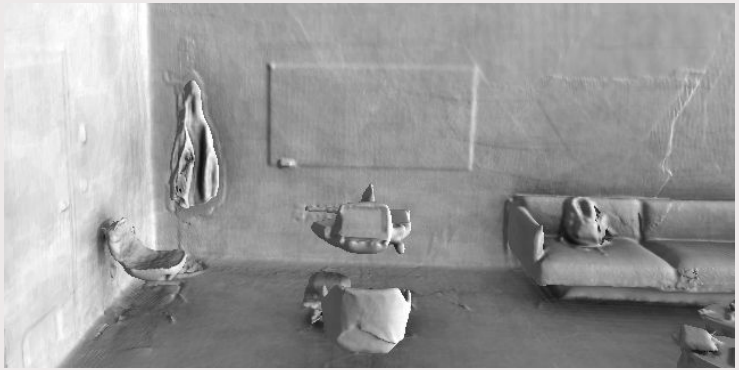
Semantic Querying of Reconstructed 3D Environments Using Pre-trained 2D Foundation Models

Master Degree in Computer Science - Artificial Intelligence Curriculum
Candidate: Paolo Fasano
Supervisors: Daniela Giorgi, Fabio Carrara, Gianpaolo Palma
Anno Accademico: 2023-2024

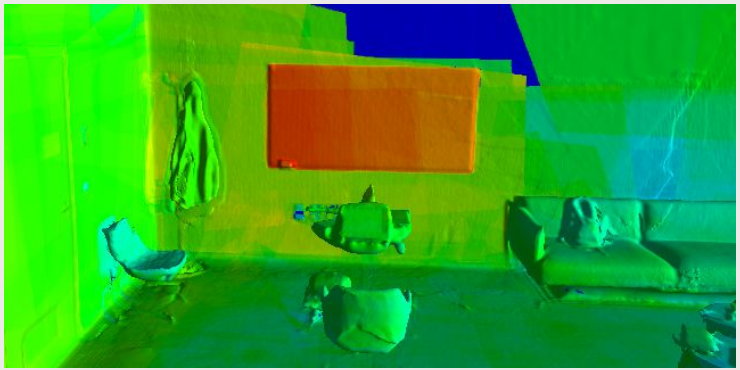
Thesis objective

Rationale: Understanding and interacting with complex 3D environments is increasingly important in robotics, VR, and autonomous systems.

Contribution: A system for reconstructing 3D environments enriched with semantic features, enabling zero-shot querying using pre-trained 2D models like SAM, CLIP and DinoV2.

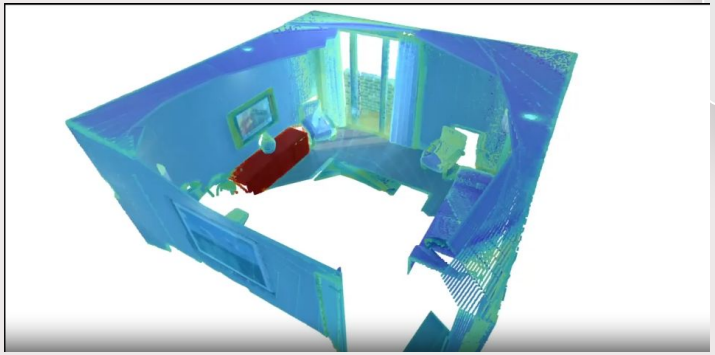


Query: Where is the whiteboard?



Related work - ConceptFusion [1]

Concept-Fusion Objective: semantically enhance a mesh for multi-modal querying



Concept-Fusion aim

- Multimodal Data Fusion
- Open-Set Recognition
- 3D Semantic Mapping

Input:

- Images
- Segmentation Masks

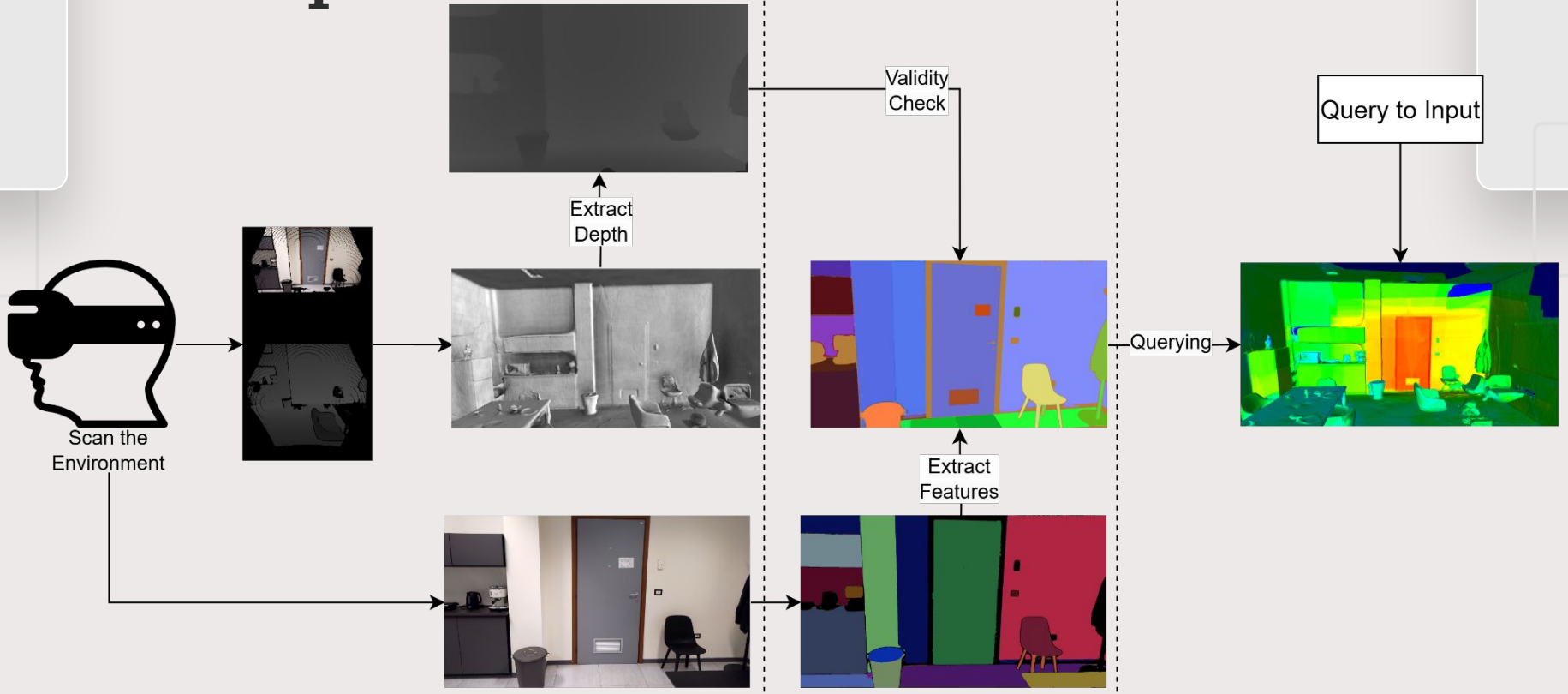
Issues:

- Works fine on artificial datasets, has issues working with real world dataset
- The real world dataset can only be constructed in very low resolution
- Activation on real world data is very sparse

We built our system to address such issues and work on real world data

[1] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. **Conceptfusion: Open-set multimodal 3d mapping.** Robotics: Science and Systems (RSS), 2023

