CVPR Nashville JUNE 11-15, 2025

CrossOver: 3D Scene Cross-Modal Alignment

Sayan Deb Sarkar[1] Ondrej Miksik[2] Marc Pollefeys[2, 3] Daniel Barath[3, 4] Iro Armeni[1]

1 Stanford University, Gradient Spaces 2 ETH zürich, CVG Computer Vision and Geometry Lab 3 Microsoft 4 HUN REN SZTAKI

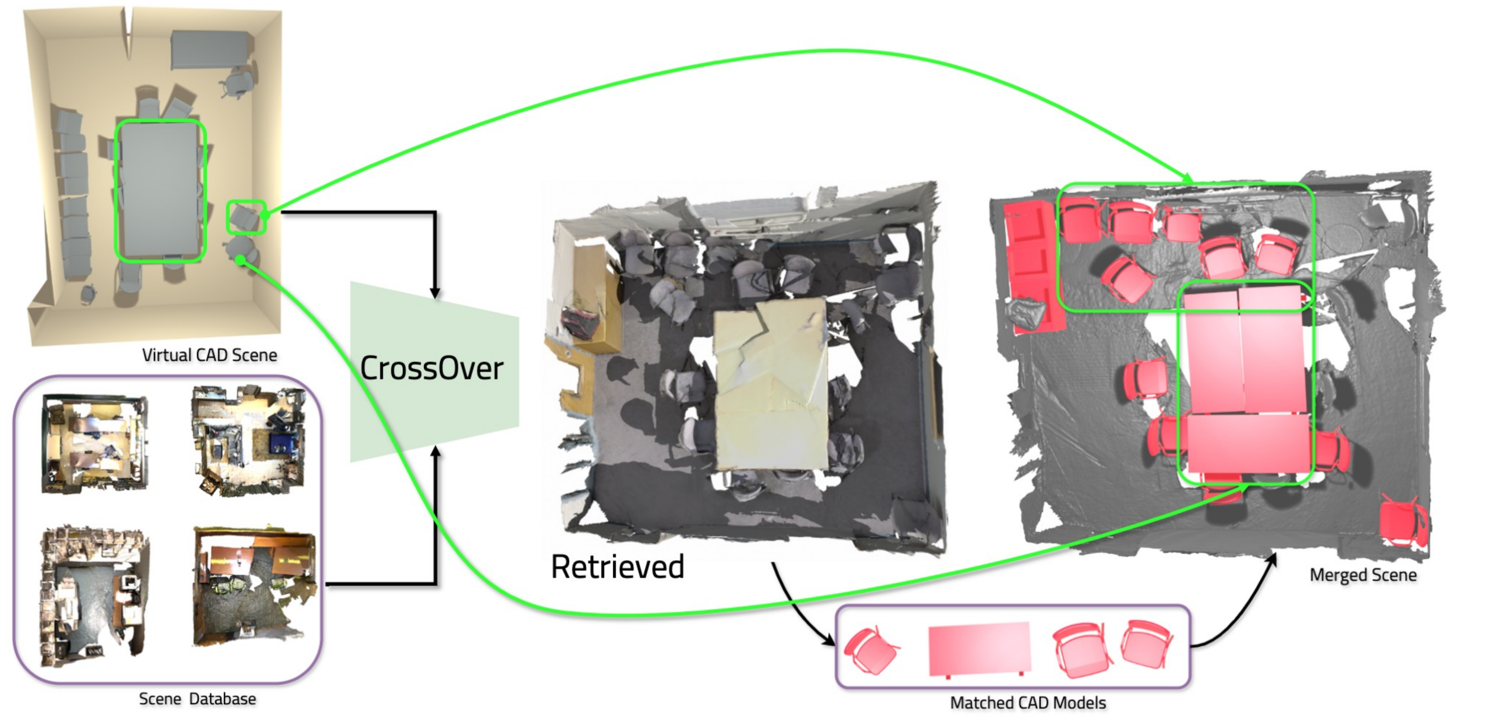
1. Problem Statement

Input Multi-modal scene representations including images, point clouds, CADs, floorplans & text.

Goal Cross-modal object-level and/or scene-level alignment.

