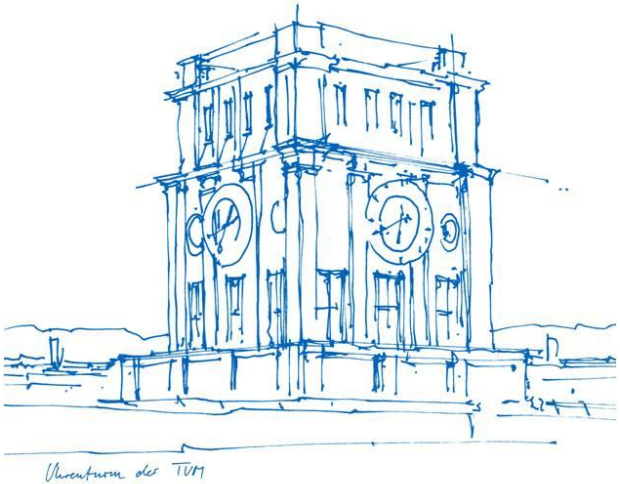
Final Presentation: Feature Gaussian SLAM

Vitus Becker, Nick Deng, Ahmed Kadri
Technische Universität München
TUM School of Computation, Information & Technology
Chair of Computer Aided Medical Procedures
Garching, 07. February 2025

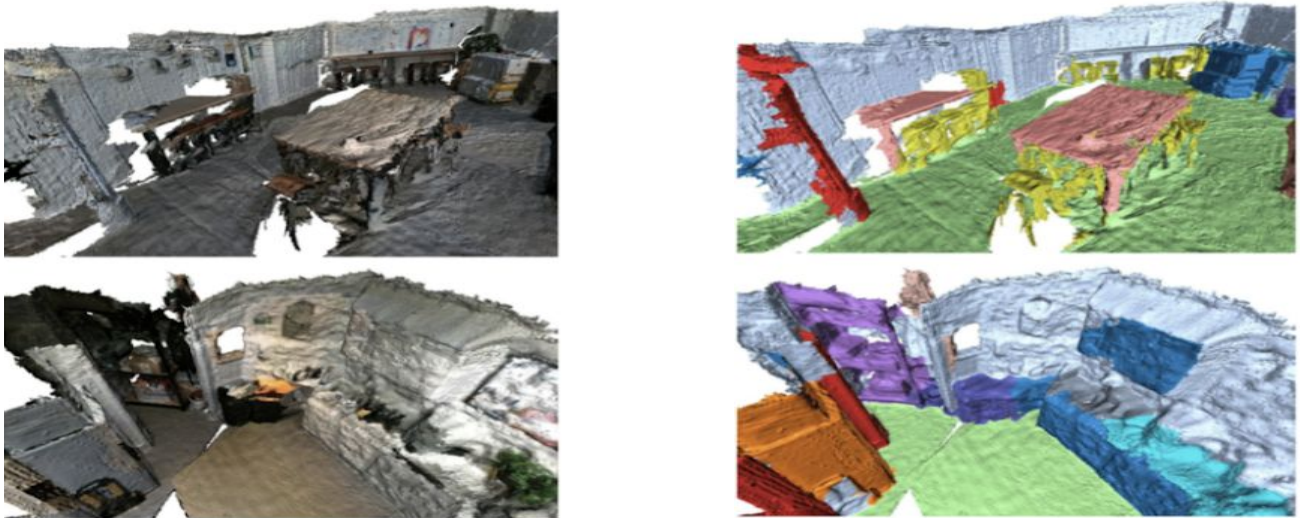


Overview

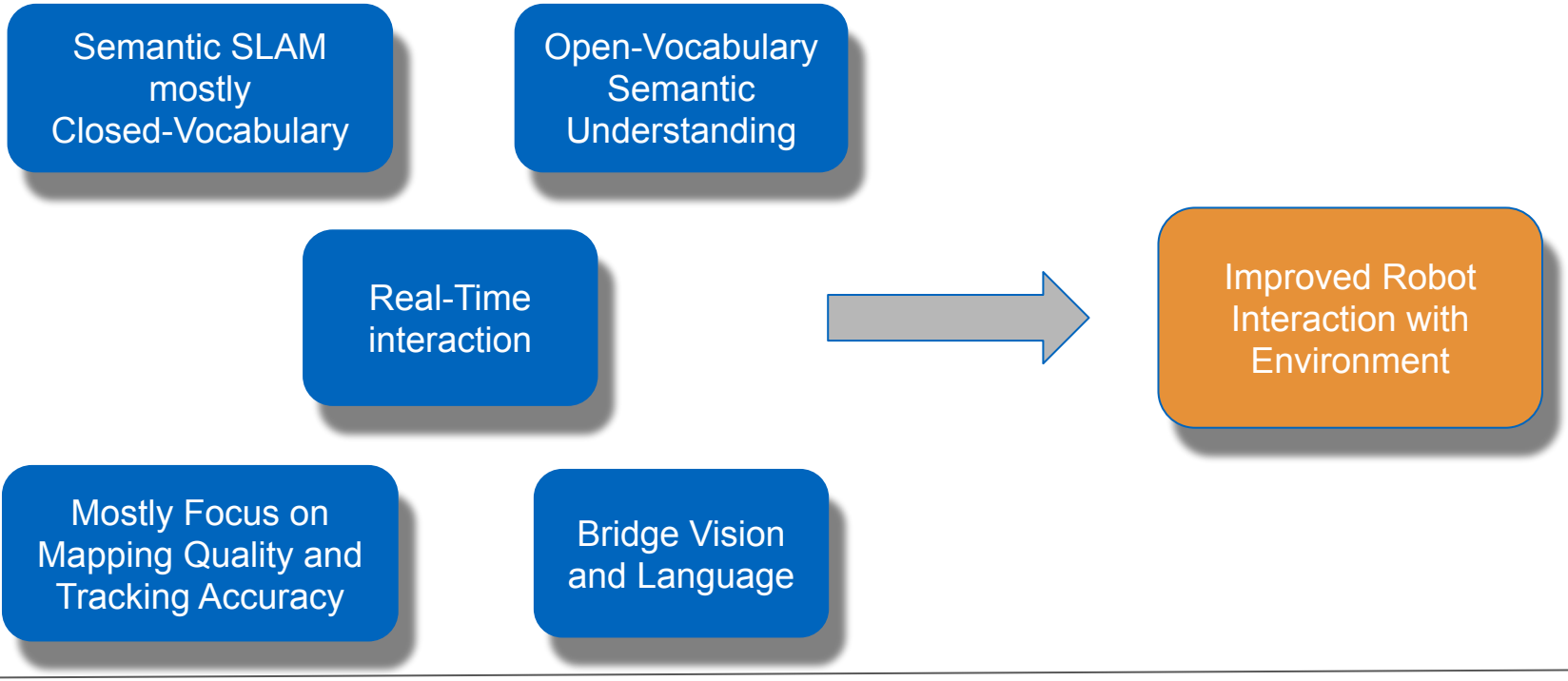
1. **Introduction & Motivation**
2. **Related Work**
3. **Initial Approach**
4. **Problems**
5. **New Approach**
6. **Future Work**
7. **Conclusion**
8. **Q&A**

Introduction

- Traditional SLAM: Geometric maps and localization
  → Lacks semantic understanding of environment
- Semantic SLAM: Integration of object-level understanding



