Language-driven Scene Understanding with 3D Scene Graphs

Sebastian Koch
Ulm University and Bosch Center for AI

Huawei Munich Research Center

Feb 06, 2024

Who Am I?

- **PhD student since April 2022**
  - Timo Ropinski
  - Pedro Hermosilla
- **Sponsored by Bosch**
  - Narunas Vascevikius
  - Mirco Colosi

kochsebastian.github.io

Motivation

AR/VR

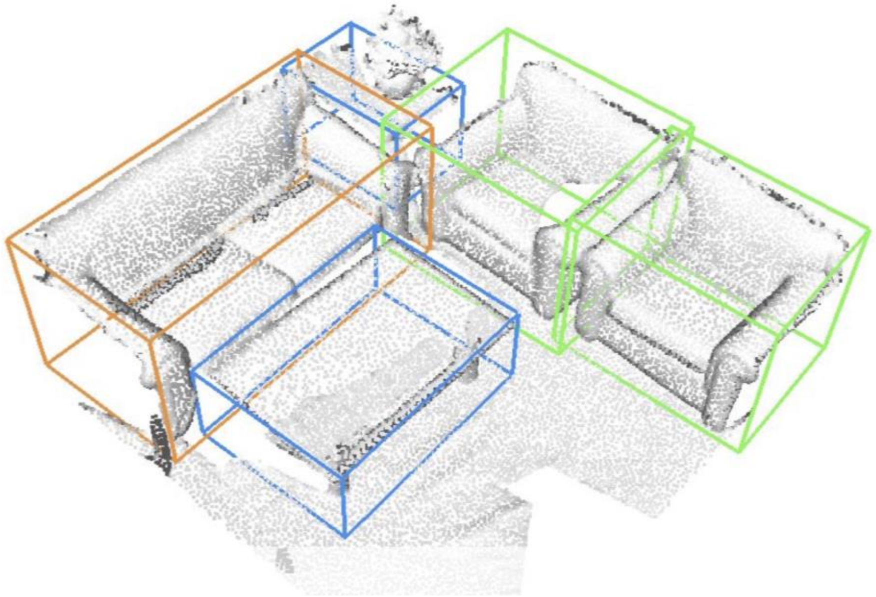
Robotics

Language-driven 3D scene understanding enhances AR/VR and robotics with richer context and actionable interaction.

3D Scene Representations

3D Semantic Instance Segmentation

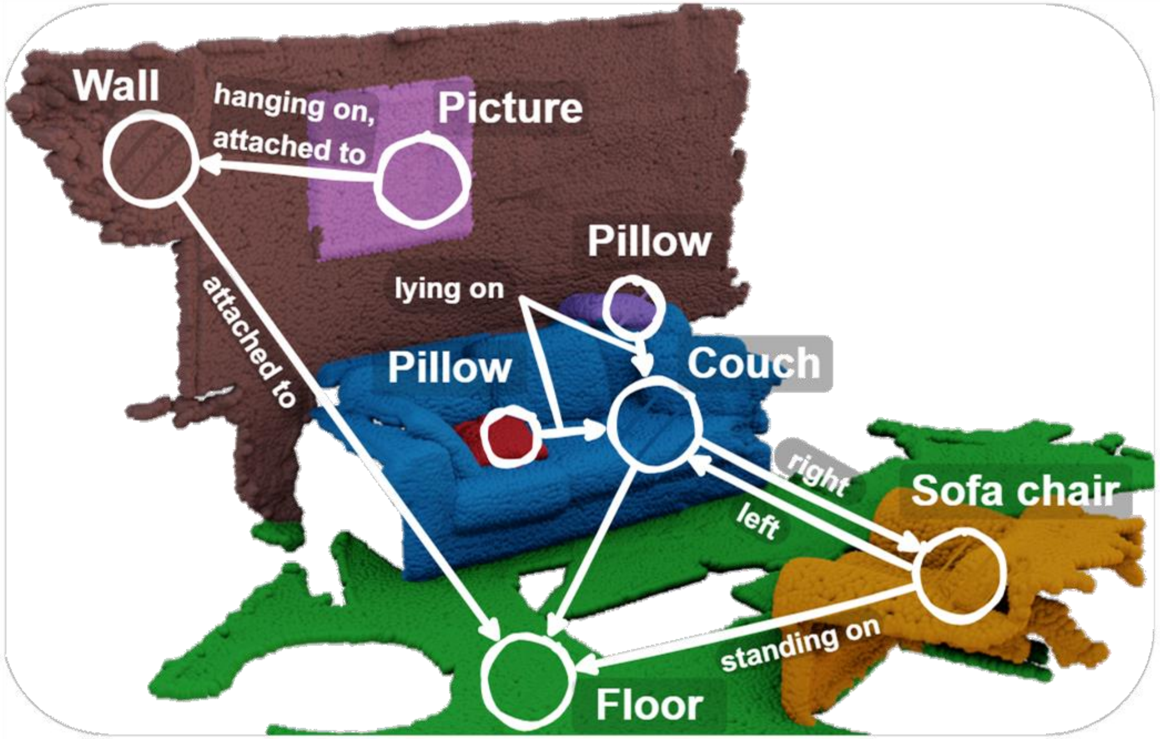
3D Object Detection

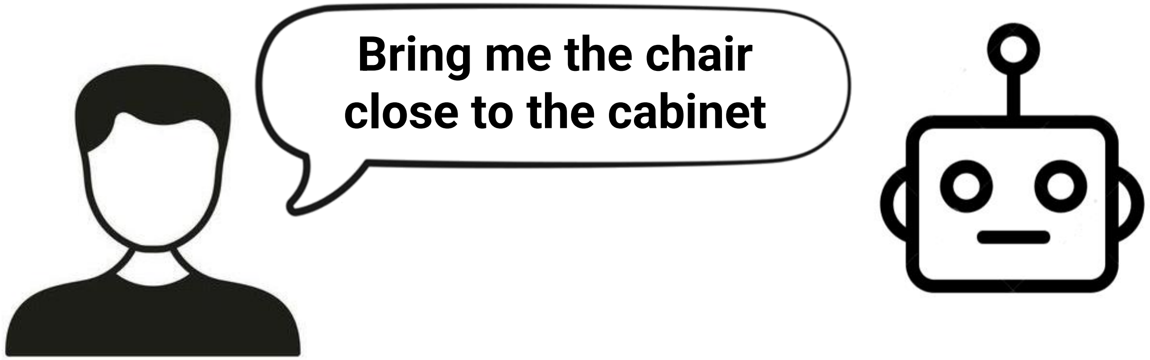


- Relationships and object interactions are often disregarded
- Expensive to store and difficult to directly use for downstream tasks like planning

[1] Misra et al.: An End-to-End Transformer Model for 3D Object Detection, ICCV'2021
[2] Schult et al.: Mask3D: Mask Transformer for 3D Instance Segmentation, ICRA'2023

Why do relationships matter?

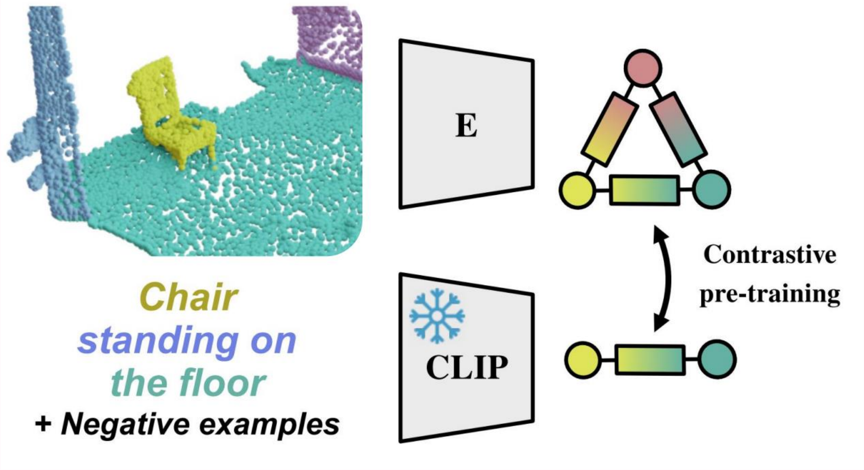




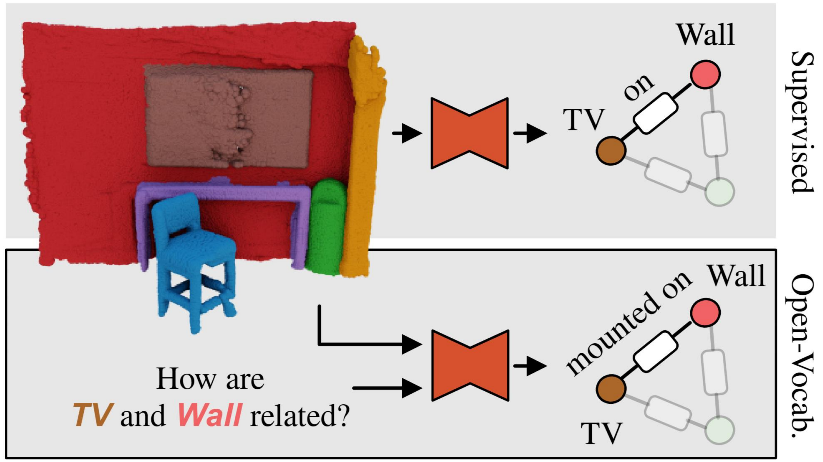
🥺 3D Scene Graphs can model

- Objects
- Relationships
- Affordances
- Attributes
- Etc.

