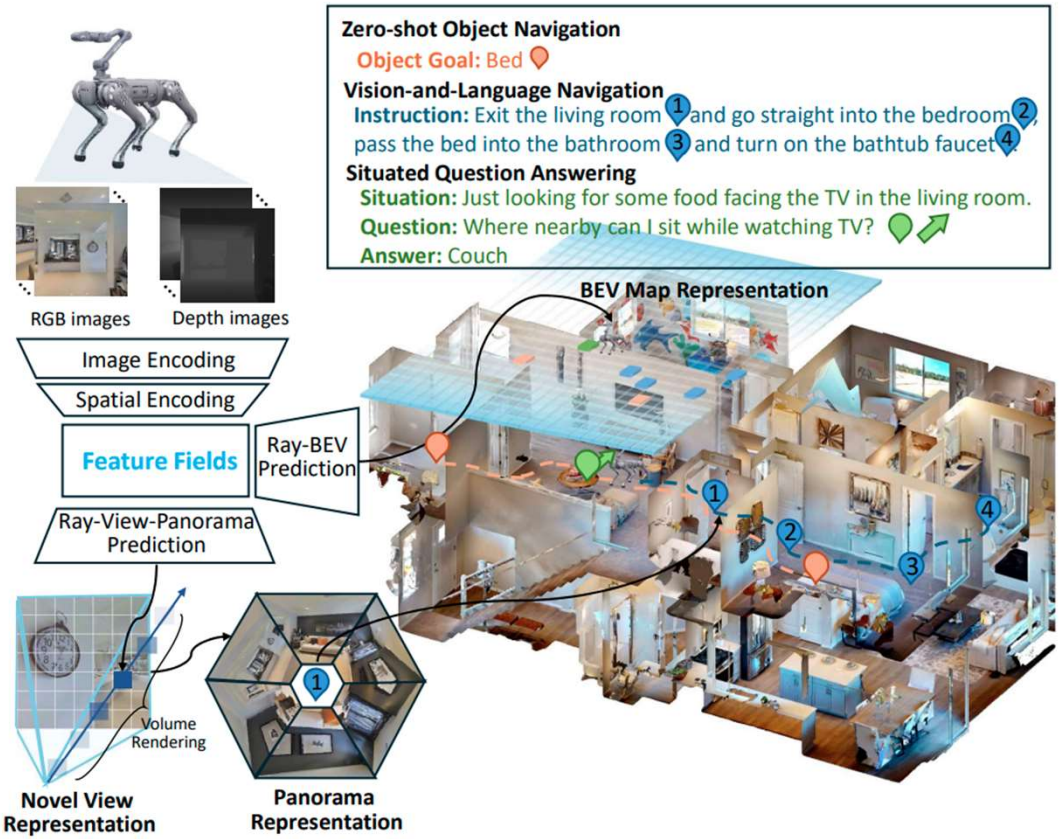
g3D-LF: Generalizable 3D-Language Feature Fields for Embodied Tasks



What 3D representation model is suitable for Embodied AI?

1) **Generalizable** to unseen scenes

2) Construct and **update** representations **in real time**

3) **Open-vocabulary** semantic space

Generalizable Feature/Semantic Fields maybe a possible way, but previous works…

1) Always supervised by 2D models (CLIP, DINOv2) lacks **3D spatial understanding**