Motivation


Semantic SLAM mostly Closed-Vocabulary
Open-Vocabulary Semantic Understanding
Real-Time interaction
Improved Robot Interaction with Environment
Mostly Focus on Mapping Quality and Tracking Accuracy
Bridge Vision and Language

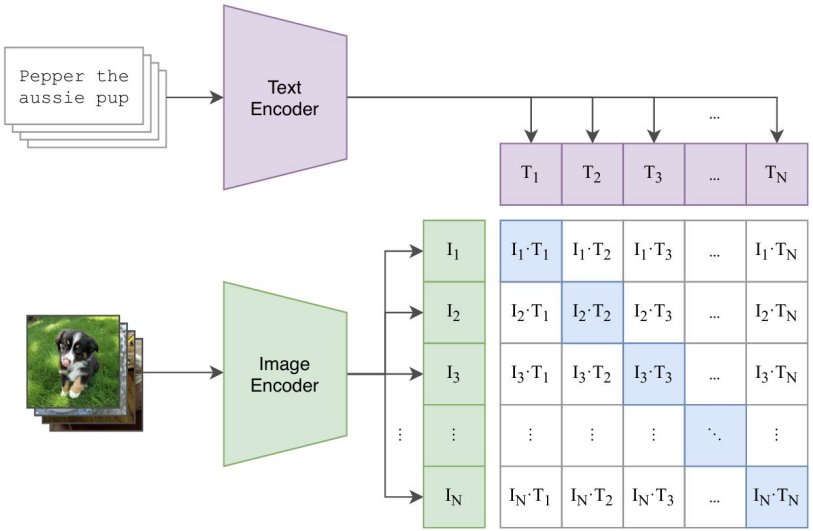
Problem Statement

01	**Goals**	• Open-Vocabulary semantic SLAM system extending existing ones • promptable inference opportunities
02	**Challenges**	• Integration Challenges • Computational Challenges • Robustness Challenges
03	**Benefits**	• Improve Robotic Navigation • No pre-defined labels necessary • Use in autonomous systems/AR/VR

Related Work

Foundation Models

Contrastive Language-Image Pre-Training [1]

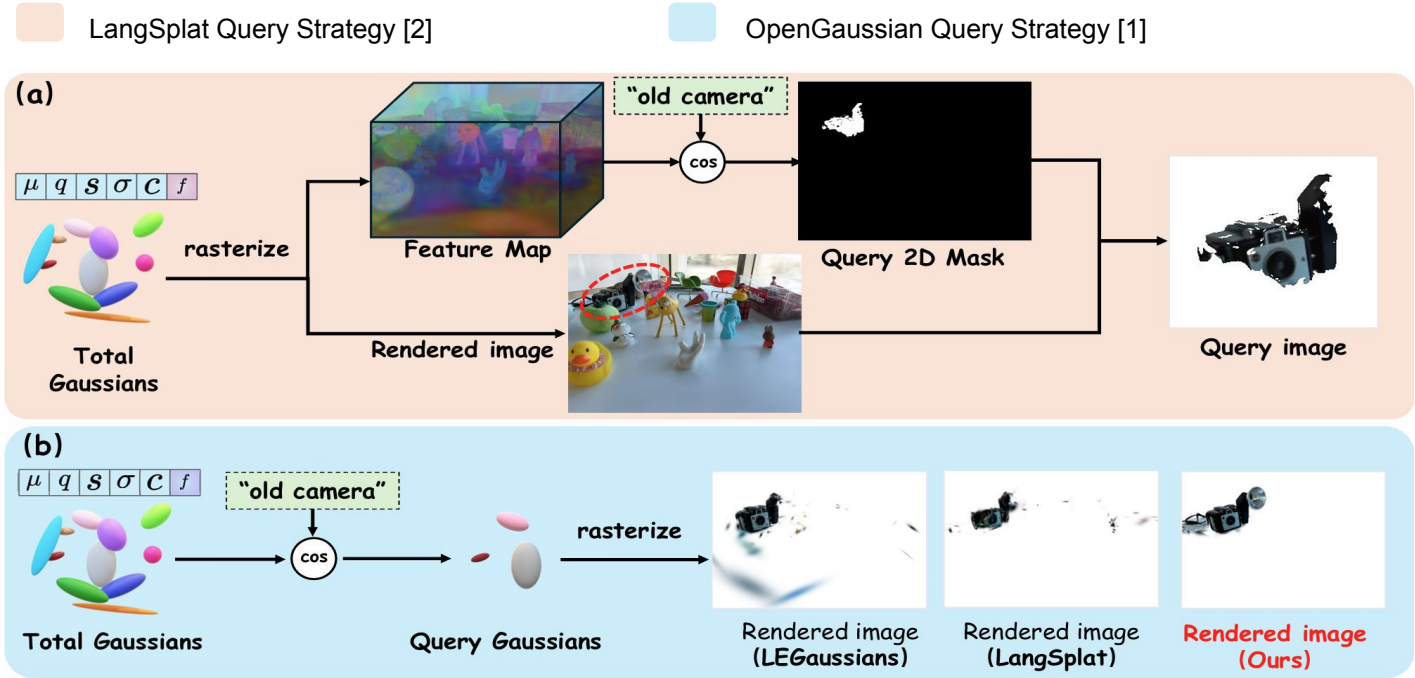


Segment Anything [2]

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
[2] Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Related Work

3D Reconstruction



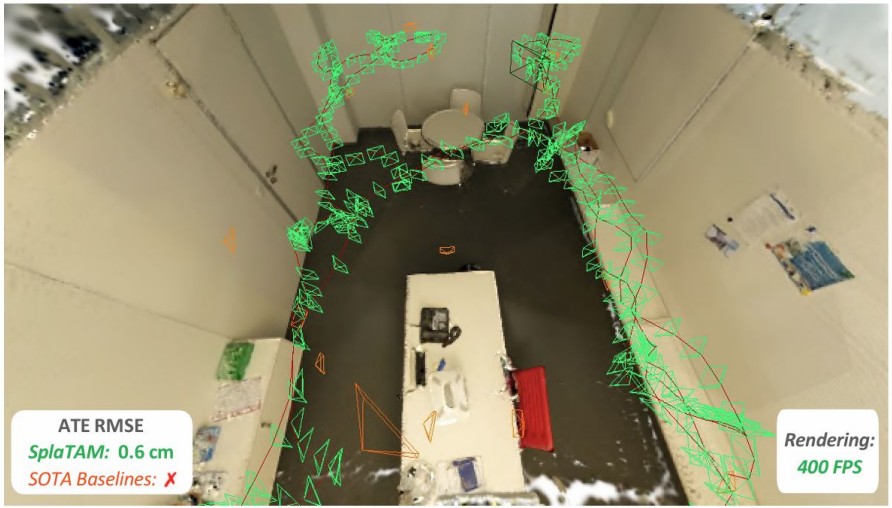
[1] Wu, Yanmin, et al. "OpenGaussian: Towards Point-Level 3D Gaussian-based Open Vocabulary Understanding." arXiv preprint arXiv:2406.02058 (2024).