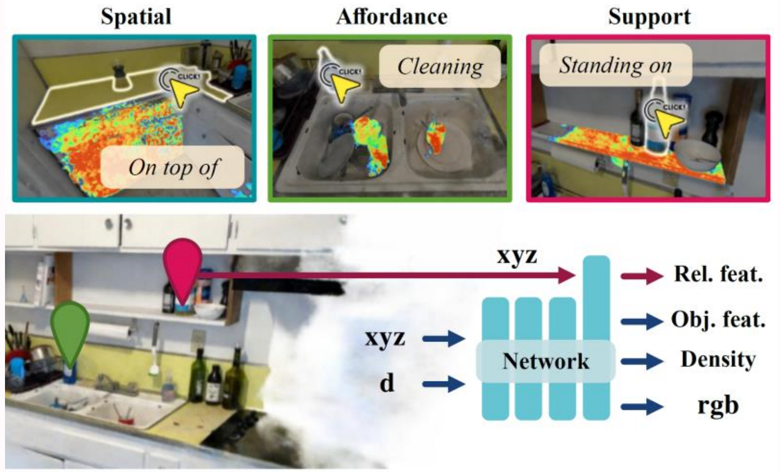
Talk outline



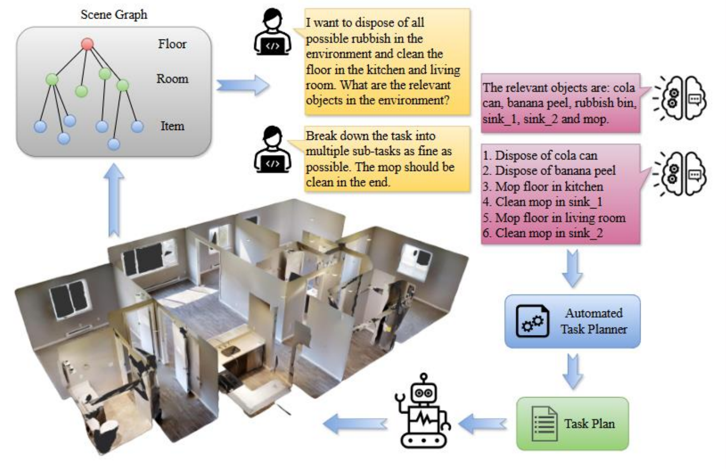
Lang3DSG (3DV 2024)



Open3DSG (CVPR 2024)



RelationField (under review)



DELTA (ICRA 2025)

Lang3DSG

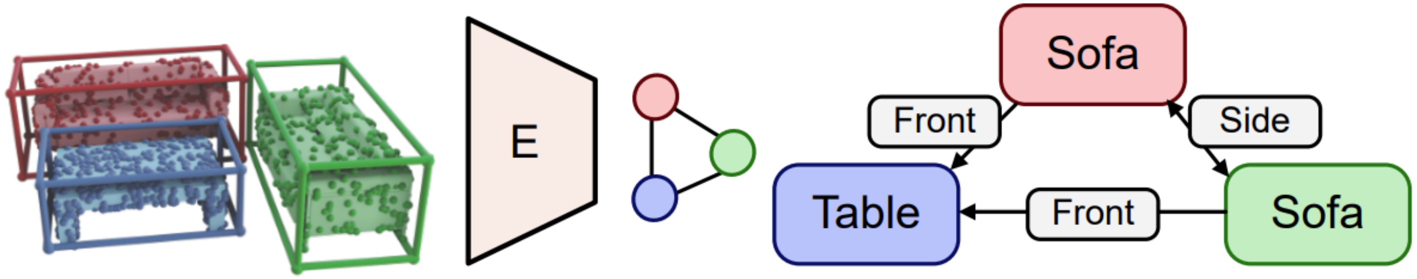
Language-based contrastive pre-training for 3D Scene Graph prediction

3DV 2024

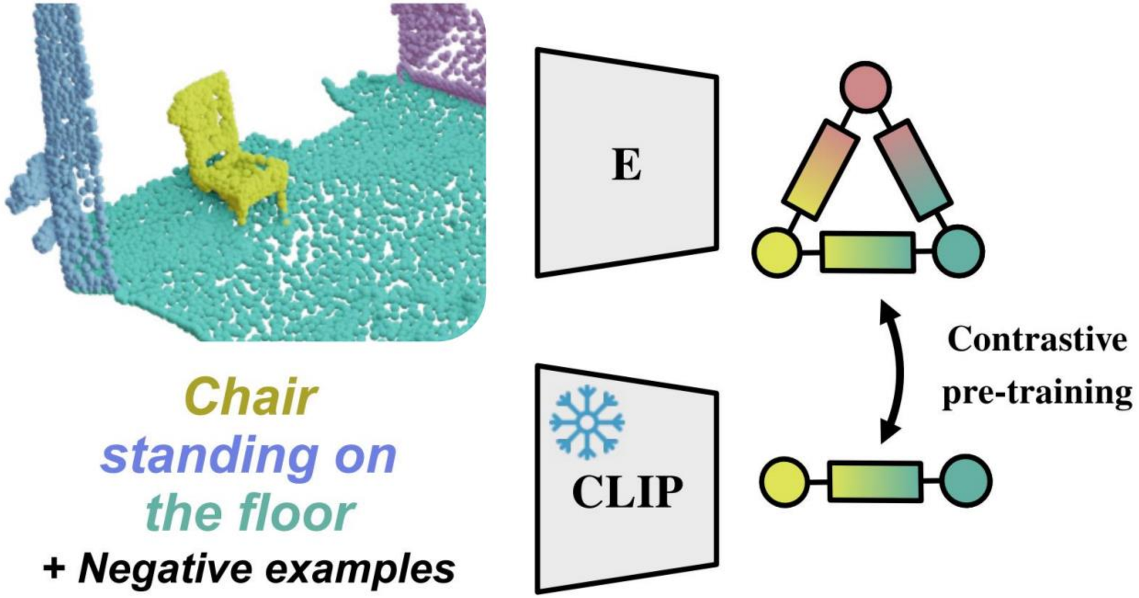
Sebastian Koch Pedro Hermosilla Narunas Vascevicius
Mirco Colosi Timo Ropinski

Language & 3D Scene Graphs

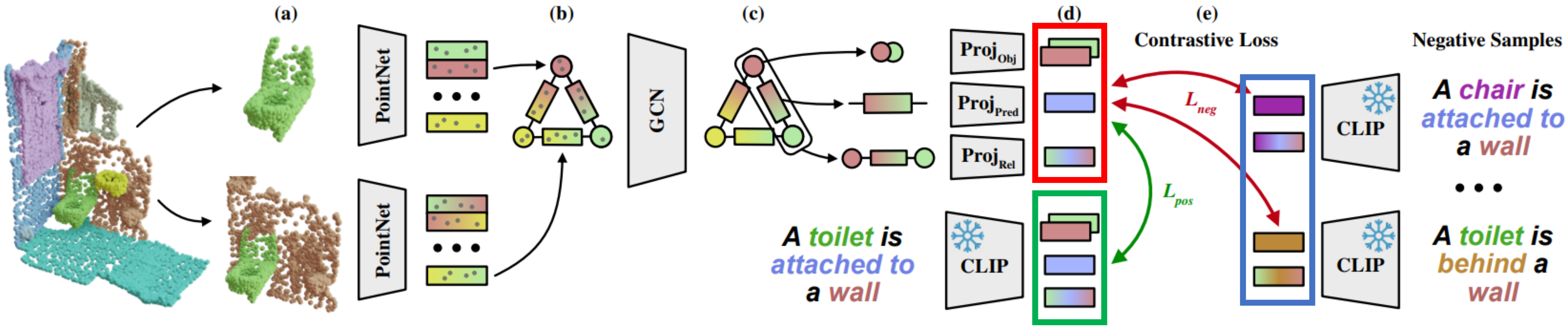
Challenge: Learning 3D Scene Graphs needs a lot of annotated data



Key Idea: Leverage natural similarity between 3D Scene Graphs & Language



Lang3DSG Approach

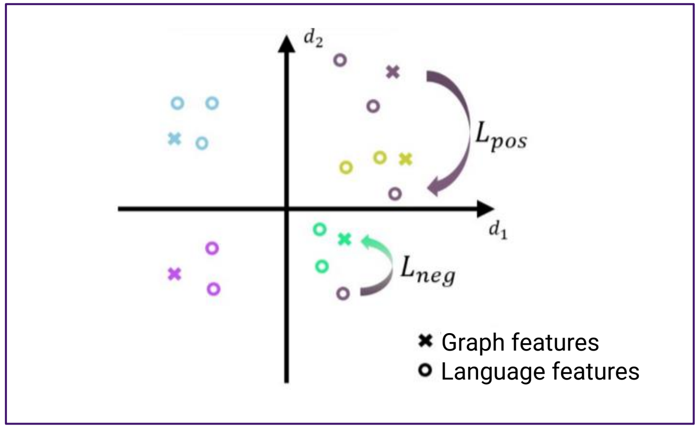


Losses

$$\mathcal{L}_{pos} = \sum_{i=1}^{N} \frac{1}{|K|} \sum_{j \in K} 1 - \cos(\boxed{f_i}, \boxed{f^t_{h(j)}})$$

$$\mathcal{L}_{neg} = \sum_{i=1}^{N} \frac{1}{|M|} \sum_{j \in M} \max\left(0, \cos(\boxed{f_i}, \boxed{f^t_{h(j)}}) - \tau\right)$$

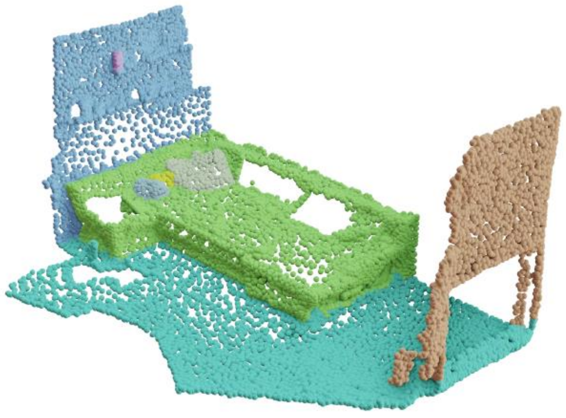
Feature space

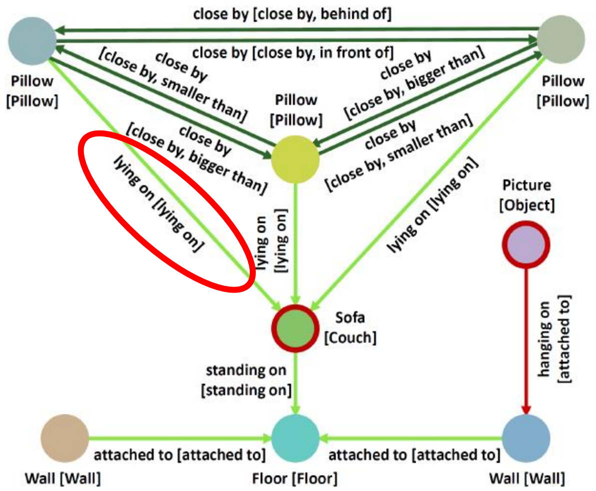


Fine-tuning on **pre-defined** classed needed!