Our Pipeline

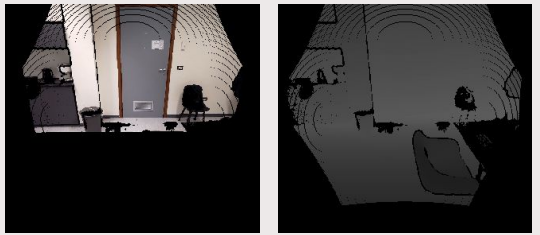


Phase 1: Environment Reconstruction

PHASE 1.1
Scan the
Environment

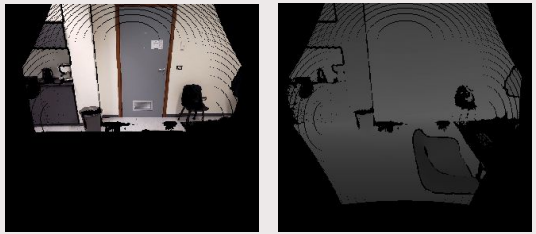
PHASE 1.2
Environment
Reconstruction

PHASE 1.3
Calculate
Depth Images



N-th RGB-D value

Phase 1: Environment Reconstruction



N-th RGB-D value

Point-Cloud 1

Point-Cloud 2

Point-Cloud N

Phase 1: Environment Reconstruction

PHASE 1.1
Scan the Environment

PHASE 1.2
Environment Reconstruction

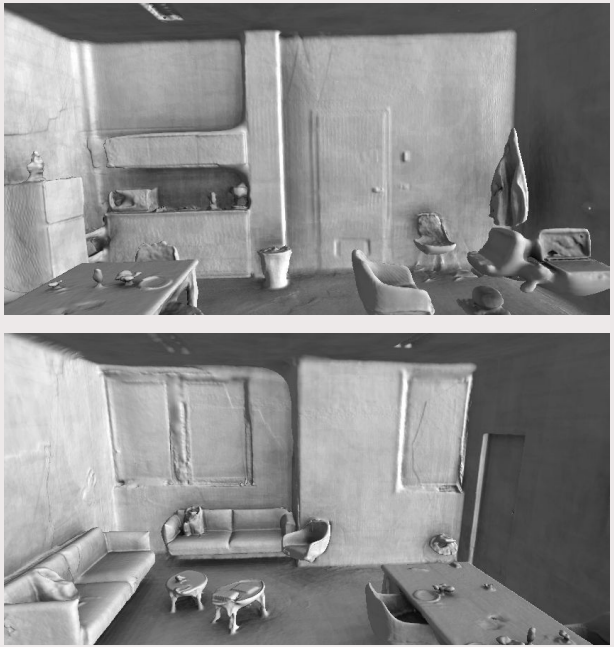
PHASE 1.3
Calculate Depth Images

Point-Cloud 1

Point-Cloud 2

Point-Cloud N

Poisson Surface Reconstruction

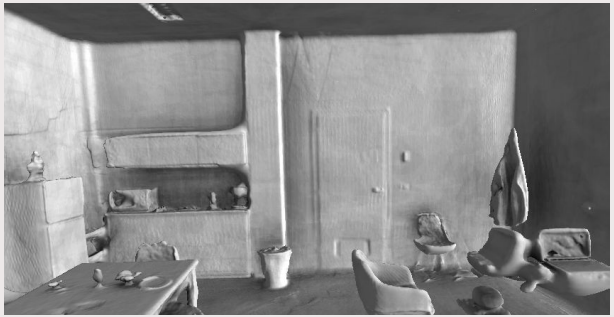


Phase 1: Environment Reconstruction

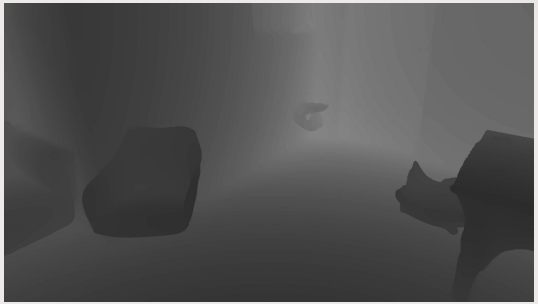
PHASE 1.1
Scan the
Environment

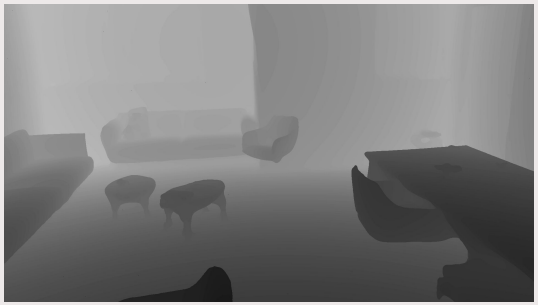
PHASE 1.2
Environment
Reconstruction

PHASE 1.3
Calculate
Depth Images



Calculate Depth
via Raytracing





Phase 2: Feature Extraction

PHASE 2.1
Segmentation
via SAM

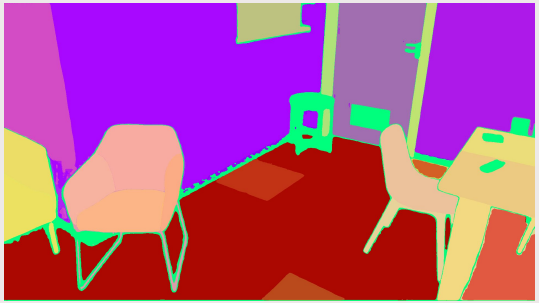
PHASE 2.2
Semantic Feature
Extraction

PHASE 2.3
Feature Projection
and Integration

Segment Anything (SAM)

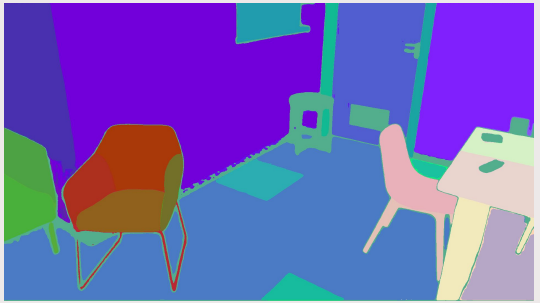
- SAM automates segmentation of images into meaningful regions
- Generates masks for image regions using grid sampling of query points.

Contrastive Language–Image Pre-training (CLIP)



- Pretrained on extensive datasets (e.g., LAION-2B) for generalization.
- Outputs embeddings where image and text semantics align.

Distillation with No Labels (DINO)



- Extracts semantic representations from images using self-supervised learning
- Encodes global and region-level image features.

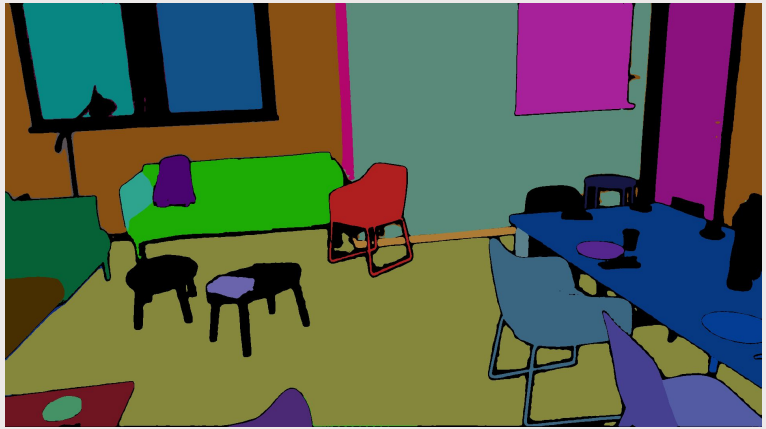
Phase 2: Feature Extraction

PHASE 2.1
Segmentation
via SAM

PHASE 2.2
Semantic Feature
Extraction

PHASE 2.3
Feature Projection
and Integration

Segmentation
via SAM



Phase 2: Feature Extraction

PHASE 2.1
Segmentation
via SAM

PHASE 2.2
Semantic Feature
Extraction

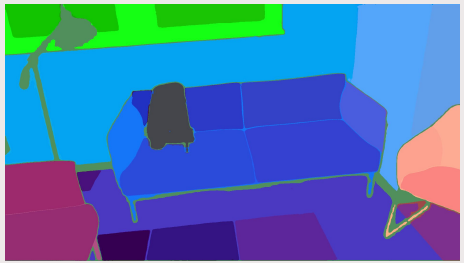
PHASE 2.3
Feature Projection
and Integration

