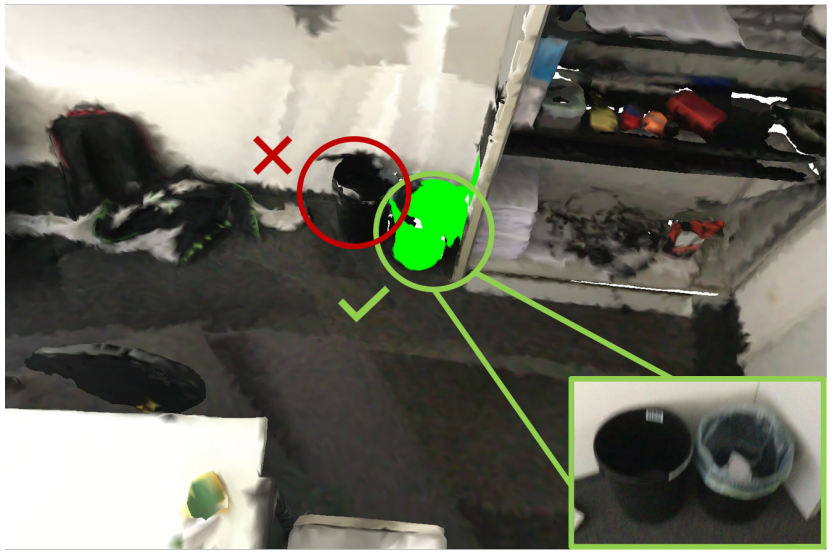
# OV3D-CG:
# Open-vocabulary 3D Instance Segmentation with Contextual Guidance

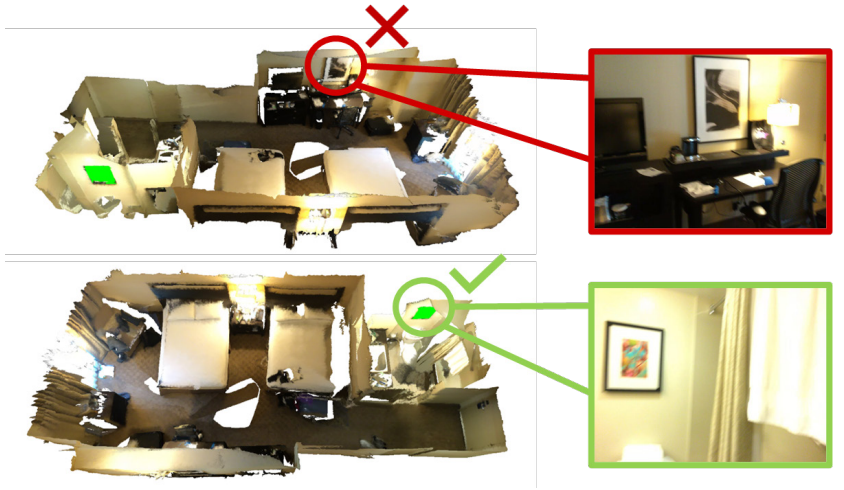
Mingquan Zhou[1,2], Chen He[1], Ruiping Wang[1,2 *], Xilin Chen[1,2]
[1]Institute of Computing Technology, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China

**Prompt**: *"A blue pillow on the bed"*

**Prompt**: *"The trashcan with a plastic bag on it"*



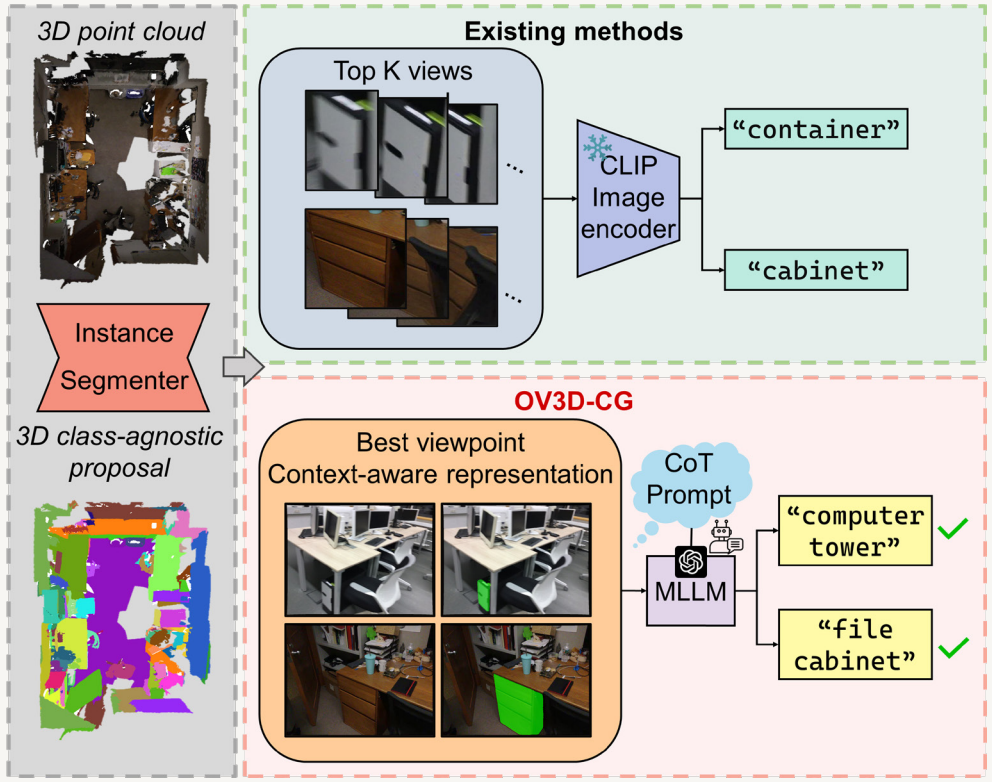
**Prompt**: *"Find the picture in the bathroom."*