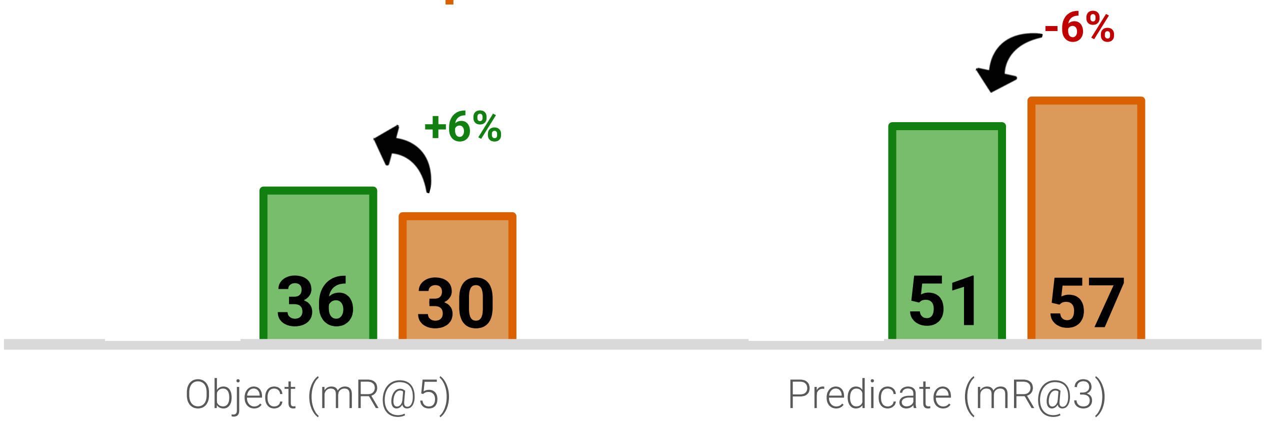
Lang3DSG Results

3D Scene Graph prediction with fine-grained labels

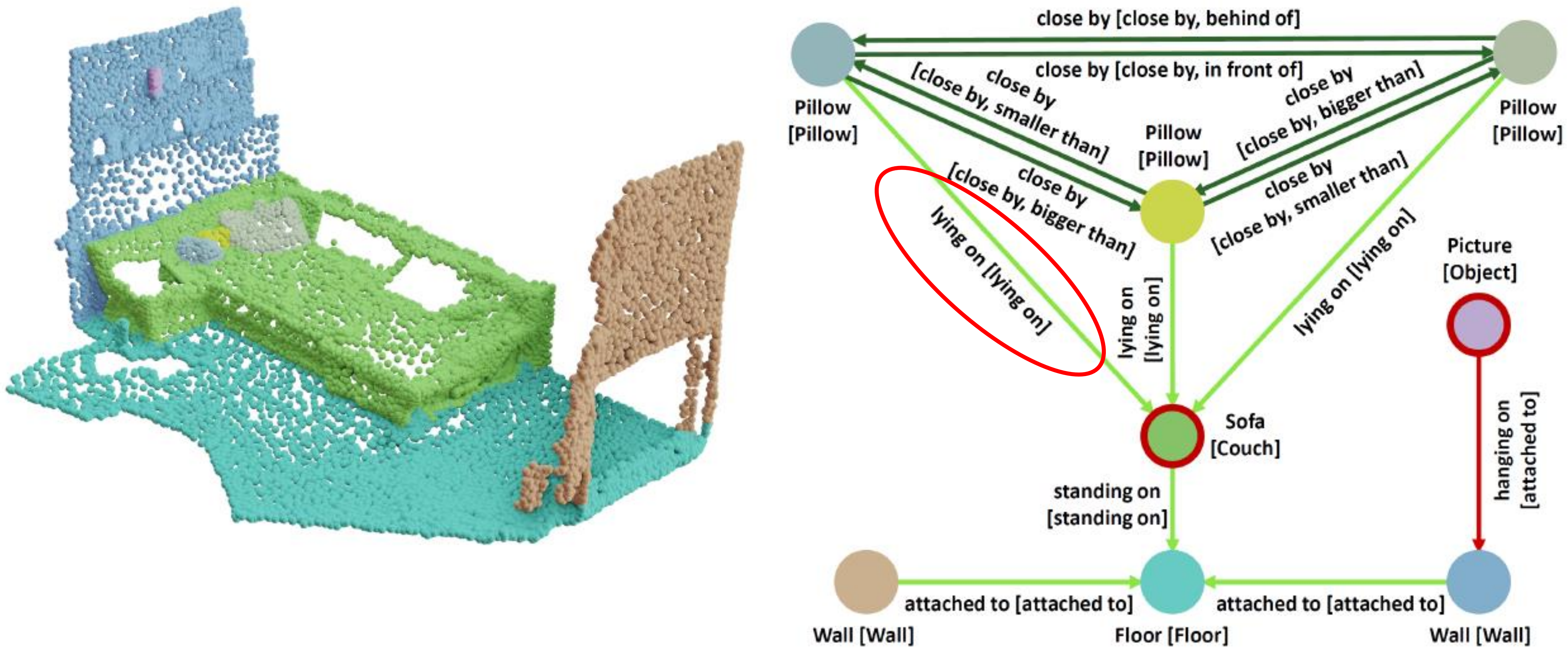




How does **point cloud pre-training** compare to a **supervised** method?

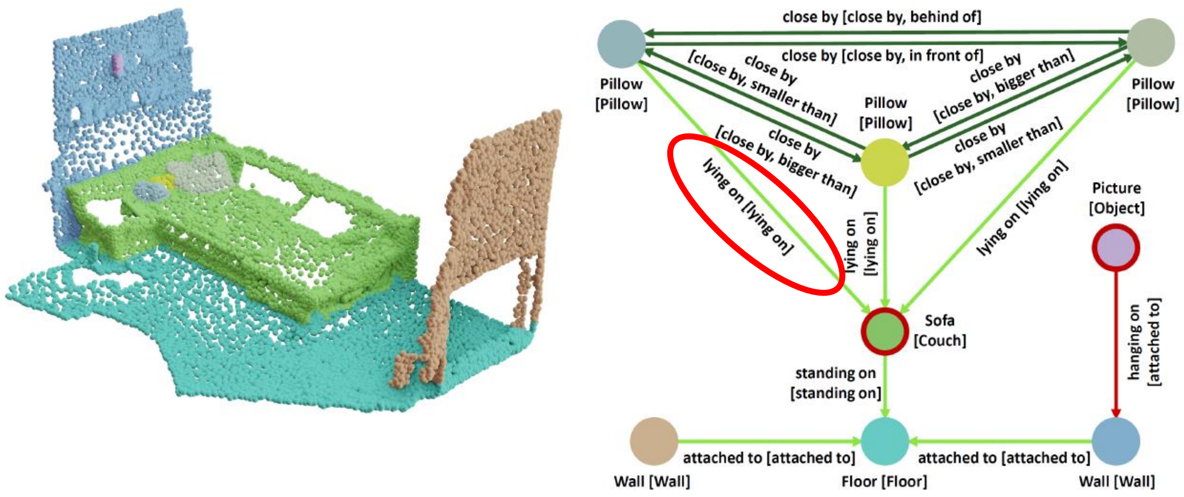


Lang3DSG Results

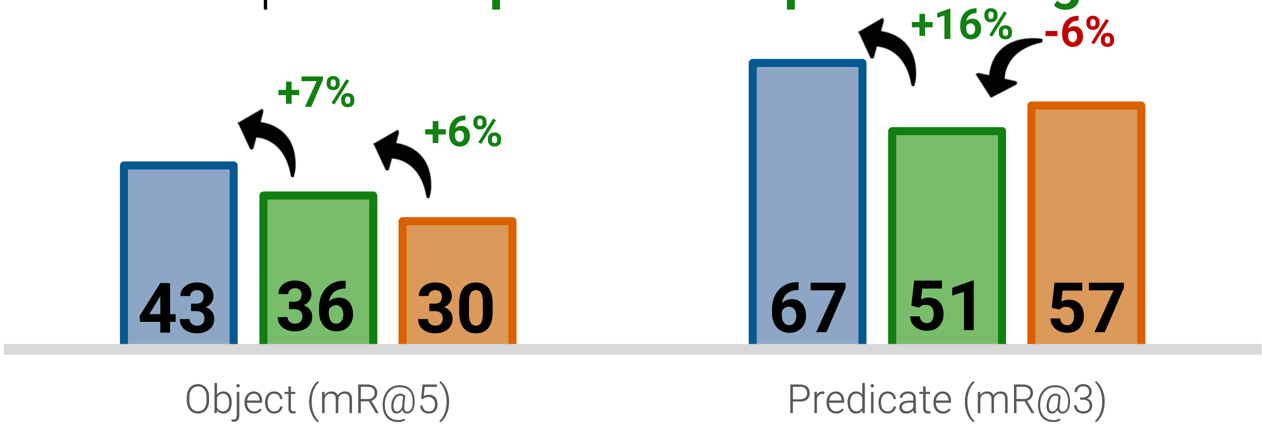


Lang3DSG Results

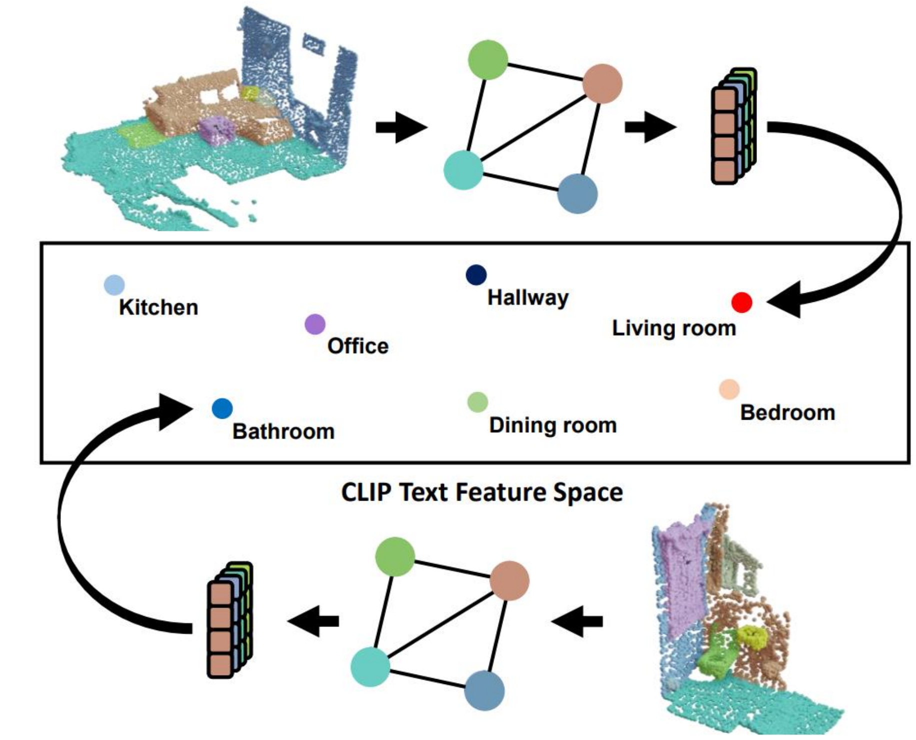
3D Scene Graph prediction with fine-grained labels

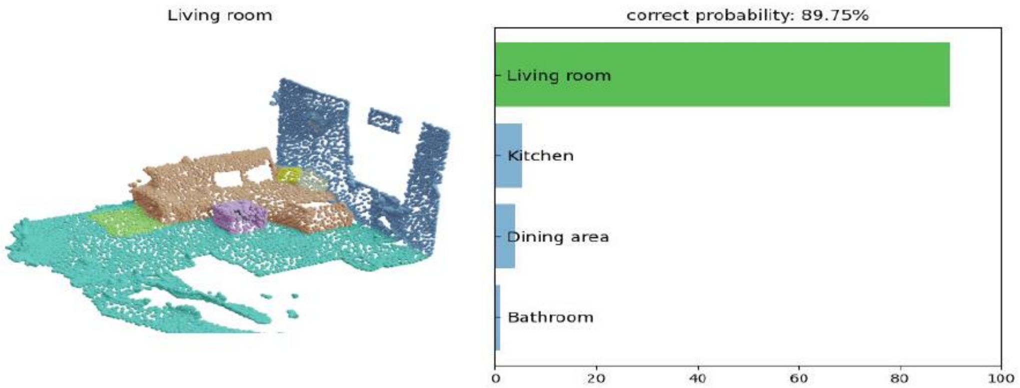


How does **language-based SG pre-training** compare to a **point cloud pre-training**?