Current Challenges [1, 2]

- Assume all data modalities are perfectly aligned and complete.
- Designed for isolated objects, not real-world 3D scenes, with incomplete or misaligned data.



2. Key Points

Research Questions:

- ➢ How can we align diverse 3D scene modalities without requiring complete or tightly matched data across modalities?
- ➢ Can we enable cross-modal understanding by leveraging scene context without relying on semantic annotations?
- ➢ Is it possible to learn cross-modal relationships that emerge naturally, even when certain modality pairs are never seen together during training?

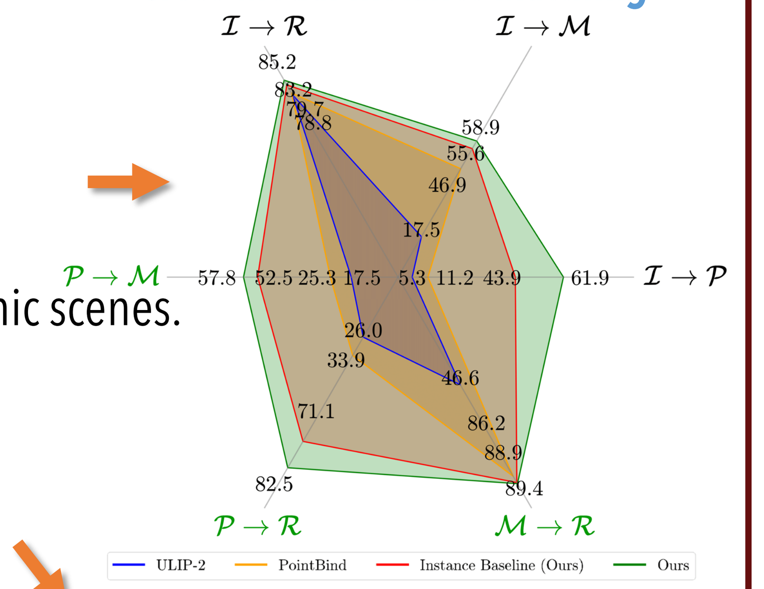
Contributions:

- ✓ Learns a unified feature space across RGB, point cloud, CAD, floorplan, and text without needing every modality to be present during training.
- ✓ 1D/2D/3D encoders tailored to each modality's dimensionality, removing the need for explicit 3D scene graphs or semantic labels during inference.
- ✓ Progressive training builds from object-level to scene-level embeddings, promoting emergent cross-modal behaviour.

