BOSCH [1]
universität ulm [2]
TU WIEN TECHNISCHE UNIVERSITÄT WIEN [3]

# Open3DSG

## Open-Vocabulary 3D Scene Graphs from Point Clouds with Queryable Objects and Open-Set Relationships

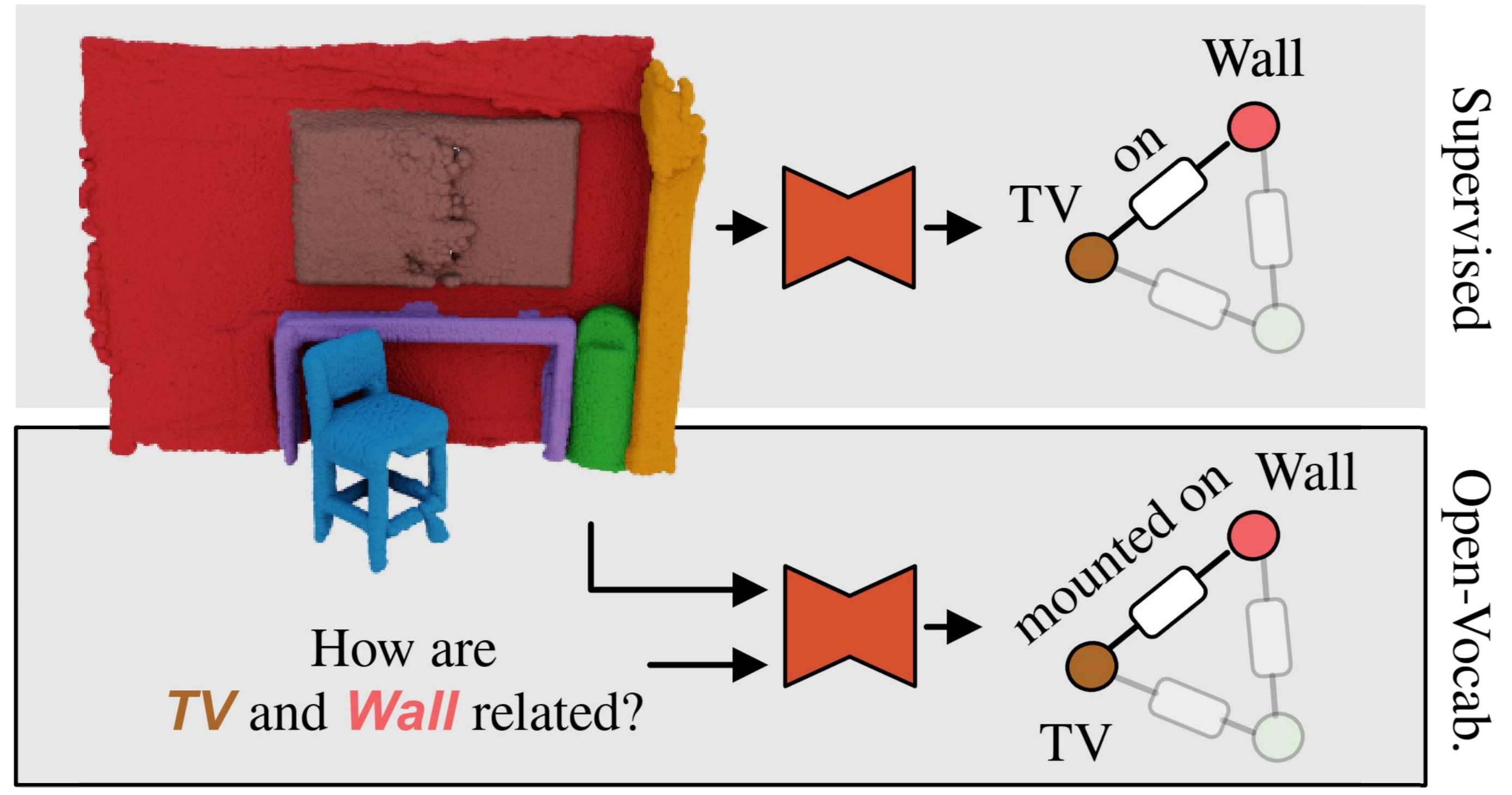
Sebastian Koch[1,2] Narunas Vaskevicius[1] Mirco Colosi[1] Pedro Hermosilla[3] Timo Ropinski[2]

