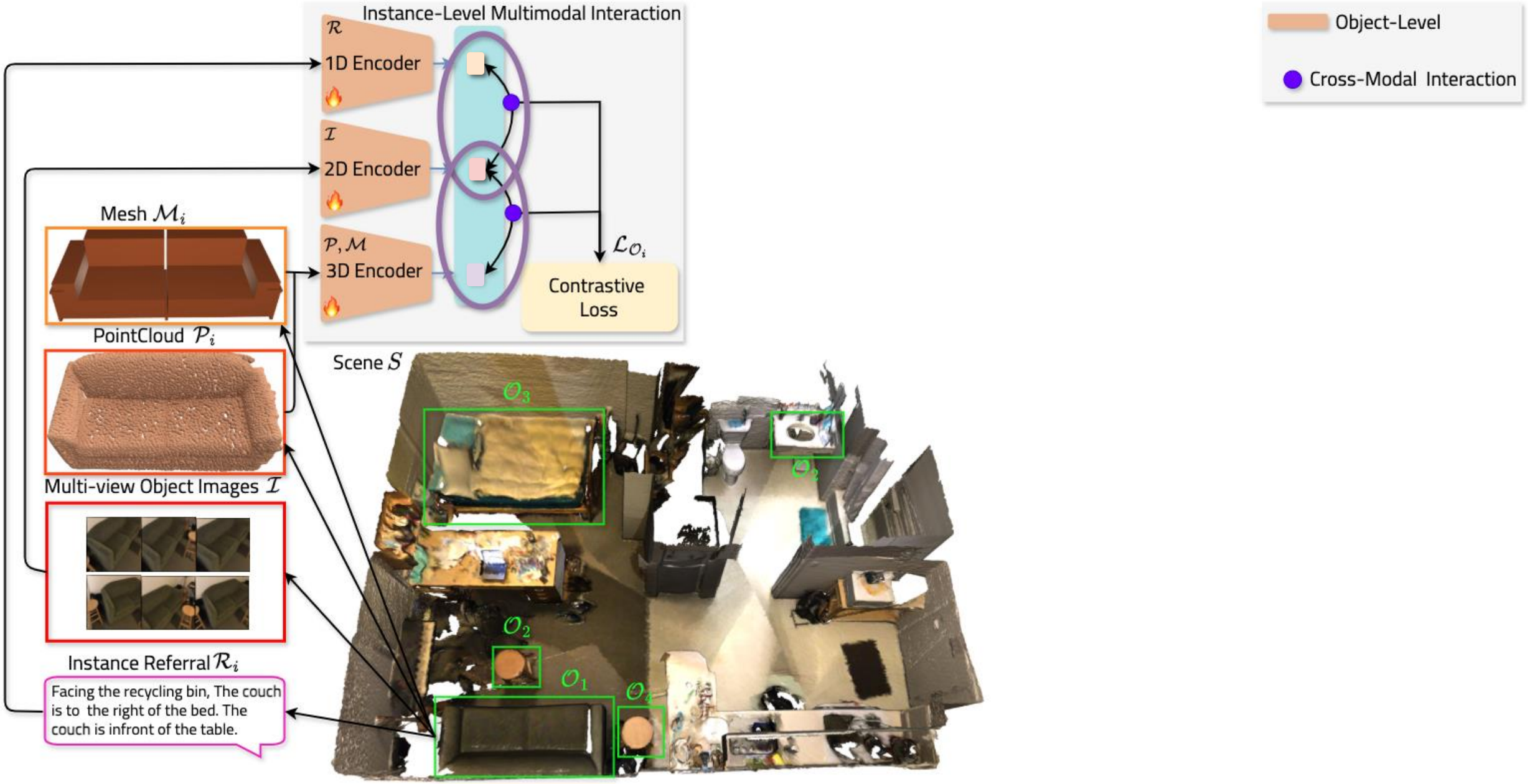
Overview of CrossOver



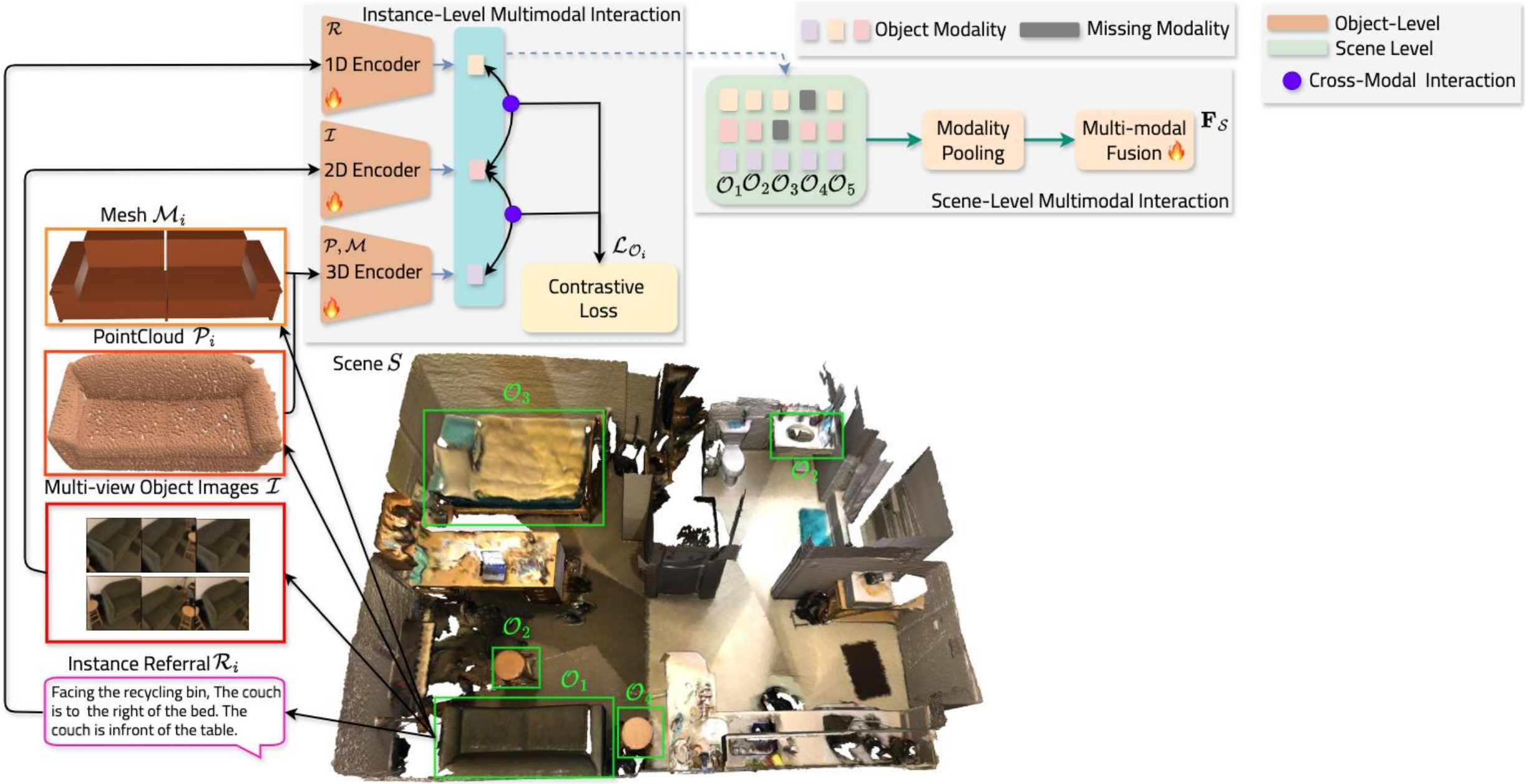
Overview of CrossOver



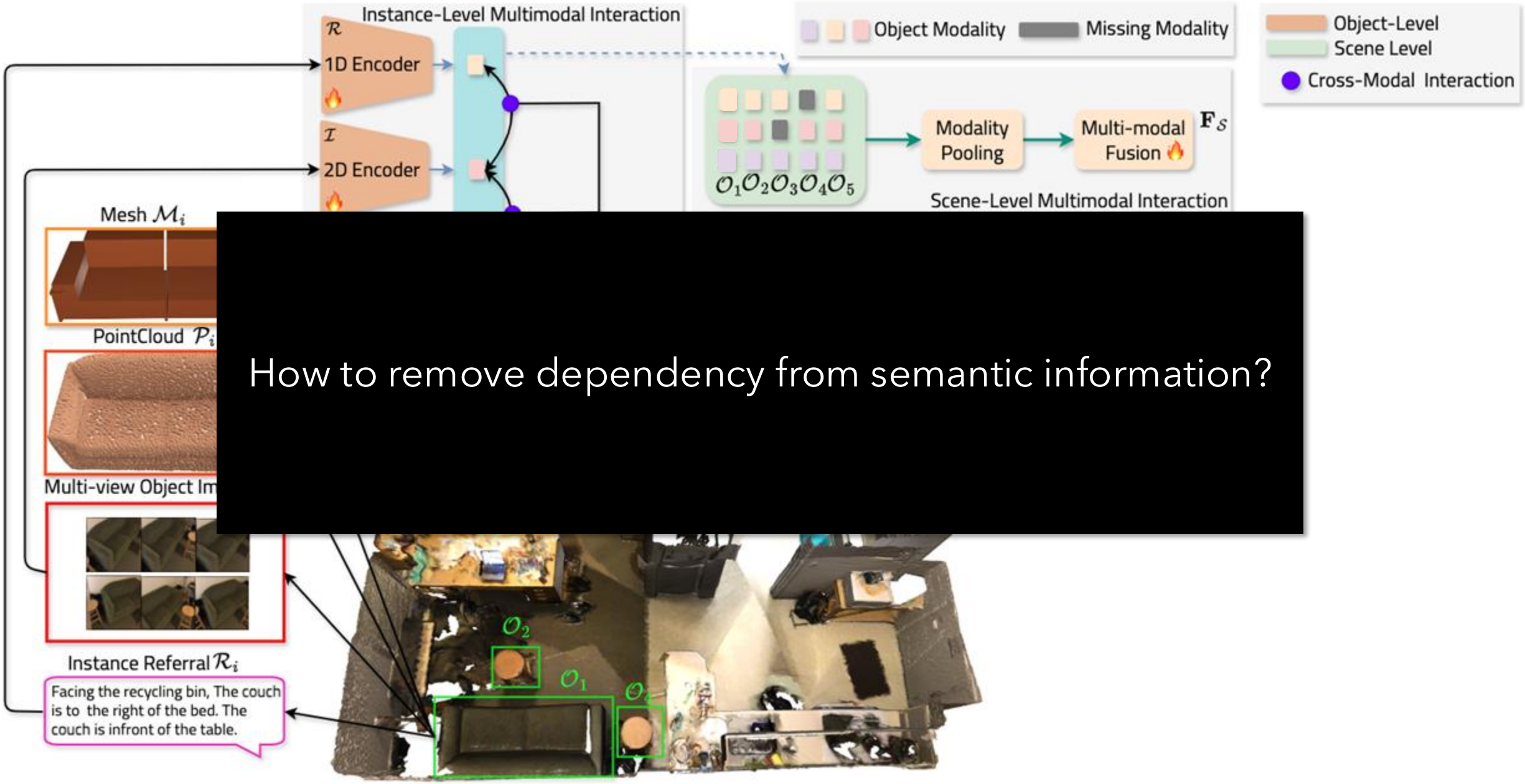