# FOCUS

- Enable AI to see and understand any 3D object
- Handle objects never seen during training
- Key for robotics, autonomous driving, and augmented reality

# MOTIVATION



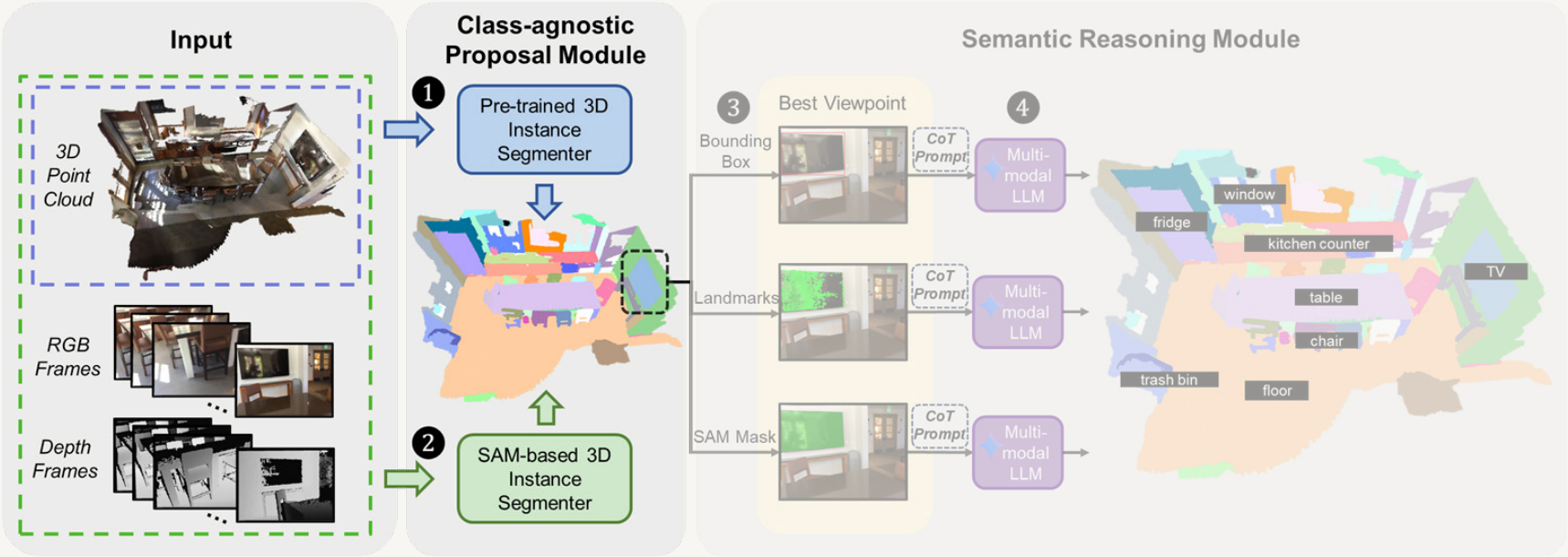

- Existing methods rely on **fixed categories**
- Fail to handle unseen or **rare objects**
- **No context reasoning** → ambiguous predictions

