NUS National University of Singapore | School of Computing

CVRP

Learning to Perceive and Interact in Open-World 3D Environments

China3DV

Gim Hee Lee

11 April, 2025

Towards Open-World and Interactive Intelligence

Traditional Paradigm

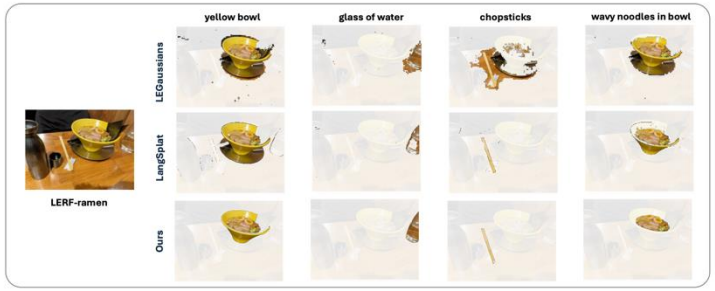
- Closed-set Environment
  - Fixed number of classes
  - Static scenes
- Passive observation
  - Agent observes
  - No agency or influence over environment

The New Frontier

- Open-world setting
  - New & unseen entities appear
  - Dynamic scenes
- Interactive learning
  - Agent acts & explores
  - Receives feedback & adapts
  - Language

Towards Open-World and Interactive Intelligence

Open-Vocabulary Semantic Segmentation

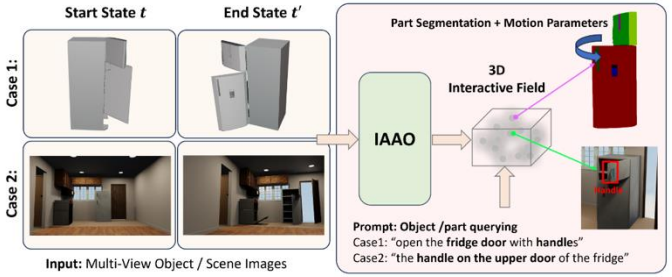


Can Zhang, Gim Hee Lee, **econSG: Efficient and Multi-view Consistent Open-Vocabulary 3D Semantic Gaussians**, ICLR 2025

Vocabulary-free Instance Segmentation

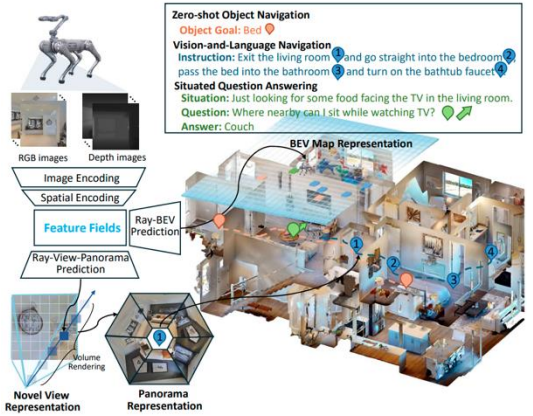
Seungjun Lee, Yuyang Zhao, Gim Hee Lee, **Segment Any 3D Object with Language**, ICLR 2025

Learning Affordance & Articulations



Can Zhang, Gim Hee Lee, **Interactive Affordance Learning for Articulated Objects in 3D Environments**, CVPR 2025

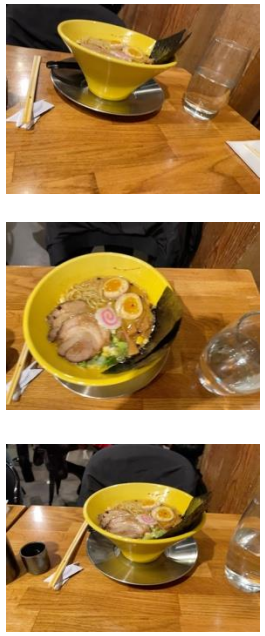
Embodied AI



Zihan Wang, Gim Hee Lee, **g3D-LF: Generalizable 3D-Language Feature Fields for Embodied Tasks**, CVPR 2025

Open-Vocabulary 3D Semantic Segmentation

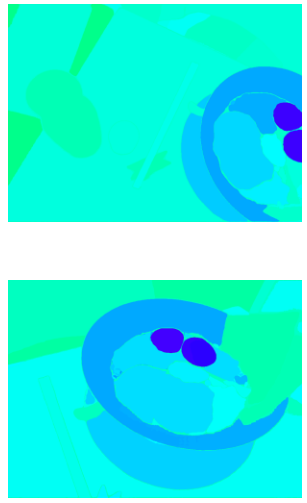
- **Given:** The posed images and the corresponding open-vocabulary queries from the frozen text encoder of a VLM
- **Goal:** Synthesize the semantic masks from novel view renderings



Posed Images

`Prompt: egg`

Rendered Images

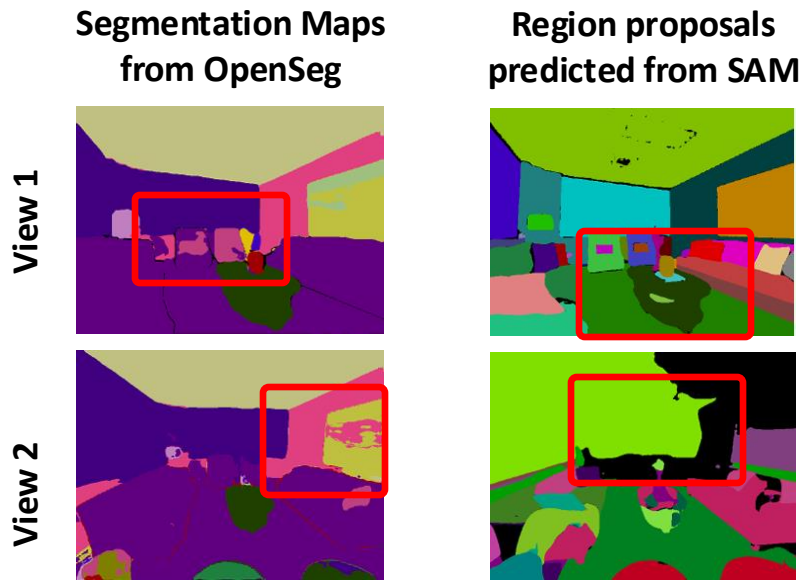


Mask prediction

Feature Visualization

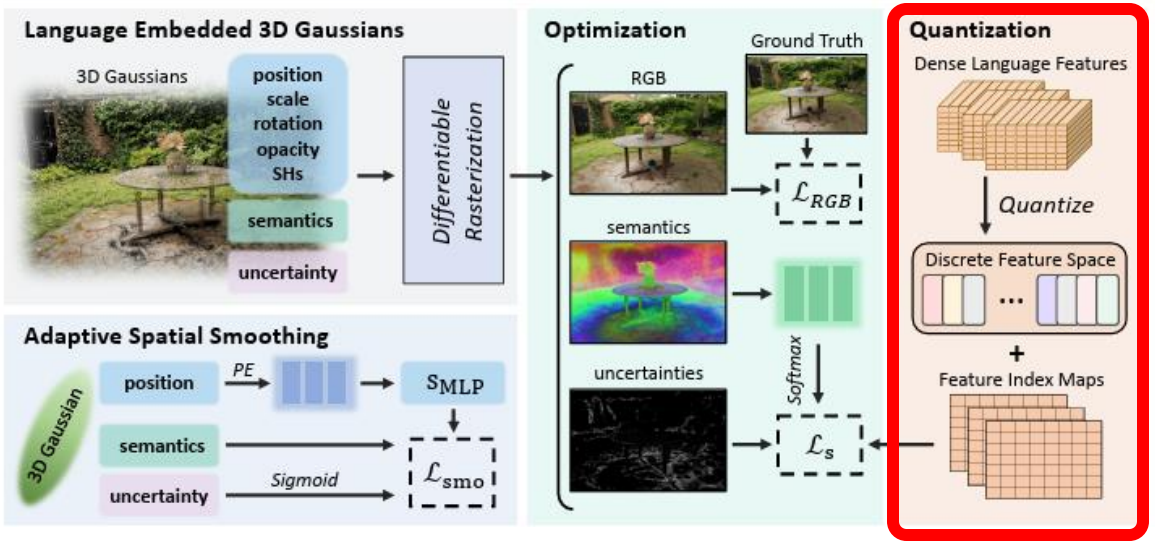
Limitations of Existing Works

1. Existing works either use CLIP with imprecise boundaries or over-trusted DINOv2 or SAM with imperfect regional masks.
2. Reduction of feature dimension in the 2D before lifting into the 3D can lead to multi-view inconsistency that hurts performance.