Overview of CrossOver



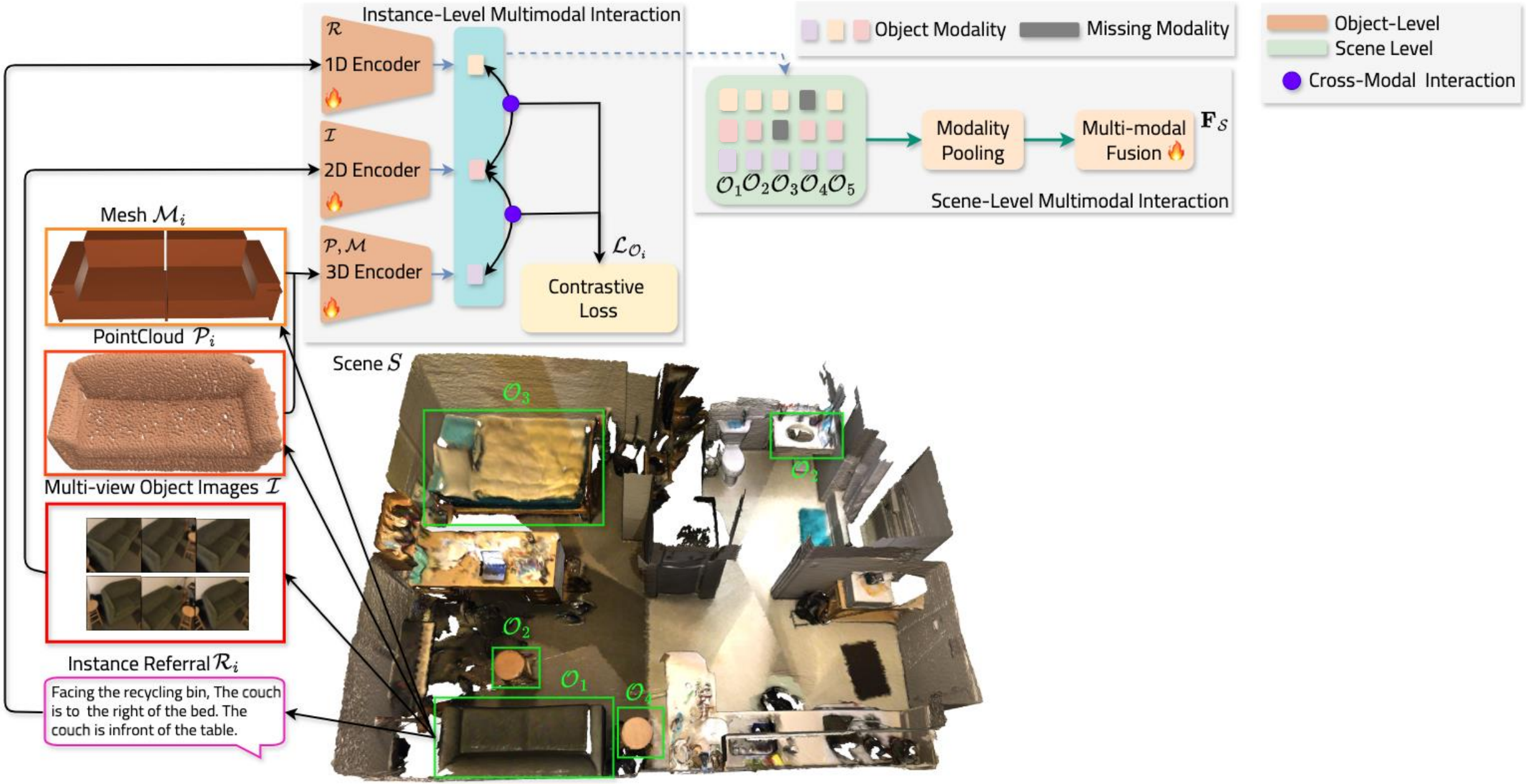
Overview of CrossOver



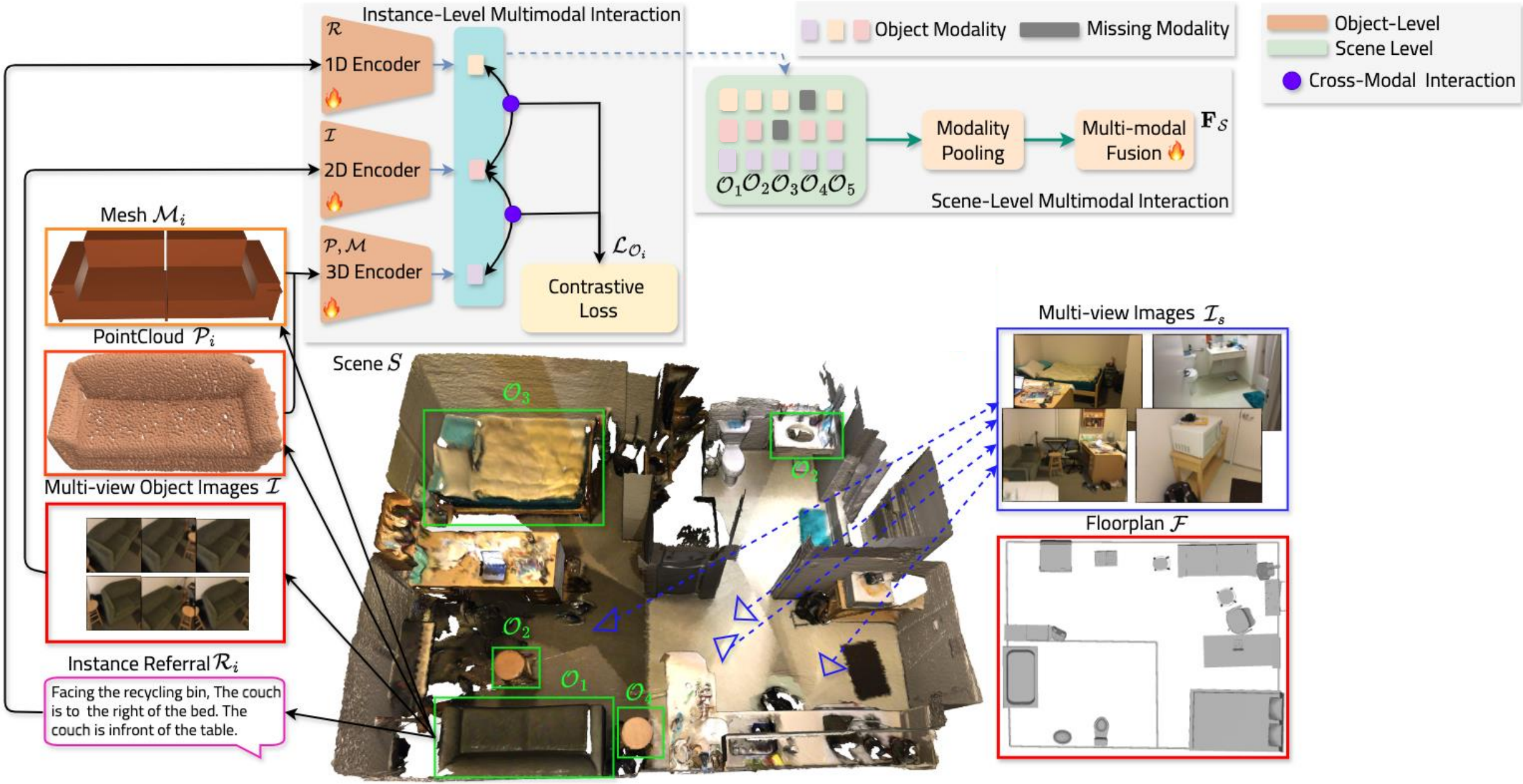
Overview of CrossOver



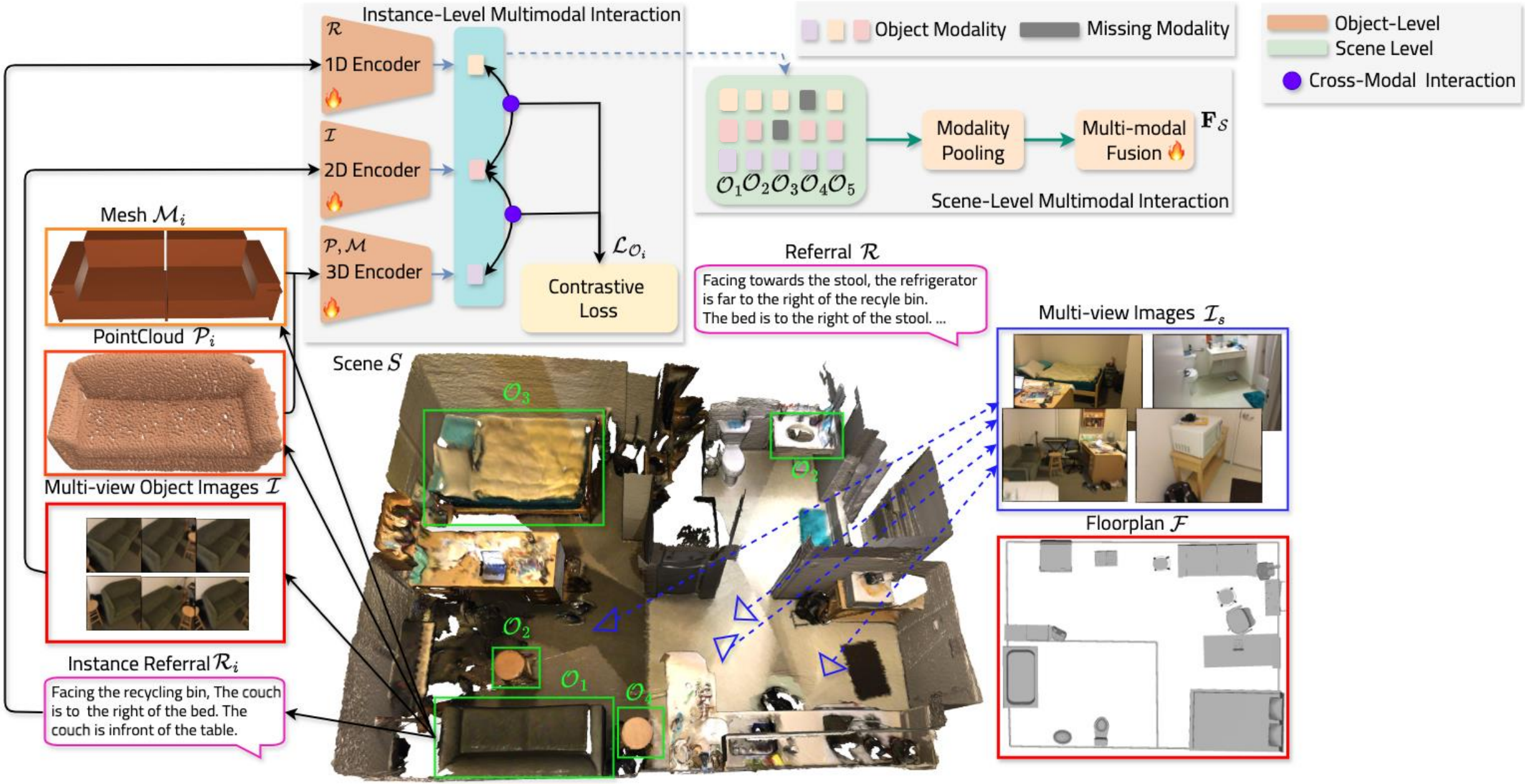