**We argue**: Context is key for open-vocabulary 3D segmentation

# OV3D-CG FRAMEWORK

## ➢ Class-agnostic Proposal Module

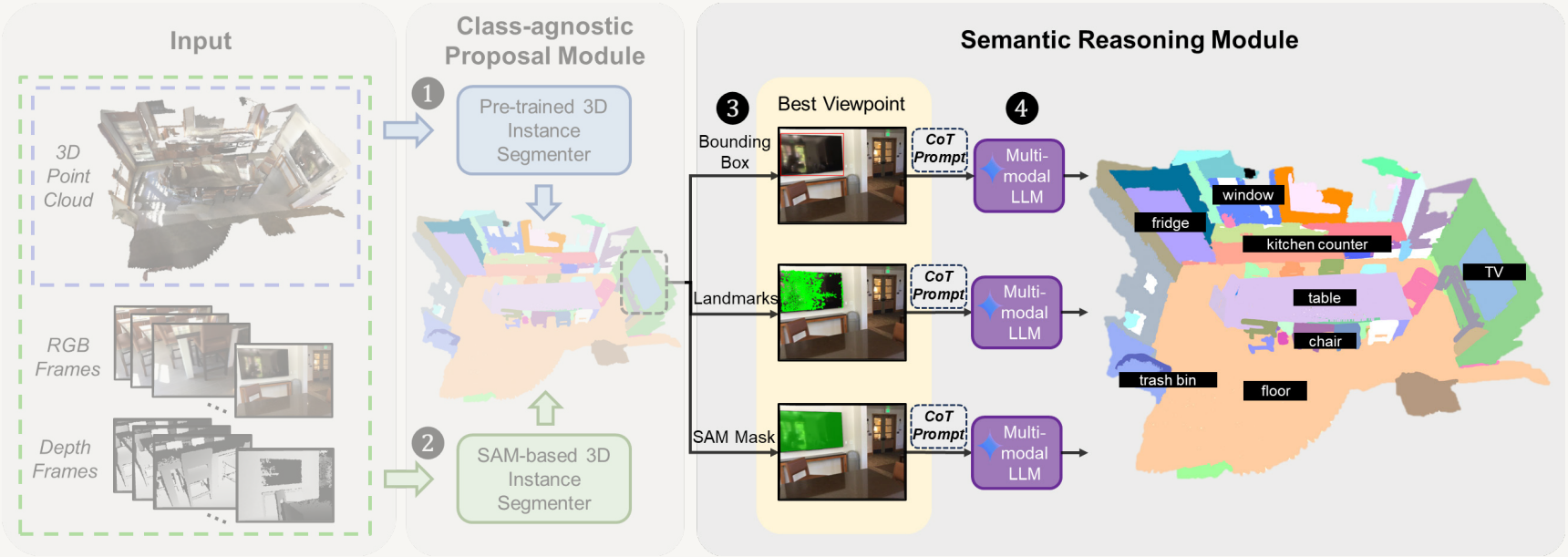
- Generate 3D object masks without category labels
- Combine pre-trained 3D segmenter + SAM-guided segmenter
- Robust proposals for small & rare objects



# OV3D-CG FRAMEWORK

## ➢ Semantic Reasoning Module

- Select best viewpoint for each object
- Build context-aware representations: Bounding Box, Landmarks, SAM mask
- Use multimodal LLM + Chain-of-Thought prompts



# QUANTITATIVE RESULTS

ScanNet200: 25.4% AP → surpass SOTA works

| Method | $AP$ | $AP_{50}$ | $AP_{25}$ | $AP_{head}$ | $AP_{comm}$ | $AP_{tail}$ |
|---|---|---|---|---|---|---|
| SAM3D (arXiv 2023) | 6.1 | 14.2 | 21.3 | 7.0 | 6.2 | 4.6 |
| OVIR-3D (CoRL 2023) | 13.0 | 24.9 | 32.3 | 14.4 | 12.7 | 11.7 |
| OpenMask3D (NeurIPS 2023) | 15.4 | 19.9 | 23.1 | 17.1 | 14.1 | 14.9 |
| OpenIns3D (ECCV 2024) | 8.8 | 10.3 | 14.4 | 16 | 6.5 | 4.2 |
| SAI3D (CVPR 2024) | 12.7 | 18.8 | 24.1 | 12.1 | 10.4 | 16.2 |
| Open3DIS (CVPR 2024) | 23.7 | 29.2 | 33.1 | 26.4 | 22.5 | 21.9 |
| **Ours** | **25.4** | **32.9** | **37.0** | **28.2** | **24.8** | **22.9** |