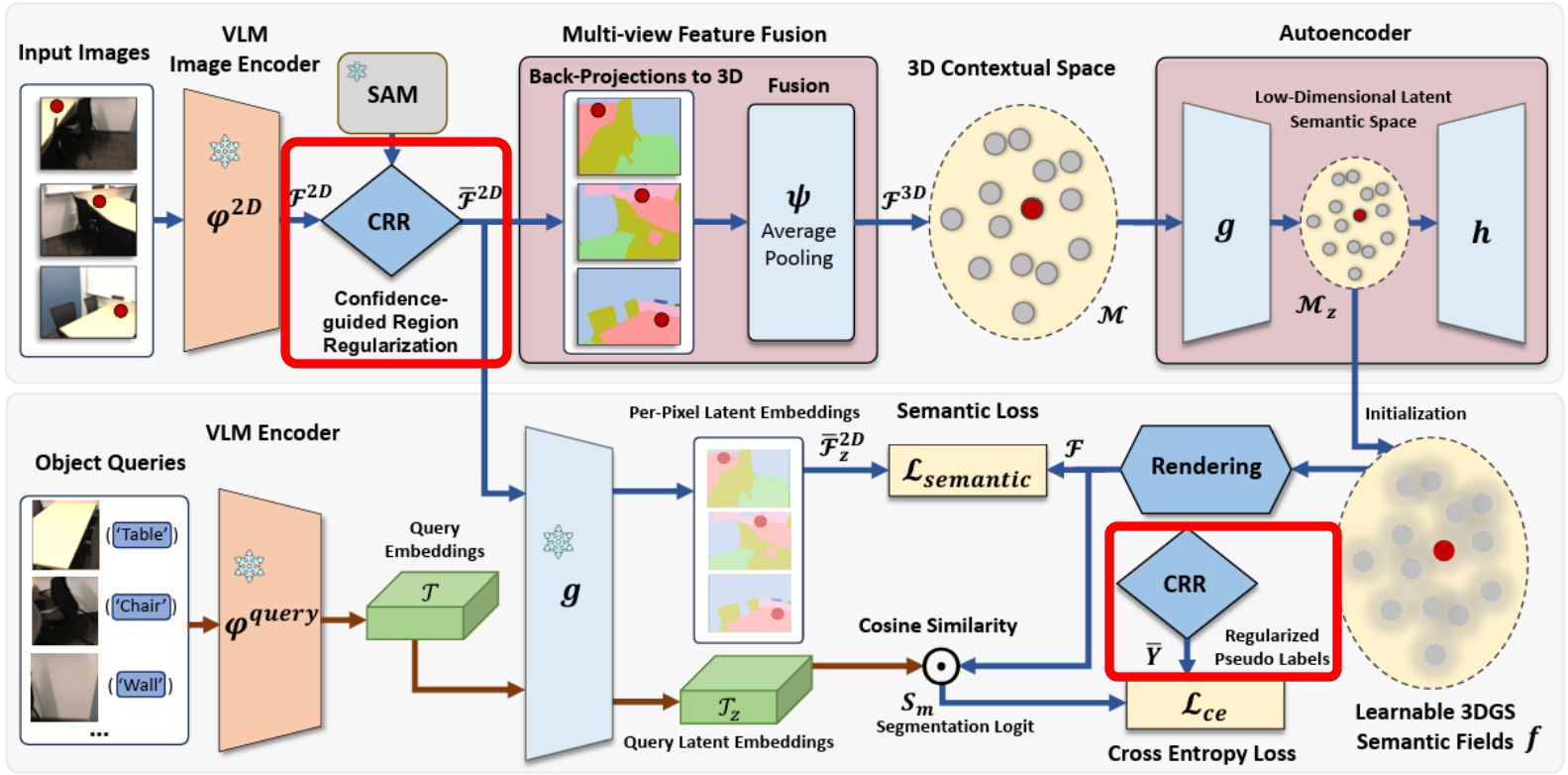
Imprecise boundaries

Feature dimension reduction in 2D



Shi et al, **Language Embedded 3D Gaussians for Open-Vocabulary Scene Understanding**, CVPR 2024

econSG: Our Framework

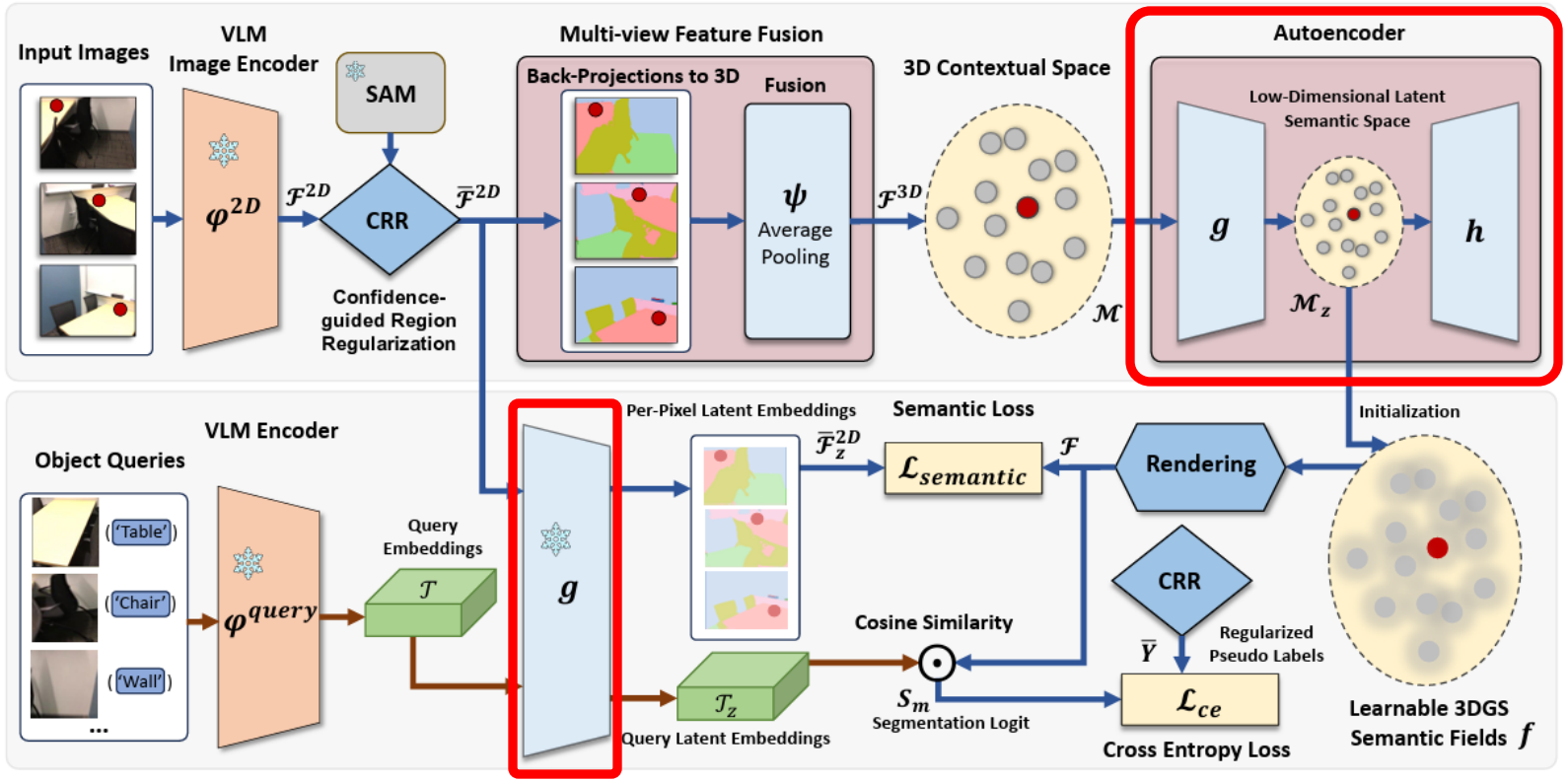


Confidence-Guided Region Regularization (CRR)

1. 3D Consistency: Multiview fusion of OpenSeg features
2. Better Prompt (SAM): Use reprojected fused features as prompt to SAM

Can Zhang, Gim Hee Lee, **econSG: Efficient and Multi-view Consistent Open-Vocabulary 3D Semantic Gaussians**, ICLR 2025

econSG: Our Framework



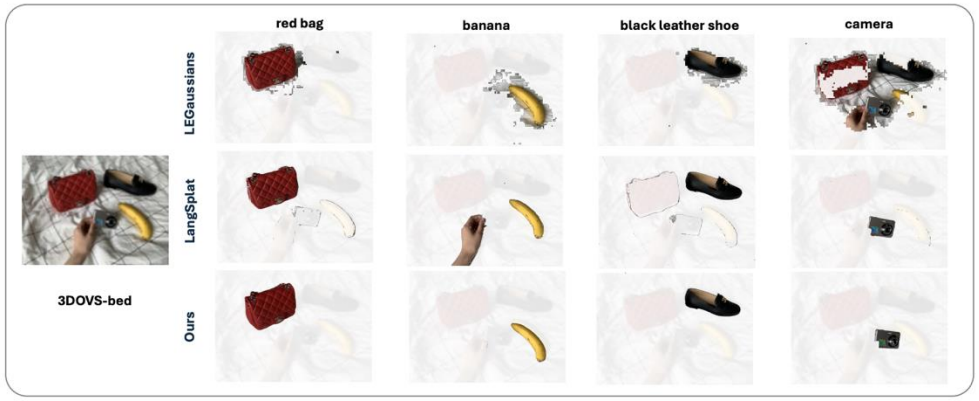
Low-Dimensional 3D Contextual Space

1. 3D Consistency: Fuse the backprojected multi-view 2D features
2. Computational efficiency: Pre-train an autoencoder to get the low-dimensional latent semantic space.

Can Zhang, Gim Hee Lee, **econSG: Efficient and Multi-view Consistent Open-Vocabulary 3D Semantic Gaussians**, ICLR 2025

econSG: Results

Figure 7: Qualitative 3D segmentation results of our econSG on the Scannet dataset.