Overview of CrossOver



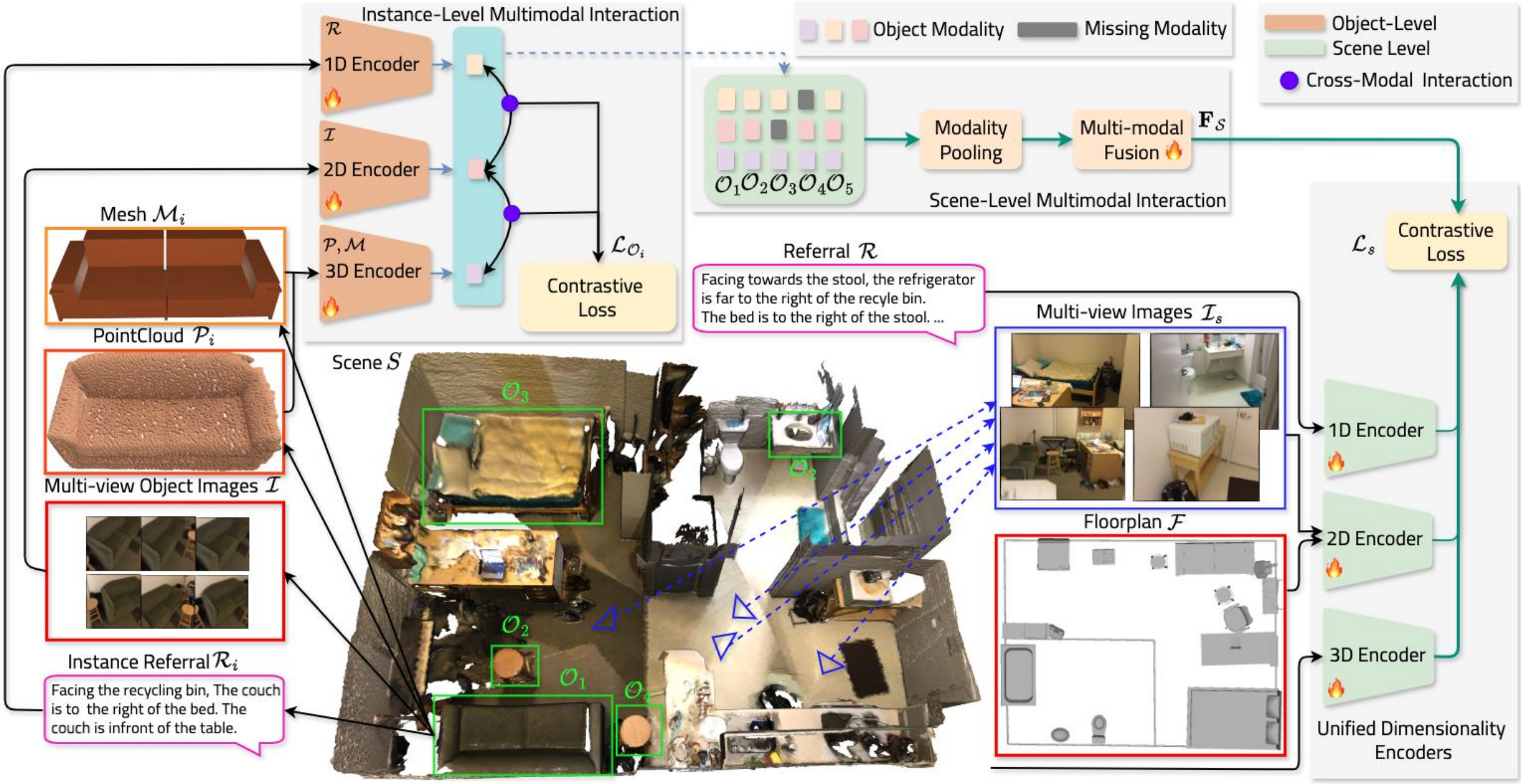
Overview of CrossOver



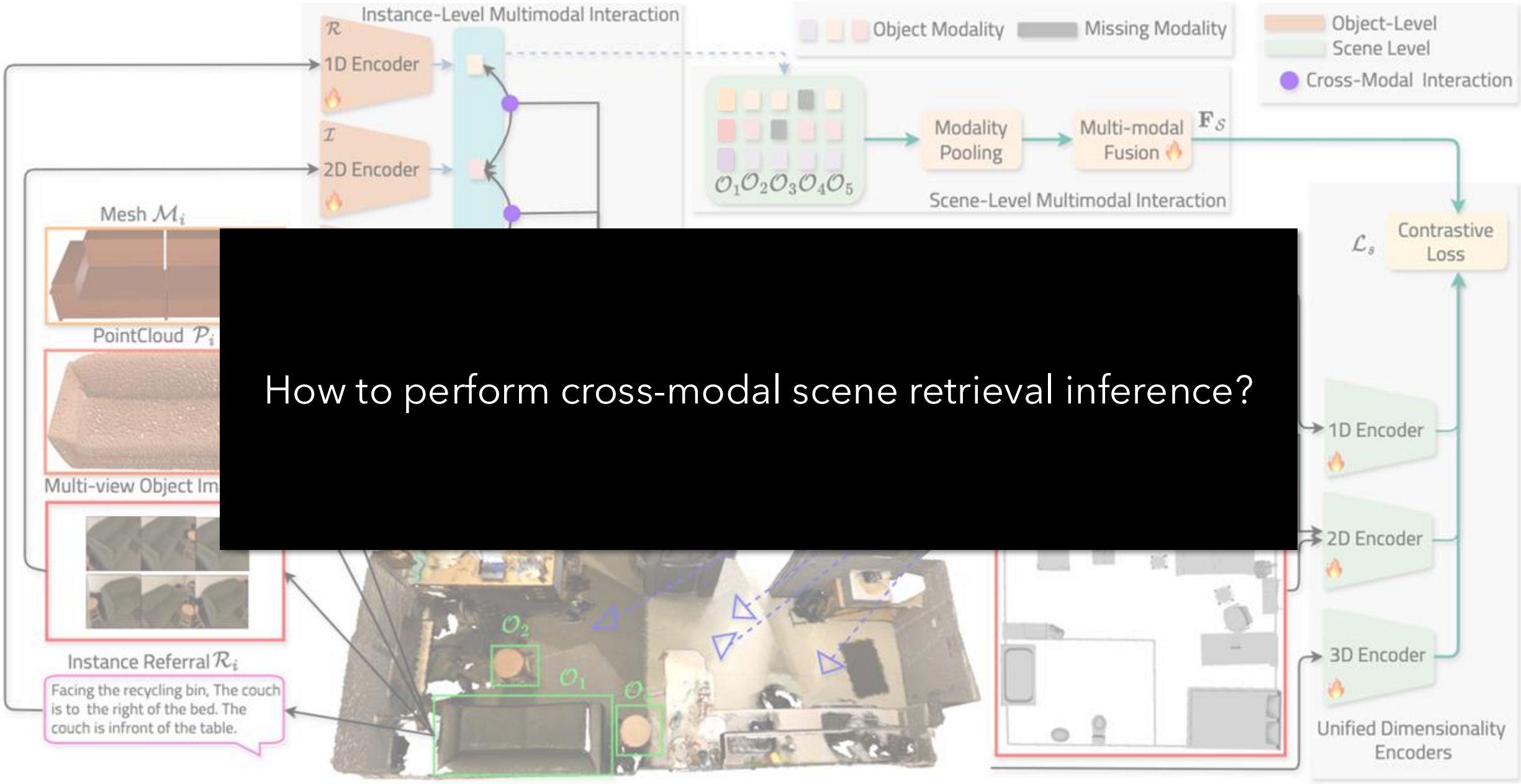
Overview of CrossOver