SEATTLE, WA CVPR JUNE 17-21, 2024

Paper, Code & additional results:
**kochsebastian.com/open3dsg**

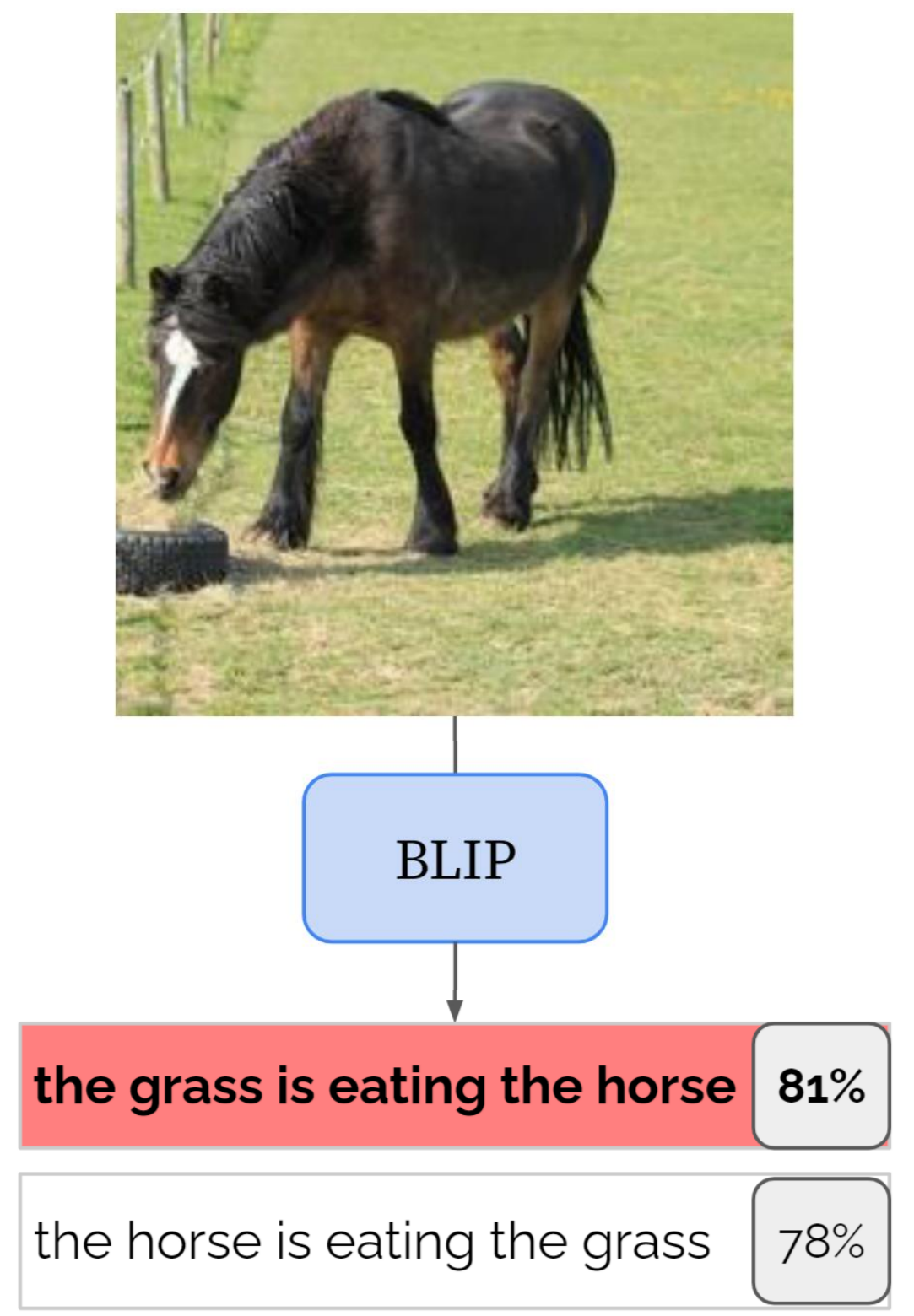
## 1. Overview

Introduction of Open-Vocabulary 3D Scene Graph Prediction Task



- ✓ No need for labeled 3D scene graph training data
- ✓ Interactive & not limited to pre-defined labels sets
- ! Challenging because VLMs struggle with compositionality