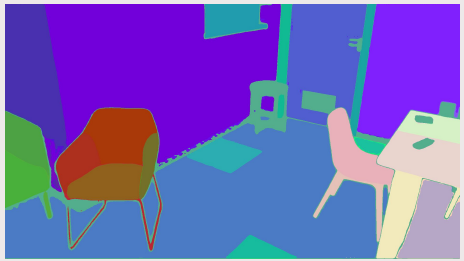
Feature Extraction
via CLIP

Feature Extraction
via DinoV2

Phase 2: Feature Extraction

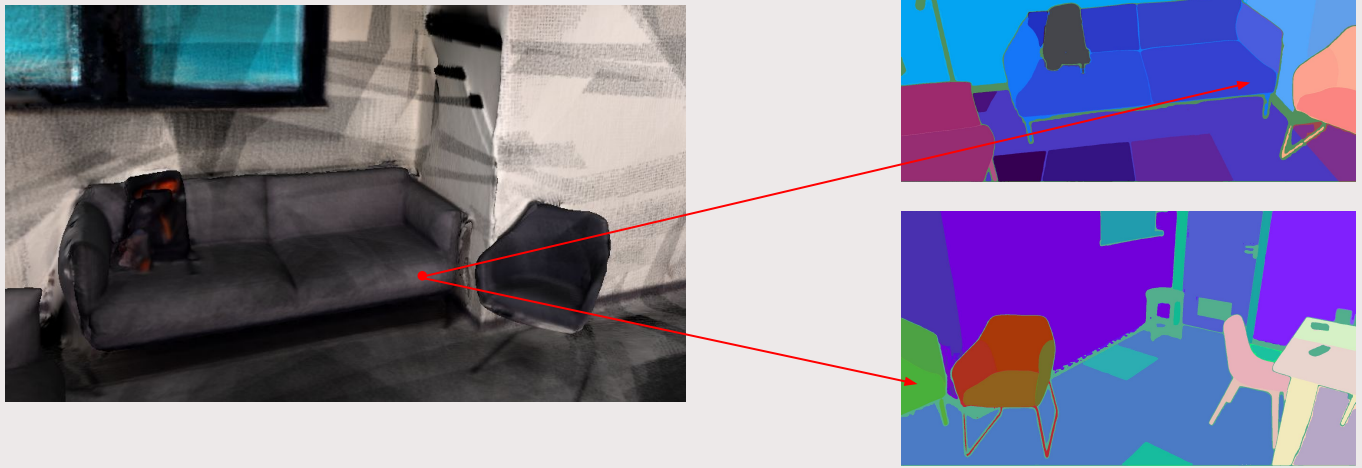
- Same vertex falls on pixel of different images
- Sum and L2-norm the features from different images that falls on same vertex



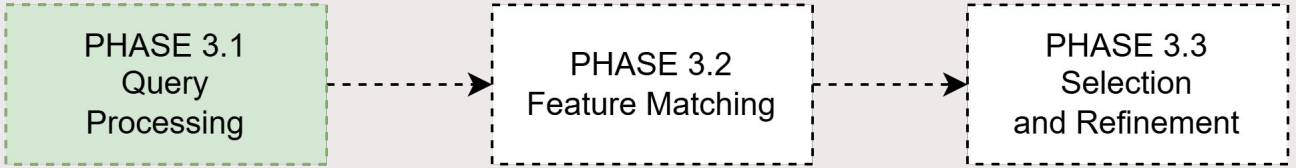


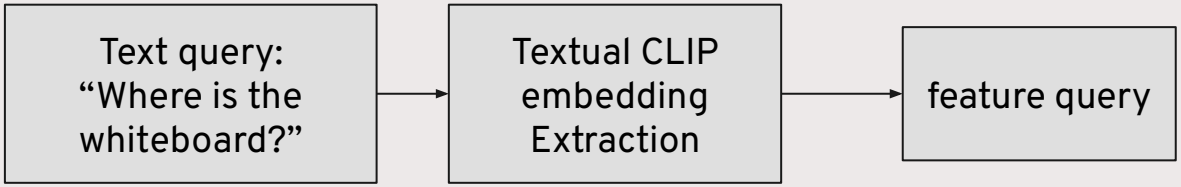
Phase 2: Feature Extraction

- Same vertex falls on pixel of different images
- Sum and L2-norm the features from different images that falls on same vertex