Overview of CrossOver

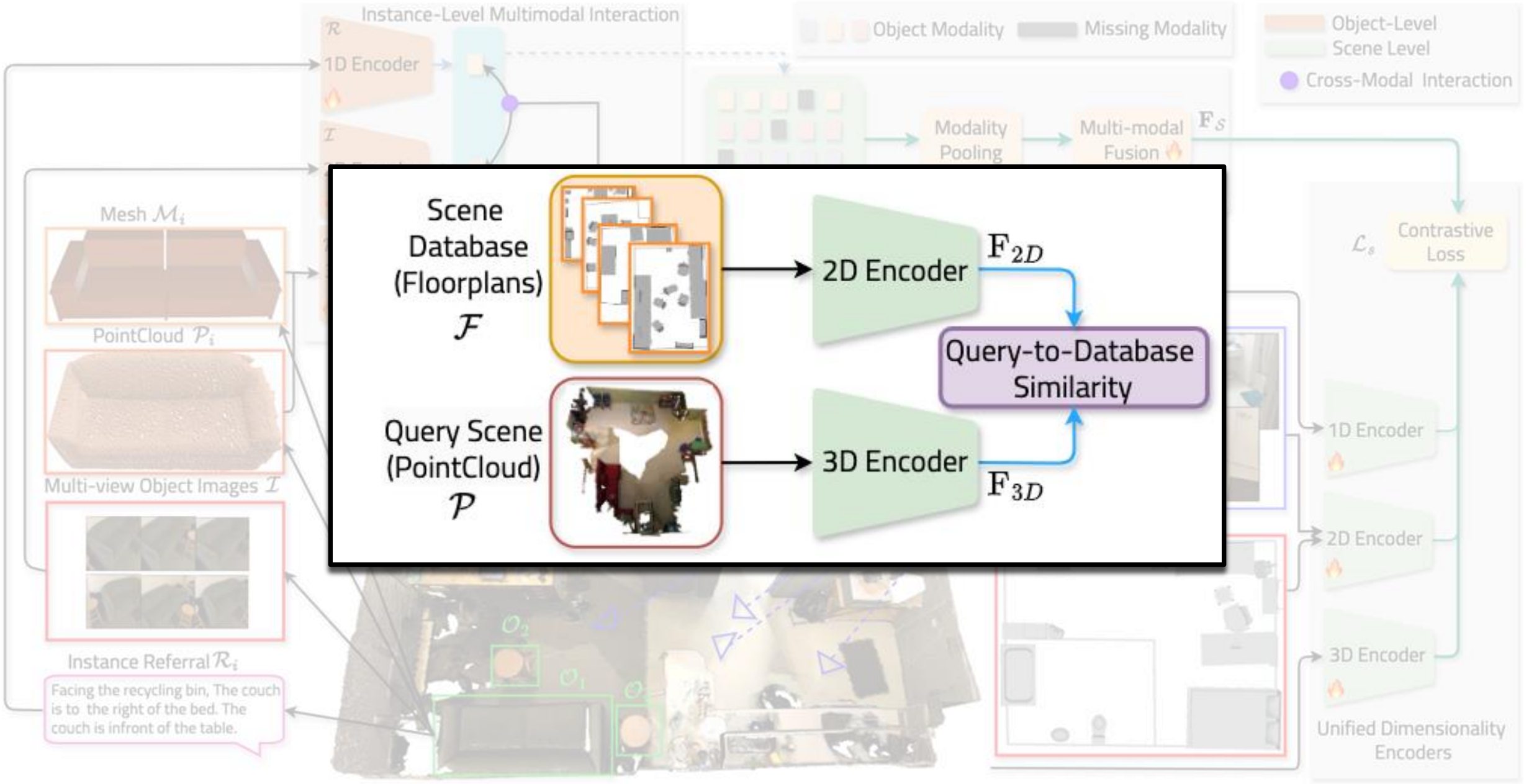


Overview of CrossOver



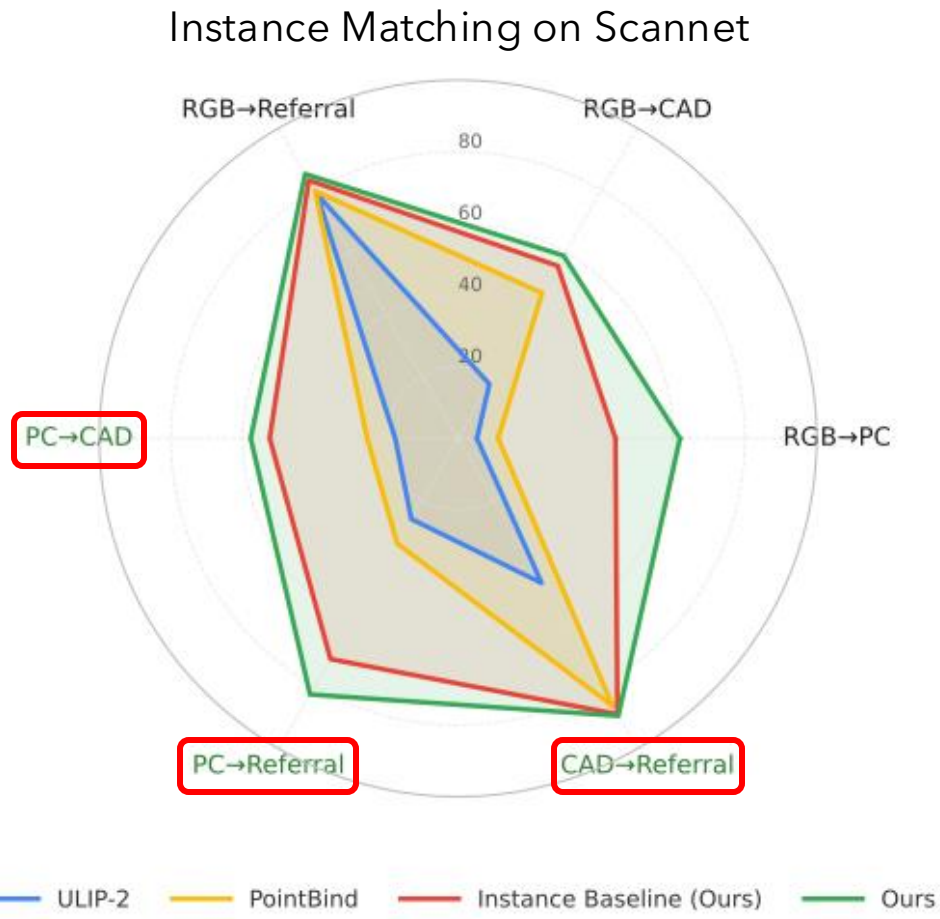
How to perform cross-modal scene retrieval inference?