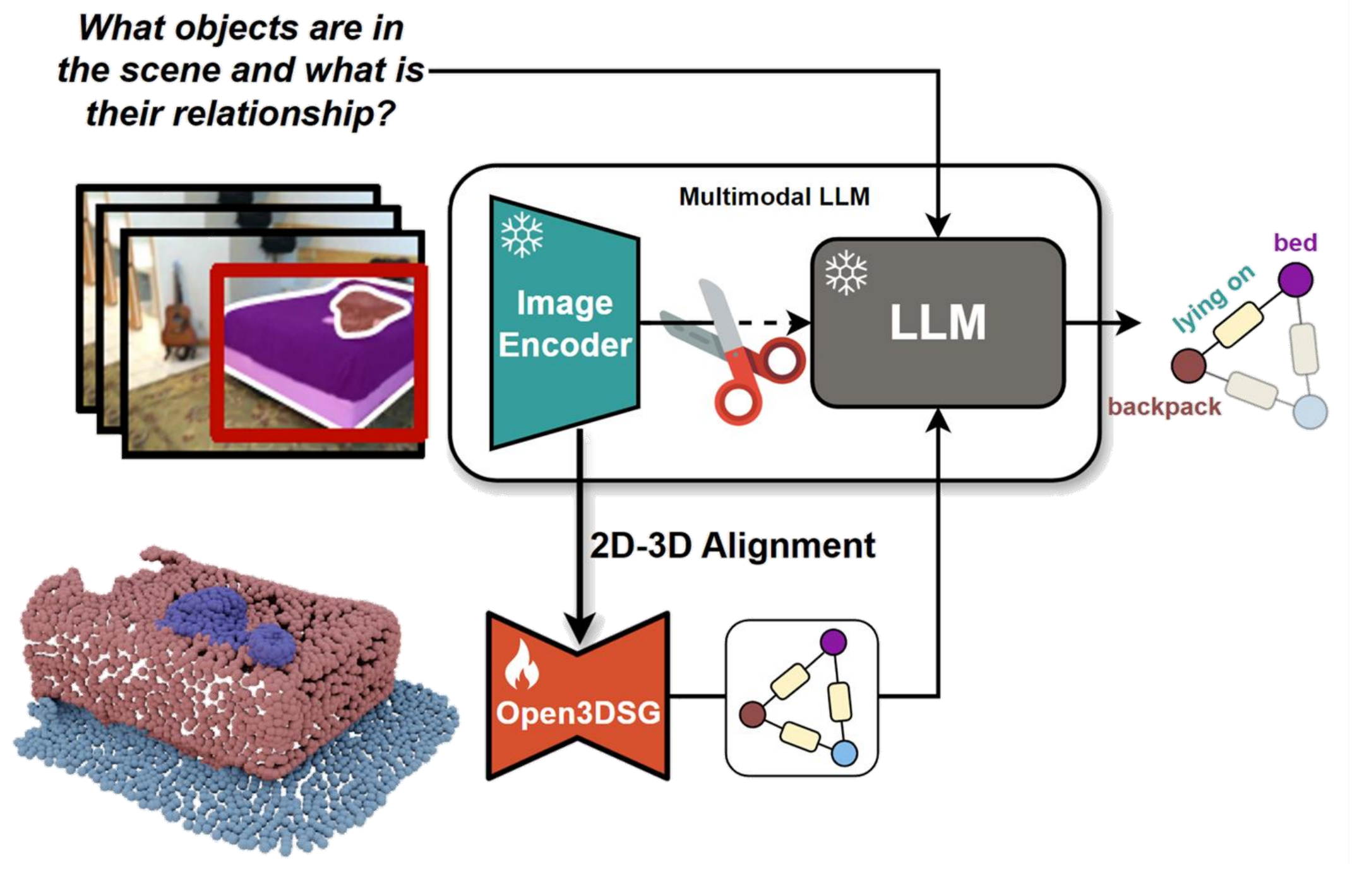
## 2. VLMs are BoWs


BLIP
the grass is eating the horse 81%
the horse is eating the grass 78%

*When and why vision-language models behave like bags-of-words, and what to do about it? – ICLR 2023*