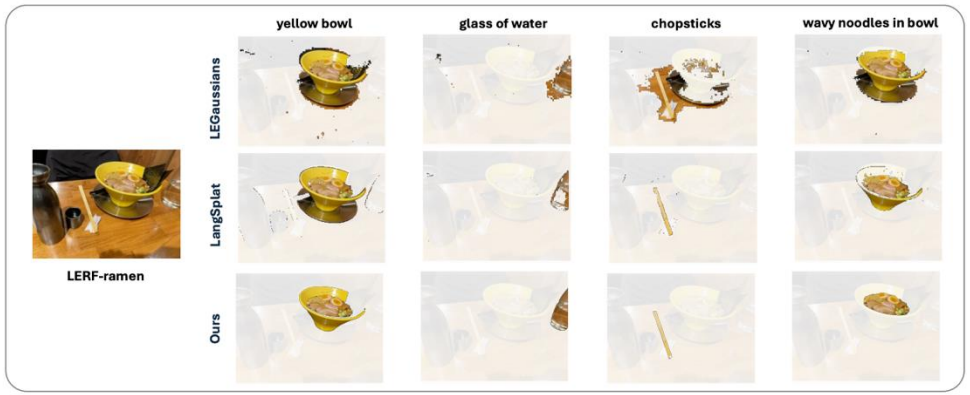


Figure 10: Qualitative comparison of our econSG with baselines on the LERF dataset. We show the visualization of the retrieved objects in the scene. The quantitative results are in Table 2 in the main paper.

Table 1: Comparisons of open-vocabulary segmentation on 3DOVS dataset. Best results in **bold**.

	Dataset	3DOVS											
	Method	bed		sofa		lawn		room		bench		overall	
		mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
2D	LSeg	56.0	87.6	4.5	16.5	17.5	77.5	19.2	46.1	6.0	42.7	20.6	54.1
3D	LERF	73.5	86.9	27.0	43.8	73.7	93.5	46.6	79.8	53.2	79.7	54.8	76.7
	3DOVS	89.5	96.7	74.0	91.6	88.2	97.3	92.8	98.9	89.3	96.3	86.8	96.2
	Feature3DGS	56.6	87.5	6.7	12.4	37.3	82.6	20.5	36.7	6.2	43.0	25.5	52.4
	LEGaussians	45.7	-	48.2	-	49.7	-	44.7	-	47.4	-	47.1	-
	LangSplat	73.5	89.7	82.3	**98.7**	89.9	95.6	95.0	**99.4**	70.6	92.6	82.3	95.2
	econSG (Ours)	**94.9**	**97.4**	**91.6**	**98.7**	**96.3**	**98.5**	**95.8**	**99.4**	**93.0**	**97.6**	**94.3**	**98.3**

Table 2: Comparisons of localization accuracy on LERF dataset. Best results in **bold**.

	Dataset	LERF				
	Method	ramen	figurines	teatime	waldo_kitchen	overall
2D	LSeg	14.1	8.9	33.9	27.3	21.1
3D	LERF	62.0	75.0	84.8	72.7	73.6
	LangSplat	73.2	80.4	88.1	95.5	84.3
	SemanticGaussian	76.8	83.1	89.8	90.9	85.2
	LEGaussians	78.6	73.7	85.6	90.1	82.0
	econSG (Ours)	**83.2**	**89.3**	**93.4**	**96.2**	**90.5**

Table 3: Comparison with other methods on segmentation of novel views from Scannet and Replica. Best results highlighted in **bold**.

Dataset	FPS	Replica				Scannet			
		sparse-view		multi-view		sparse-view		multi-view	
		mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
LERF	0.2	4.312	17.080	8.285	22.125	14.059	38.734	15.349	40.294
3DOVS	0.3	4.553	19.356	9.081	23.938	14.227	40.584	17.802	42.532
Feature3DGS	2.5	9.584	38.245	10.634	36.520	17.552	48.686	18.069	54.101
econSG (Ours)	156	**25.513**	**70.716**	**33.869**	**78.564**	**39.018**	**74.805**	**48.205**	**86.178**

econSG: Results

Table 4: Training efficiency analysis on the sofa scene of the 3DOVS dataset.

Methods	LERF	3DOVS	Langsplat	Feature3DGS	Ours			Ours (remove autoencoder)
Feature dimension	512	512	3	128	6	16	32	512
mIoU (%)	27.0	74.0	82.3	6.7	91.6	91.8	91.8	OOM
Training time (min)	19.4	78	66	87	29	32	43	OOM
Inference (s)	121.4	6.6	401.9	6.0	4.9	5.2	5.3	OOM

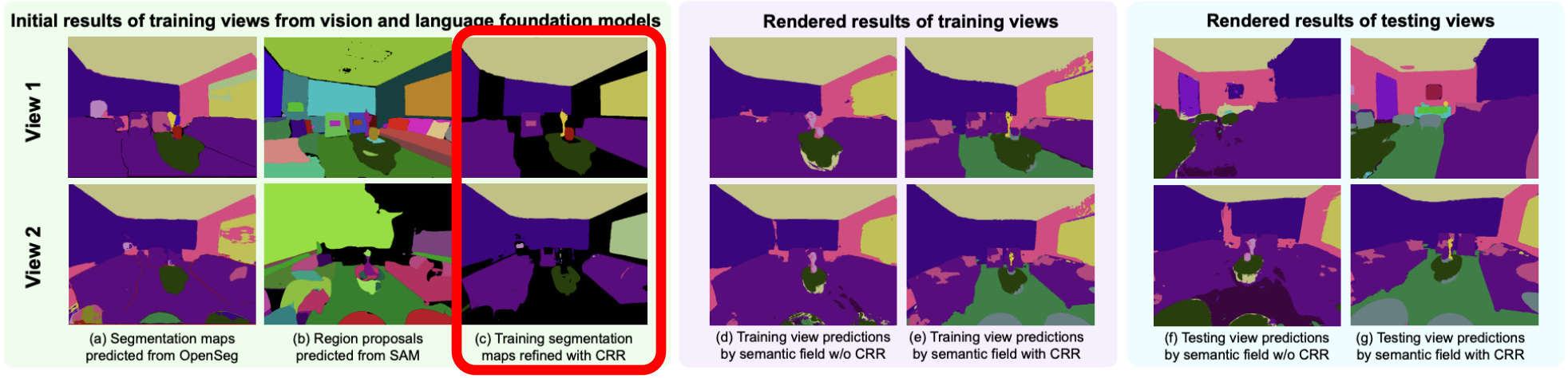


Figure 3: Ablation on confidence-guided region regularization (CRR) with qualitative results of our econSG on Replica. Panels (a)-(e) are from training views, and panels (f)-(g) are from testing views.

Vocabulary-free Instance Segmentation

- **Goal:** Given 3D point cloud of a scene, the goal is to detect and segment unseen classes based on language instructions.

Open-Set 3D Instance Segmentation (OS-3DIS)

(a) "Can I wash my hands?"

Visual question

(b) "Brown Furnitures"

Attributes description

(c) "Device to play game"

Functional description

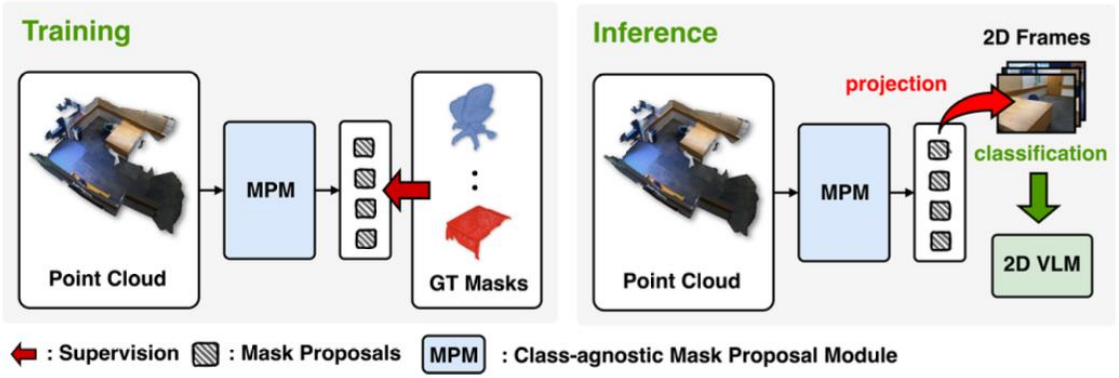
(d) "Bathroom"

Space-level description

Limitation of Existing Methods

Two-stage Approach:

1. Extracting 3D masks from class-agnostic 3D mask proposal module
2. Classifying 3D masks with 2D foundation model



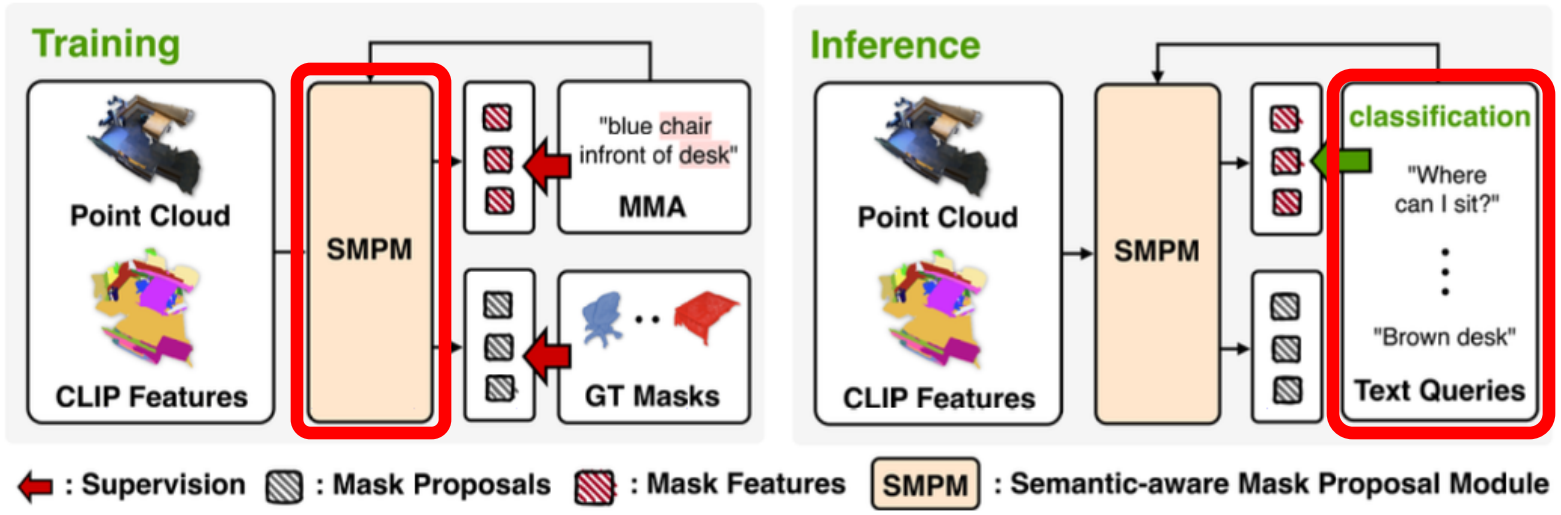
Examples:
OpenMask3D (NeurIPS'24), OpenIns3D (ECCV'24), Open3DIS (CVPR'25)

Limitation:
No semantic information in the mask generation stage → low semantic generalizability

SOLE: Our Framework

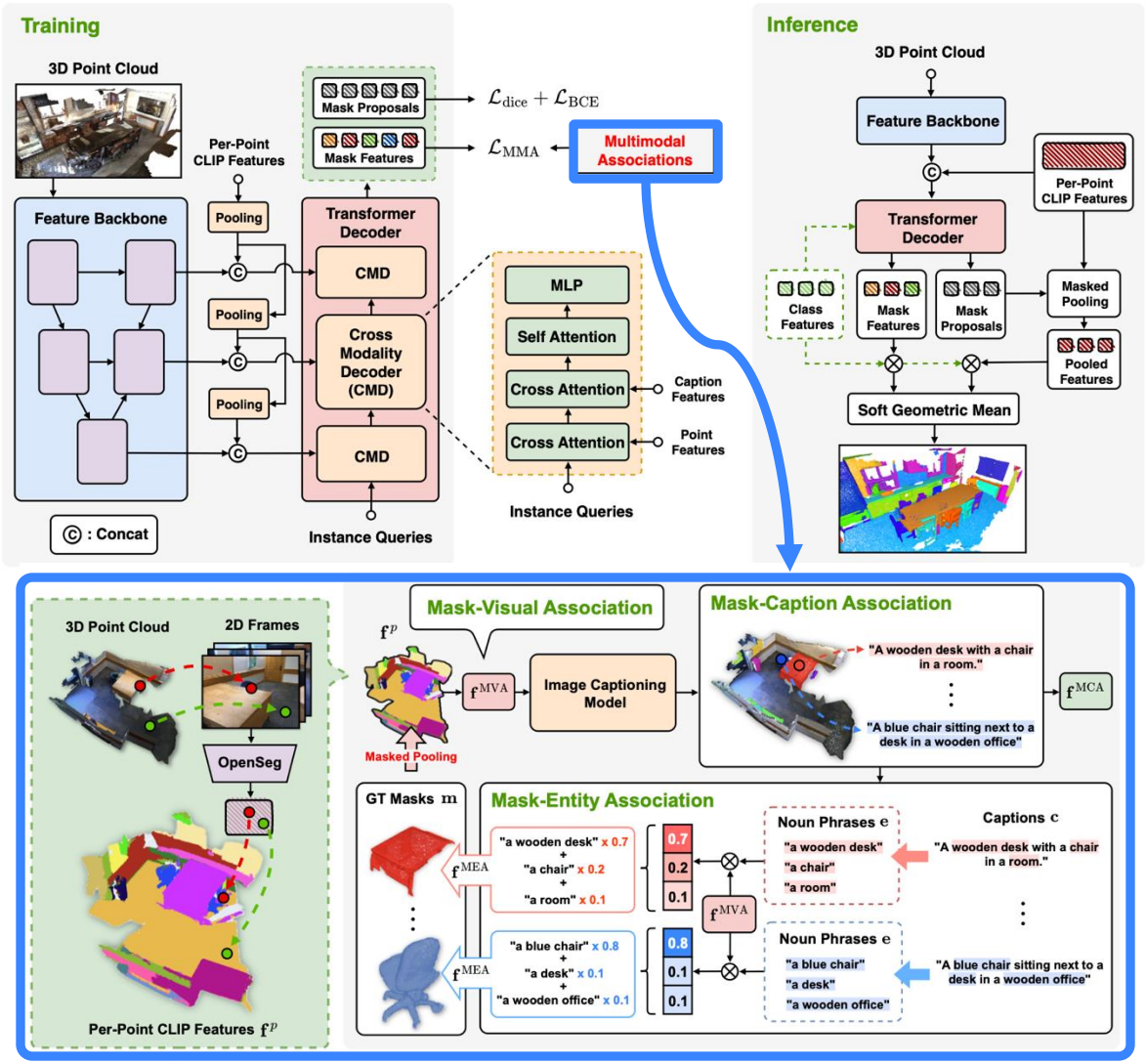
Our SOLE improves semantic generalizability by:

1. Generating semantic-aware 3D masks.
2. Classifies the 3D masks by being responsive to free-form language beyond the noun-level descriptions



Seungjun Lee, Yuyang Zhao, Gim Hee Lee, **Segment Any 3D Object with Language**, ICLR 2025

SOLE: Our Framework



Multimodal Association

Three types of 3D mask-to-language alignment:

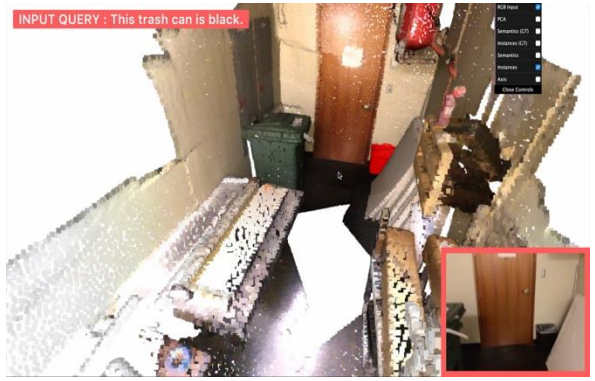
1. 3D masks – Image Features
2. 3D masks – Sentence-level Captions Features
3. 3D masks – Noun-level Captions Features

SOLE: Results

Functional Description

👦💻: I want to heat the food.

Attributes Description

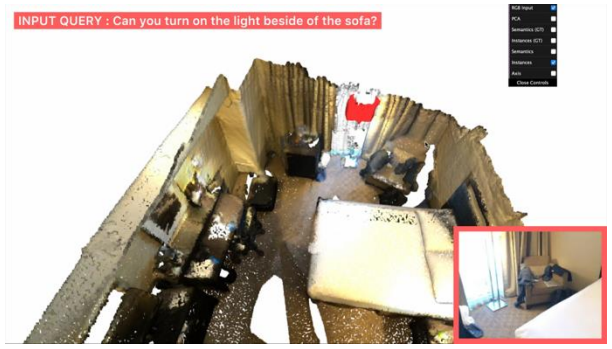


👦💻: This trash can is black.

Visual Grounding with Multiple Attributes Description

👦💻: This is a small black office chair.

Relationship with Other Objects



👦💻: Turn on the light beside of the sofa.