Scene Retrieval Inference Pipeline



Experimental Results

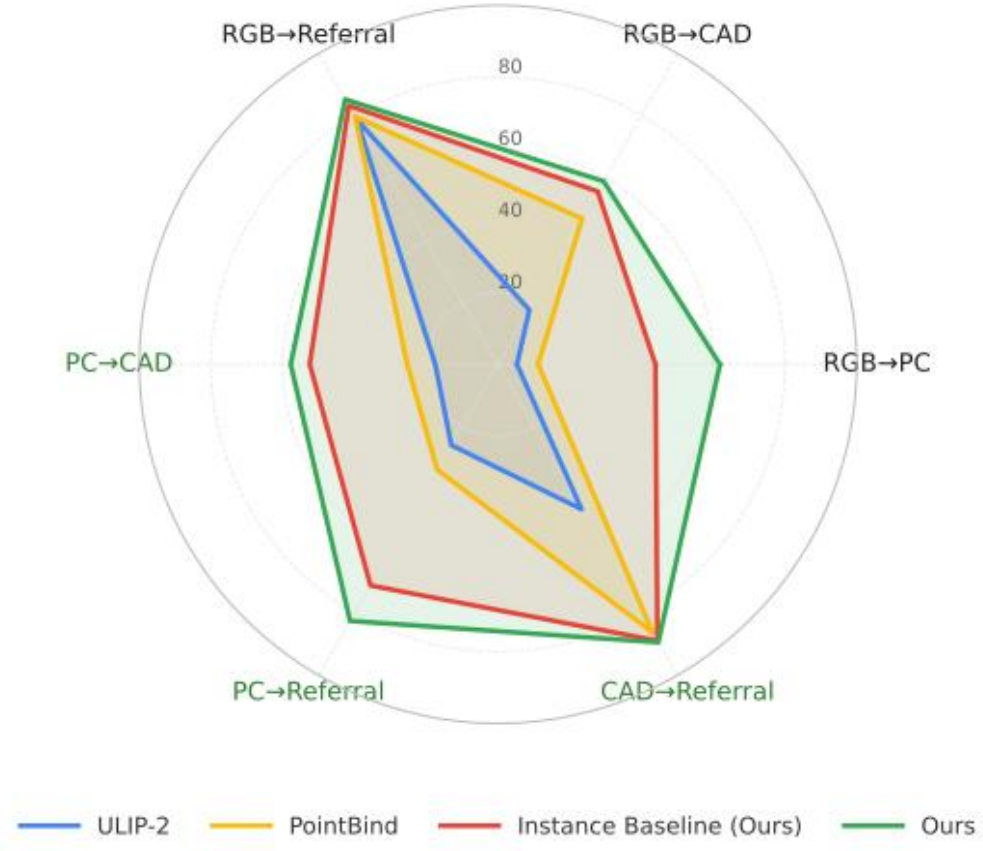
Cross-Modal Instance Matching



Experimental Results

Cross-Modal Instance Matching

Instance Matching on Scannet



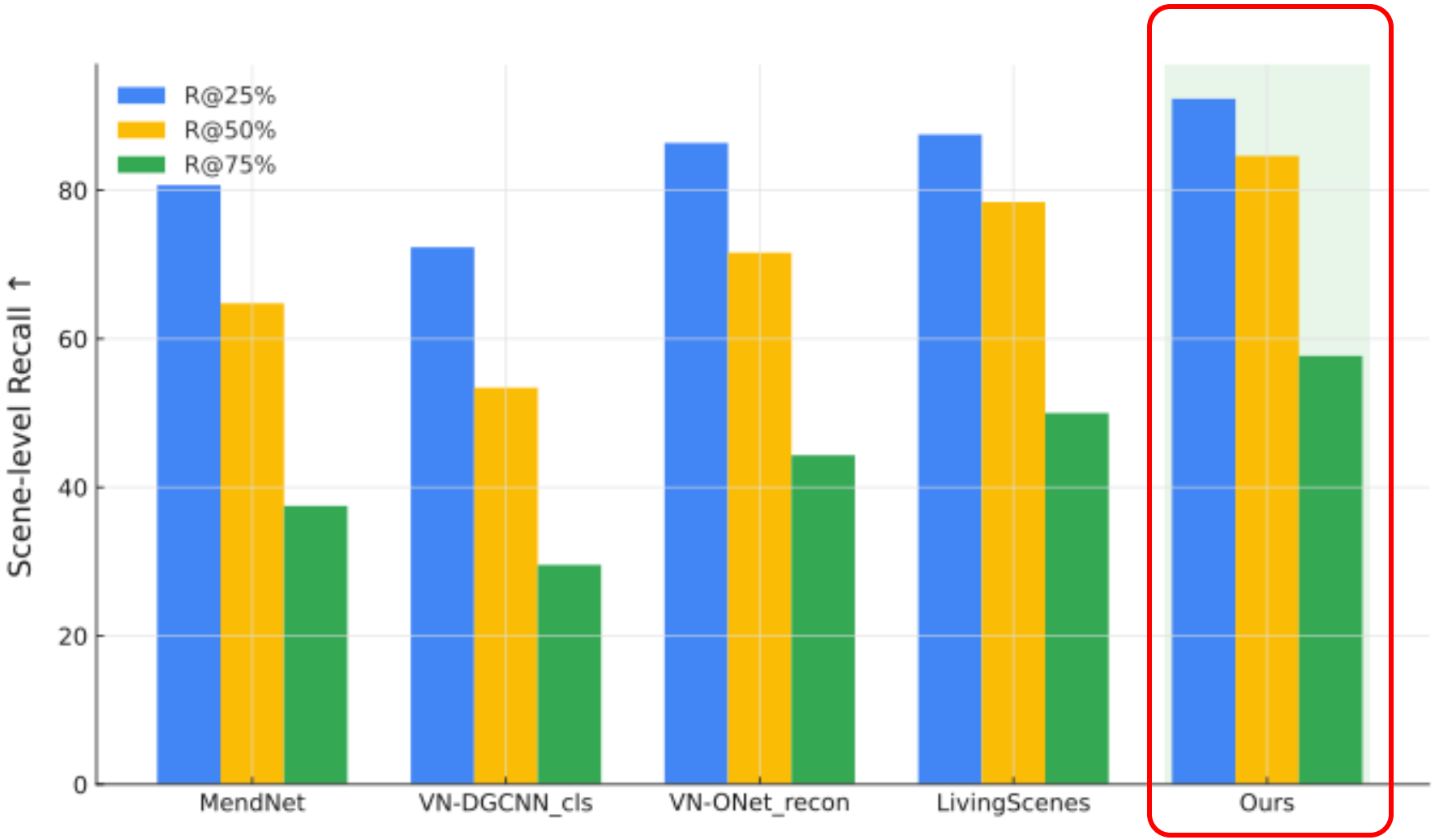
Scene Level Instance Matching

	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**

Emergent cross-modal understanding **without explicit pairwise supervision**.

Experimental Results

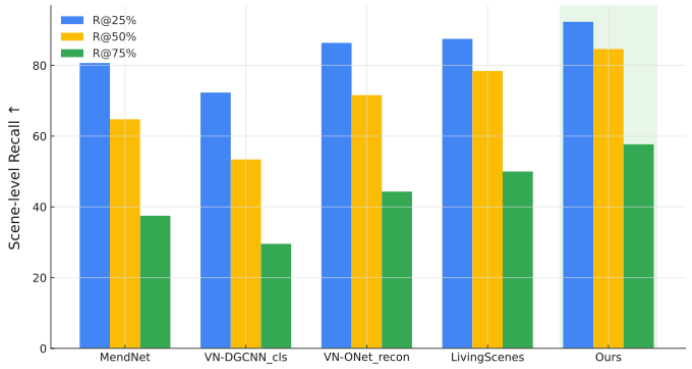
Temporal Instance Matching



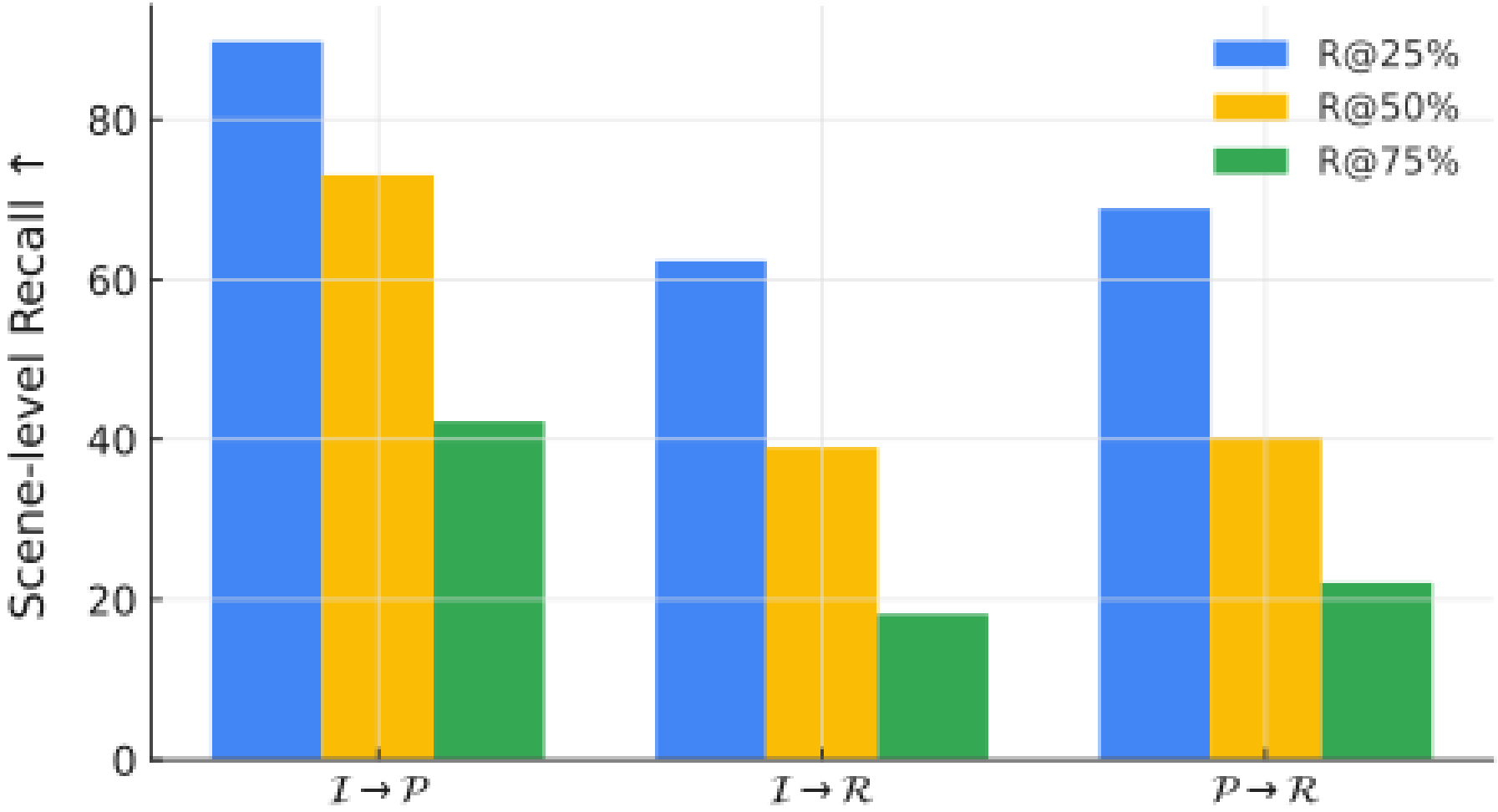
Better performance on instance matching, even without **explicit temporal training**.