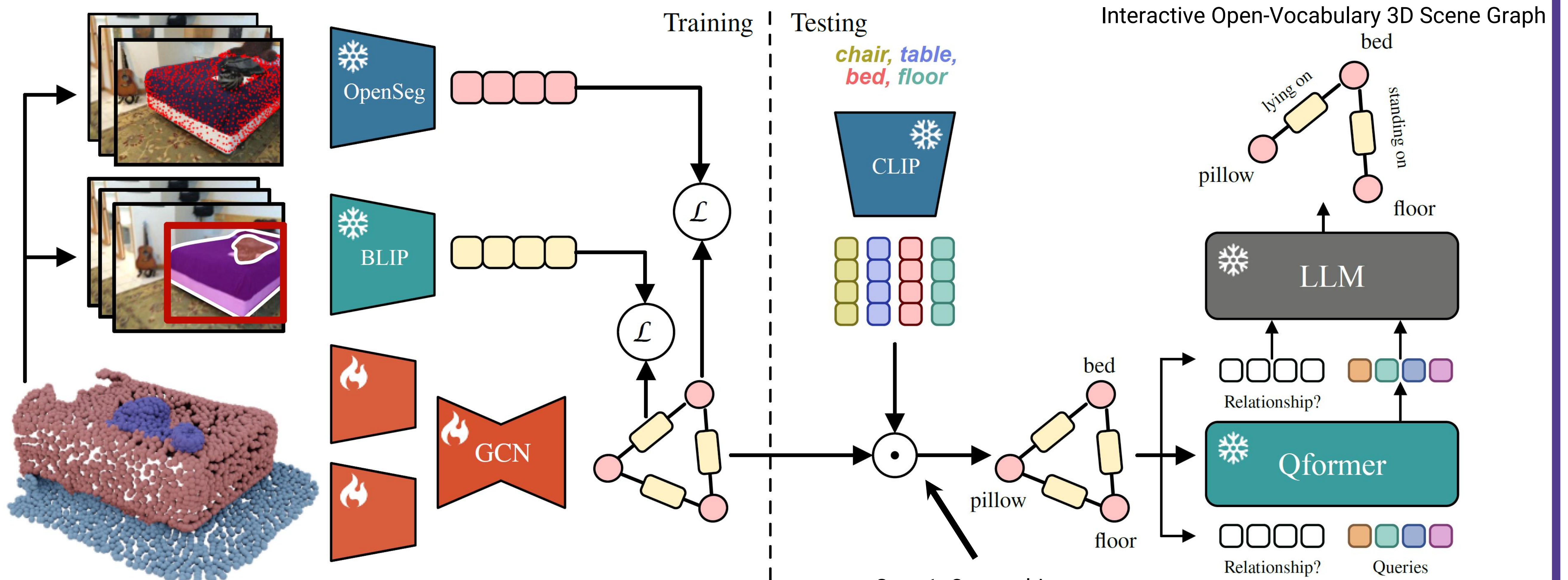
## 3. Key Idea

Unlike contrastive VLMs, multi-modal LLMs contain strong world knowledge but are limited to 2D representations



Distill an aligned vision encoder of an LLM into a 3D GNN for structured open-vocabulary 3D scene graph predictions by LLM prompting