Comparison of closed-set 3D instance segmentation on ScanNetv2

Method	AP	AP_{50}	AP_{25}	voxel size
OpenIns3D (Huang et al., 2023b)	-	28.7	38.9	2cm
OpenMask3D (Takmaz et al., 2023)	31.0	39.5	44.0	2cm
Open3DIS (only 3D) (Nguyen et al., 2024)	31.3	42.4	47.8	2cm
SOLE *w 4cm voxel size*	30.8	52.5	70.9	4cm
SOLE *w/o text sup*	35.0	50.2	60.2	2cm
SOLE (*ours*)	**44.4** (+13.1)	**62.2** (+19.8)	**71.4** (+23.6)	2cm
Mask3D (Schult et al., 2022) (*fully sup*)	55.2	73.7	83.5	2cm

Comparison of closed-set 3D instance segmentation on ScanNet200

Method	AP	AP_{50}	AP_{25}	AP_{head}	AP_{com}	AP_{tail}
OpenIns3D (Huang et al., 2023b)	8.8	10.3	14.4	16.0	6.5	4.2
OpenMask3D (Takmaz et al., 2023)	15.4	19.9	23.1	17.1	14.1	**14.9**
Open3DIS *(only 3D)* (Nguyen et al., 2024)	18.6	23.1	27.3	24.7	16.9	13.3
SOLE (*ours*)	**20.1** (+1.5)	**28.1** (+5.0)	**33.6** (+6.3)	**27.5** (+2.8)	**17.6** (+0.7)	14.1 (-0.8)
Mask3D *(fully sup)* (Schult et al., 2022)	26.9	36.2	41.4	39.8	21.7	17.9

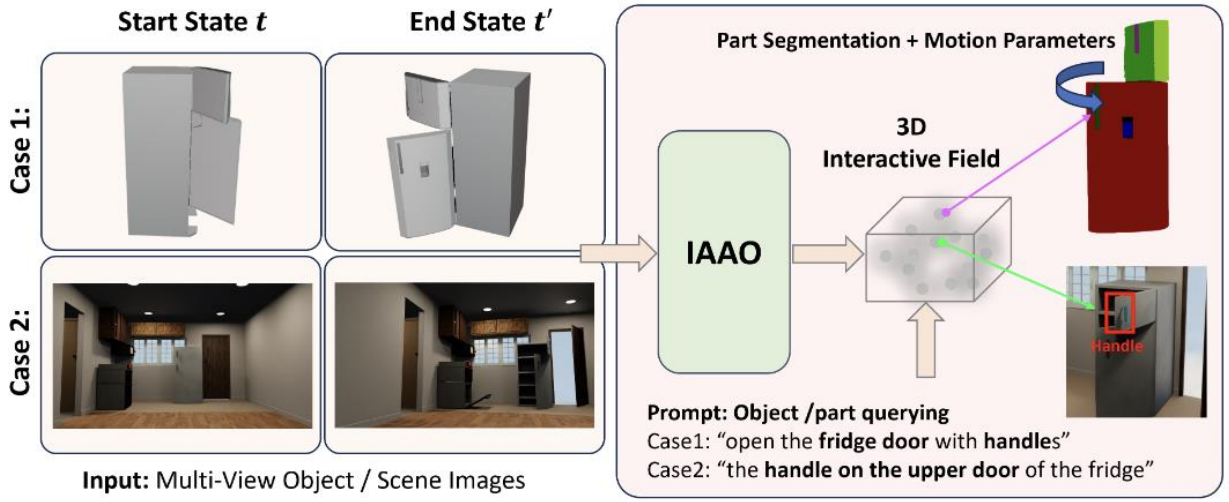
Comparison of open-set 3D instance segmentation on ScanNet200 → Replica

Method	Mask Training	AP	AP_{50}	AP_{25}
OpenIns3D (Huang et al., 2023b)	ScanNet200	13.6	18.0	19.7
OpenMask3D (Takmaz et al., 2023)	ScanNet200	13.1	18.4	24.2
Open3DIS *(only 3D)* (Nguyen et al., 2024)	ScanNet200	14.9	18.8	23.6
SOLE (*ours*)	ScanNet200	**24.7** (+9.8)	**31.8** (+13.0)	**40.3** (+16.1)

Learning Affordance & Articulations

Given: Multi-view images of the object or indoor scene from two different joint states (left figure).

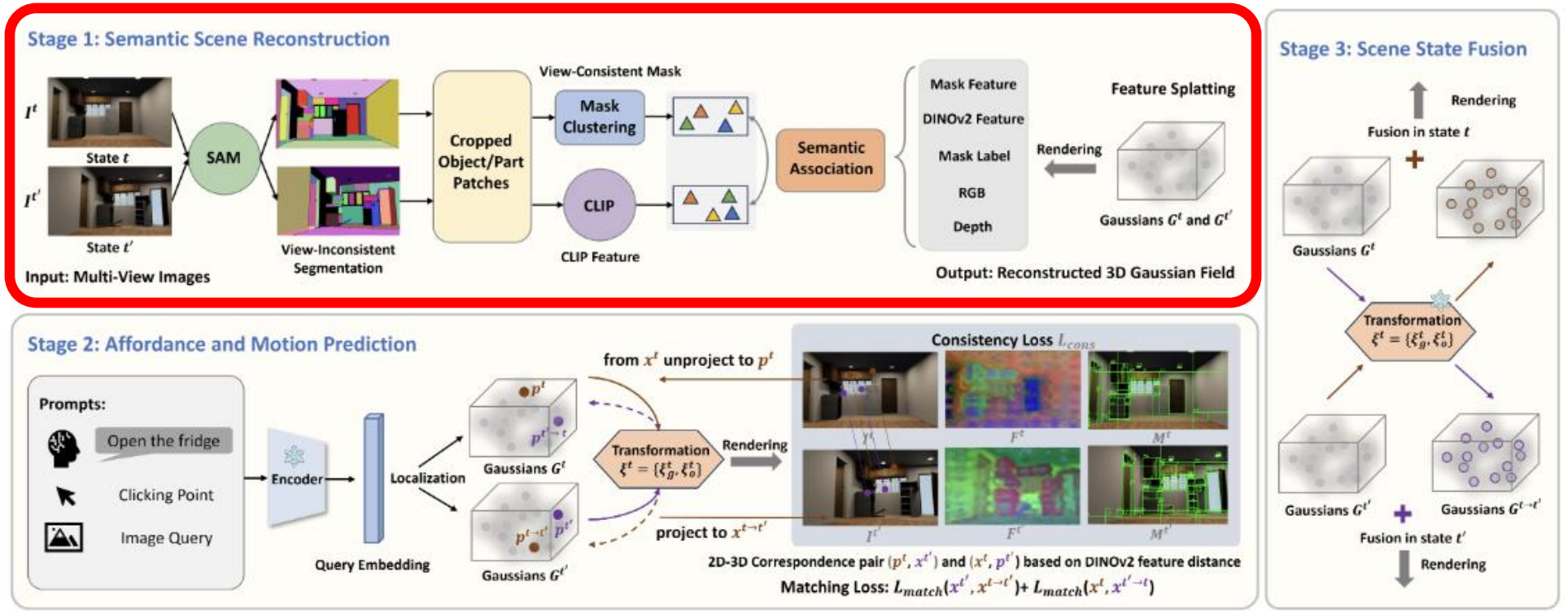
Find: 3D interactive field which supports interactions with multiple movable parts for fine-grained segmentation and articulation reconstruction



IAAO: Our Framework

Stage 1: Constructing 3D Gaussian fields in each state

- Optimize 3DGS fields with hierarchical mask features
- Incorporate geometry information from depth images into 3D Gaussians

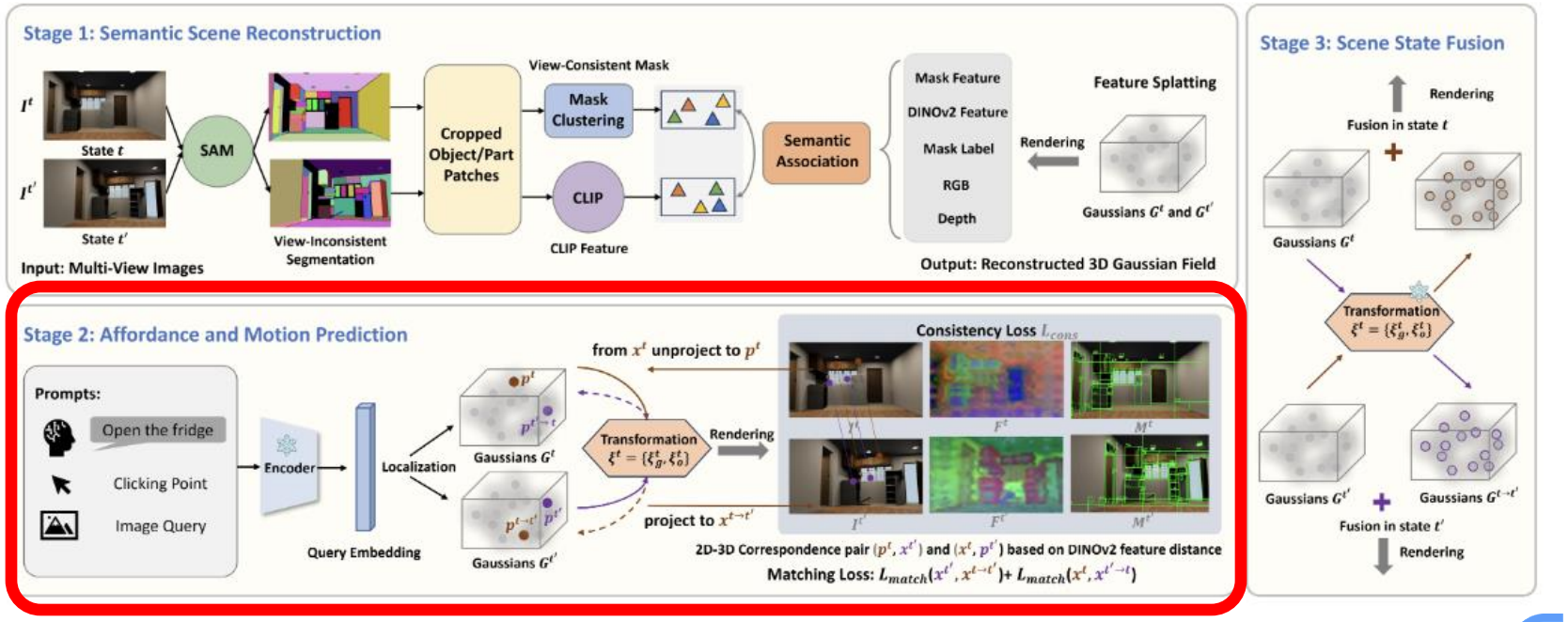


Can Zhang, Gim Hee Lee, **Interactive Affordance Learning for Articulated Objects in 3D Environments**, CVPR 2025