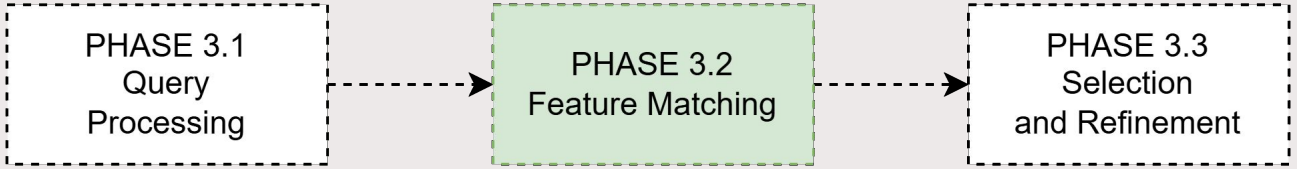


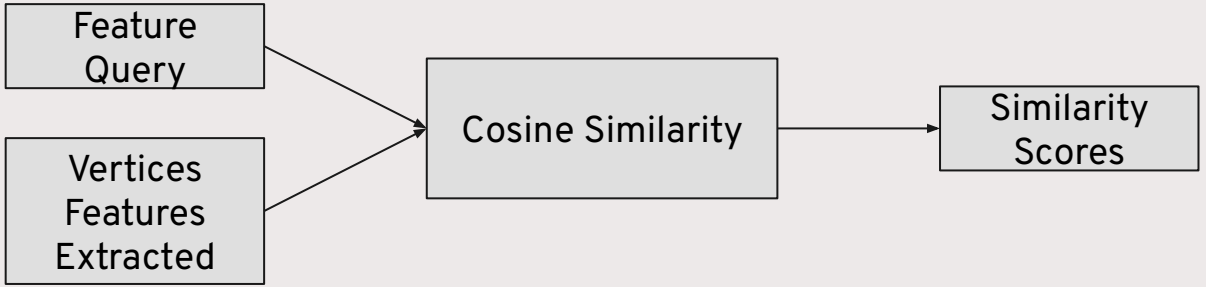
Phase 3: Querying



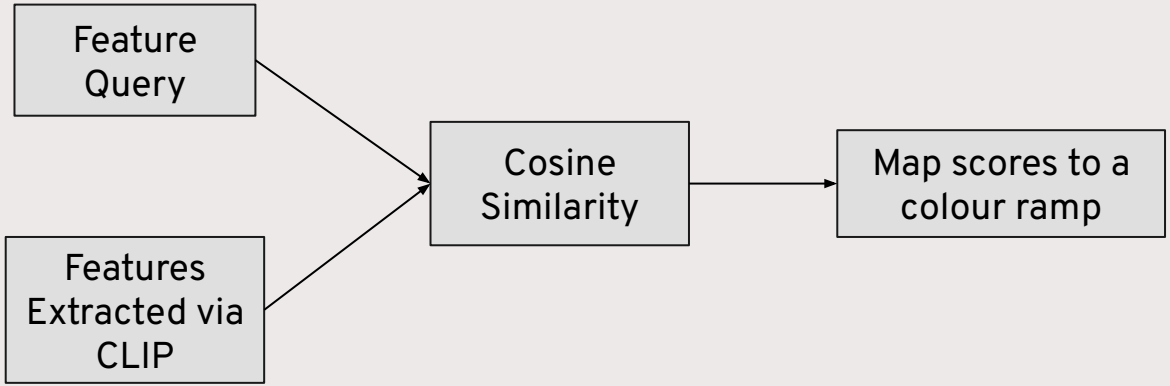


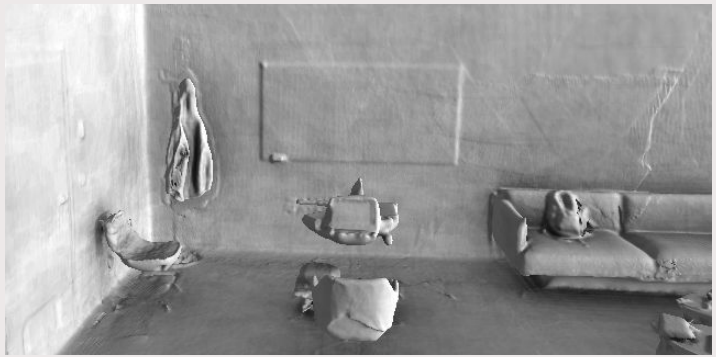
Phase 3: Querying



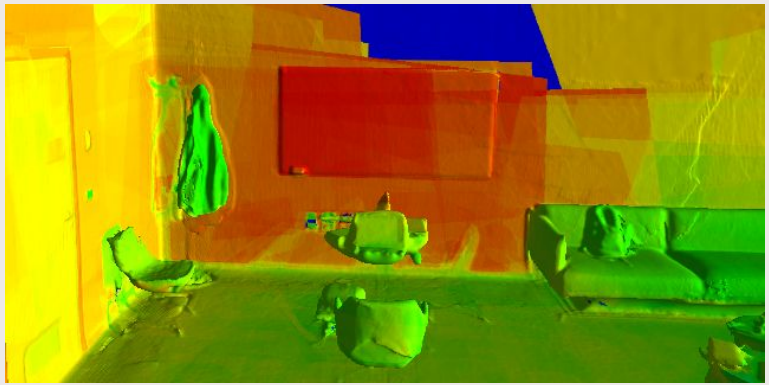


Phase 3.2a: Feature Matching via CLIP

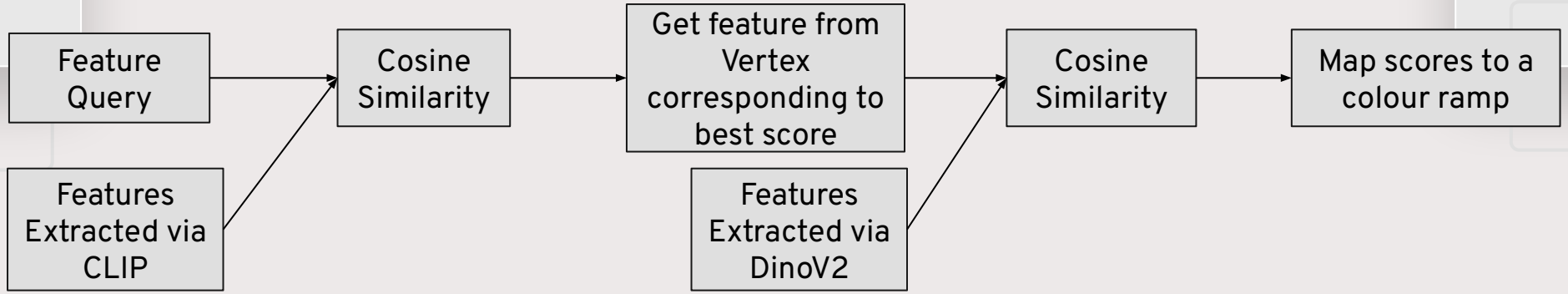


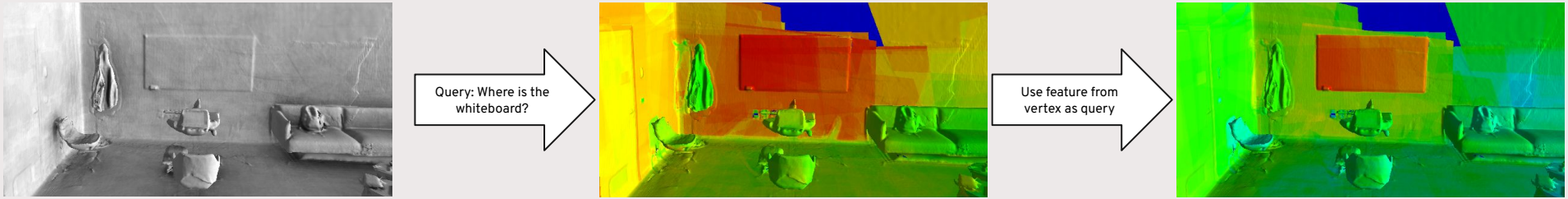


Query: Where is the whiteboard?

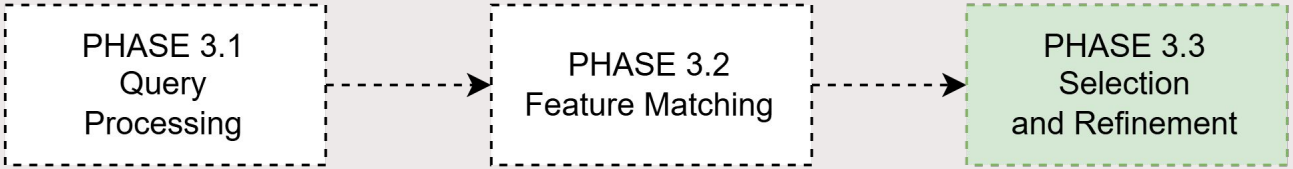


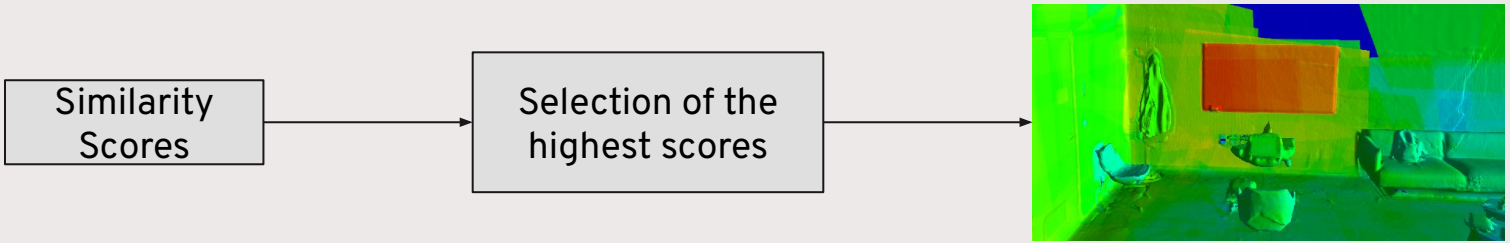
Phase 3.2b: Feature Matching via CLIP-DinoV2 Fusion