Language-based downstream applications





Take aways

- **Lang3DSG pre-training** achieves **SOTA** 3D Scene Graph prediction.
- **Long-tail relationships** are recognized **exceptionally well**.
- **Language alignment** enables **zero-shot** applications.
- **Fine-tuning on predefined classes** is still needed!

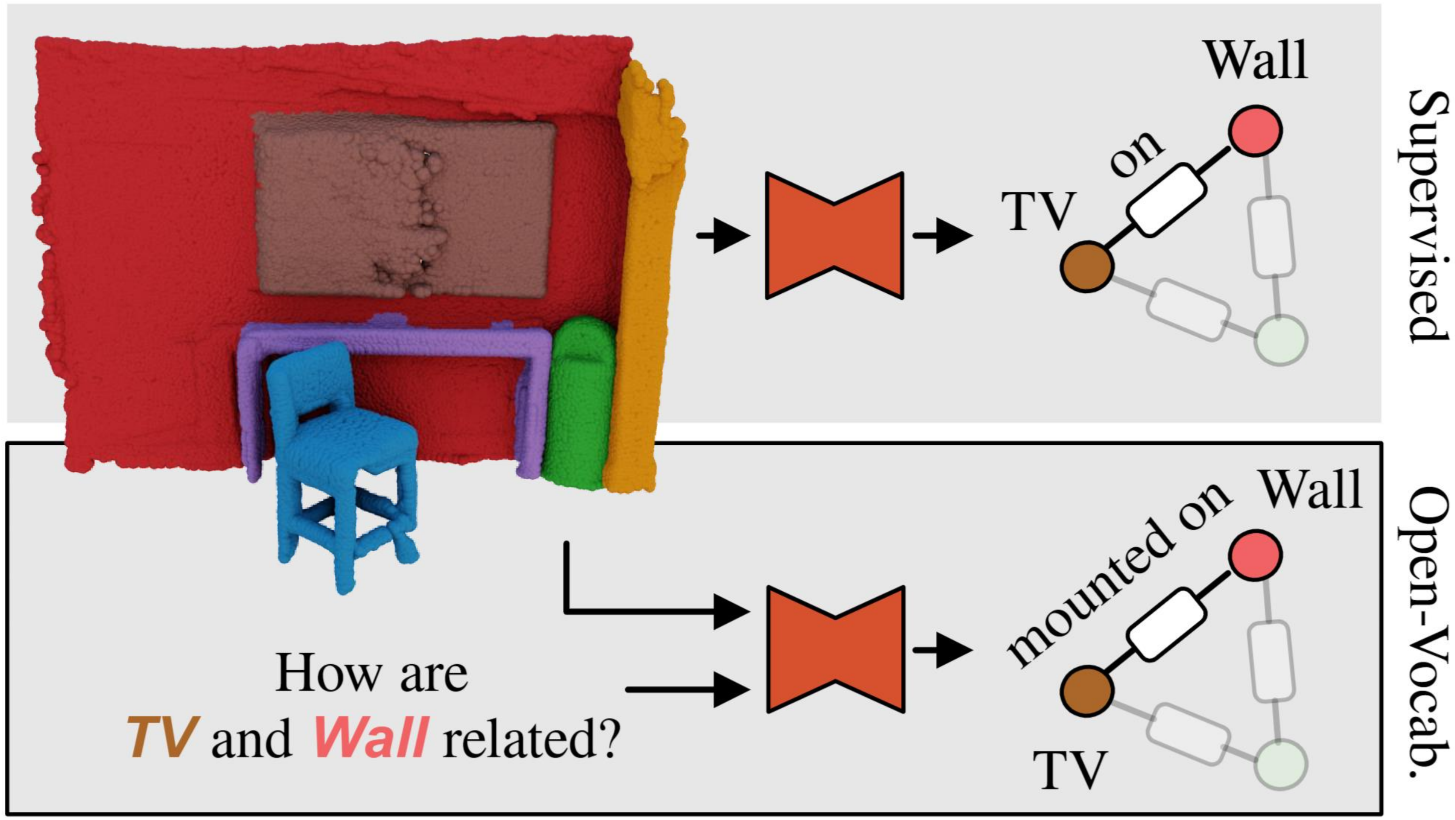
Open3DSG

Open-Vocabulary 3D Scene Graphs with Queryable Objects & Open-Set Relationships

CVPR 2024

Sebastian Koch Narunas Vascevicius Mirco Colosi
Pedro Hermosilla Timo Ropinski

Motivation



Limited/Pre-defined Object & Relationships labels

Research Questions

🤔 Can we use 2D foundation models for 3D relationship reasoning?

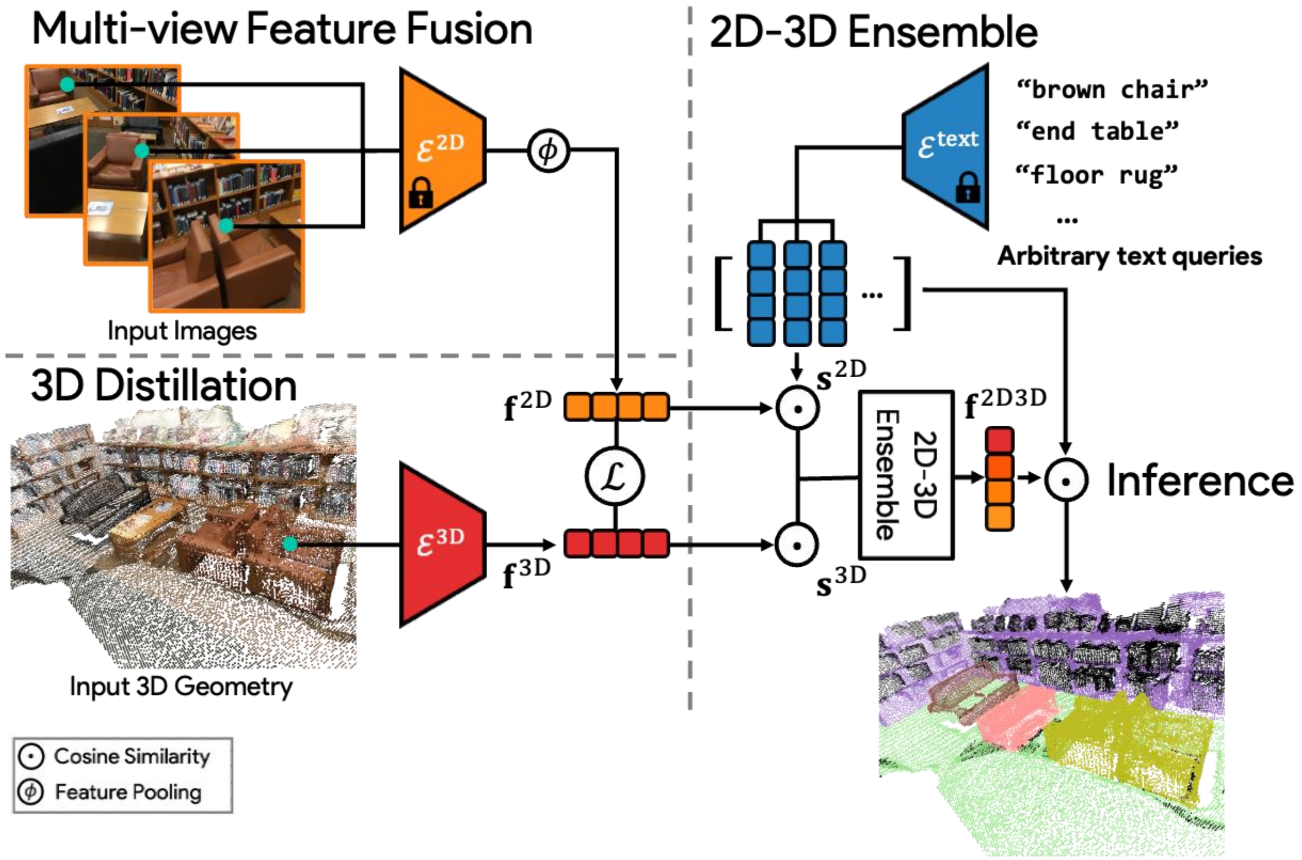
🤔 How can we distill knowledge from a 2D model into a 3D model?

Open-Vocabulary 3D Understanding

Goal
3D Open-Vocabulary Semantic Segmentation

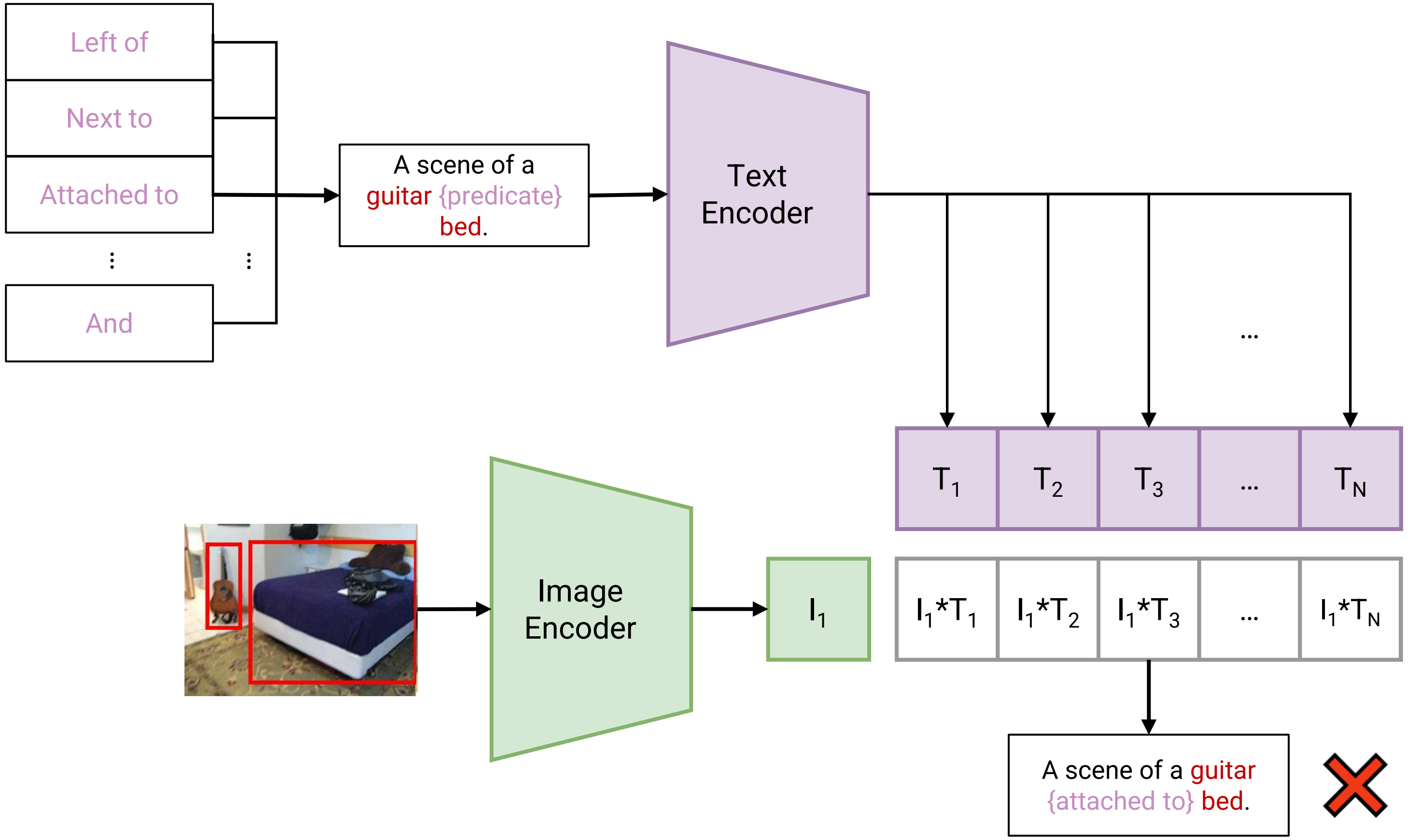
Requirements
- 3D point cloud
- Multi-View Images
- Depth + Pose

