Project Proposal

Kevin Zhou (03760454), Jakob Günther (03738122)

December 13, 2025

1 Proposal List:

OpenScene: 3D Scene Understanding with Open Vocabularies

2 Own Proposal:

Paper

LERF: Language Embedded Radiance Fields

Dataset

Replica or handheld captures

Modifications (2 of 3, depending on time)

1. SAM-supervised Boundary Improvement

Complement CLIP supervision using Segment-Anything (SAM) to receive clearer object boundaries and prevent the heatmap from bleeding outside of the object. Evaluation using localization accuracy (high-relevancy pixel within hand-labeled bounding box) and/or IoU between generated relevacny maps and SAM-generated masks from different test-views.

2. Higher-Resolution Feature Distillation

Combine DINO with upsampling models (e.g. FeatUp, LiFT) to receive per-pixel embeddings resulting in higher resolution for heatmaps and querying of smaller, more detailed objects. Evaluation using success-rate on queries for objects smaller than e.g. x% of the image.

3. 3D-to-Text Captions

Enable 3D-to-Text functionality, e.g. selecting an area in 3D and receiving a textual description. Implement some "volumetric selector" that either: 1. retrieves the learned CLIP embeddings from a 3D region (aggregate or highest relevance) and feed the embeddings to e.g. ClipCap (might require changing the CLIP-model to receive ClipCap compatible embeddings), 2. or take "Virtual Camera" to render image crop for VLM (e.g. LLaVa or BLIP) to caption