[2] Qin, Minghan, et al. "Langsplat: 3d language gaussian splatting." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

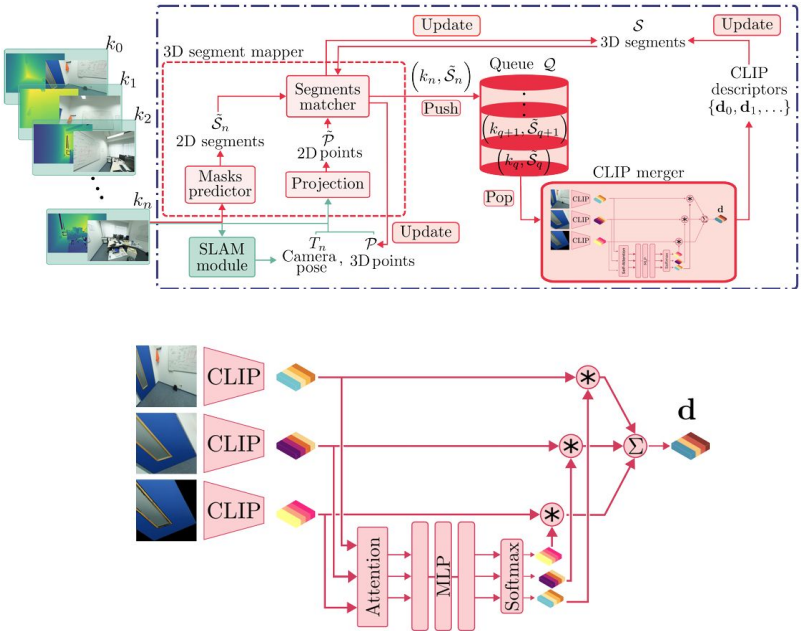
Related Work

SLAM

SplaTAM [1]



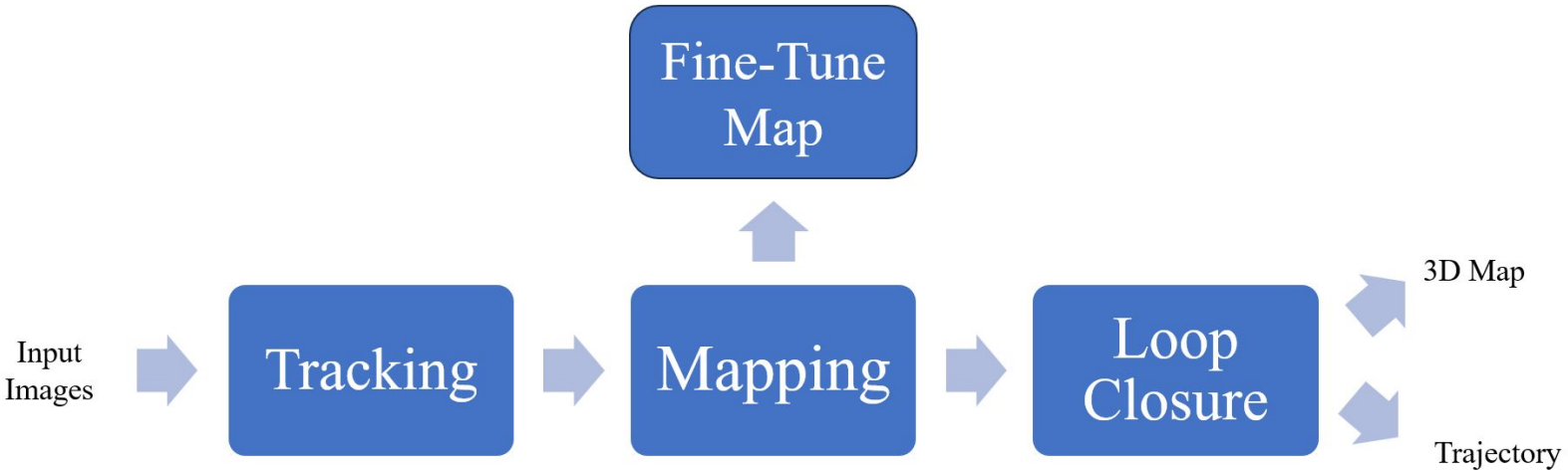
Ovo-SLAM [2]



[1] Keetha, Nikhil, et al. "SplaTAM: Splat Track & Map 3D Gaussians for Dense RGB-D SLAM." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[2] Martins, Tomas Berriel, Martin R. Oswald, and Javier Civera. "OVO-SLAM: Open-Vocabulary Online Simultaneous Localization and Mapping." arXiv preprint arXiv:2411.15043 (2024).

Initial Approach

SplaTAM + OpenGaussian

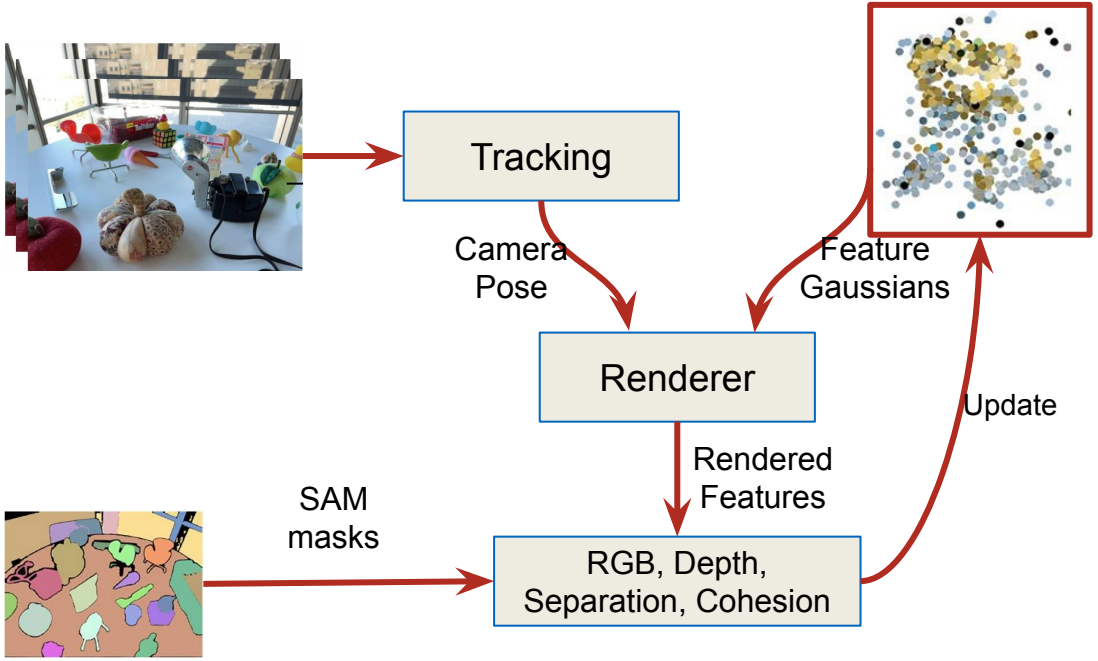


Initial Approach

Methodology

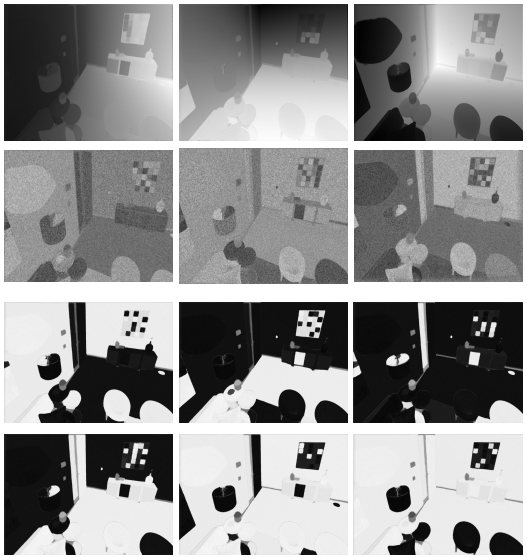
→ 6 dimensional feature for each gaussian