Insight
2D CLIP features transferable using projection & cosine-similarity distillation

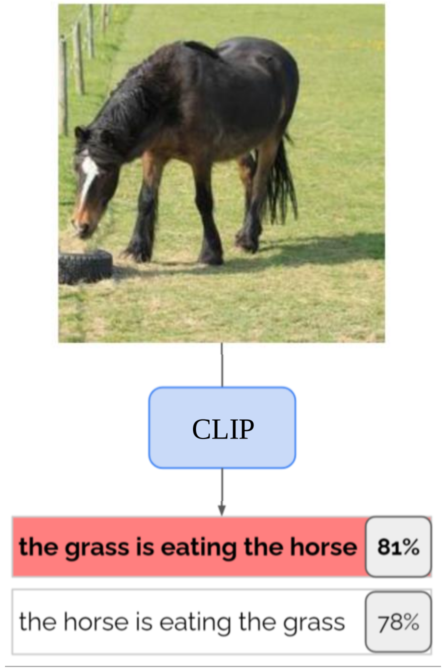


[1] Peng et al.: OpenScene: 3D Scene Understanding with Open Vocabularies, CVPR'2023

CLIP = Bag-of-words representation



Extensive Study here:



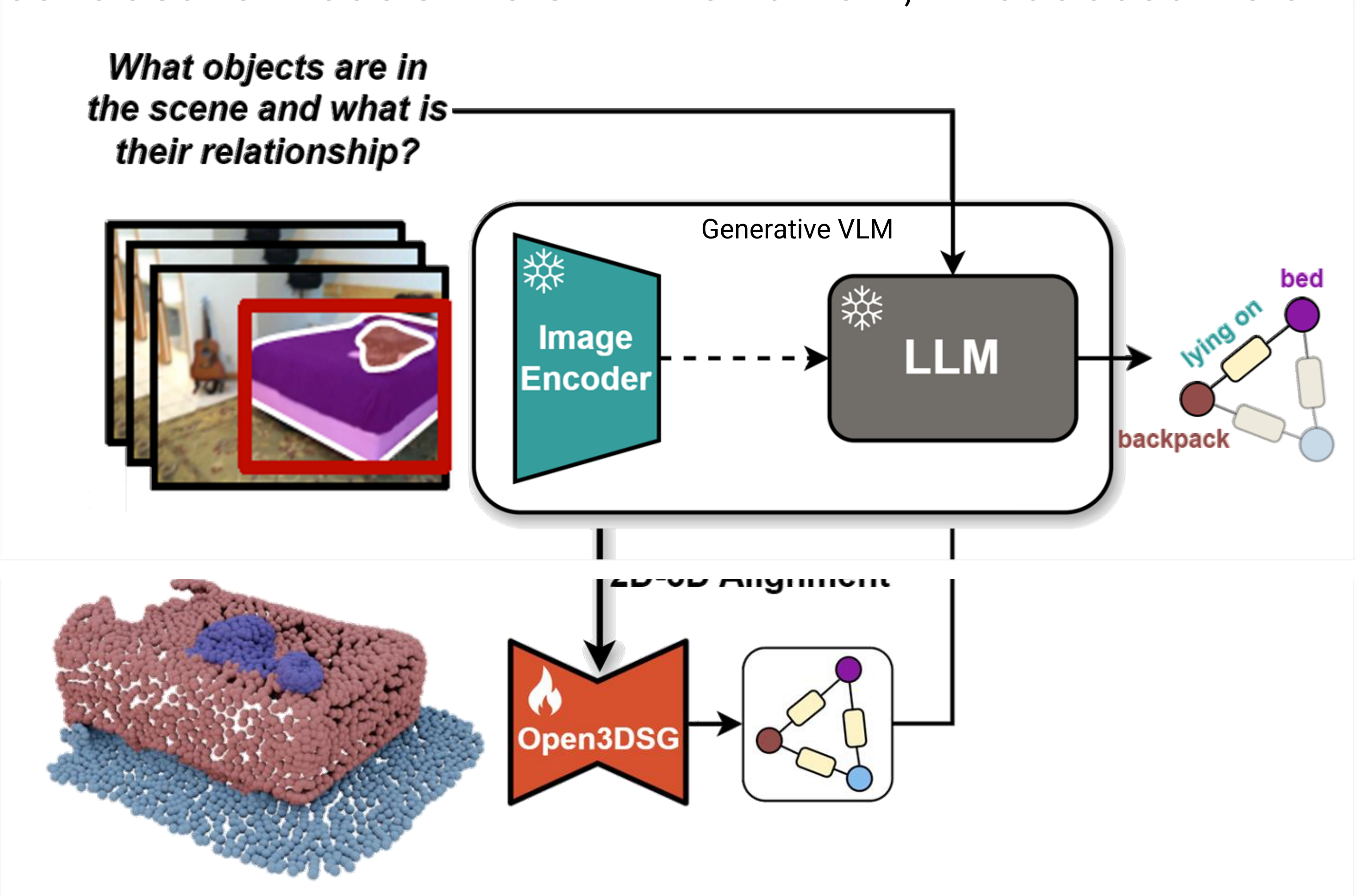
When and why vision-language models behave like bags-of-words, and what to do about it? – ICLR 2023

Insight: While good for object classification, CLIP does not understand relationships

[1] Radford et al.: Learning Transferable Visual Models From Natural Language Supervision, ICML'2021

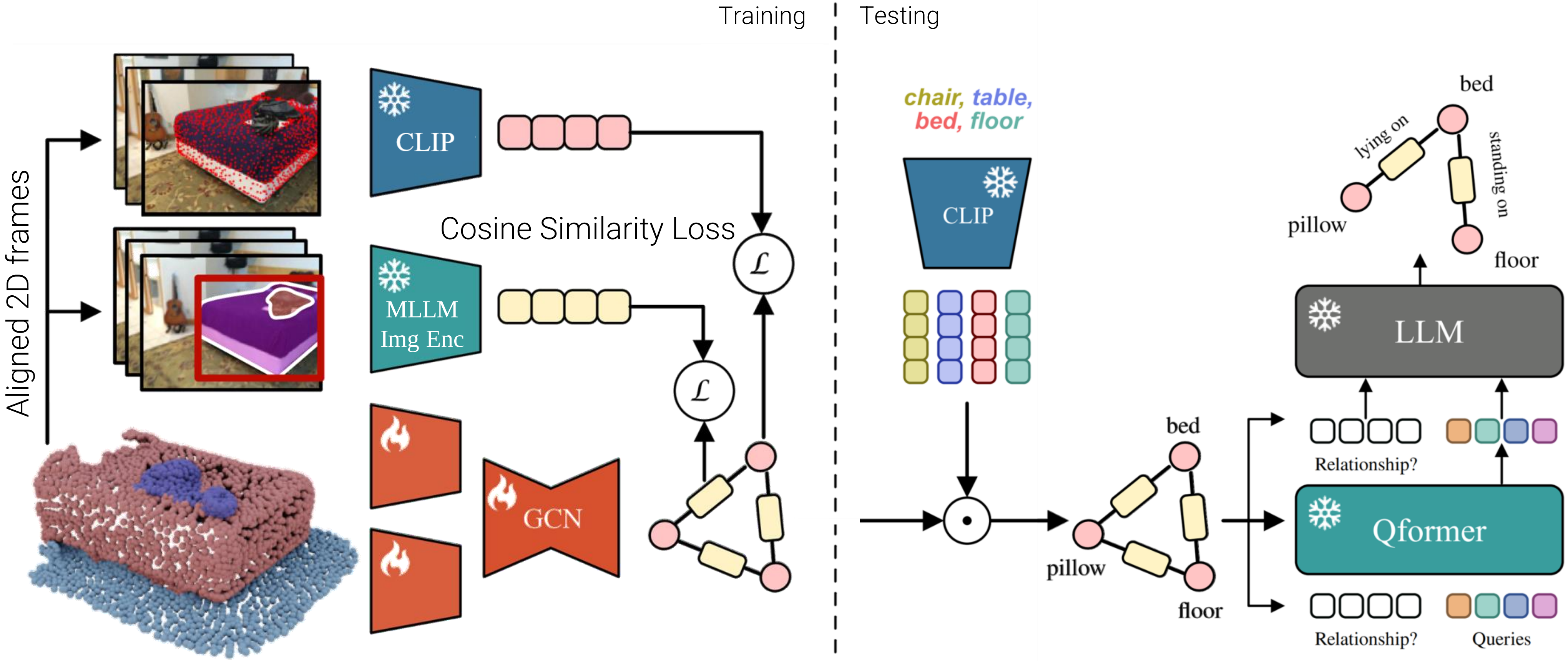
Core Idea

Question: When contrastive models like CLIP won't work, what about multi-modal LLMs?