## 4. Method

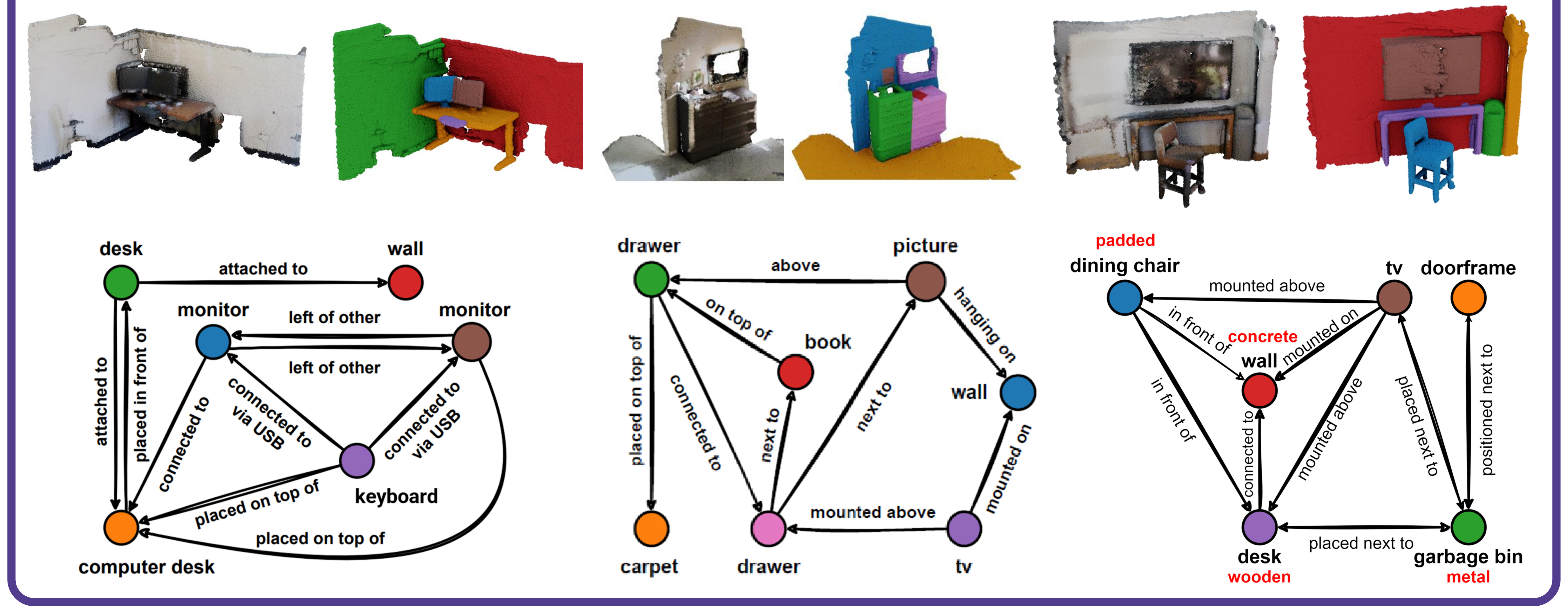


$$\mathcal{L} = 1 - \cos(\mathbf{F}_o^{2D}, \mathbf{F}_o^{3D}) + 1 - \cos(\mathbf{F}_r^{2D}, \mathbf{F}_r^{3D})$$

Step 1: Query objects based on CLIP similarity

Step 2: Prompt LLM for relationships between objects

## 5. Predicted Open-Vocabulary 3D Scene Graphs



## 6. Results

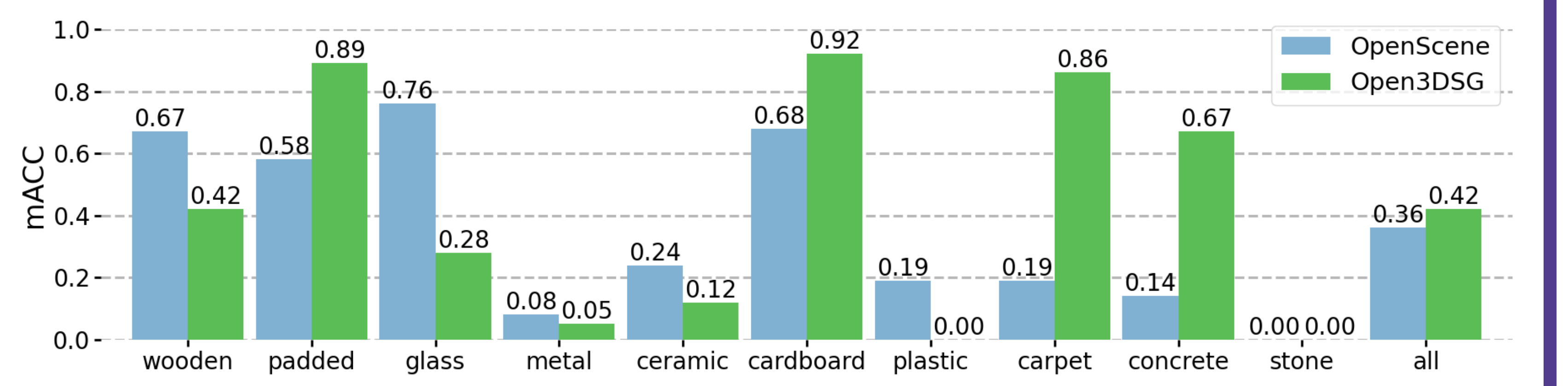
| | Object | | Predicate | | Relationships | |
|---|---|---|---|---|---|---|
| | R@5 | R@10 | R@3 | R@5 | R@50 | R@100 |
| *Fully Supervised* | | | | | | |
| 3DSSG | 0.68 | 0.78 | 0.89 | 0.93 | 0.40 | 0.66 |
| VL-SAT | **0.78** | **0.86** | **0.98** | **0.99** | **0.90** | **0.93** |
| *Zero-shot open-vocabulary* | | | | | | |
| CLIP (naïve) | 0.35 | 0.42 | 0.09 | 0.19 | 0.02 | 0.04 |
| OpenSeg+NegCLIP | 0.38 | 0.45 | 0.10 | 0.20 | 0.05 | 0.08 |
| **Open3DSG** | **0.51** | **0.62** | **0.62** | **0.70** | **0.63** | **0.65** |