→ Discretize and cluster gaussians based on features



Initial Approach - Initial Results

→ 6-dimensional Features



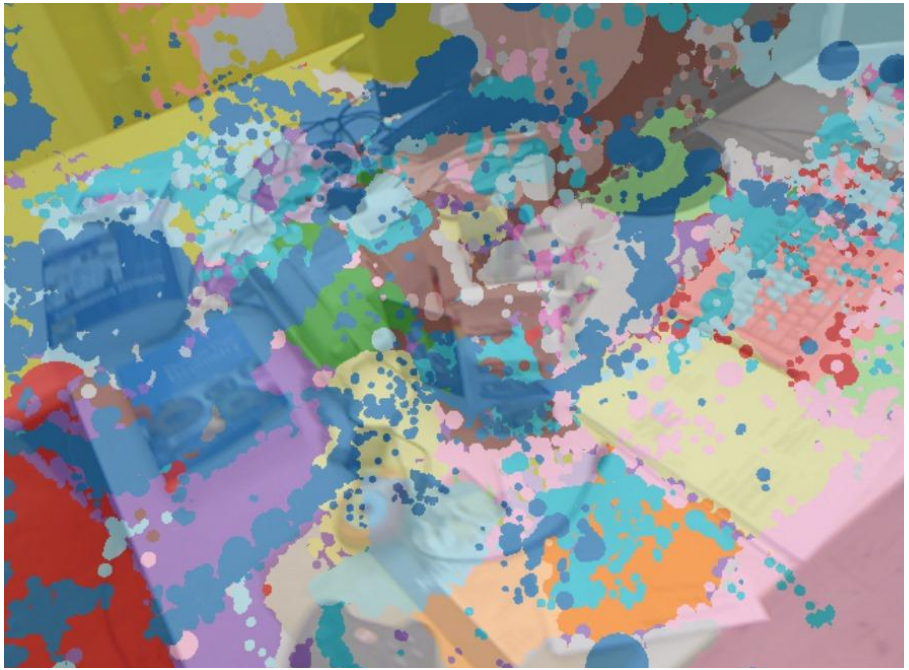
→ Clustered Gaussians

Initial Approach

Problems: Clustering

- Fixed number of clusters
- Overlapping clusters
- Slow clustering

Possible Solutions:
→ Adaptive/dynamic number of clusters
→ Clustering optimization

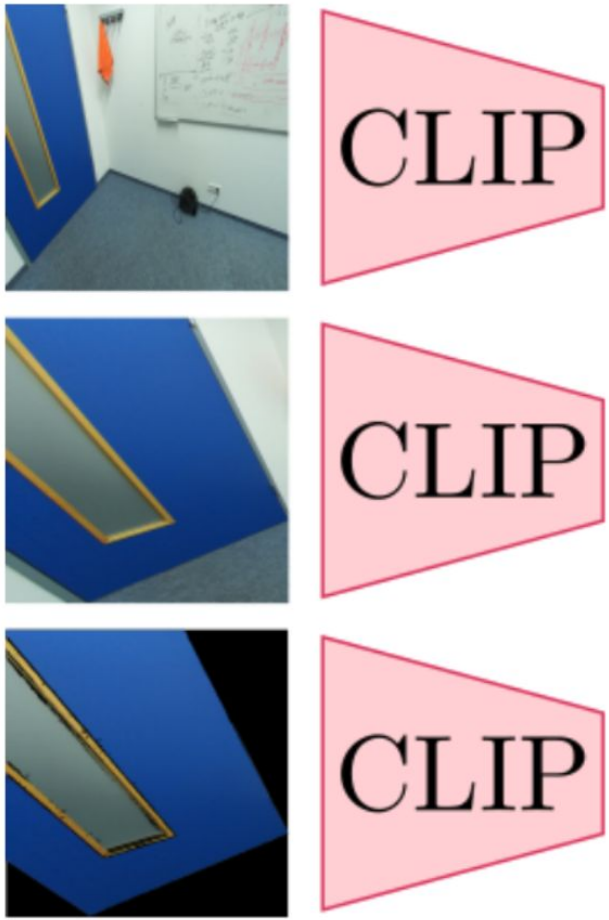


Initial Approach

Problems: Weak Language Features

- Idea 1: CLIP extraction on cropped object
  → No context

- Idea 2: CLIP extraction on different scopes of cropping
  → More context
  → Need for sophisticated CLIP merging

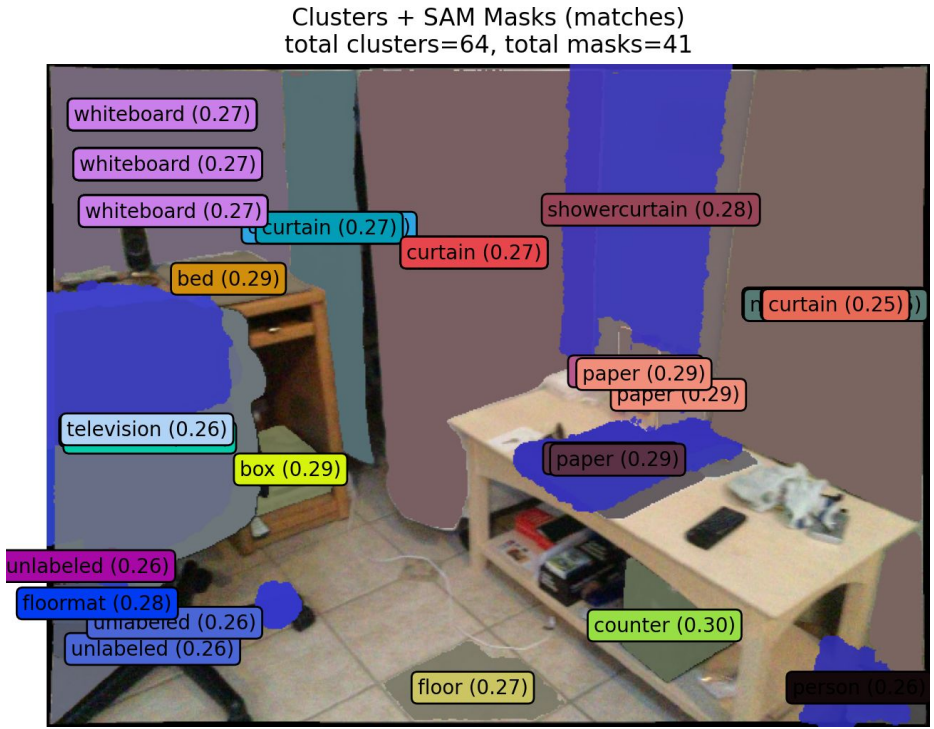


Initial Approach

Problems: Weak Language Features

- Contextless segmentation
  - Missing parts
  - Unrecognized objects

→ Bad CLIP features
→ Misclassification of objects



Recap - Approach

- **Before**: Integrate OpenGaussian's techniques in SplaTAM
  - Train and cluster 6 dimensional semantic feature for each Gaussian
  - Associate SAM Mask to Cluster (IoU)
  - Extract CLIP from SAM Mask (no context of picture)