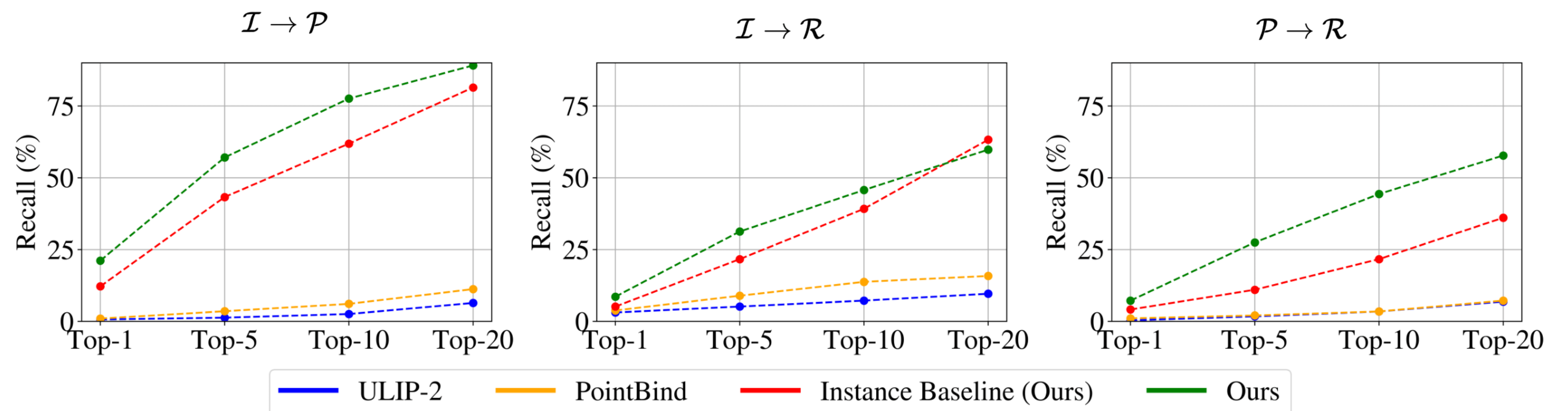
4. Experimental Results

- Datasets: ScanNet [3] and 3RScan [4], RGB-D video sequences of ~2.9K scans with *CAD+floorplans* from Scan2CAD [5] & *text referrals* from SceneVerse [6].
- Excels at **emergent cross-modal instance matching** across unseen modality combinations.
- Performs well on **temporal instance retrieval** capturing spatial and geometric relationships in dynamic scenes.
- Consistently better at **cross-modal scene retrieval using unified, semantics-free encoders** robust to noisy data.