2) Have 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**

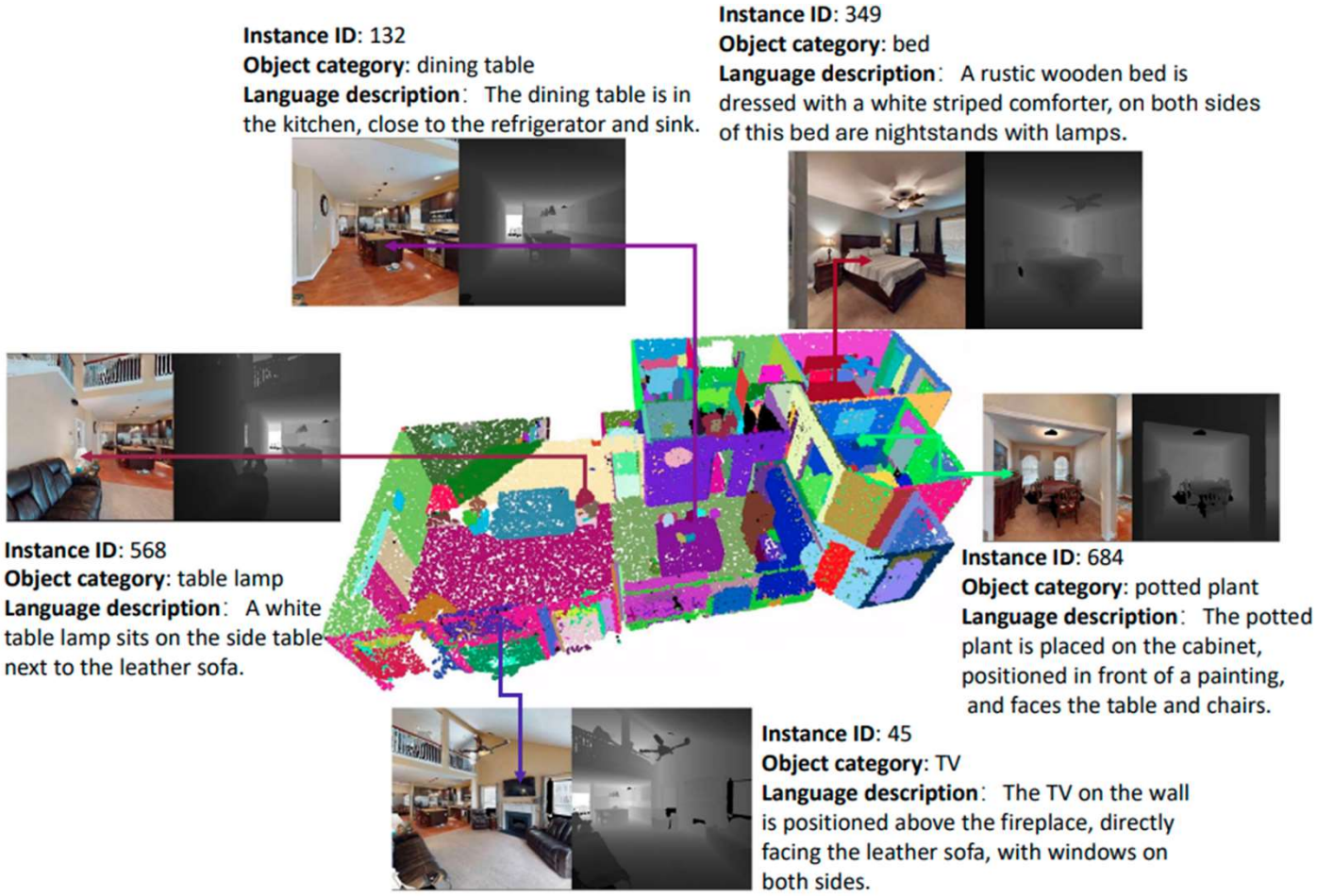
In 3D-Language Feature Fields, we…

1) Organize **a large-scale 3D-Language dataset** for Feature Fields model pre-training

2) Propose **a multi-level contrastive learning framework** to align the multi-scale representations with multi-granularity language.

3) **Adapt** the 3d-language feature fields model to **various embodied tasks**.