  ⇒ Slow mapping and weak language features

- **Now**: Reproject pixels to Gaussians
  - No features
  - Grounded SAM (SAM + Grounding DINO + RAM) for context informed labels
  - Merge clusters based on IoU and labels (CLIP)

  ⇒ Fast mapping, higher quality SAM masks and language features

New Idea - Gaussian Reprojection

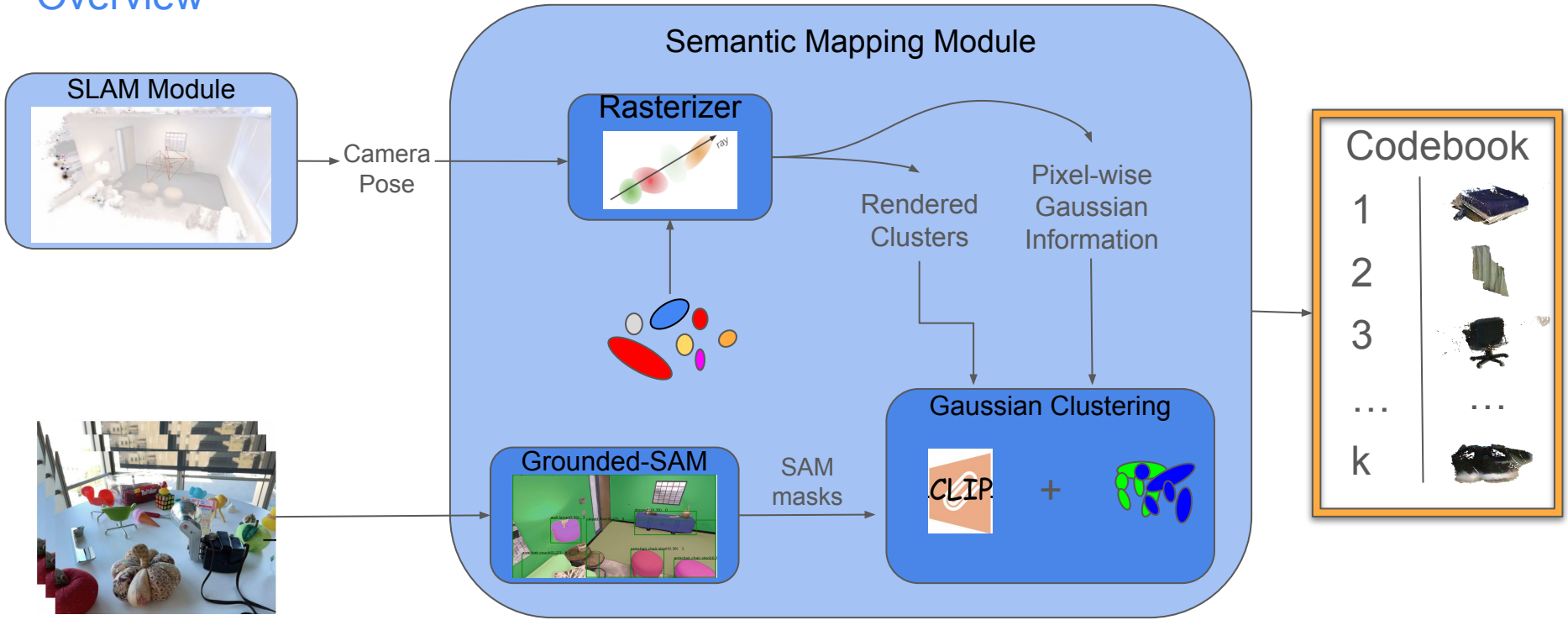
→ Pixelwise Gaussian Information

→ Pixels within Mask reproject to a Cluster

→ Cluster Gaussians for 2D sam mask

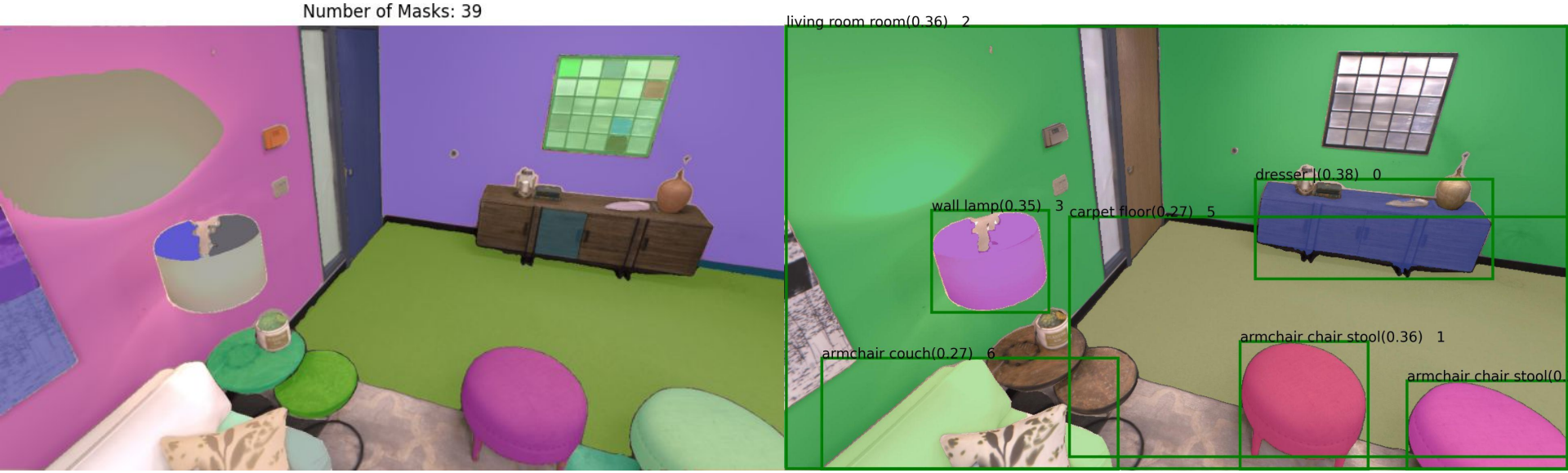
Current Approach

Overview



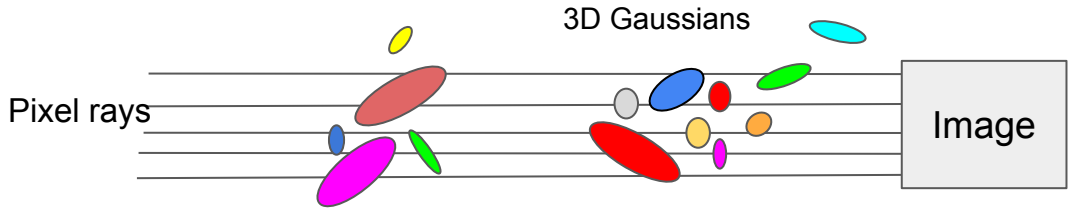
Current Approach

Grounded SAM (Grounding DINO + SAM + RAM)



Current Approach

Gaussian Rasterizer



Standard Rasterizer → Color/Depth

$$C = \sum_{i=1}^{n} (T_i \cdot c_i), \quad \text{where} \quad T_i = \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j)$$