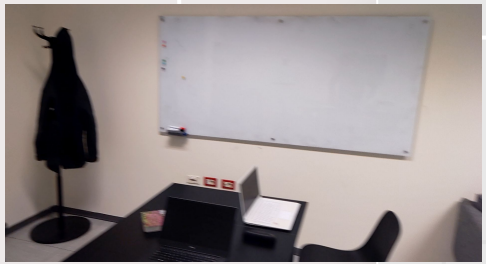
Phase 3: Querying



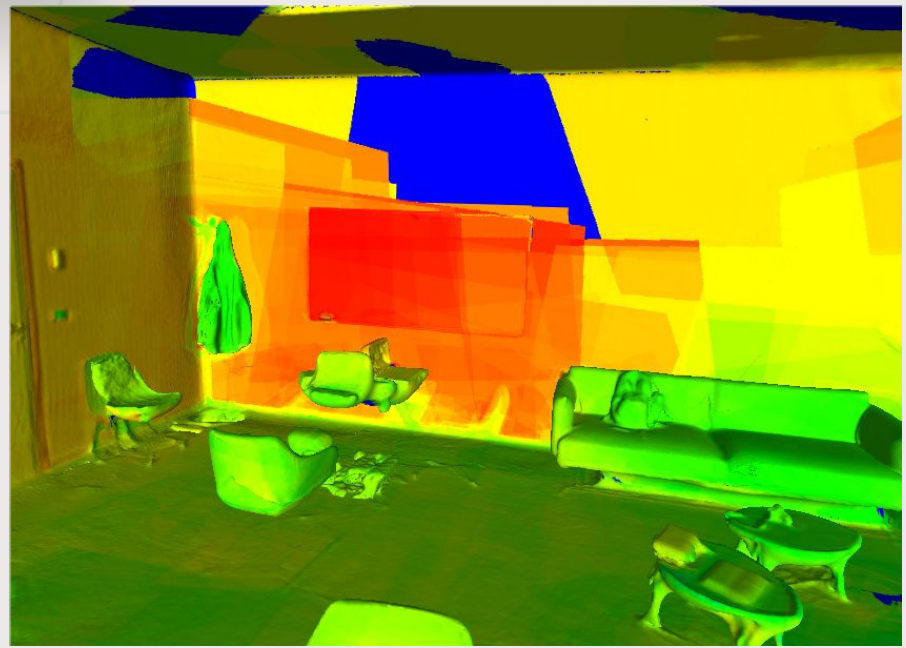


Results

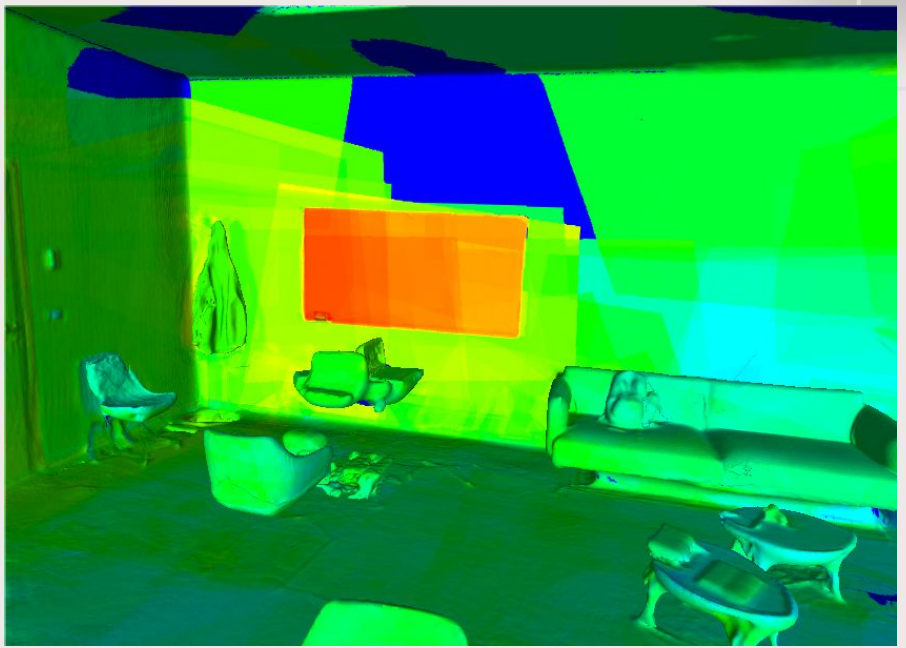
Query: Where is the whiteboard?



CLIP



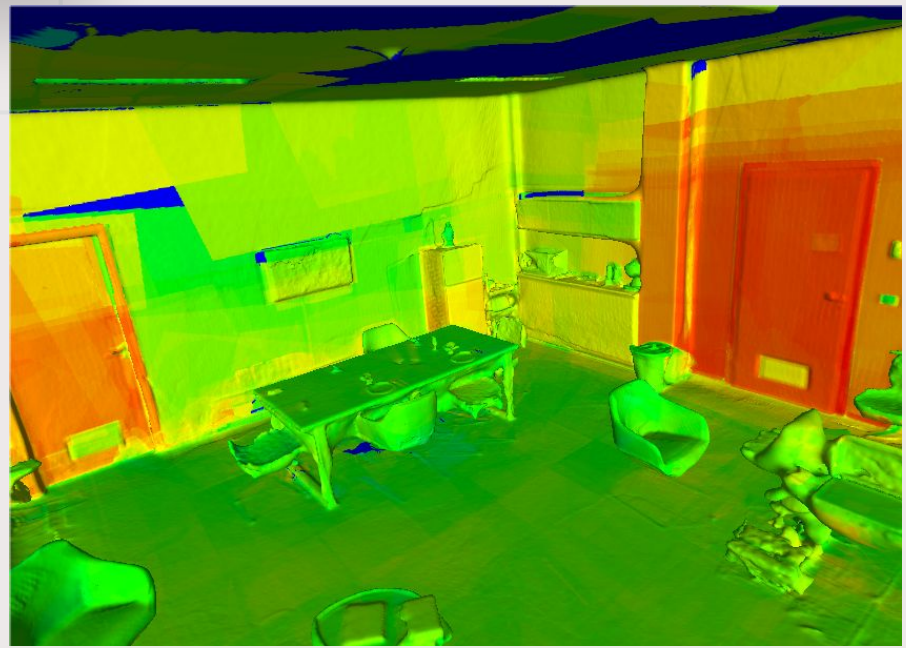
CLIP+DinoV2 Fusion



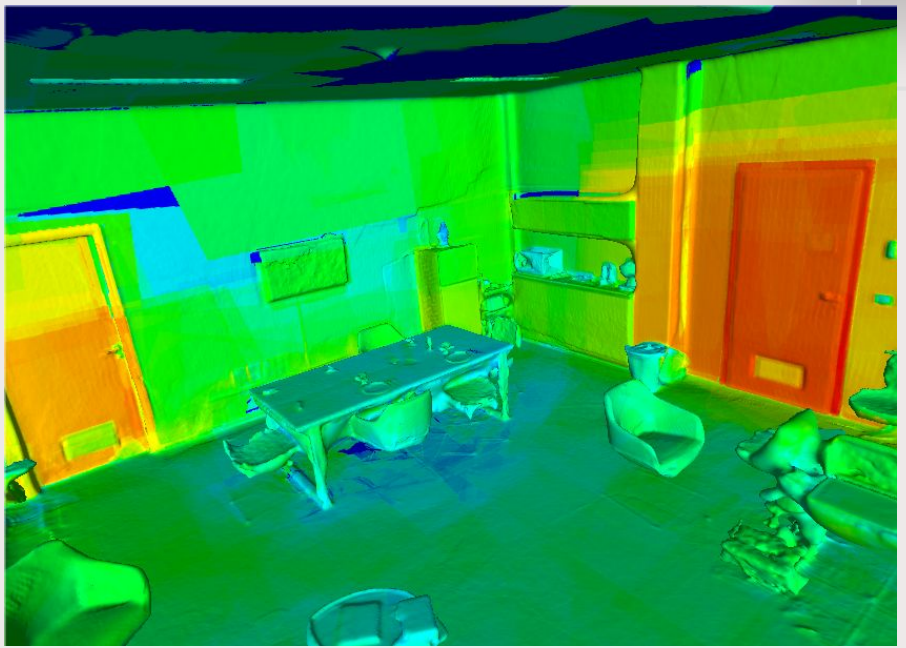
Results

Query: Where are the doors?