IAAO: Our Framework

Stage 2: Affordance and motion prediction

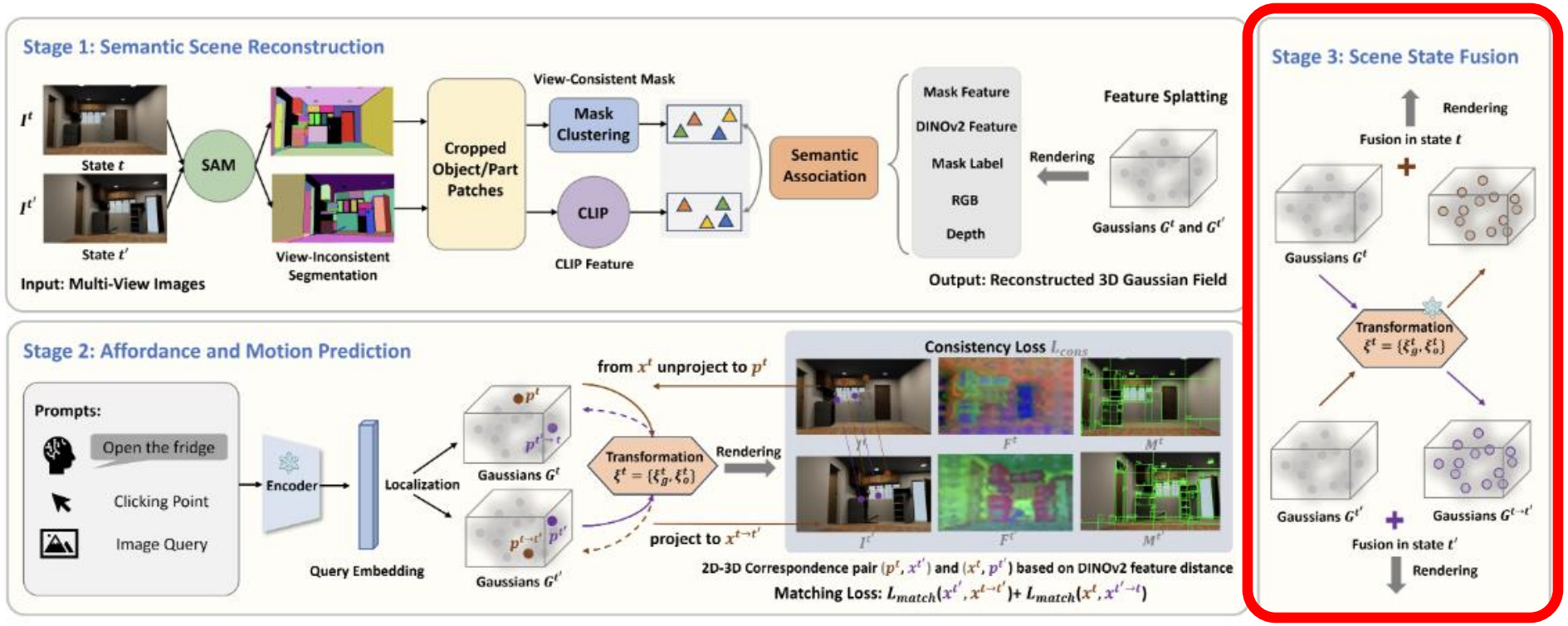
- Affordance: A query prompt is embedded using a pretrained encoder to localize relevant regions in the 3D Gaussians
- Motion: Optimize transformation with consistency and matching losses to 2D-3D correspondences between states



IAAO: Our Framework

Stage 3: Scene fusion

- Using the estimated transformations, we merge reconstructed 3DGS models from both states, aligning static and articulated elements



IAAO: Results

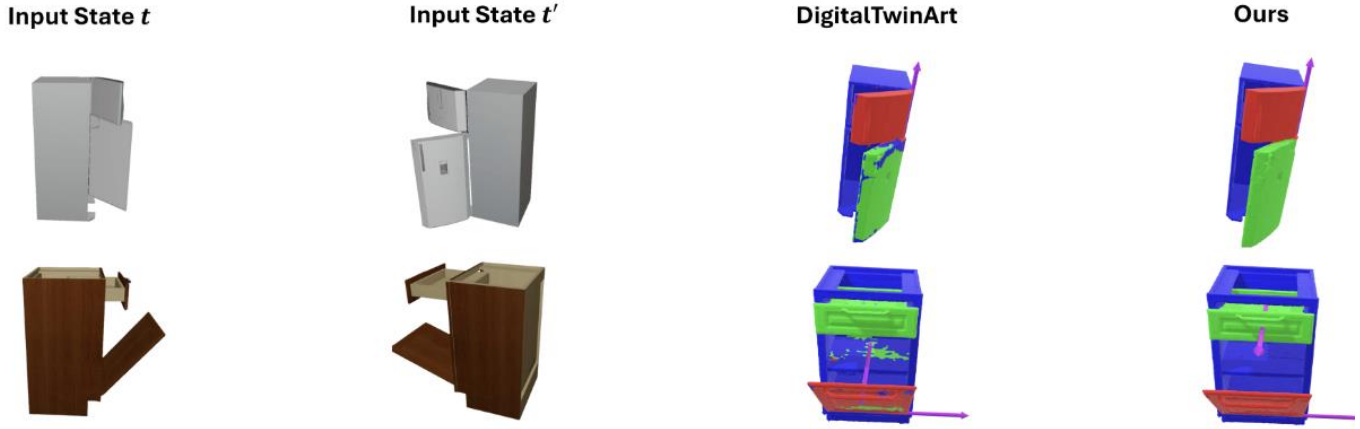


Figure 3. Qualitative analysis of shape reconstruction, part segmentation, and joint prediction results on multi-part object d

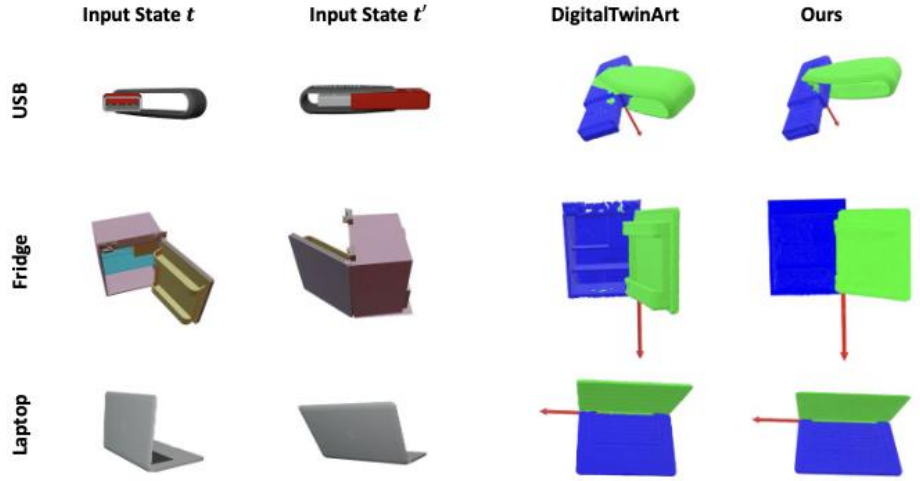


Figure 4. Qualitative results of shape reconstruction, part segmentation, and joint prediction on PARIS.

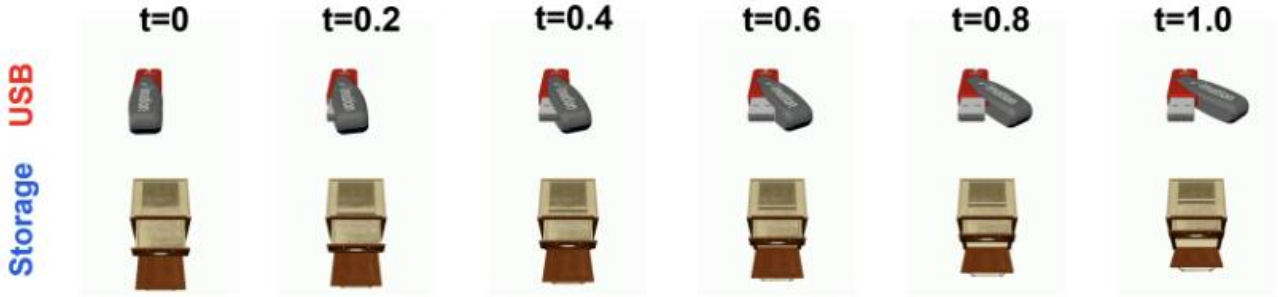


Figure 5. Motion snapshots on PARIS & multi-part object.