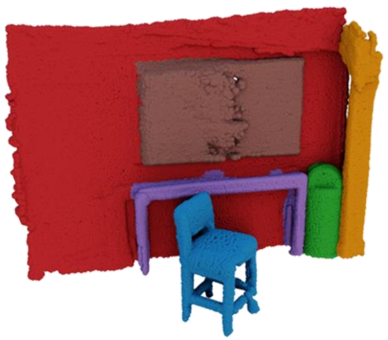
Idea: Condition the LLM output with a 3D Scene Graph backbone

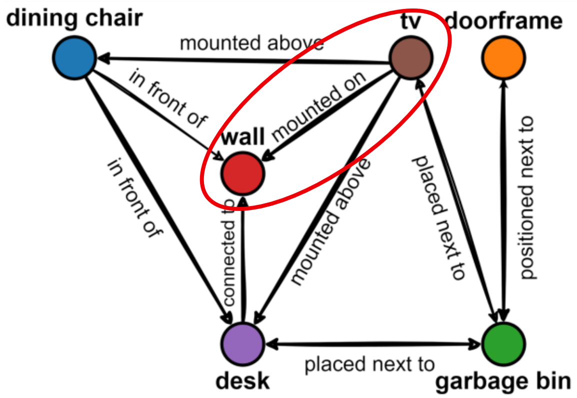
Open3DSG: A closer look



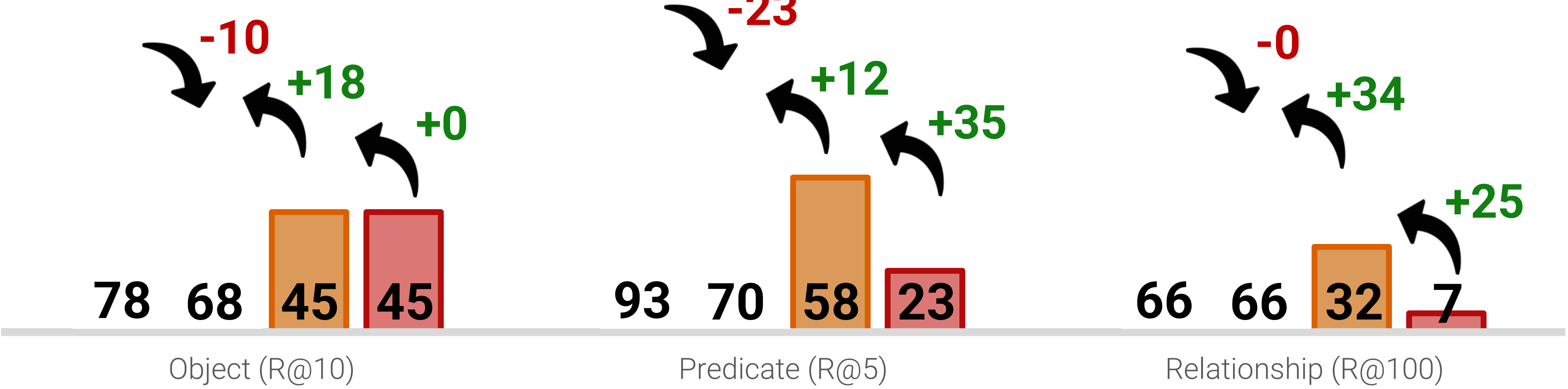
$$\mathcal{L} = 1 - \cos(\mathbf{F}_o^{2D}, \mathbf{F}_o^{3D}) + 1 - \cos(\mathbf{F}_r^{2D}, \mathbf{F}_r^{3D})$$

Open-Vocabulary 3D Scene Graphs



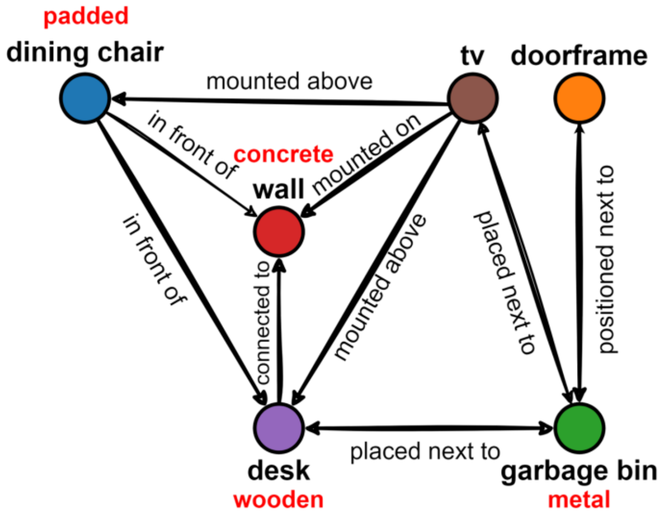


How does our **open-vocabulary** method compare to a **supervised** method?

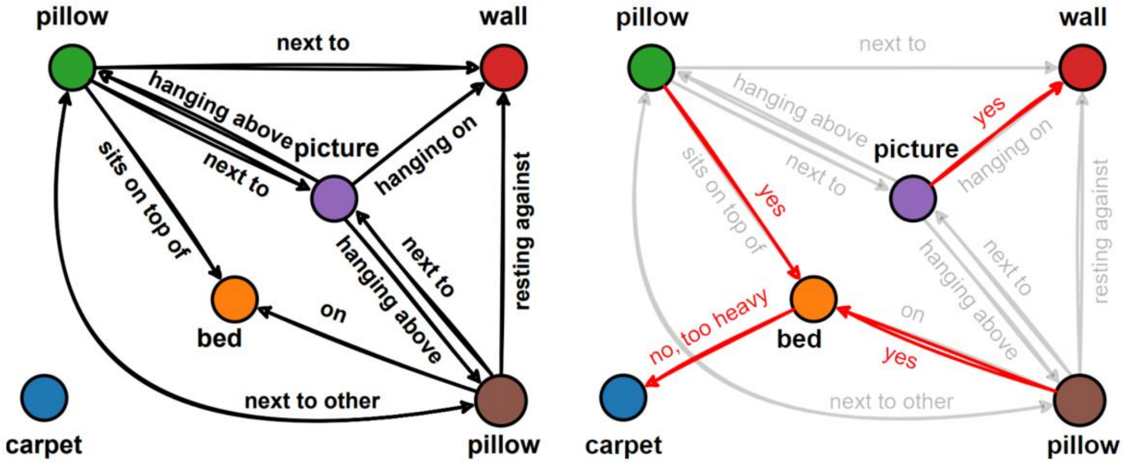


Scene Graph Scene Reasoning

Attribute Querying



Affordance Prompting



Can you lift [x] from [y]?

Take aways

- **Open3DSG enables open-vocabulary reasoning** for objects and relationships in 3D scenes.
- **LLM-based predictions outperform CLIP-based queries**, enabling more accurate and flexible scene understanding.
- **Zero-shot inference supports attributes, affordances, and task-specific interactions**, without requiring manual annotations.
- **No labeled data is needed for training**, reducing annotation costs and improving scalability.
- **Requires 2D-3D aligned datasets** for effective training and scene grounding.

RelationField

Relate Anything in Radiance Fields

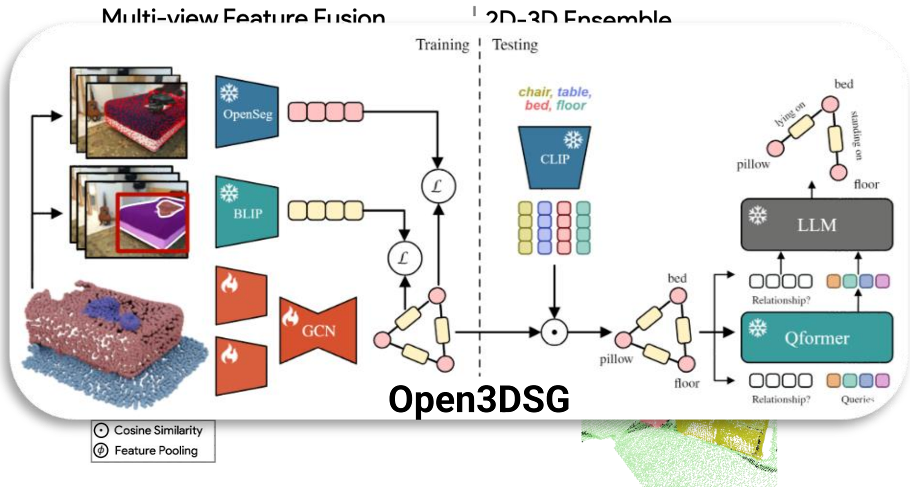
under review

Sebastian Koch Johanna Wald Narunas Vascevicius Mirco Colosi
Pedro Hermosilla Federico Tombari Timo Ropinski

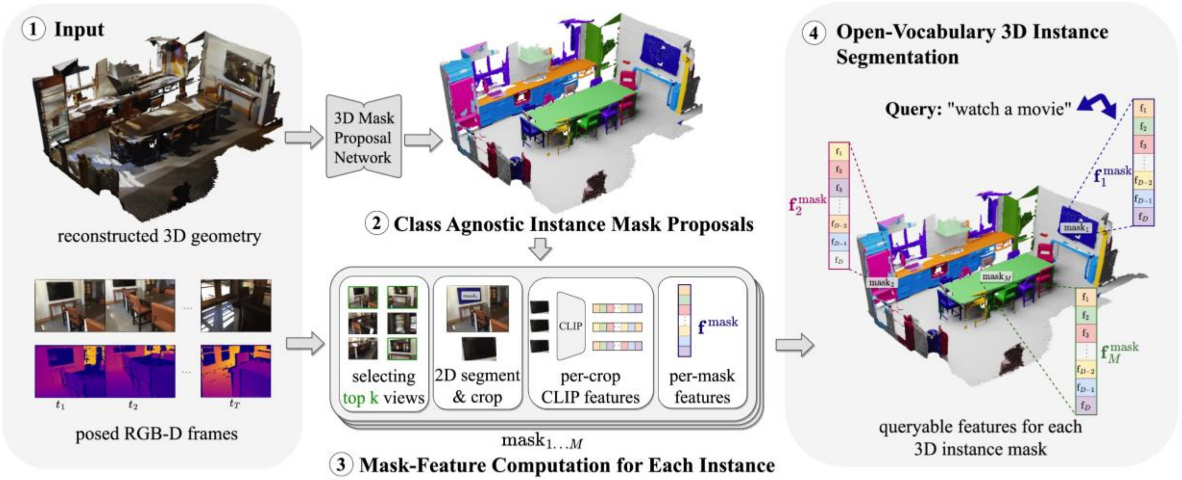
universität ulm BOSCH Google TECHNISCHE UNIVERSITÄT WIEN Vienna | Austria

Motivation

Distillation (OpenScene)



Mask-Lifting (OpenMask3D)



😟 Training requires aligned 2D-3D

😊 Inference can be done using 3D alone

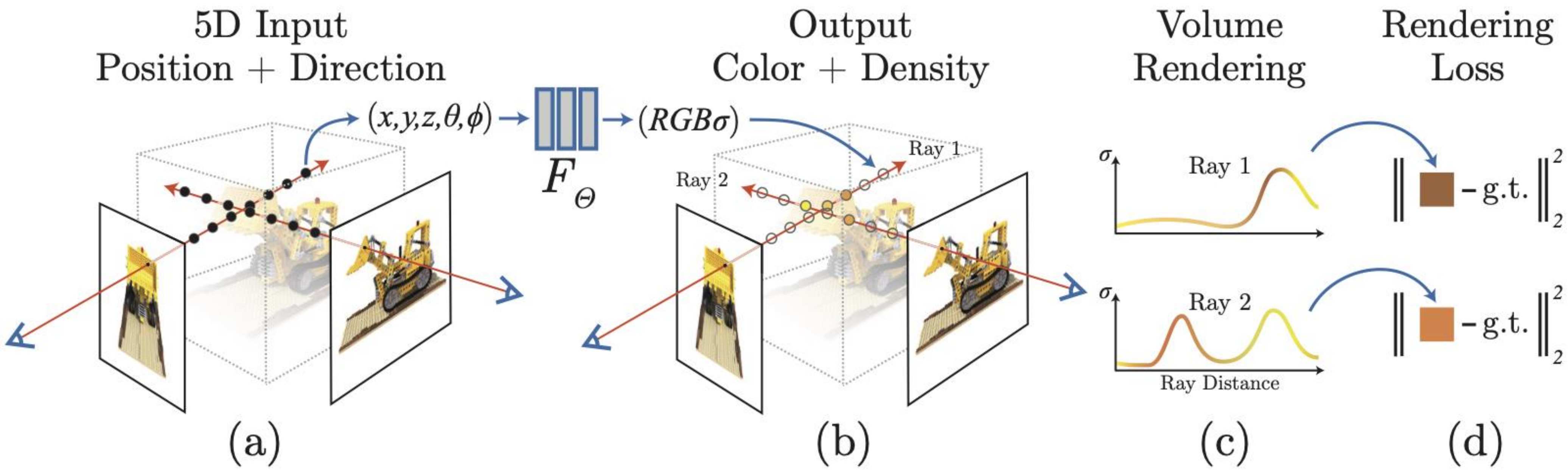
😊 Separate training of 2D & 3D backbones

😟 Inference needs aligned 2D & 3D data

🤔 **Question:** Can we train on 2D images alone but reason about 3D scene graphs & relationships?