Pre-training data



Input should be 2D, simple, easily obtainable

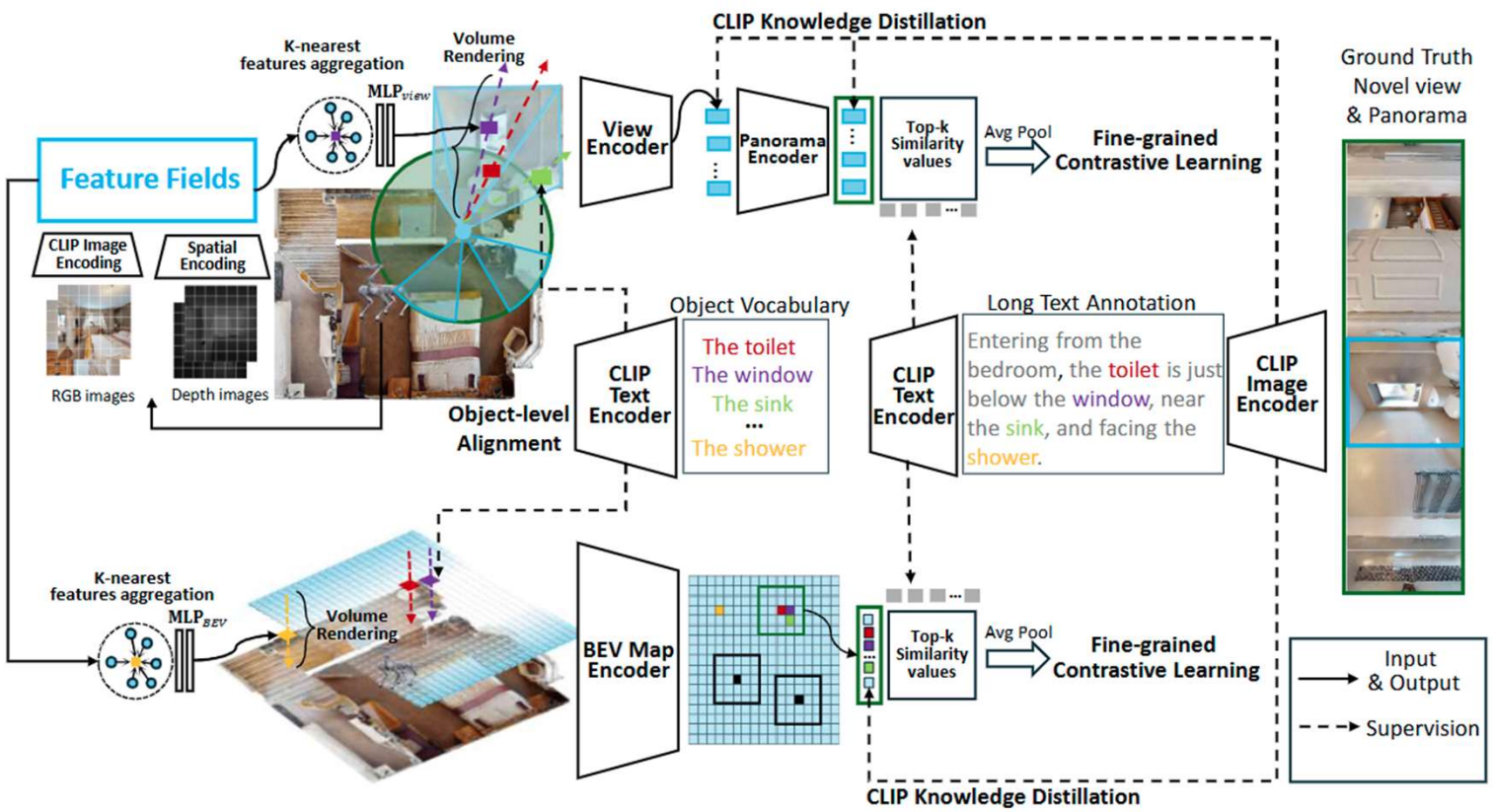
e.g., posed RGB and depth images

Supervision should be 3D, fine-grained, multi-level

e.g., instance point clouds, multi-level language annotations

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

Framework



Multi-scale Representation

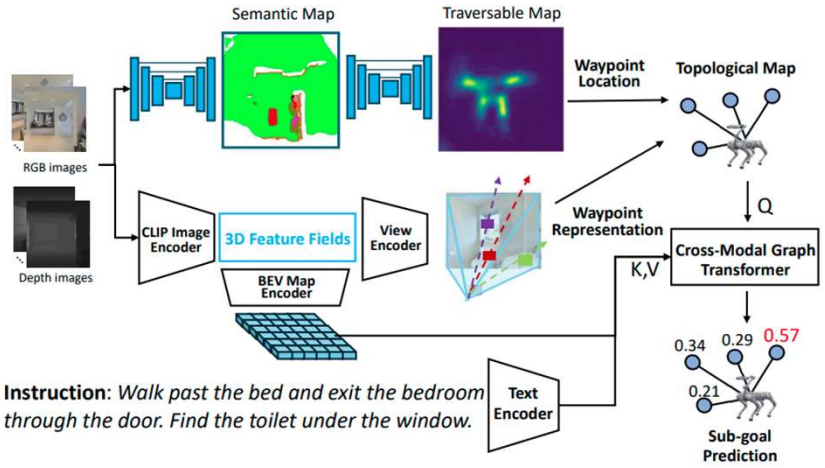
- Render the **ray**-level representation
- Combine the rays into the **view** representation
- Encode the **panorama** with multi-views
- Encode the top-down rays for large-scale **BEV map**