CLIP

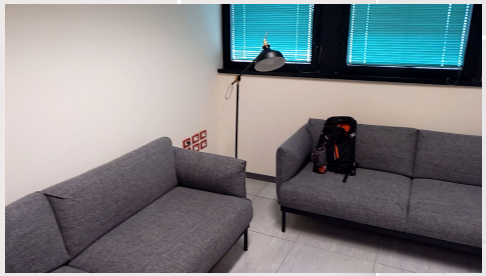


CLIP+DinoV2 Fusion

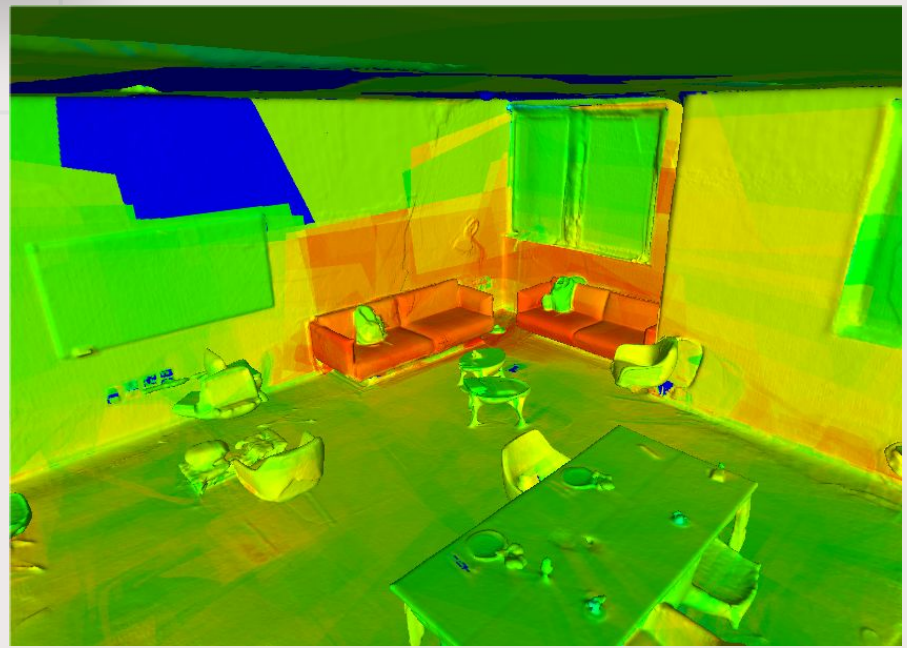


Results

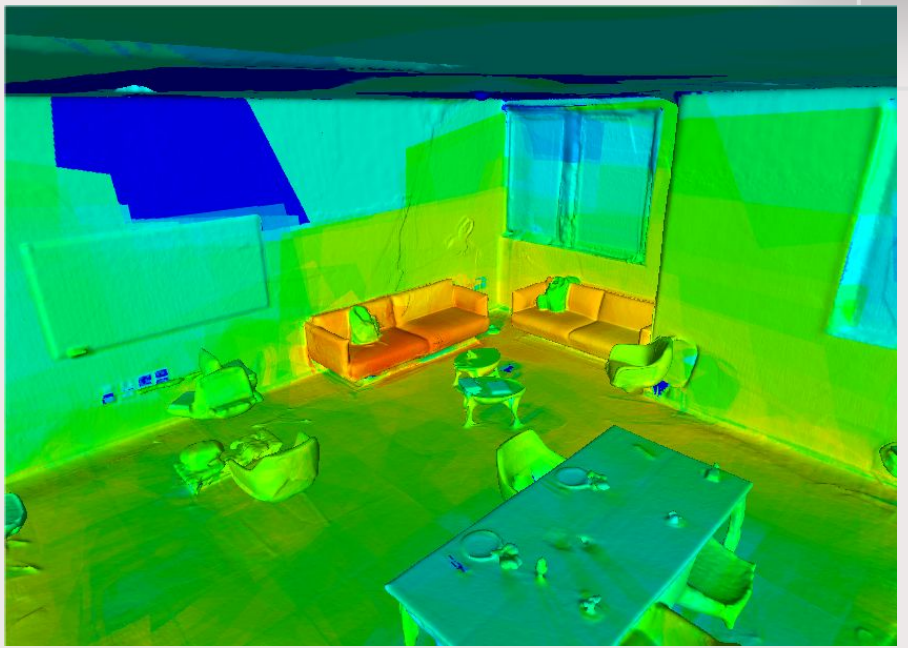
Query: Where are the couches?