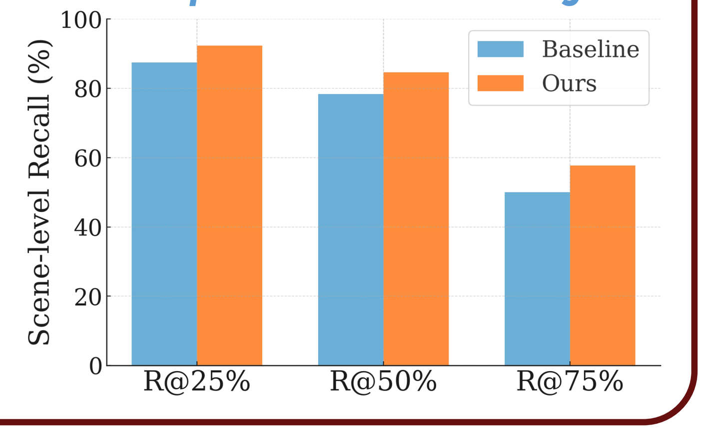
Cross-Modal Instance Matching



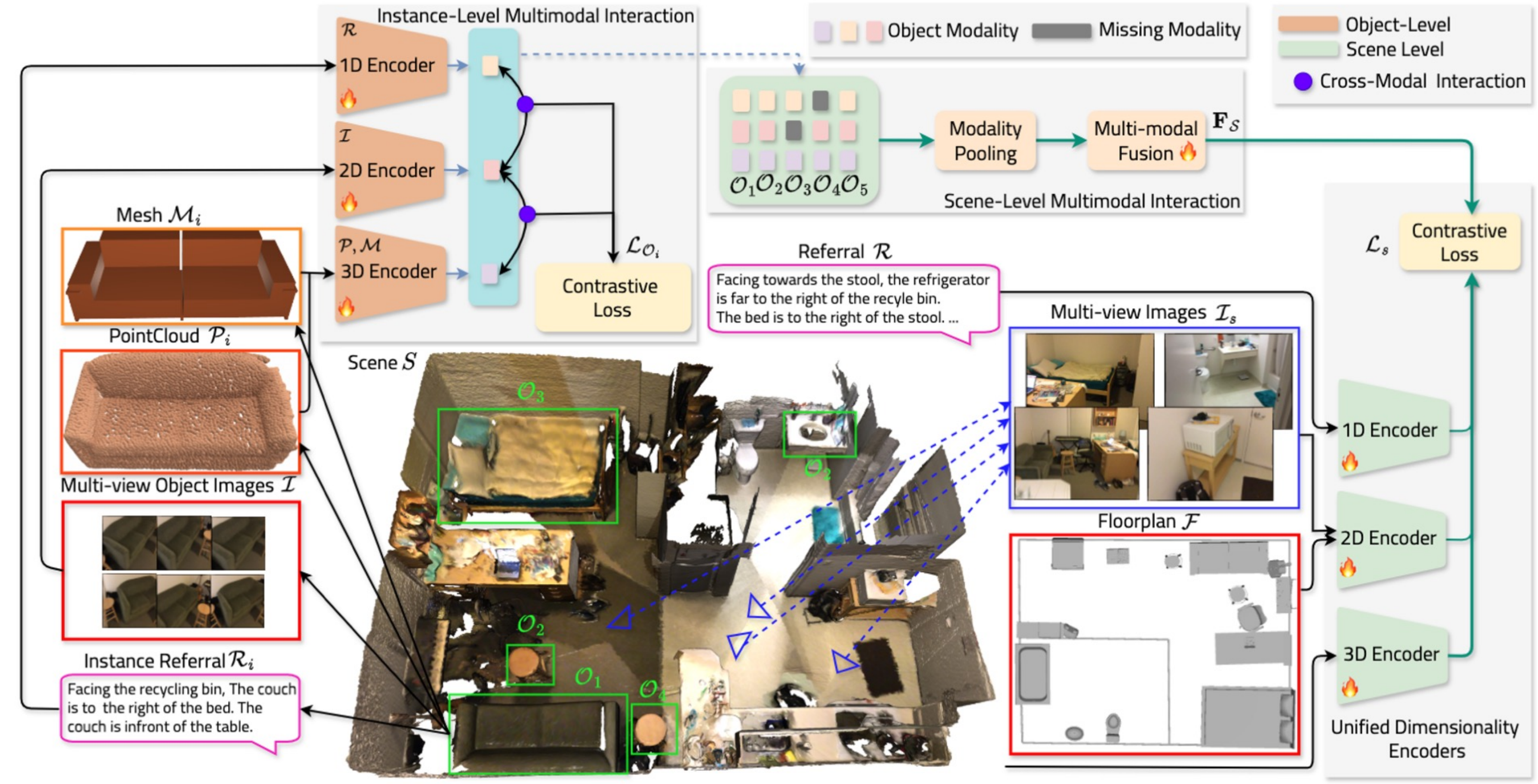
Scene Retrieval Across Modalities



Temporal Instance Matching



3. Method Overview

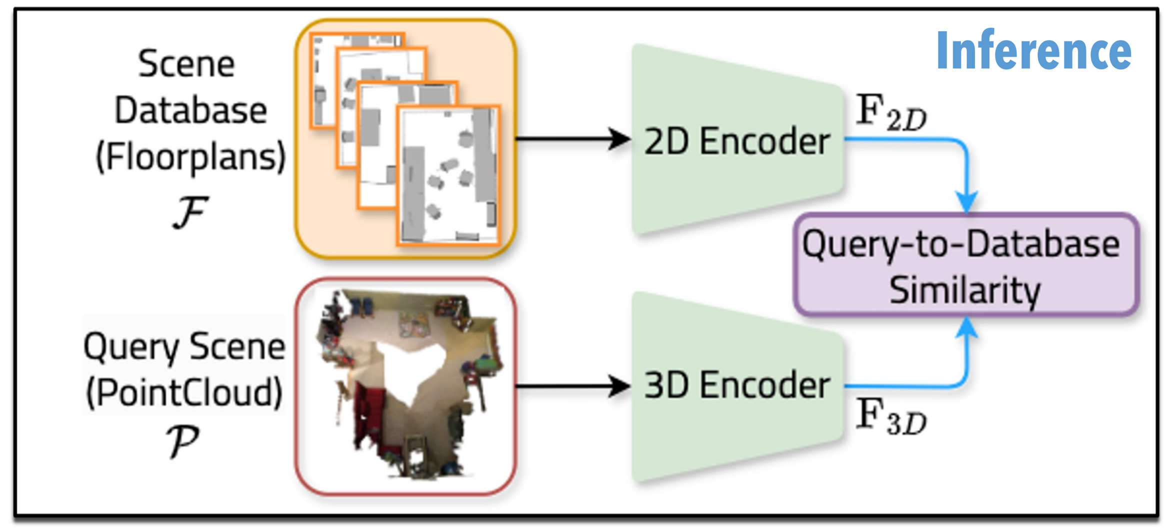


Our Building Blocks:

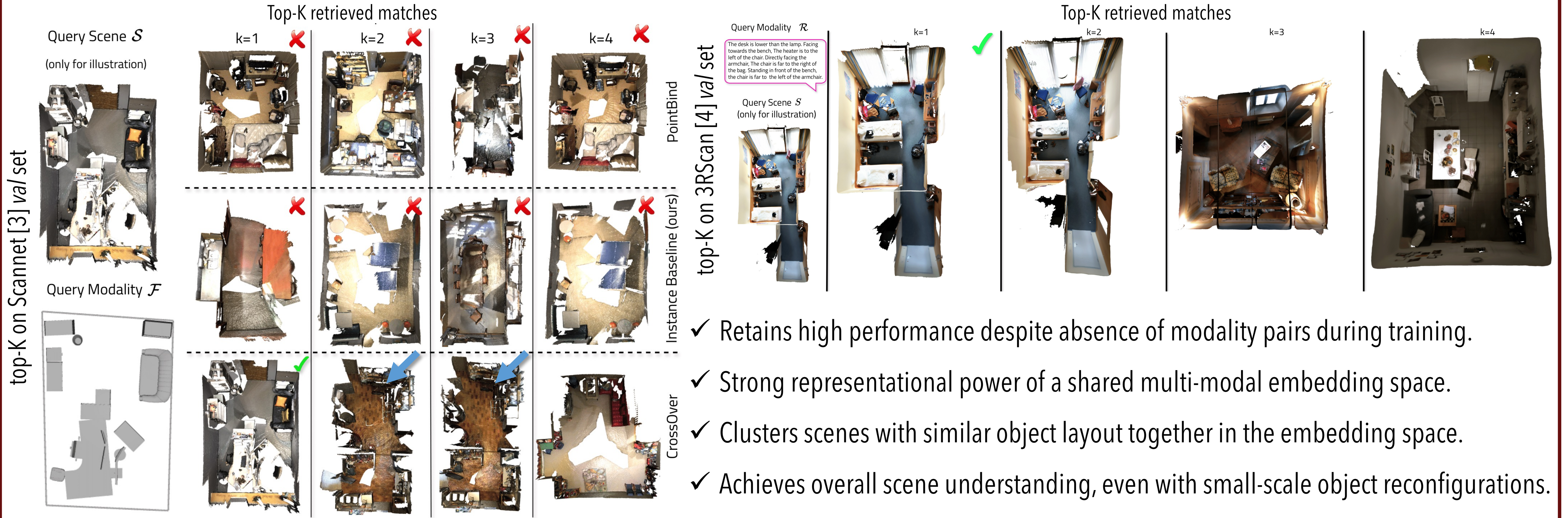
- **Instance-Level Multimodal Interaction** module learns embeddings for object instances by capturing cross-modal interactions and spatial relationships within a scene.
- **Scene-Level Multimodal Interaction** module jointly processes all instances to represent the scene with a single feature vector.
- **Unified Dimensionality Encoders** learn to handle each modality independently while interacting with a shared scene representation, eliminating reliance on semantic annotations.

✓ **How to remove dependency from semantic information?** Transfer knowledge from instance encoders to a unified encoder that directly process raw scene inputs - enabling modality-agnostic scene understanding.

✓ **How to perform cross-modal scene retrieval inference?** Given a scene represented in query modality, we extract its feature using the corresponding unified dimensionality encoder and retrieve the closest match from the target modality in the shared embedding space.



5. Cross-Modal Scene Retrieval



- ✓ Retains high performance despite absence of modality pairs during training.
- ✓ Strong representational power of a shared multi-modal embedding space.
- ✓ Clusters scenes with similar object layout together in the embedding space.
- ✓ Achieves overall scene understanding, even with small-scale object reconfigurations.

Key Takeaways

- ➢ End-to-End Framework for flexible, scene-level cross-modal alignment without the need for semantic annotations or perfectly aligned data.
- ➢ Enables seamless scene matching to anchor virtual content in real-world scenes; direct application(s) in robotics, gaming and AR/VR.

[1] Guo et. al, Point-Bind & Point-LLM, 2023; [2] Xue et. al., ULIP-2, 2024; [3] Dai et. al., Scannet, 2017; [4] Wald et. al., RIO, 2019; [5] Avetisyan et. al., Scan2CAD, 2019; [6] Jia et. al., SceneVerse, 2024