Rebuttal for
A Unified Framework for 3D Scene Understanding

Anonymous Author(s)
Affiliation
Address
email

Figure-R 1: Qualitative results of UniSeg3D on the ScanNet200 dataset. We show open-vocabulary segmentation results using open-set text prompts. Red prompts mean categories that are not present in the ScanNet200 labels, while blue prompts describe the attributes of various objects, such as affordances and color. For each scene, we show the instance with the highest similarity score to the query embedding. These results highlight the model's open-vocabulary capability.

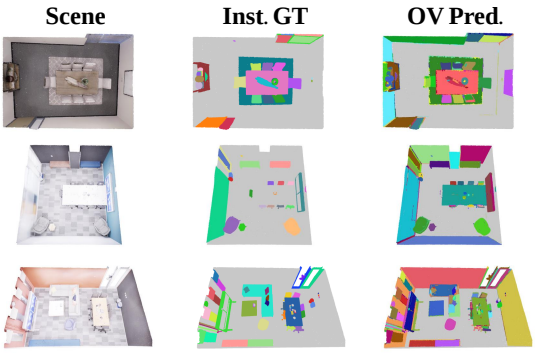

Figure-R 2: Visualization of open-vocabulary segmentation results obtained by UniSeg3D on the Replica dataset. **Our model has never been trained on the Replica dataset,** yet it is still able to produce effective segmentation results. This demonstrates the strong zero-shot capability of UniSeg3D.

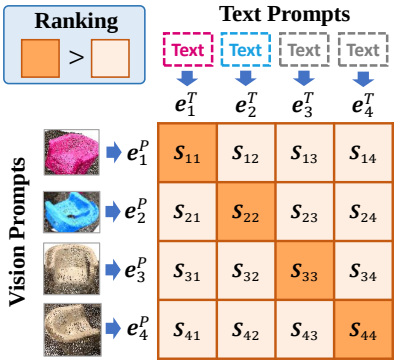

Figure-R 3: A contrastive learning matrix for the vision-text pairs. The ranking rule is designed to suppress the incorrect pairs.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.