CLIP



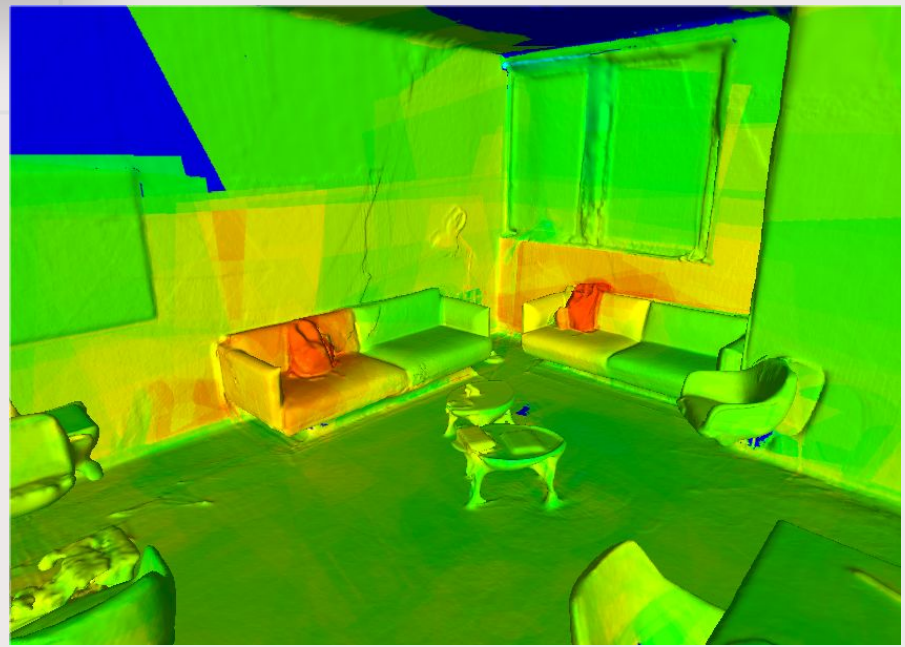
CLIP+DinoV2 Fusion



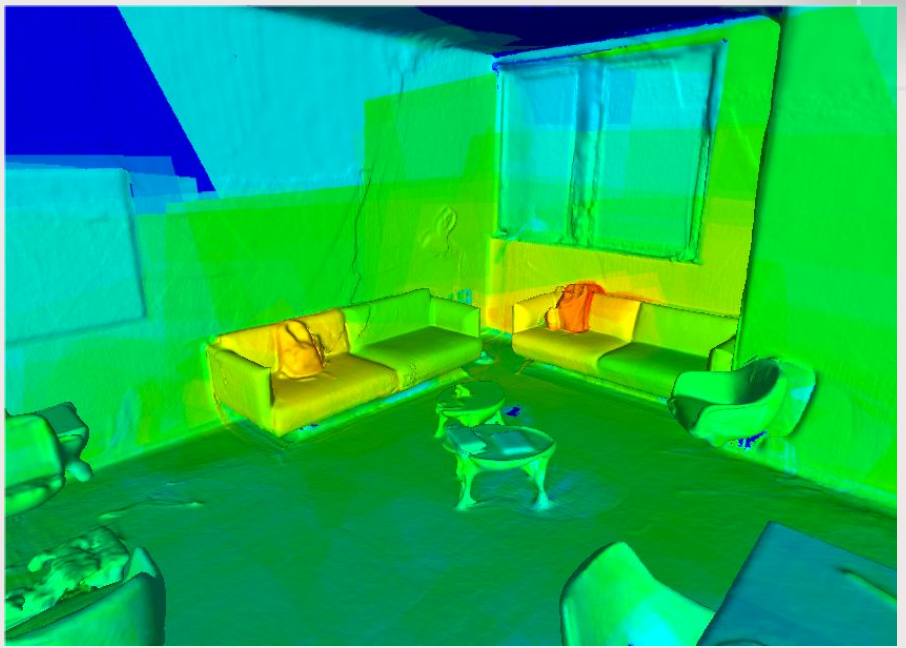
Results

Query: Can I find my backpack?

CLIP

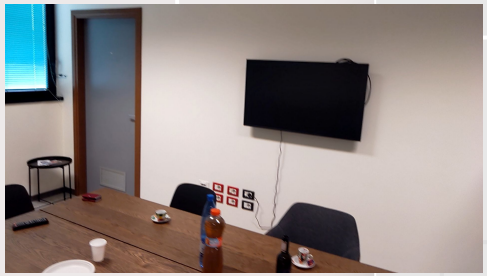


CLIP+DinoV2 Fusion

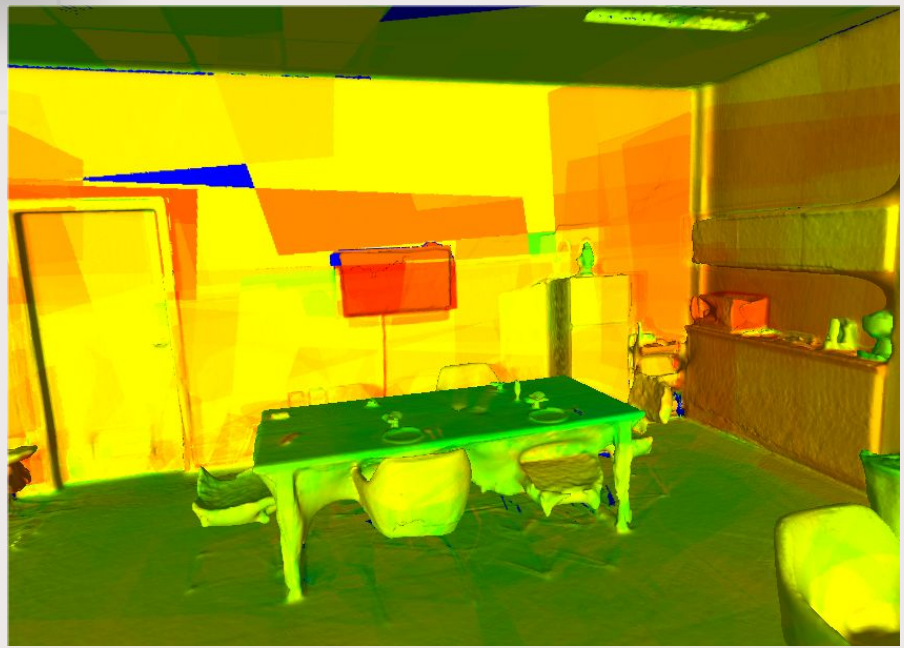


Results

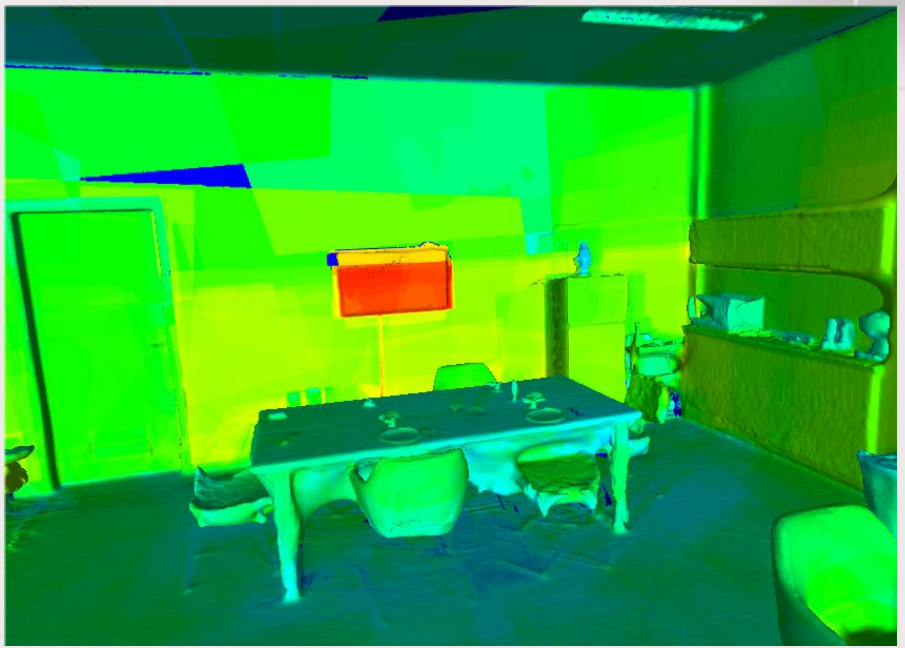
Query: Is there a TV?