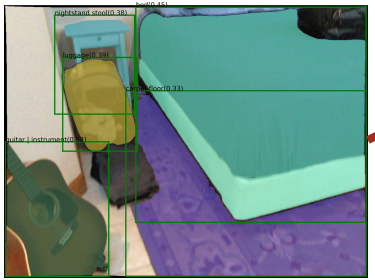
Modified Rasterizer → Color/Depth, Pixelwise Gaussian Contribution

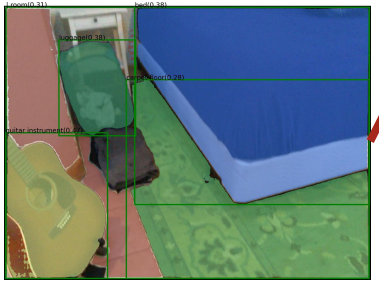
37	32	12	3	98
21	74	89	42	57
1	92	56	2	98
63	13	16	17	83

Current Approach

2D-Mask to 3D-Gaussians

1. Get pixels of mask
2. Get closest Gaussian on Ray
3. Group Gaussians





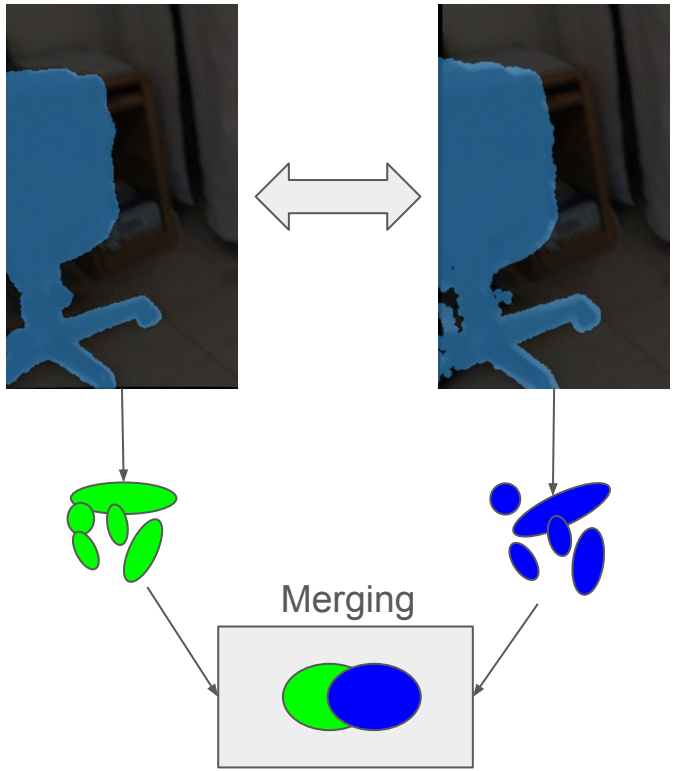
Current Approach

Clustering Module

- Challenge: Match clusters/masks between frames
- Given additionally: language feature (mask label)

- Solution: Merge clusters
  - Case: small camera movement
    → Merge if high rendered overlay (IoU)

  - Case: big camera movement
    → Merge if low rendered overlay (IoU)
    and strong CLIP cosine similarity



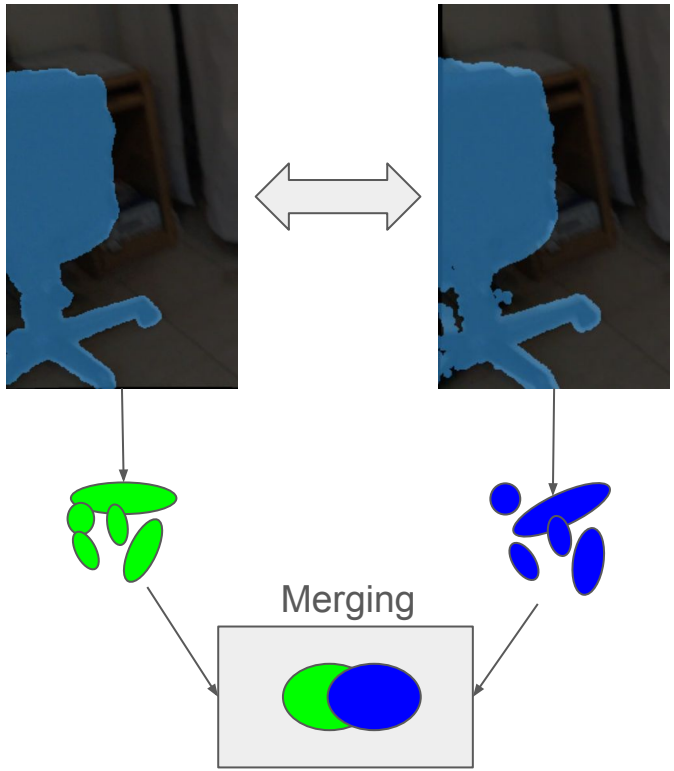
Current Approach

Clustering Module

- Challenge: Match clusters/masks between frames
- Given additionally: language feature (mask label)

- Solution: Merge clusters
  - Case: small camera movement
    → Merge if high rendered overlay (IoU)

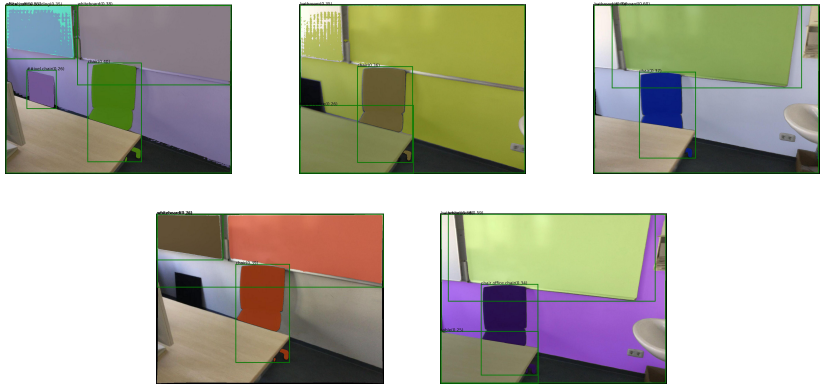
  - Case: big camera movement
    → Merge if low rendered overlay (IoU)
    and strong CLIP cosine similarity



Current Approach

Possible Extensions

→ Regular Re-clustering of Current Clusters



Chair visible in all views

→ Postprocess: Merging of Clusters

Evaluation

Datasets & Metrics

Quantitative Analysis

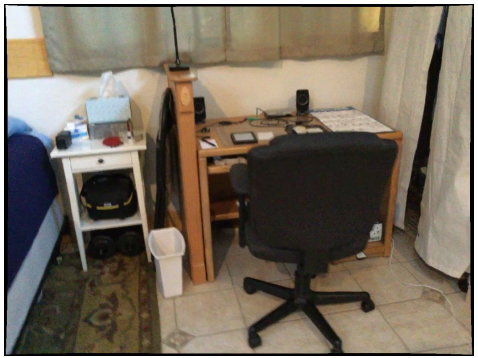
Metrics: mIoU, Accuracy, ATE

Dataset: Replica (synthetic Dataset)