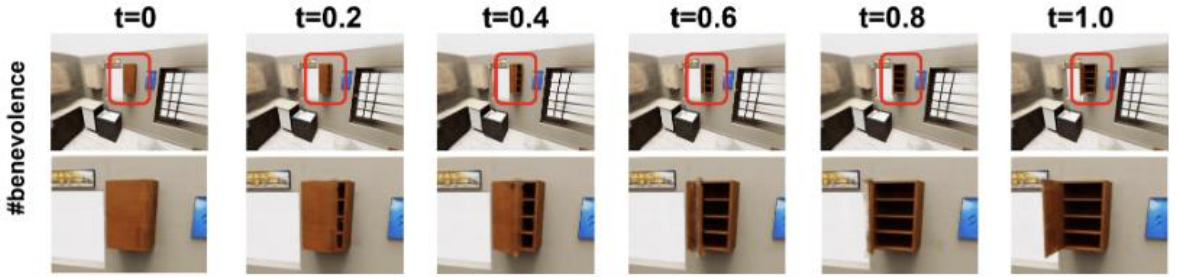
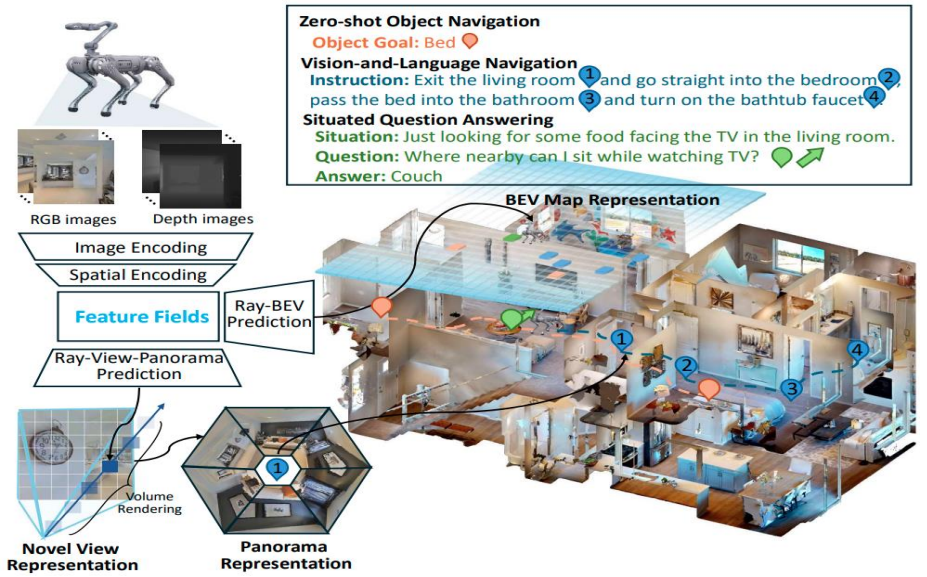


Figure 6. Motion snapshots on scene-level OmniSim dataset.

Embodied AI

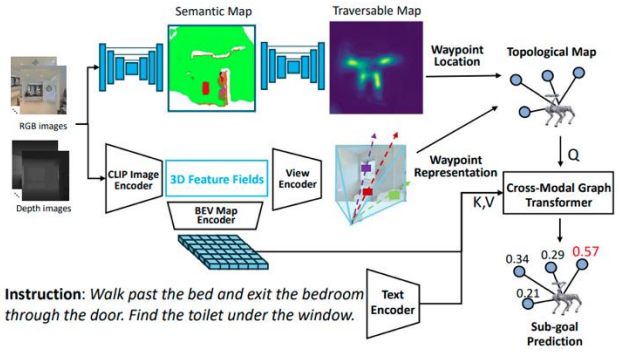
Objective:

Learn a model for 1) Novel view representation; 2) Generations of agent-centric BEV maps; 3) Querying targets using multi-granularity language.

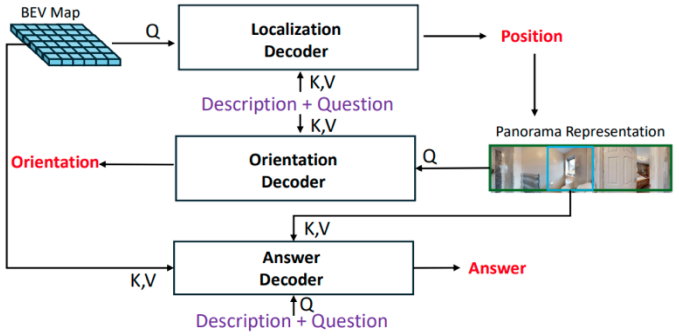


Tasks:

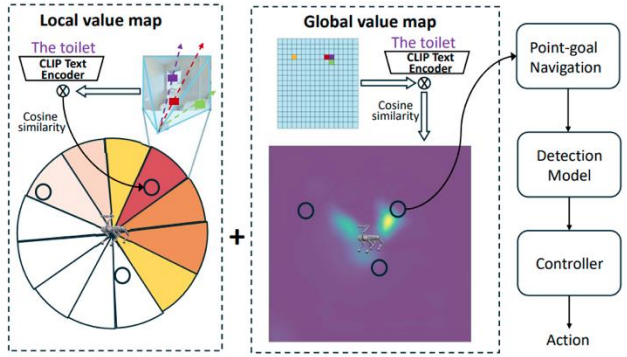
Monocular Vision-Language Navigation



Situated Question Answering



Zero-shot Object Navigation



3D Representation Model for Embodied AI

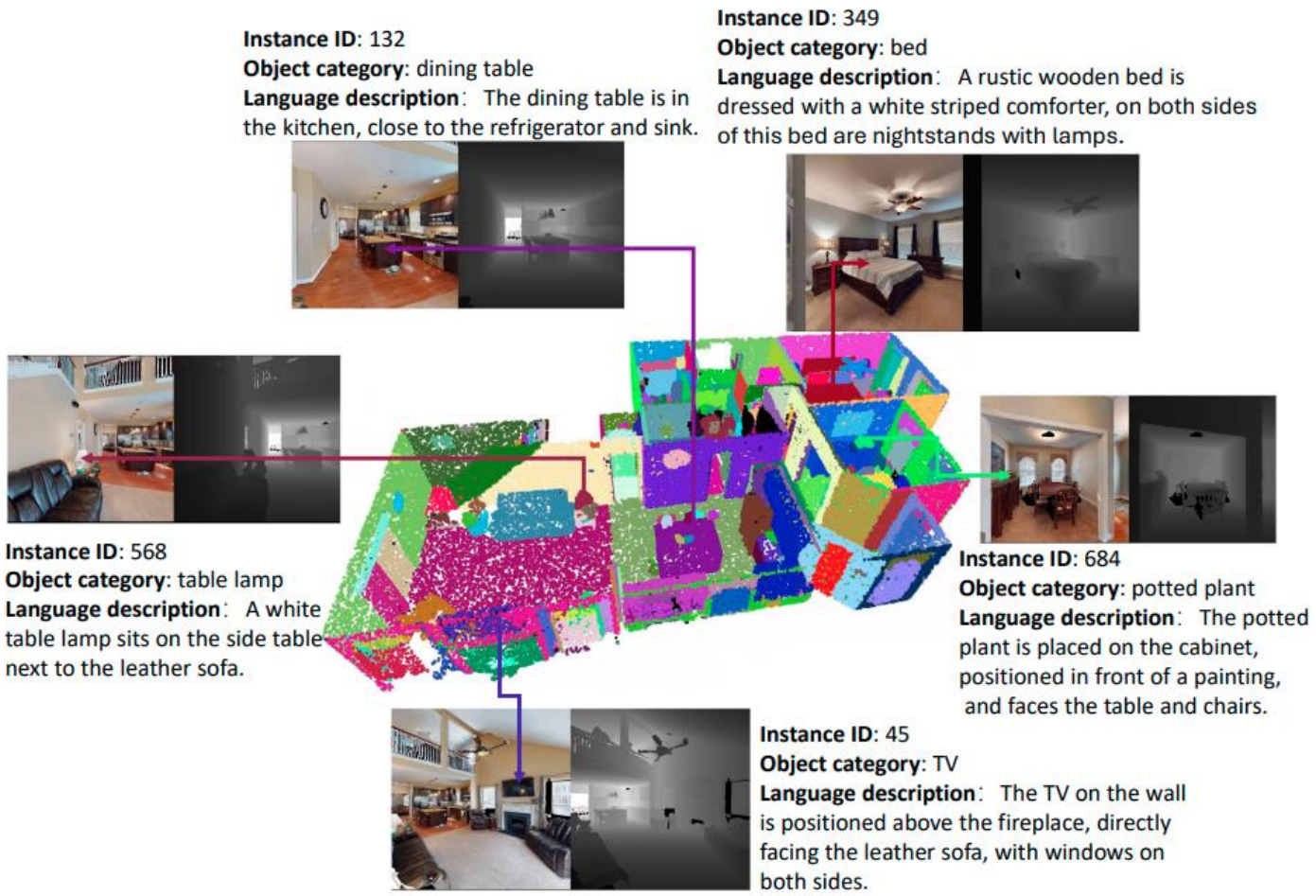
What makes a good 3D representation model for Embodied AI?