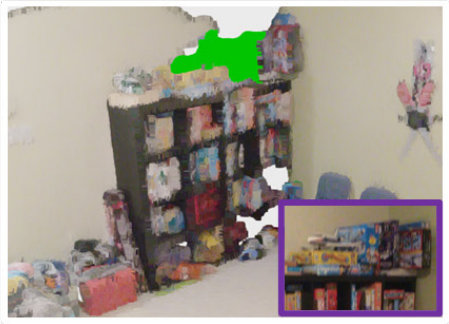
# QUANTITATIVE RESULTS

Replica: strong performance → proves generalization to rare objects

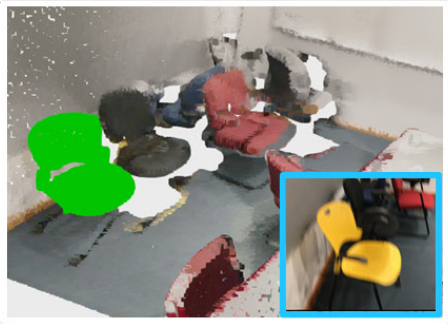
| Method | $AP$ | $AP_{50}$ | $AP_{25}$ |
|---|---|---|---|
| OpenScene + Mask3D (CVPR 2023) | 10.9 | 15.6 | 17.3 |
| OpenMask3D (NeurIPS 2023) | 13.1 | 18.4 | 24.2 |
| Open3DIS (CVPR 2024) | 14.9 | 18.8 | 23.6 |
| **Ours** | **20.0** | **25.8** | **32.0** |

# QUALITATIVE RESULTS

*“potted plant”*



*“board games”*



*“yellow chair”*

*“a grey fabric chair positioned next to a white desk”*

*“fireplace”*

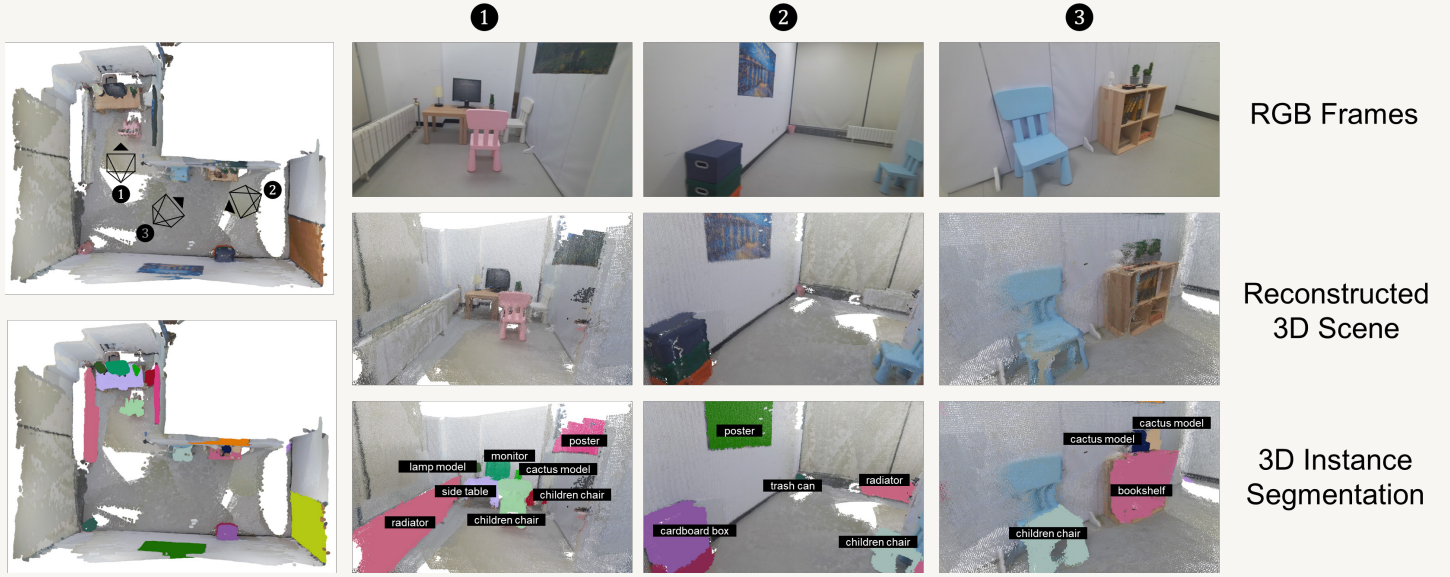
*“lampshade”*

*“purple duvet”*

*“a red cylindrical air compressor mounted on a metal rack”*

# REAL-WORLD DEMO

- Deployed on a **robot platform**
- Recognized & described objects **beyond predefined categories**
- Followed **natural instructions** like:
  - *“Find somewhere to sit”*
  - *“Navigate to where the books are”*



# KEY TAKEAWAYS

**CONTEXT MAKES A DIFFERENCE**

By reasoning about the environment, our system avoids many common errors

**FLEXIBLE AND ROBUST**

It works well in both benchmark datasets and real-world environments

**MOVES US CLOSER TO EMBODIED AI**

Robots that can understand the world in an open-ended way, just like humans do

# THANK YOU