Qualitative Analysis

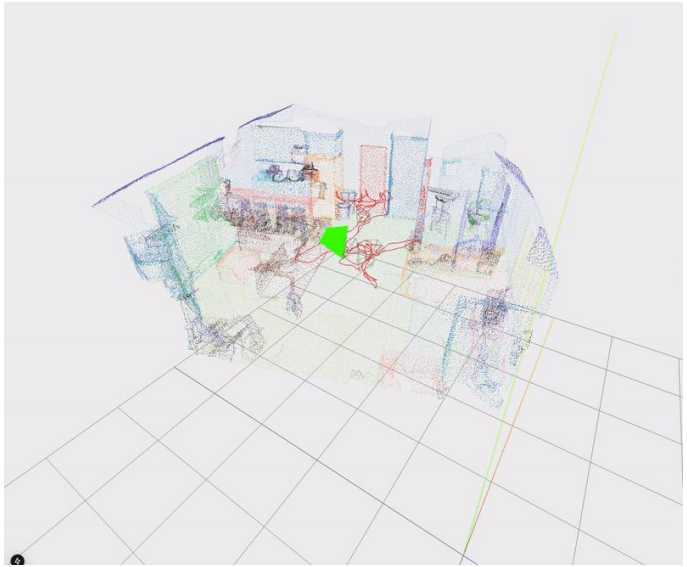
Approach: Render Single Clusters (Overlay), Single Prompts

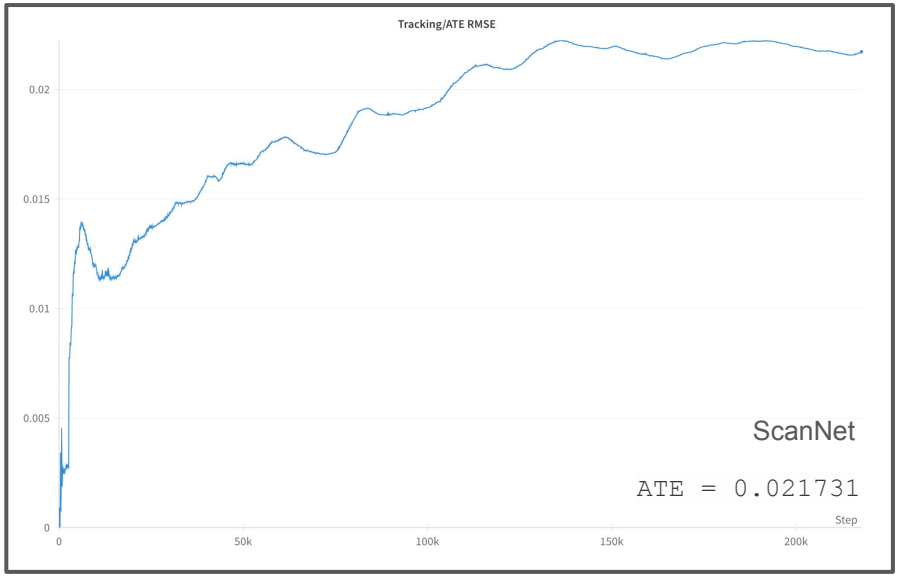
Datasets: Replica, Scannet (scene0)



Tracking Performance

Tracking Module Unmodified

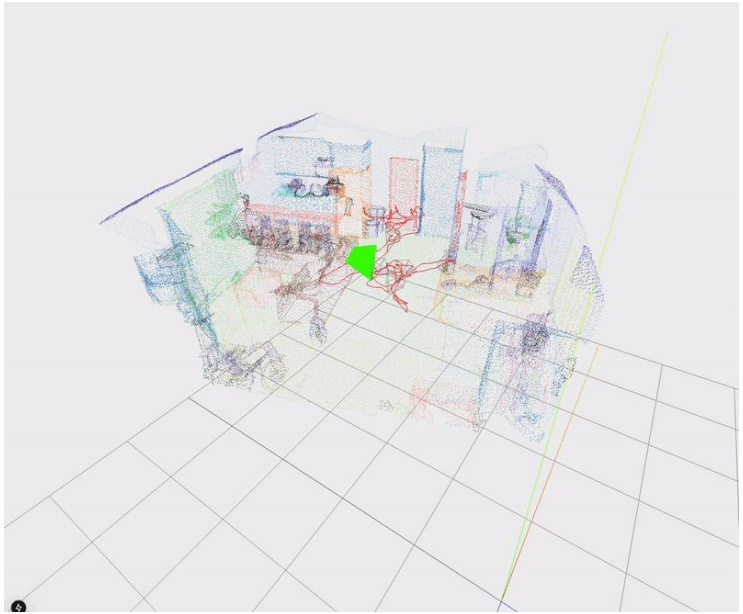


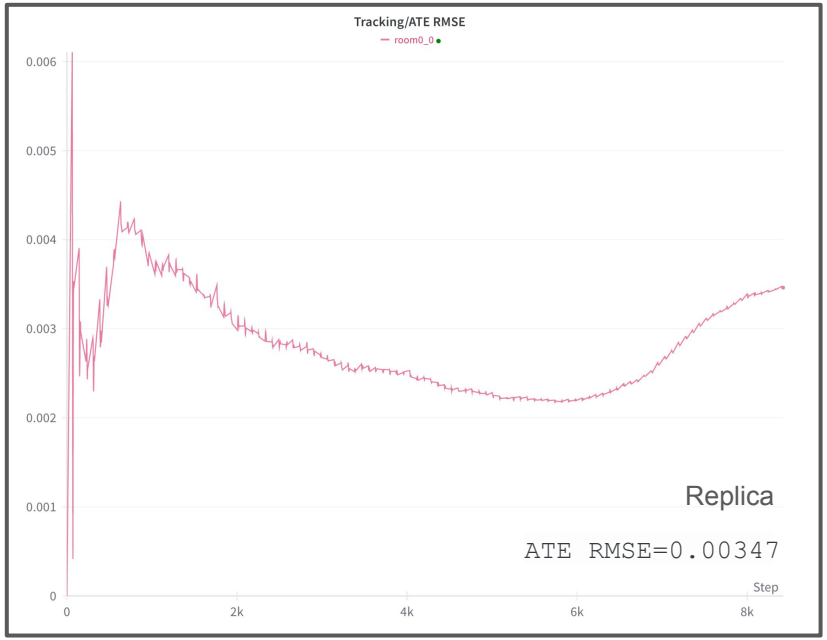


25

Tracking Performance

Tracking Module Unmodified





26

Quantitative Results

Table 1: Replica Semantic Evaluation

Scene	Only Scene Labels		51 Labels		101 Labels	
	mIoU (%)	Accuracy (%)	mIoU (%)	Accuracy (%)	mIoU (%)	Accuracy (%)
Room0	27.65	45.13	26.54	44.68	26.44	44.62
Room1	35.55	36.03	33.87	35.78	33.88	35.78
Room2	30.98	41.58	29.63	38.63	29.04	38.65
Office0	23.75	31.47	23.25	31.75	23.75	31.47
Office1	28.21	41.03	25.29	23.53	16.85	18.64
Office2	34.76	58.46	24.77	52.85	24.77	52.85
Office3	32.05	53.90	27.86	50.00	27.86	50.00
Office4	57.11	62.14	52.72	60.84	52.57	60.81
Average	**33.86**	**46.04**	**30.49**	**42.19**	**29.16**	**41.11**
OVO-SLAM	-		27.1	38.6	-	
Open3DIS (SigLip)	-		25.6	38.7	-	
OpenNeRF	-		20.4	31.7	-	