Experimental Results

Temporal Instance Matching



Cross-Modal Temporal Recall

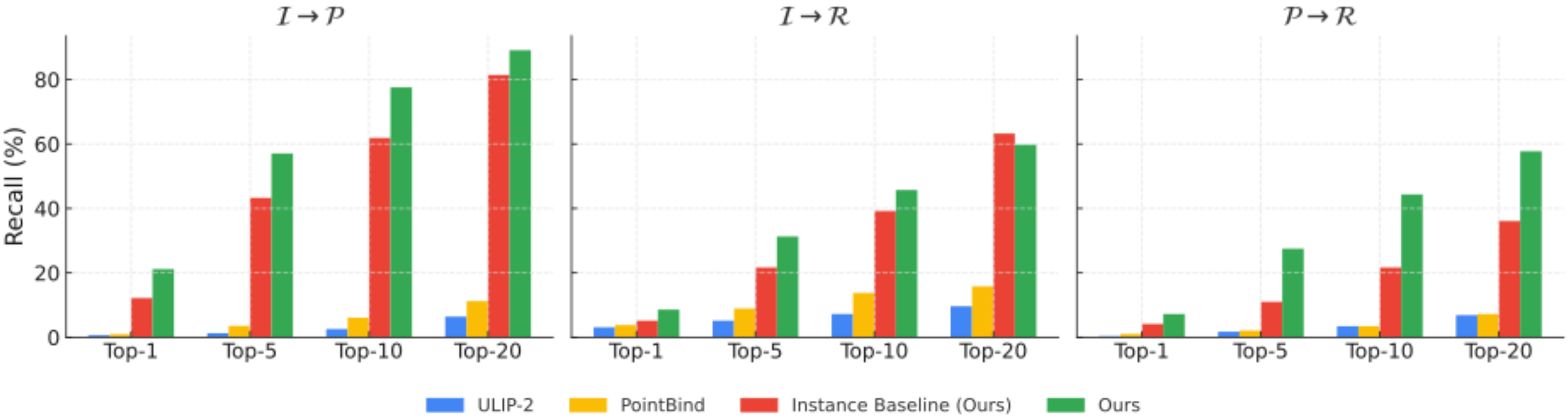


Strong cross-modal performance even with **object movements or removal across time**.

Experimental Results

Cross-Modal Scene Retrieval

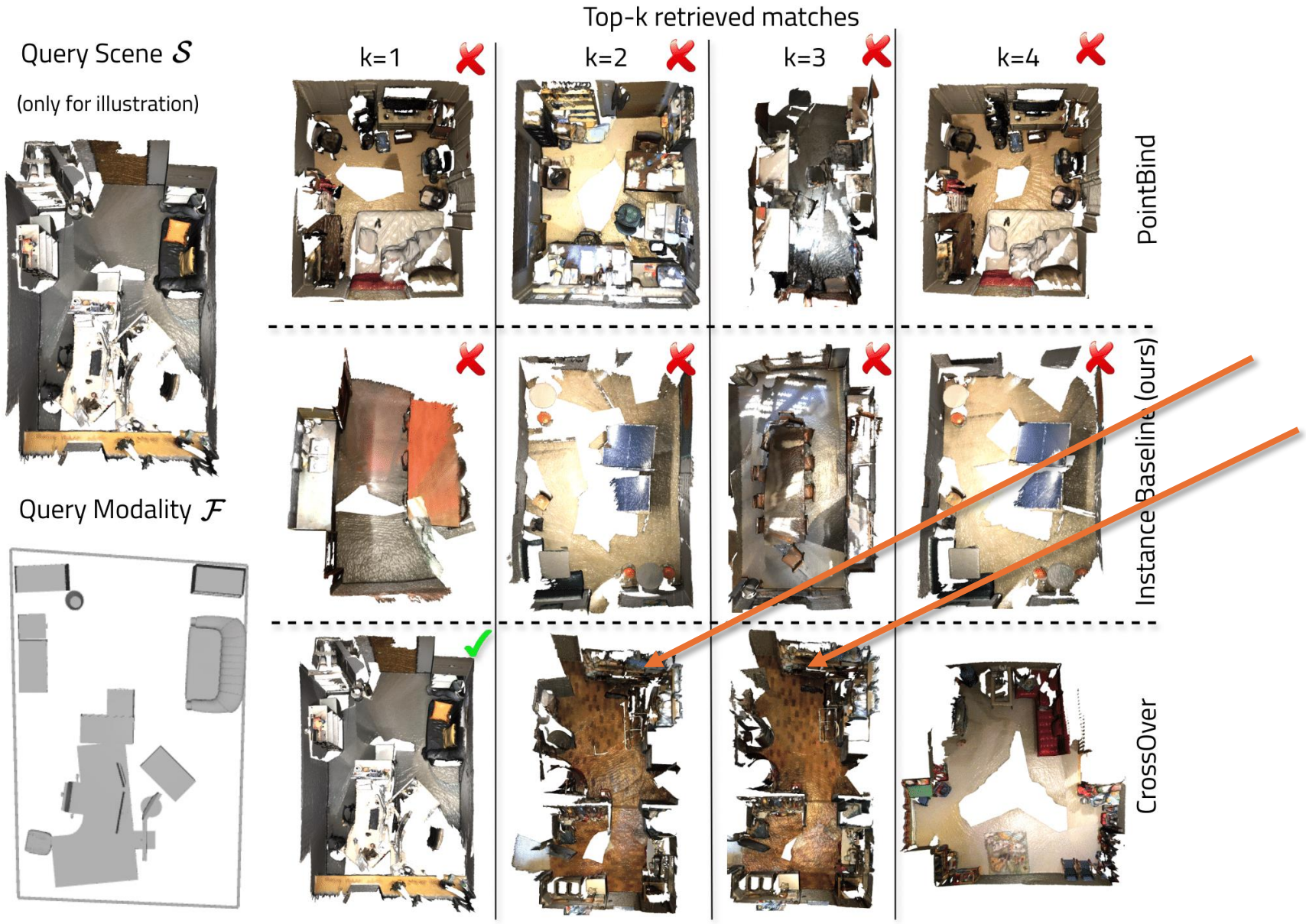
Scene matching recall of different methods on three modality pairs



Unified scene-level embeddings enable **robust cross-modal retrieval** without semantic supervision.