CLIP

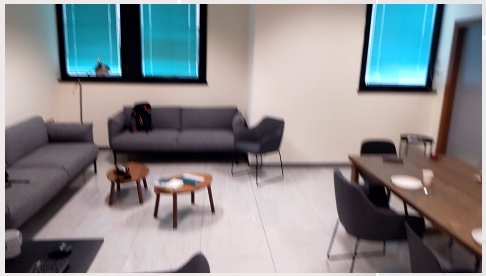


CLIP+DinoV2 Fusion

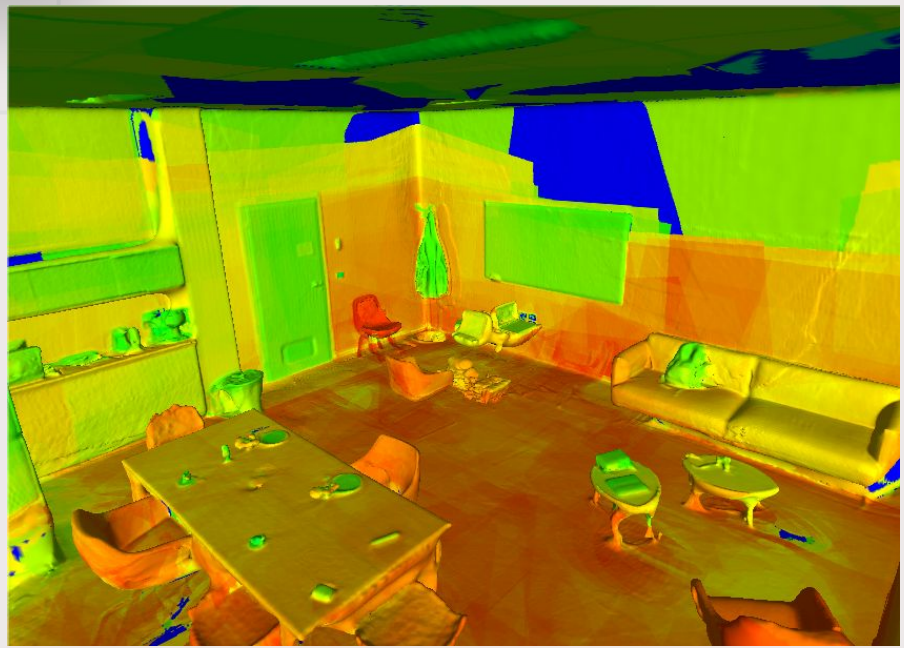


Results

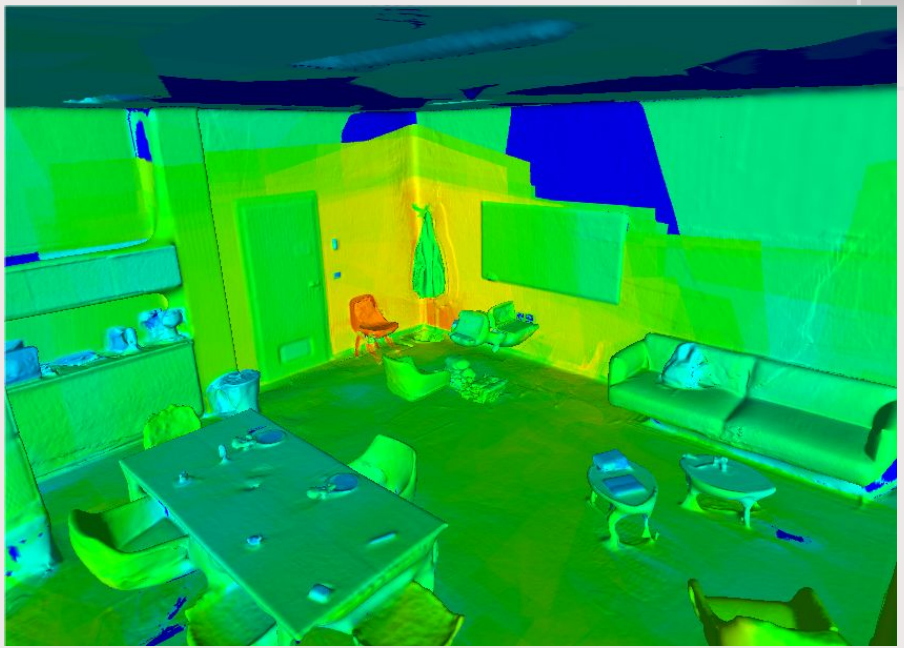
Query: Where can I sit?



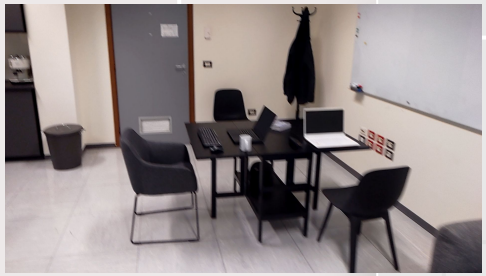
CLIP



CLIP+DinoV2 Fusion

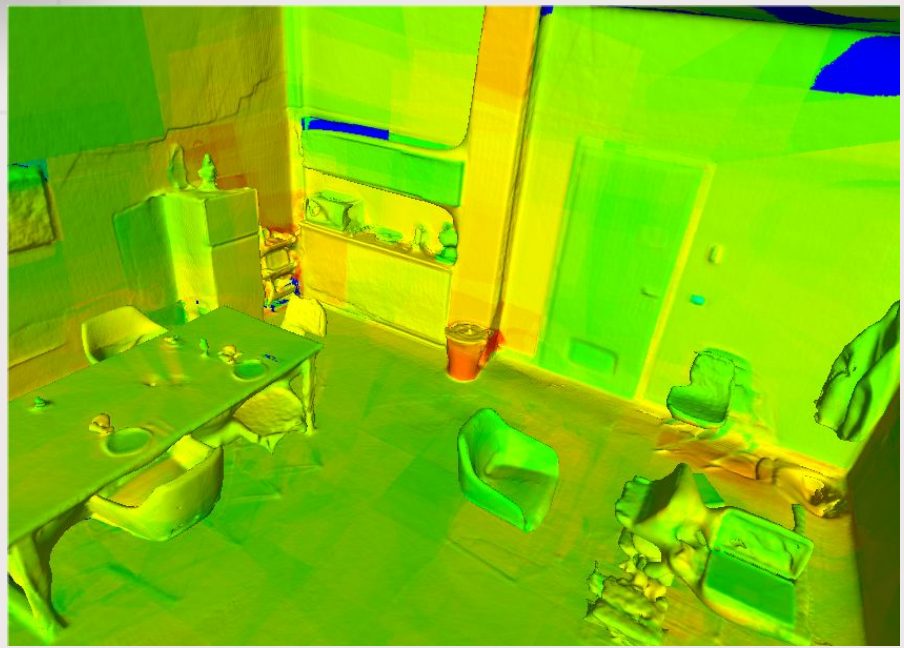


Results

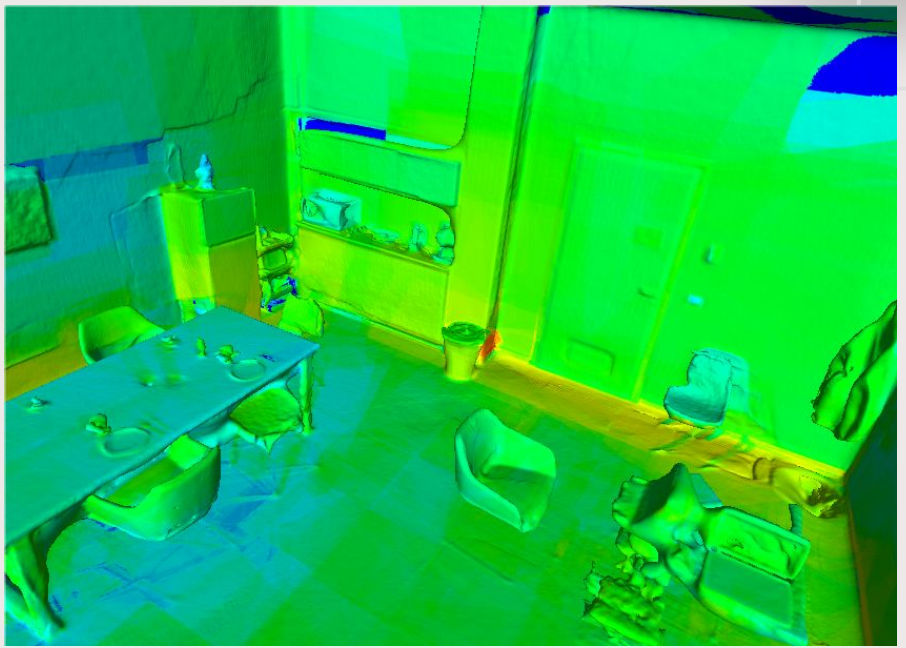


Query: Where can I put the trash?

CLIP



CLIP+DinoV2 Fusion



Conclusions

Contributions:

- The proposed approach marks a significant step towards developing a system that can effectively interact with real-world environments.
- Consistency in the querying results
- Real Time querying is reached both with only CLIP and CLIP-DinoV2 Fusion approach

Limitations:

- Handling of Vague or Ambiguous Queries
- Precision in Object Segmentation

Future work

Over the data

- Improve Multiple Object Detection in Fusion Method
- Extending the dataset with ground truth labels

Over the Ai models

- Integration of Spatial Reasoning Mechanisms
- Testing on Functional Area Recognition

THANKS FOR YOUR ATTENTION!

ANY QUESTIONS?