1. Generalizable to unseen scenes
2. Construct and update representations in real time
3. Open-vocabulary semantic space

Limitation of Existing Methods

1. Supervised by 2D models (CLIP, DINOv2) and thus lacks 3D spatial understanding
2. There is still a substantial gap with open-vocabulary language
3. The large-scale representations, e.g., panorama and BEV map is particularly challenging for long text understanding

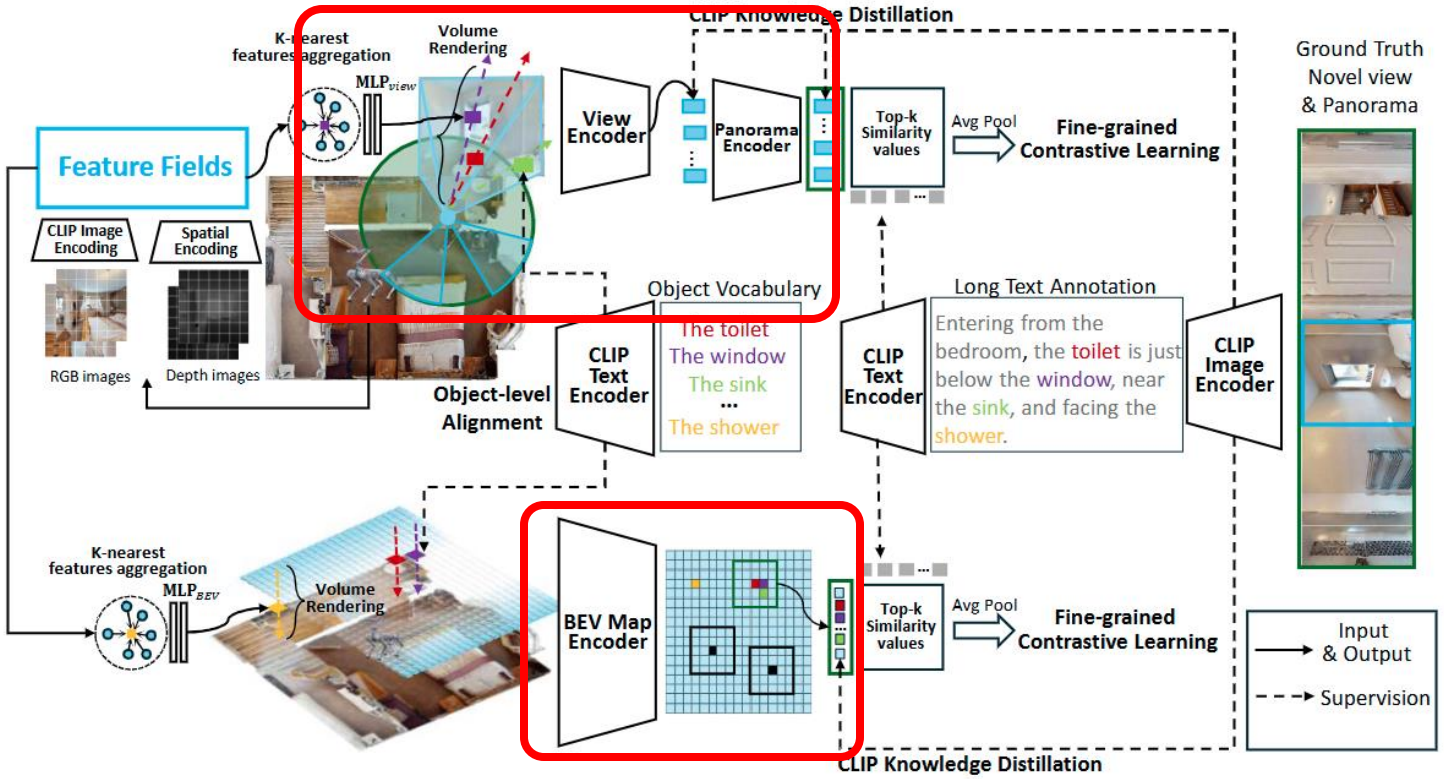
Large-Scale Language-3D Data



We curate and consolidate from various datasets:

1,883 Object categories, 5K+ 3D scenes, 1M+ language descriptions

g3D-LF: Our Framework

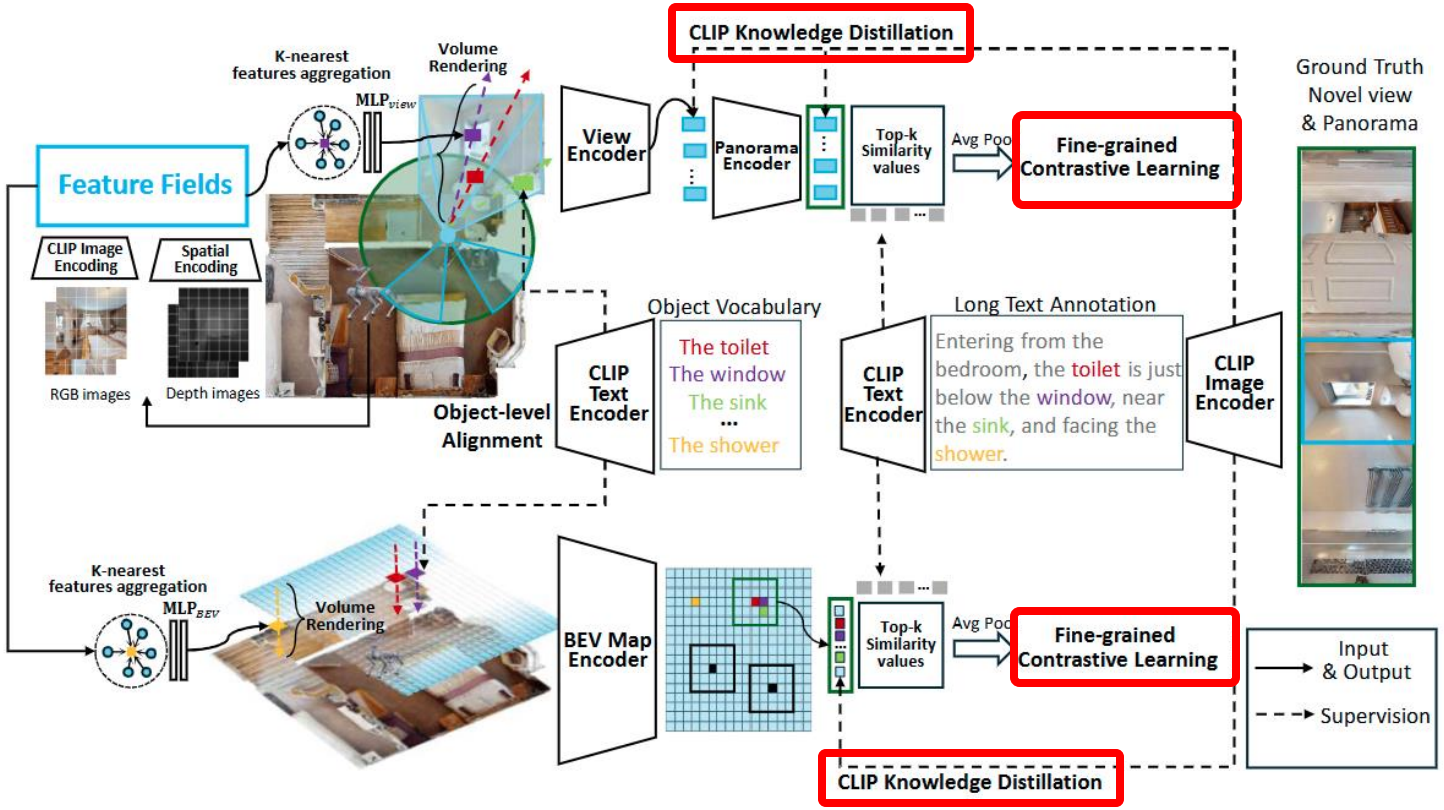


Multi-scale Representation

- Render the ray-level representation
- Combine the rays into the view representation
- Encode the panorama with multi-views
- Encode the top-down rays for large-scale BEV map

Zihan Wang, Gim Hee Lee, **g3D-LF: Generalizable 3D-Language Feature Fields for Embodied Tasks**, CVPR 2025

g3D-LF: Our Framework



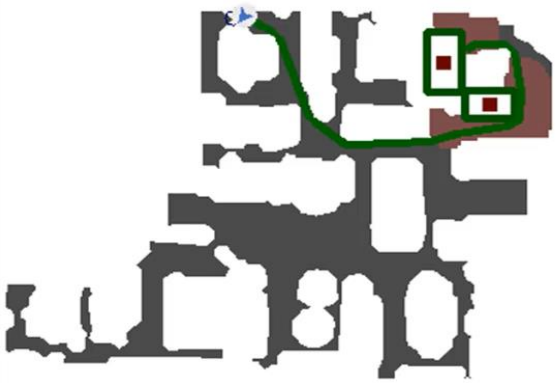
Multi-level Supervision

- For rendered rays, contrastive learning across 1,883 indoor object categories
- For novel view-panorama and BEV map, distill knowledge from 2D model
- For 3D spatial reasoning and long text understanding, use fine-grained contrastive learning

Zihan Wang, Gim Hee Lee, **g3D-LF: Generalizable 3D-Language Feature Fields for Embodied Tasks**, CVPR 2025

Object Navigation --- “Find the Couch”

Navigation Trajectory



Obstacle Map

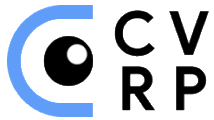
Value Map from Feature Field

Vision-and-Language Navigation

Monocular VLN

Panorama VLN


NUS
National University
of Singapore
School of
Computing


CV
RP

Thank You!