| | | Labels | Head | Body | Tail | All |
|---|---|---|---|---|---|---|
| Objects R@5 | 3DSSG | 10^5 | 0.88 | 0.45 | 0.06 | 0.30 |
| | VL-SAT | 10^5 | **0.92** | **0.73** | 0.31 | **0.46** |
| | **Open3DSG** | **0** | 0.60 | 0.50 | **0.42** | 0.45 |
| Predicates R@3 | 3DSSG | 10^5 | 0.94 | 0.83 | 0.41 | 0.57 |
| | VL-SAT | 10^5 | **0.99** | **0.94** | **0.58** | **0.75** |
| | **Open3DSG** | **0** | 0.38 | 0.29 | 0.53 | 0.37 |

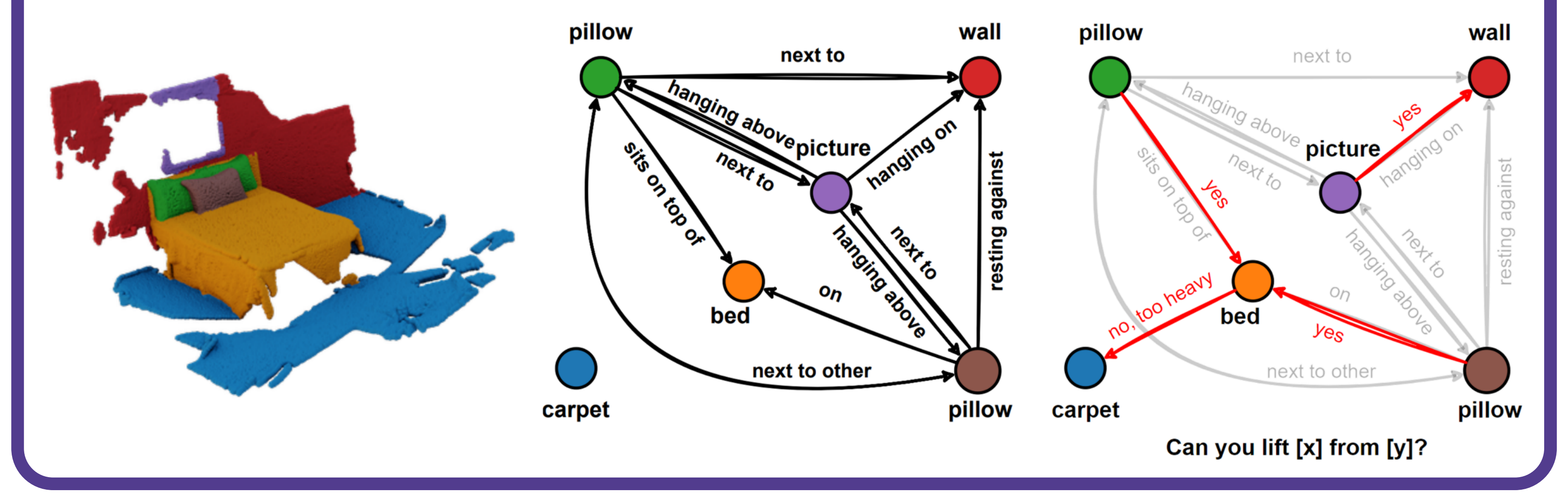
! Open3DSG excels for rare tail-distribution classes without requiring any labeled data

## 7. Zero-Shot Downstream Reasoning

### Scene Graph Attributes



### Interactive Scene Reasoning



Can you lift [x] from [y]?