[1] Peng et al.: OpenScene: 3D Scene Understanding with Open Vocabularies, CVPR’2023
[2] Takmaz et al.: OpenMask3D: Open-Vocabulary 3D Instance Segmentation, NeurIPS’2024

Radiance Fields

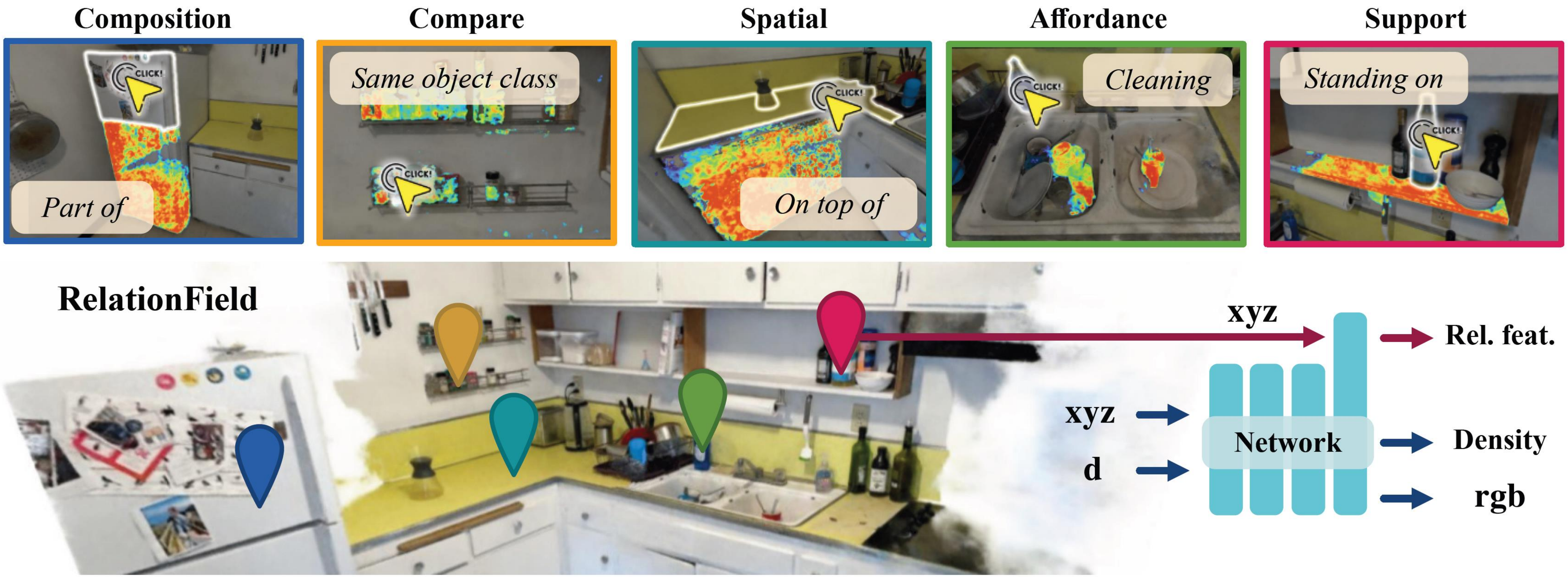


NeRF

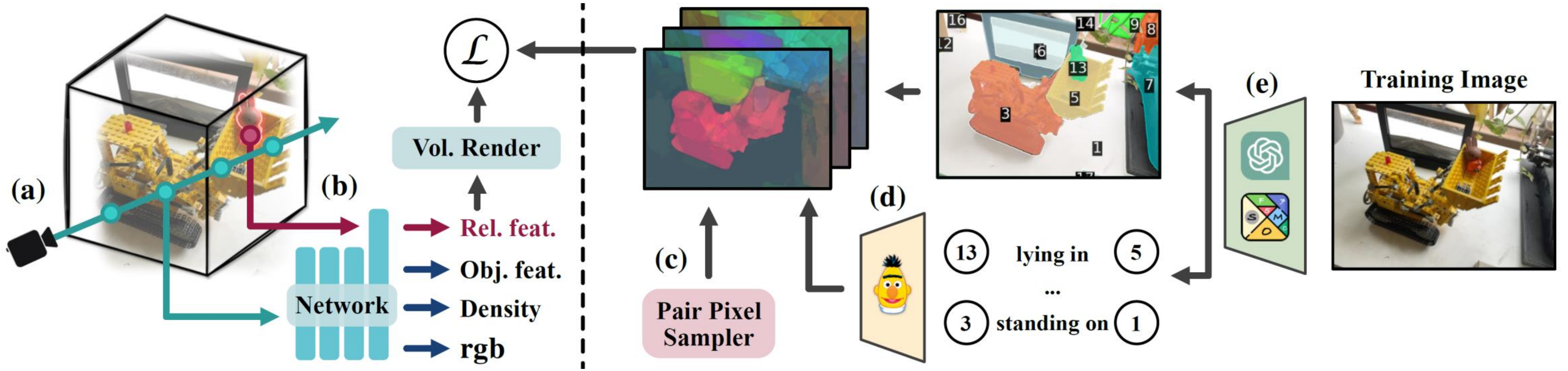
- ✅ 3D representation
- 🥳 Supervised by 2D images, perfect for 2D-3D distillation

[1] Mildenhall et al.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV'2020

Feature Fields



RelationField



Radiance field equation:

$$g_\theta(\mathbf{x}, \mathbf{d}, \mathbf{z}) \mapsto (\mathbf{c}, \sigma, \mathbf{o}, \mathbf{r})$$

Loss function:

$$\mathcal{L} = 1 - \frac{\mathbf{r}_r}{||\mathbf{r}_r||_2} \cdot \frac{\hat{\mathbf{r}}_r}{||\hat{\mathbf{r}}_r||_2}$$

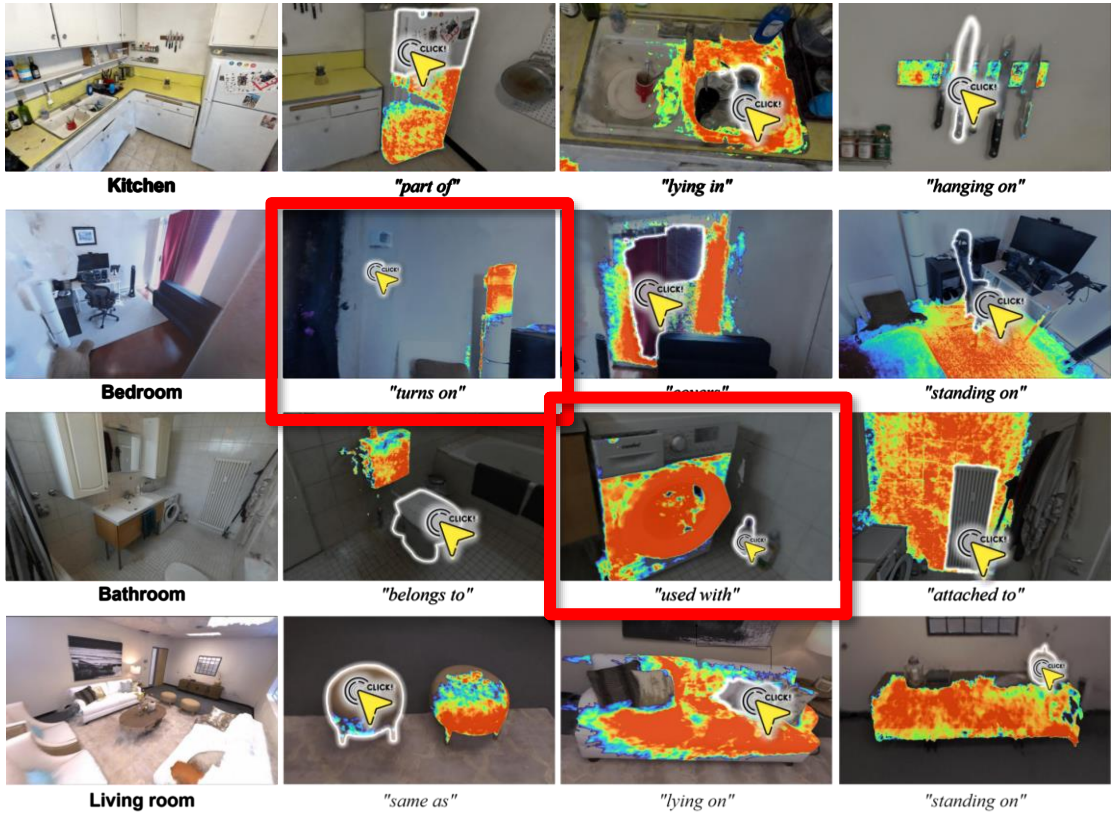
BERT embedding for feature supervision and concept generalization

GPT-4o + SoM for mask-aligned relationship captions

50 – 200 training images per scene

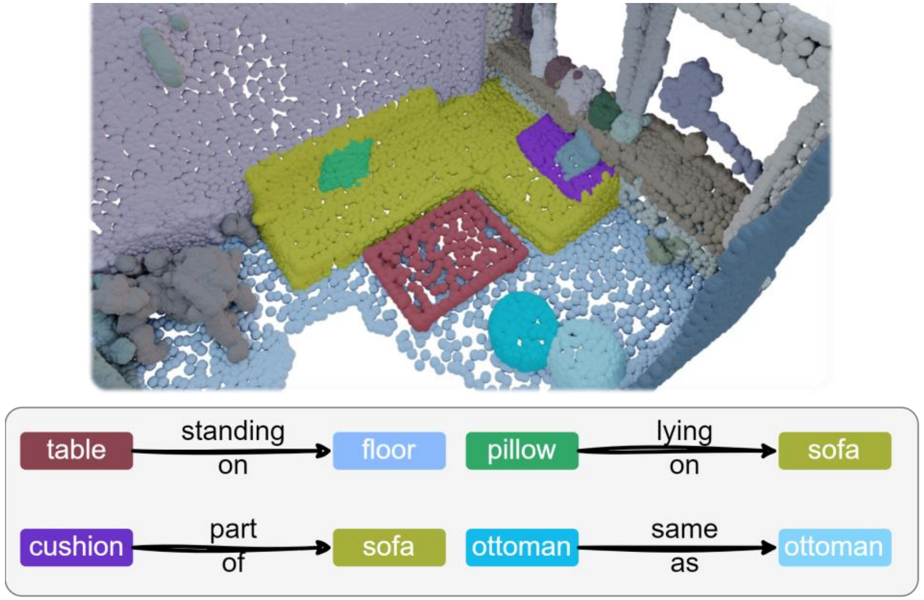
3D Relationship Reasoning

Interactive Relationship Extraction



Scene Graph Construction

Object Semantics + Relationship Semantics
= 3D Scene Graph



Take aways

- **RelationField enables 3D relationship reasoning** from 2D observations.