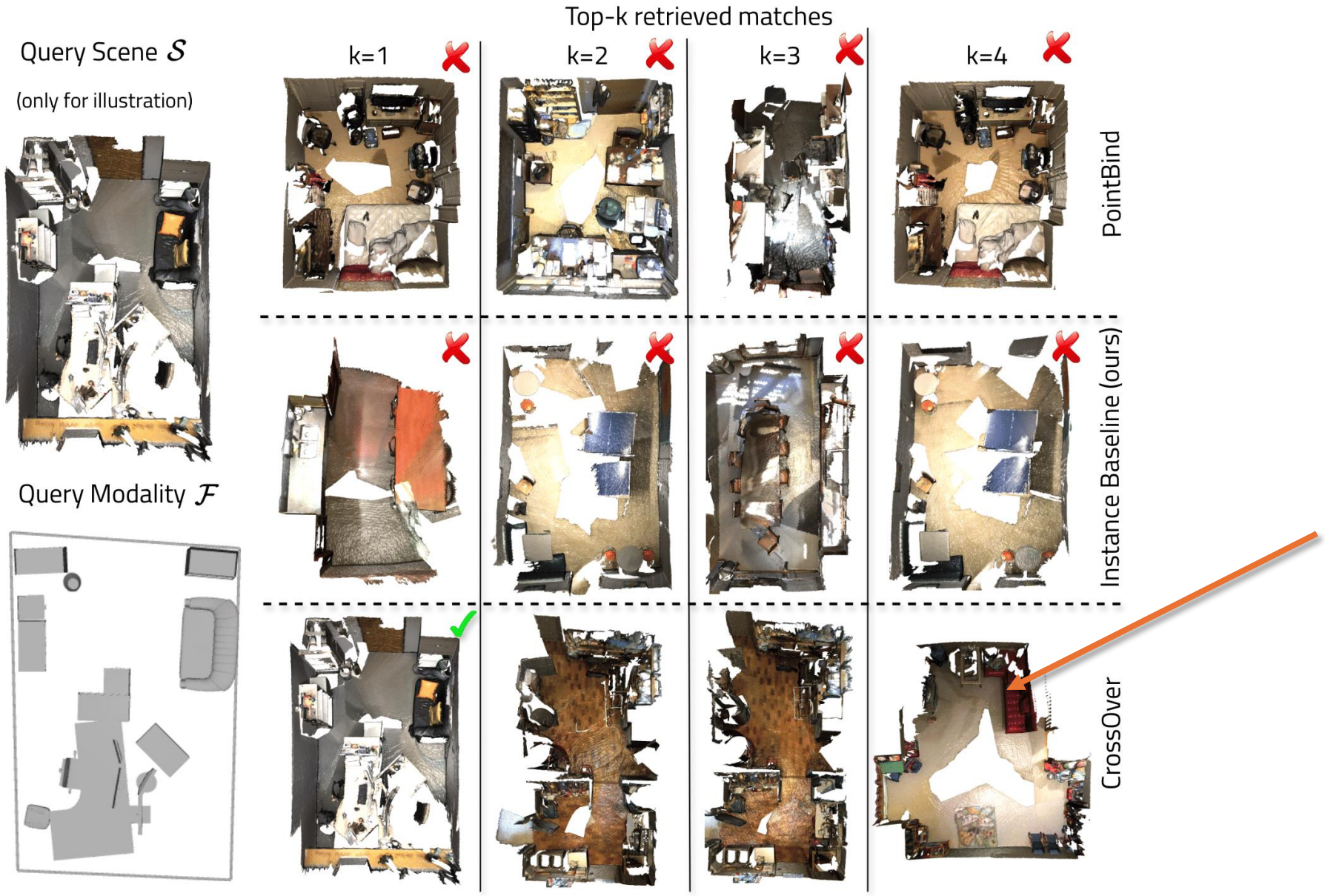
Experimental Results

Cross-Modal Scene Retrieval Visualization: Success



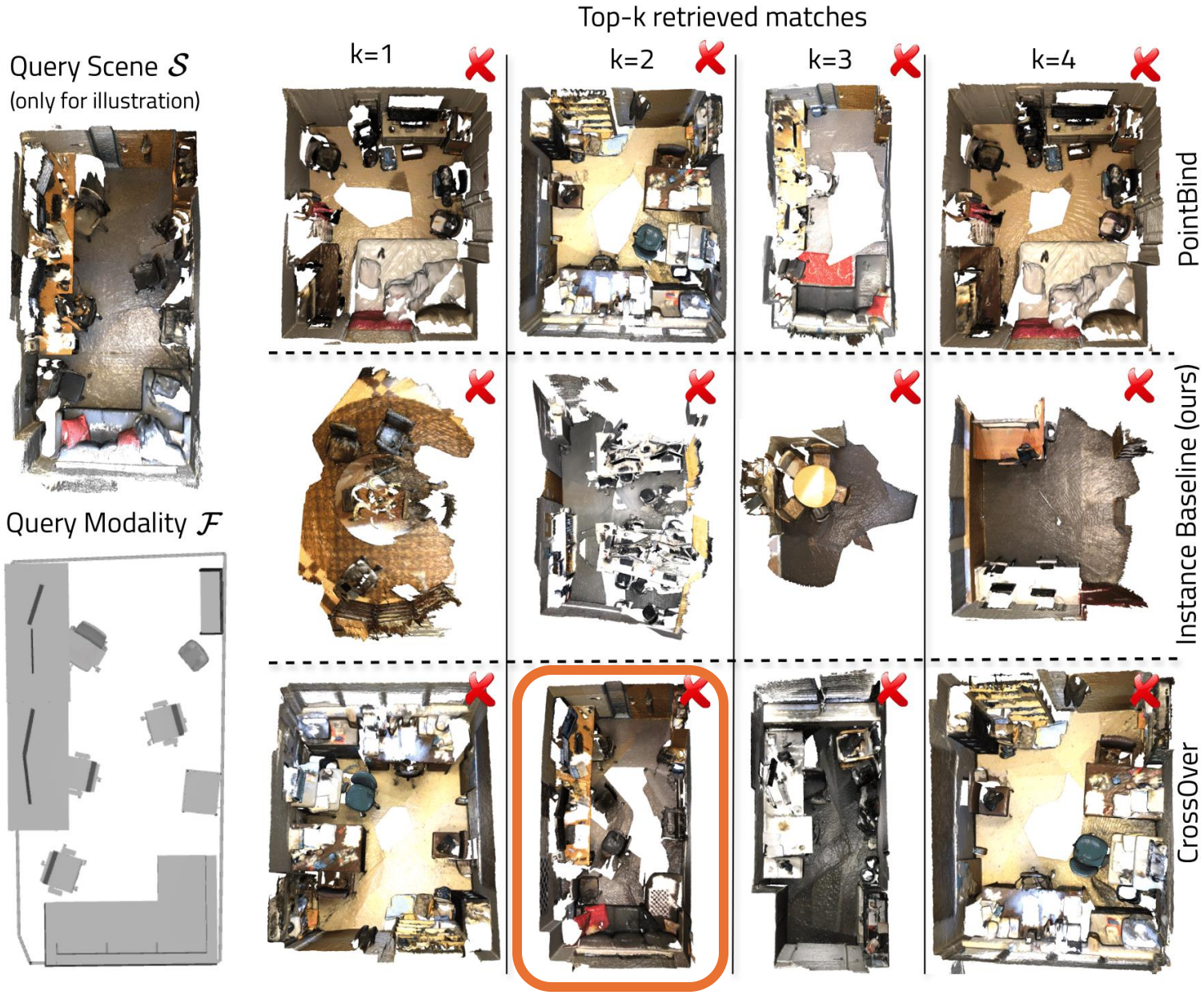
Experimental Results

Cross-Modal Scene Retrieval Visualization: Success



Experimental Results

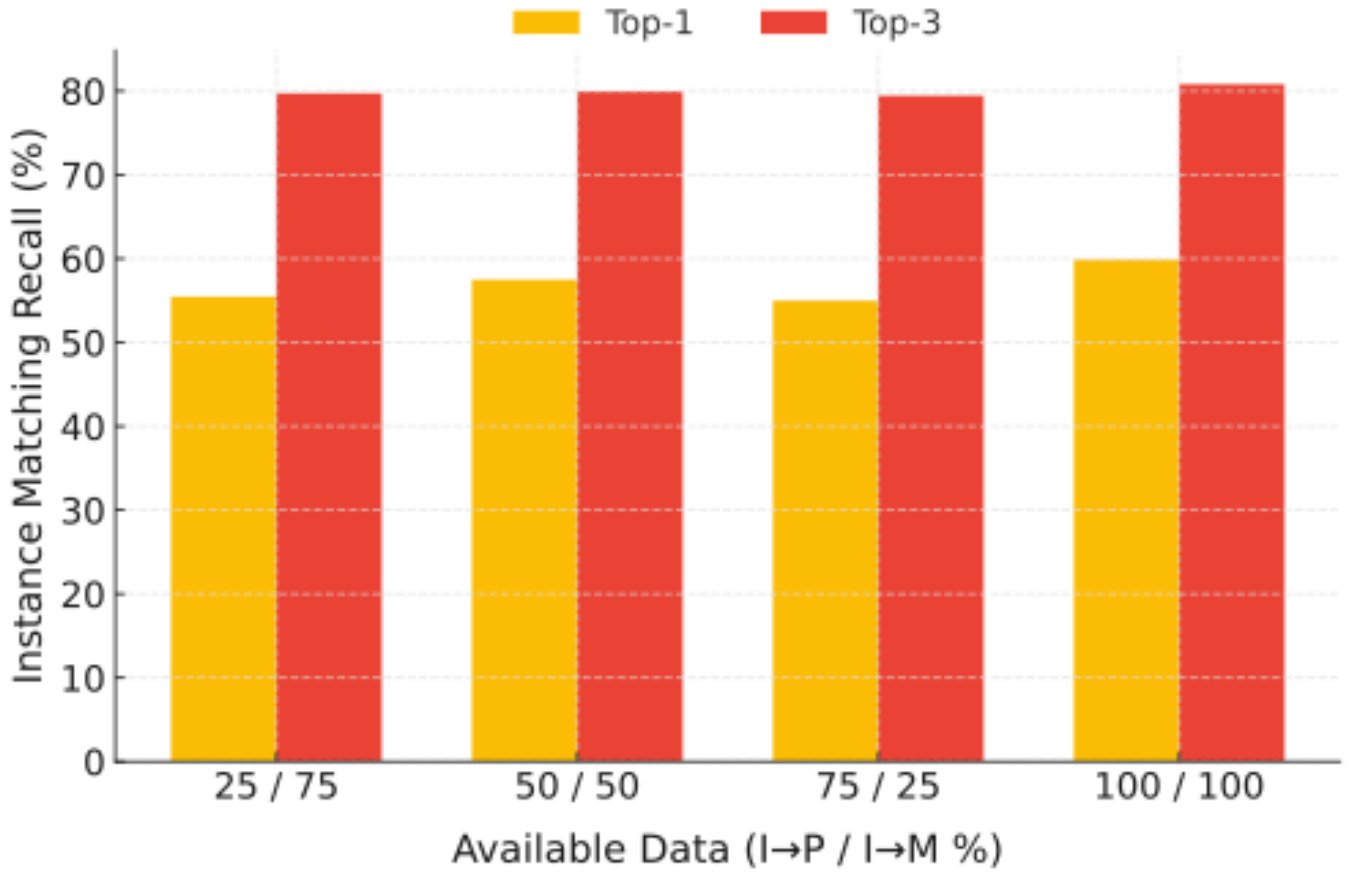
Cross-Modal Scene Retrieval Visualization: Failure



Experimental Results

Missing Modalities

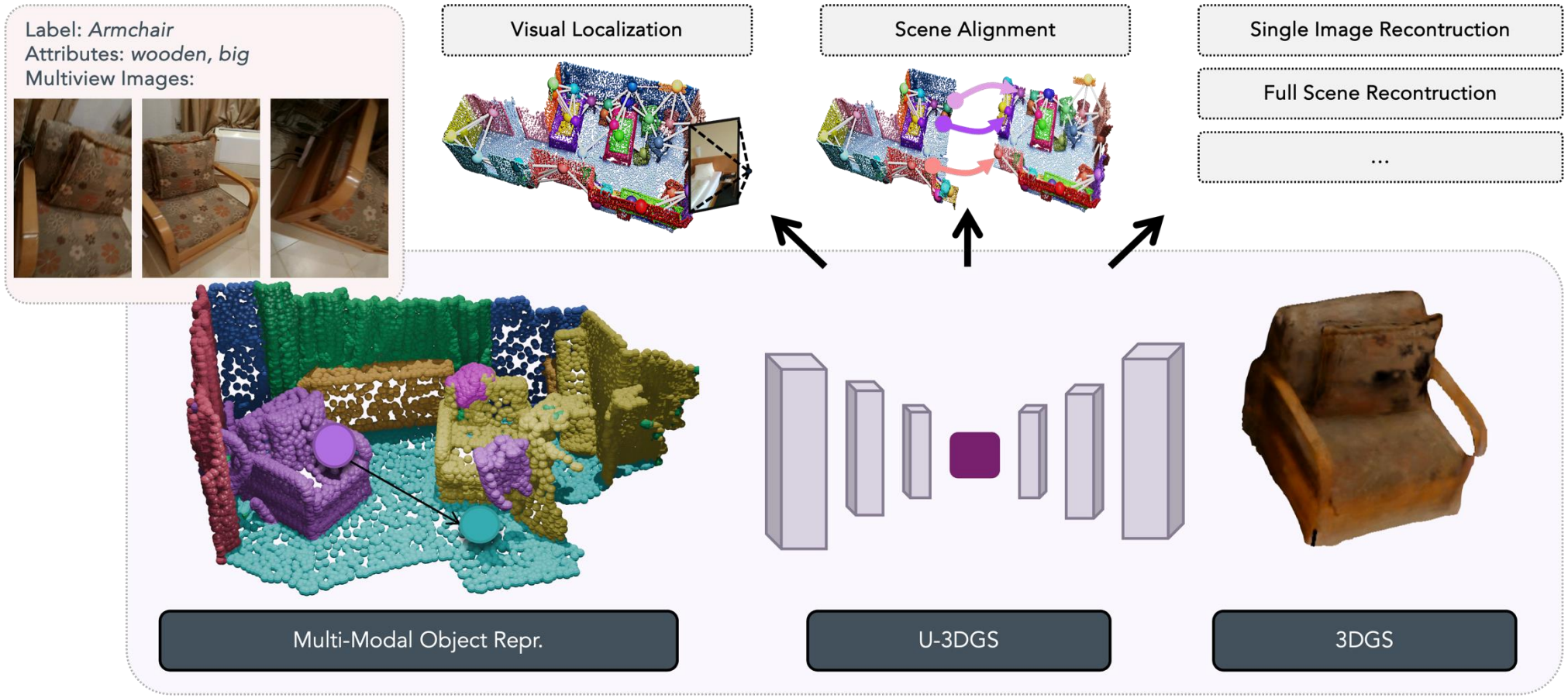
Ablation on instance matching with *non-overlapping data per modality pair*



Even with **disjoint data distribution** across modalities, CrossOver learns **robust alignment**.