Multi-level Supervision

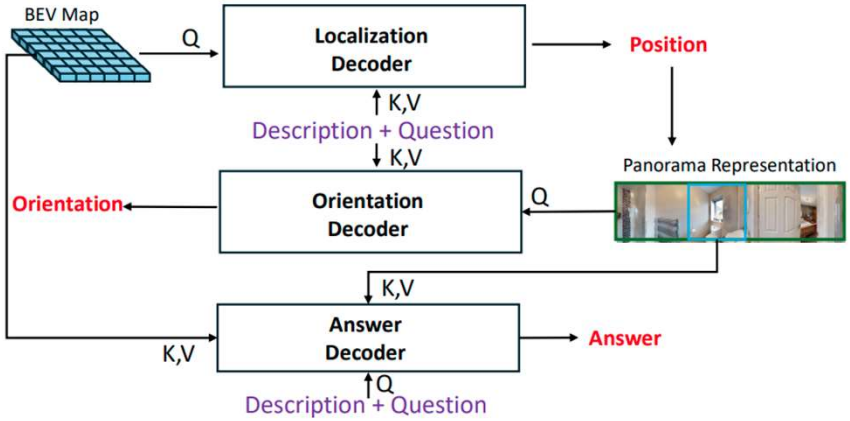
- For rendered **rays**, contrastive learning across **1,883 indoor object categories**
- For **novel view-panorama** and **BEV map**, **distill knowledge from 2D model**.
- For **3D spatial reasoning** and **long text understanding** of panorama and BEV map, use **fine-grained contrastive learning**

Embodied Tasks

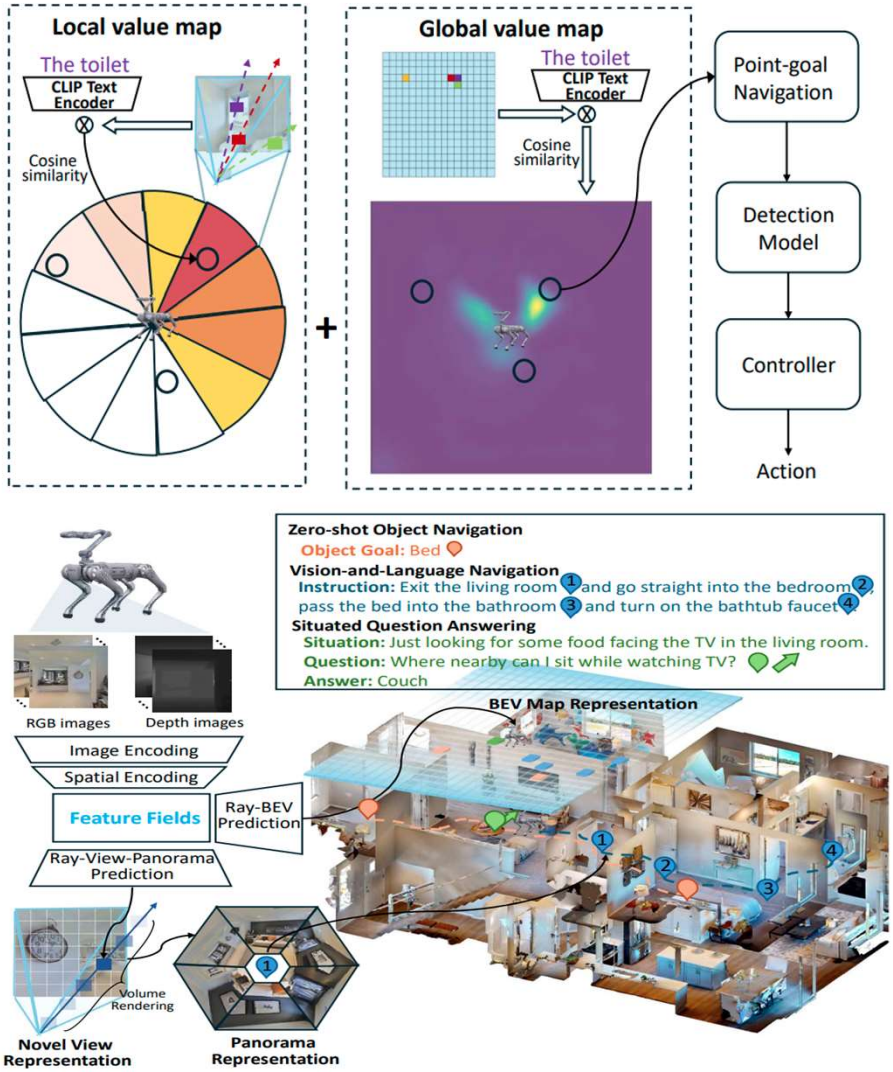
- **Monocular Vision-Language Navigation**