- **Inter-object relationships are defined as ray pairs**, capturing spatial and semantic interactions between objects.

- **RelationField encodes powerful foundation model knowledge**, making relationships **queryable in near real-time**.

- **RelationField models complex and causal relationships**, enabling diverse downstream applications.

DELTA

Decomposed Efficient Long-Term Robot Task Planning using Large Language Models

ICRA 2025

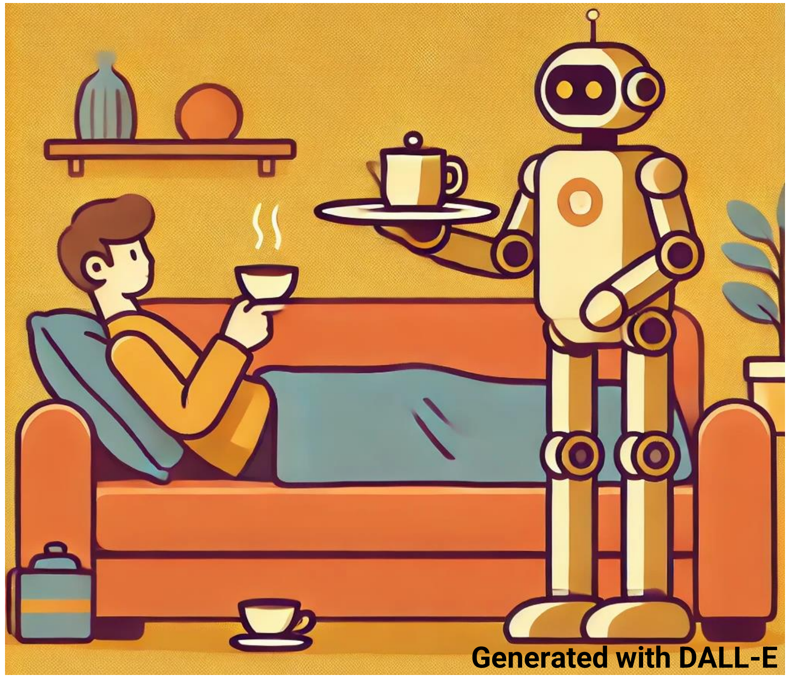
Yuchen Liu   Luigi Palmieri   Sebastian Koch
Ilche Georgievski   Marco Aiello

Universität Stuttgart   BOSCH   universität uulm

Current Challenges in Planning

Even simple task like bring me a coffee require complex planning

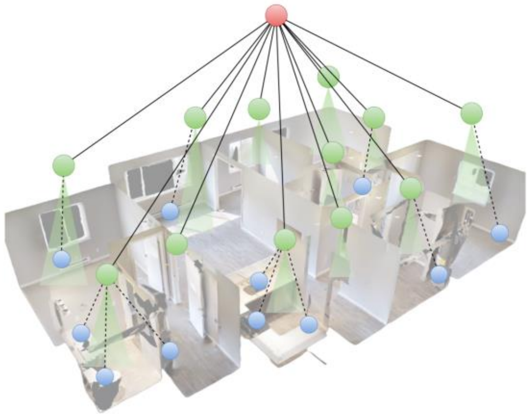
1. *Go to the kitchen*
2. *Get cup from cabinet*
3. *Turn on the coffee-machine*
4. *Make coffee*
5. *Go to living room*

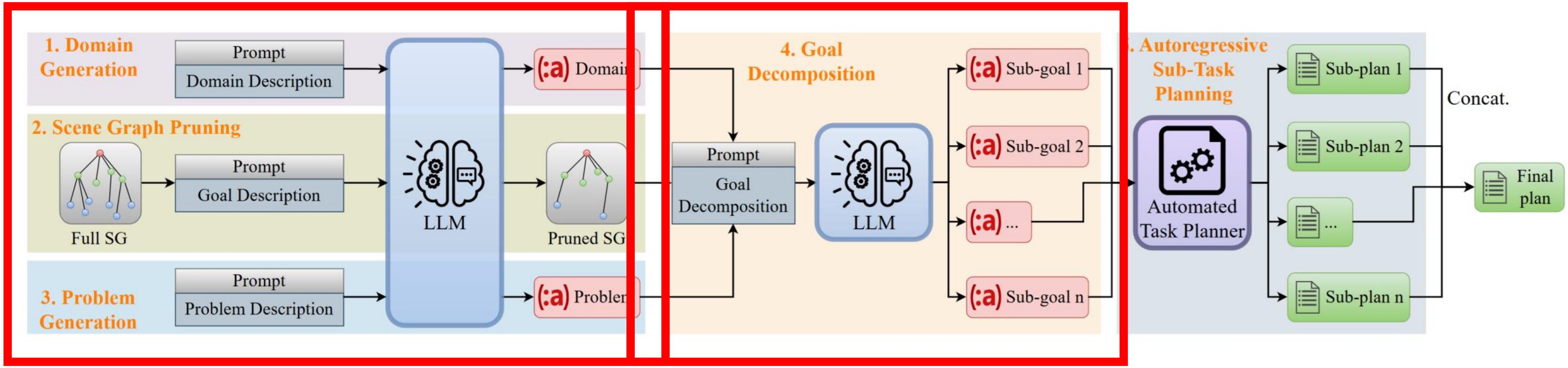


- **Symbolic Planners** often need precise information about objects, affordances, actions, etc.
- **Symbolic Planners** need a lot of planning time for complex observation & action spaces

🤔 **LLM Planners** enable efficient, intuitive planning with dynamic sub-goals and chain-of-thought reasoning but often **lack real-world grounding**.

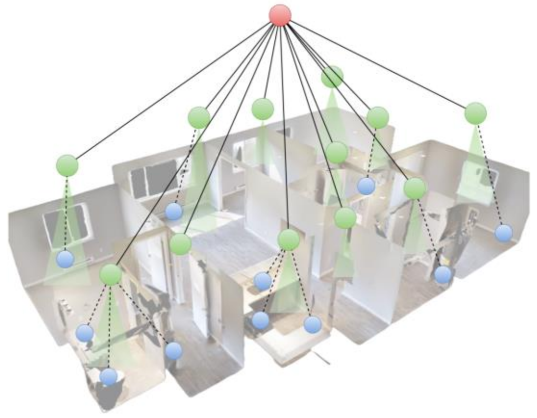
DELTA Approach



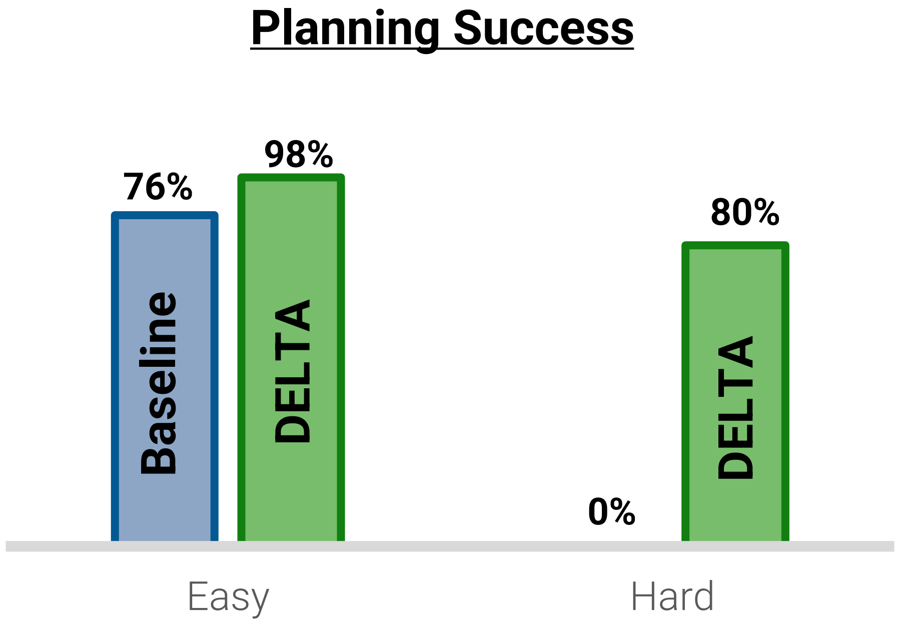


- Environment grounding using PDDL
- Structured observation from 3D Graph
- Subgoal generation with LLMs for efficient planning
- Use of strong Task Planners for successful plan execution

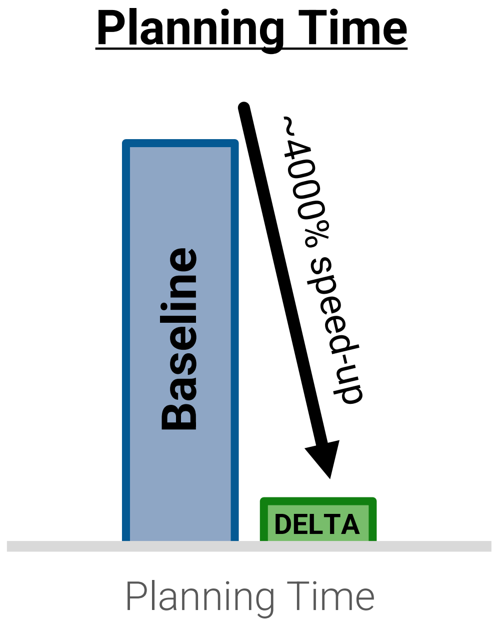
DELTA Take-aways



- **3D Scene Graphs** lead to improved planning success
- **LLM-based Sub-goals** and **SG pruning** lead to faster planning



DELTA w/ 3D scene graph representation
Baseline w/o 3D scene graph representation



Conclusion & Summery

Take-home message

- **Relationships** are very important for holistic 3D Scene Understanding.
- **3D Scene Graphs** naturally **connect 3D environments with language.**
- Open-vocabulary Scene Graphs enable flexible, **interactable representations for diverse use cases.**

What is the best way to interact with a 3D scene/interact with LLMs?

Language-driven Scene Understanding with 3D Scene Graphs

Sebastian Koch
Ulm University and Bosch Center for AI

Huawei Munich Research Center

Feb 06, 2024