Beyond CrossOver

Learning To Reconstruct Multi-modal 3D Representation



Object-X Builds on CrossOver to Achieve Unified **3D Object Reconstruction Across Modalities.**

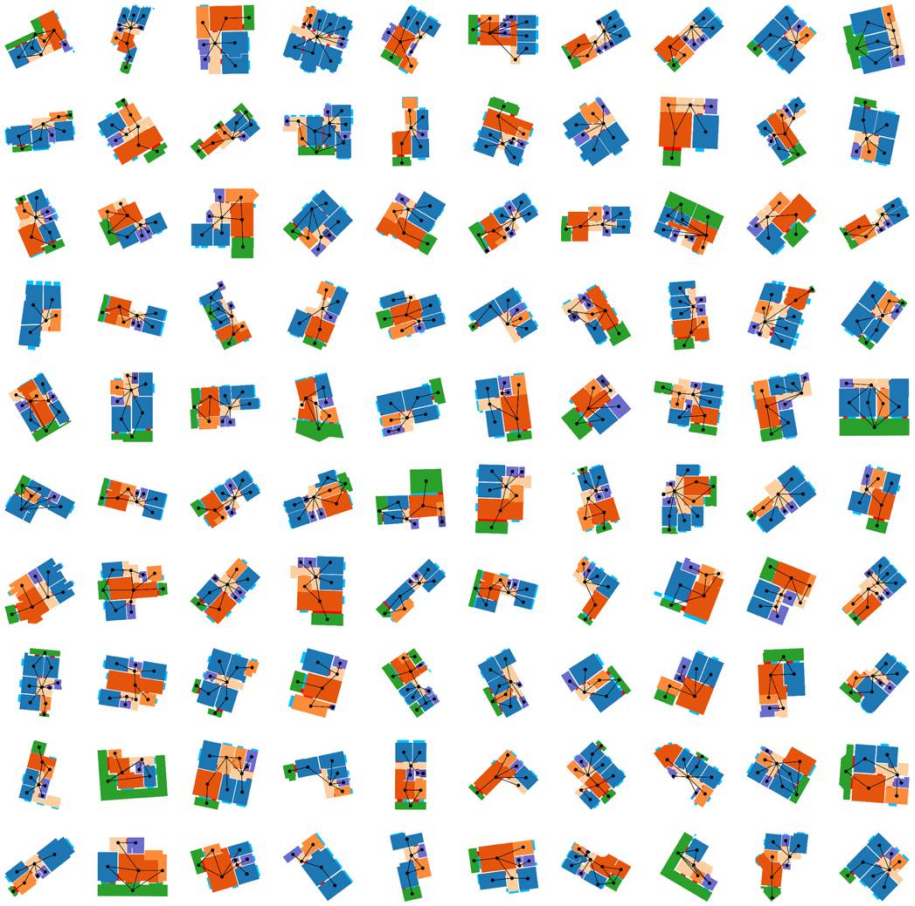
[1] Lorenzo et al, *Object-X: Learning to Reconstruct Multi-Modal 3D Object Representations*, NeurIPS 2025

Beyond CrossOver

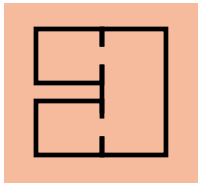
Learning Multi-modal Floorplan Alignment and Re-ranking

Work in Progress
Acknowledgement: Casper Van Engelenburg

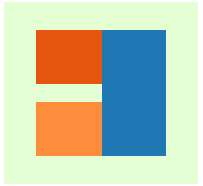
Can we leverage the same framework to learn matching + ranking?



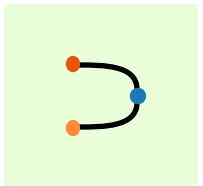
Images



Geometries



Graphs



Texts

"3 room apartment, with spacious living room …"

FloorBind

Cross modal retrieval **(matching)**
Spatially-aware search **(ranking)**

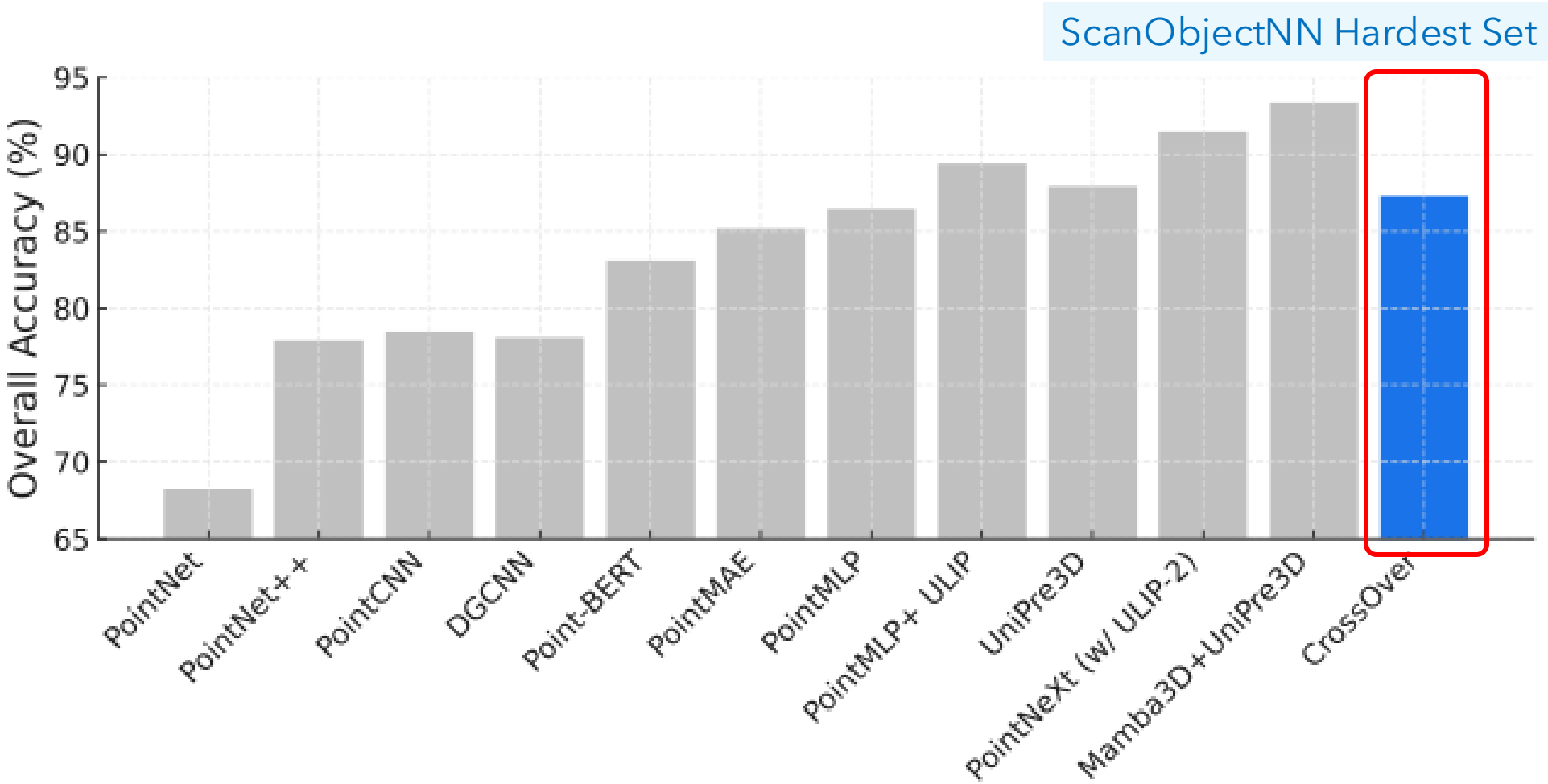
Goal: Build on CrossOver to multi-modal floorplan retrieval and re-ranking.

Beyond CrossOver

Downstream Task(s): 3D Object Recognition

Work in Progress
Acknowledgement: Gaurav Pradeep

Can we probe the embedding space to perform downstream tasks, eg, **3D object classification?**



Goal: Scaling CrossOver into a Unified 3D Multi-Modal Learning Framework

Key Takeaways

- Flexible scene-level alignment framework that connects heterogeneous 3D modalities without requiring perfect data pairing or semantic annotations.
- Leveraging dimensionality-specific encoders and a progressive training pipeline, CrossOver achieves emergent cross-modal behavior and robust generalization across unpaired modalities.

Unifying 3D scene modalities for scalable, semantic-free cross-modal alignment.

Future Directions

- How can we scale CrossOver for large-scale scene understanding under noisy and incomplete real-world data?
- How can CrossOver be extended to dynamic scene reconstruction and real-time navigation for immersive mixed-reality experiences?

Thank You!