Qualitative Results

Rendered Clusters

Scannet

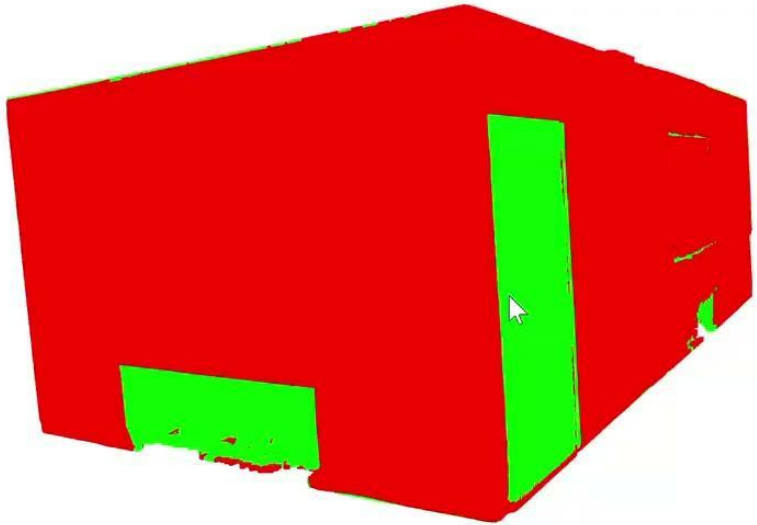
Replica

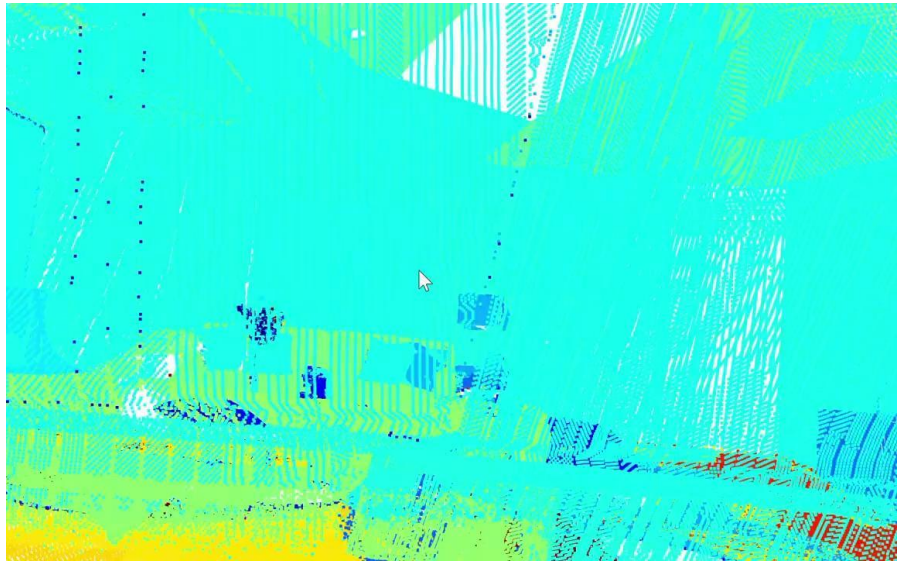
Qualitative Results

Codebook Example

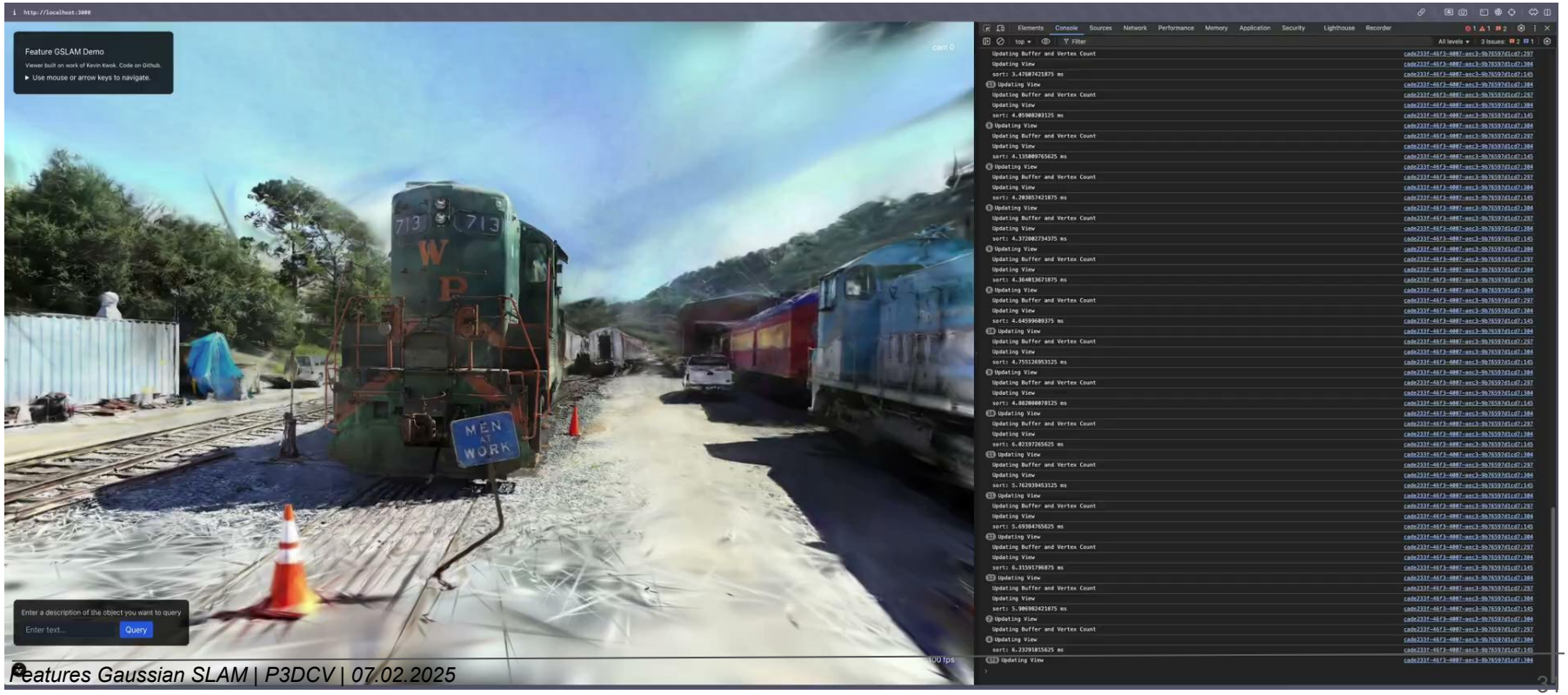
Qualitative Results

Replica Room0





Visualization & Prompting



Future Work

→ More advanced computation of language feature for cluster

→ Leverage more information coming from rasterizer

→ Better Association of clusters with SAM mask

→ Re-Clustering & Merging

→ Dependence on SAM masks → make them more reliable

Dependence on SAM masks

Lessons Learned

→ Hard to train object-wise features in SLAM

→ Approach very dependent on quality of SAM masks

→ Pixel-Gaussian correspondence delivers valid results

→ Tracking and Mapping performance very important for semantic clustering

Thank you for your attention!

If you have questions, please feel free to ask!

Additional - Merge Language Features