- **Situated Question Answering**



- **Zero-shot Object Navigation**



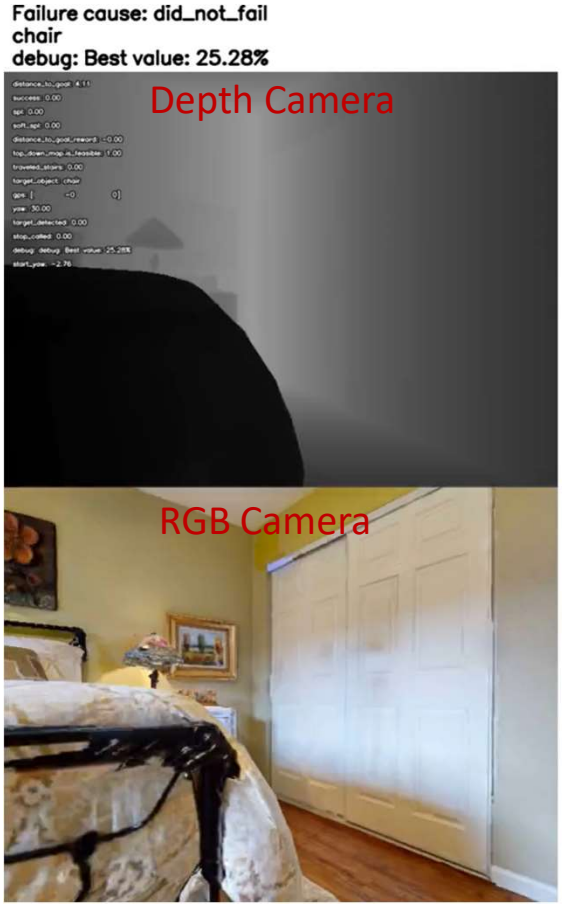
Demo for Object Navigation---find the couch

Navigation Trajectory

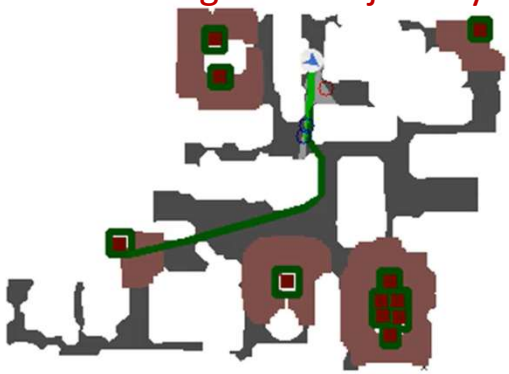
Value Map from Feature Field

Obstacle Map

Demo for Object Navigation---find the chair



Navigation Trajectory

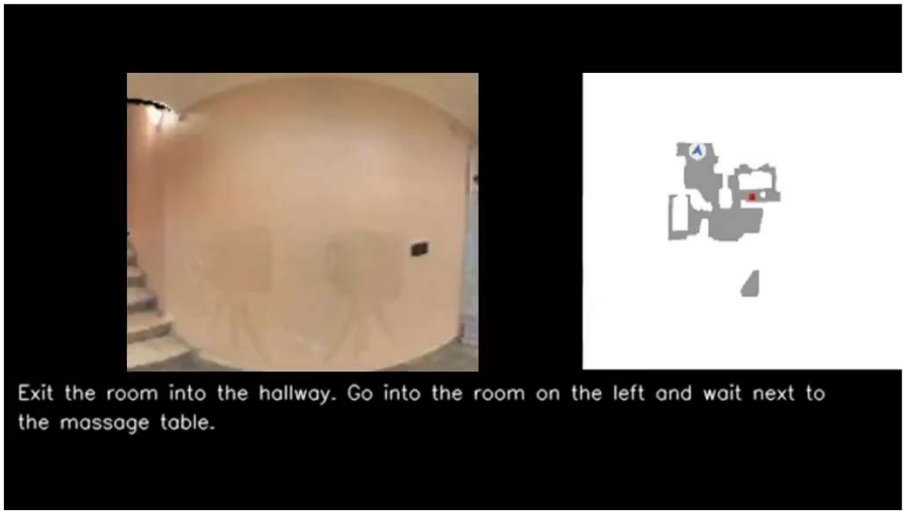


Value Map from Feature Field

Obstacle Map

Demo for Vision-and-Language Navigation

Monocular VLN



Panorama VLN

Limitations and future works

- Real-world robot
- Dynamic environments, where objects or people are moving in real time
- More fine-grained and dynamic tasks, *e.g.,* mobile manipulation
- 3D representation model with LLM
- The scale and quality of 3D-language data
- More robust input, e.g., no camera pose