Proxy3D: Efficient 3D Representations for Vision-Language Models via Semantic Clustering and Alignment

Supplementary Material

A. Implementation and Dataset Details

Training details. Table 1 shows details and hyperparameters for the proposed Proxy3D training with four stages. We employ 8× A6000 GPUs, and a complete training takes approximately 2 days. We train both the 3D embeddings including conventional positional and the proposed spatial embeddings with the language model (LM) during stages I-II and only the language model in stages III-IV.

Table 1. **Training details and hyperparameters for each stage.**

Stage	I	II	III	IV
Trainable modules	3D emb. + LM		LM	
Frozen modules	-		3D emb. + LM	
Global batch size	256	256	256	128
Accumulation steps	1	1	1	2
Number of epochs	1	3	3	1
Learning schedule	cosine with 5e-6 rate and 0.1 warmup ratio			

Dataset details. Detailed information about our SpaceSpan dataset splits and question types is reported in Table 2. Distribution of questions and dataset sources are illustrated in Figures 1-2.

Table 2. **SpaceSpan dataset details.** Text, ID, Num. and MC stands for textual, identifier, numerical and multiple choice answering formats, respectively.

Name	Text	ID	Num.	MC
ScanQA[1]	26K	0	0	0
SQA3D[8]	26K	0	0	0
Scan2Cap[3]	32K	0	0	0
ScanRefer[2]	0	32K	0	0
Multi3DRef[11]	0	39K	0	0
MMScan[7]	57K	57K	0	0
SR-91K[9]	0	0	30K	20K

Complexity analysis. As presented in Table 3, as compared to vision encoder processing time, clustering takes relatively small portion of time. The token compression, however, leads to significant reduction of 30% in inference time. This implies the effectiveness of reducing token number in improving MLLM inference efficiency.

B. Additional Qualitative Results

Visualization of attention for proxy tokens. We aggregate queries for each proxy token in the last output layer across all tokens to estimate a normalized 3D proxy attention map, as shown in Figure 3. The attention of the last layer has been selected because it contains all information in the network related to 3D reasoning task. Figure 3 shows that Proxy3D is able not only to assign attention to objects of interest but also to draw attention to related objects, indicating the emergence of spatial reasoning.

Table 3. **Latency** (sec) for batch size of 8 and 16. OOM in table refers to out-of-memory run.

	Vision Encoder		Cluster.		LLM		Latency Per Question	
Batch size	8	16	8	16	8	16	8	16
Qwen2.5-VL	9.4	16.2	-	-	9.3	OOM	2.25	OOM
Ours	9.4	16.2	2.9	5.6	1.1	1.5	**1.75**	**1.5**

Failure case analysis. Often, failure cases originate from innate problems within datasets, such as questions that have multiple candidate answers and annotation errors in ScanNet-based benchmarks as shown in Figure 4. These problems require manual corrections which is also described in other analyses [4, 5].

More visualization samples. We provide additional visualization samples in Figure 5. Proxy3D is also deft at capturing small objects, establishing strong spatial consistencies and detailed 3D vision.

Limitations and future work. Though Proxy3D performs well in 3D QA, 3D VG and spatial reasoning tasks, we still have not fully reached its potential in spatiotemporal tasks. In addition, we did not utilize post-training reward learning, which is a common way to further improve performance metrics [6, 10]. In general, spatial reasoning for outdoor scenes is an underexplored but practically important application for future work.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *CVPR*, 2022. 1

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3D object localization in rgb-d scans using natural language. In *ECCV*, 2020. 1

[3] Zhenyu Chen, Ali Gholami, Matthias Niessner, and Angel X. Chang. Scan2Cap: Context-aware dense captioning in RGB-D scans. In *CVPR*, 2021. 1

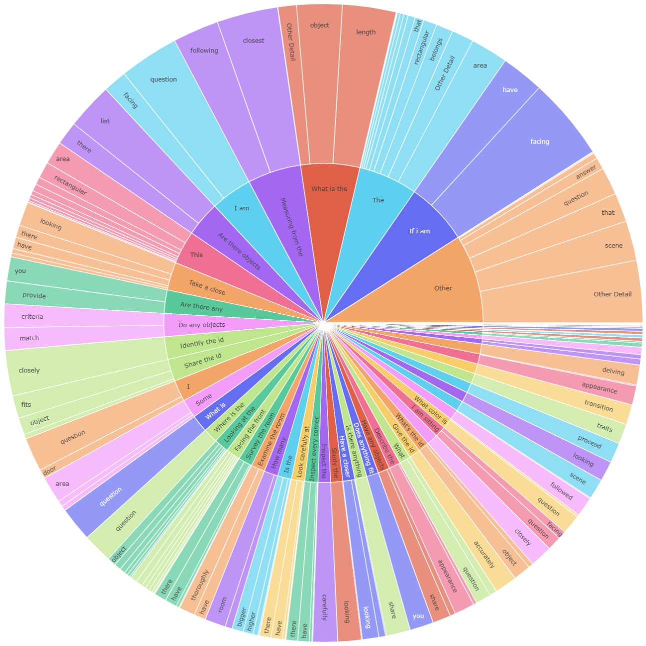

Figure 1. **Distribution of SpaceSpan questions**. SpaceSpan is a diverse dataset that consists of various question types with the goal to form 3D spatial intelligence during MLLM training.

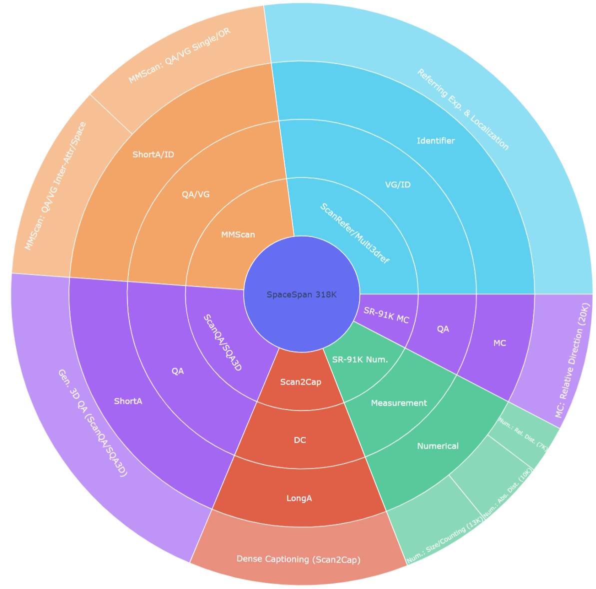

Figure 2. **SpaceSpan data sources**. SpaceSpan integrates heterogeneous datasets with visual grounding, question answering, dense captioning and measurement tasks for 3D scene understanding. LongA, ShortA stands for long answers and short answers, respectively.

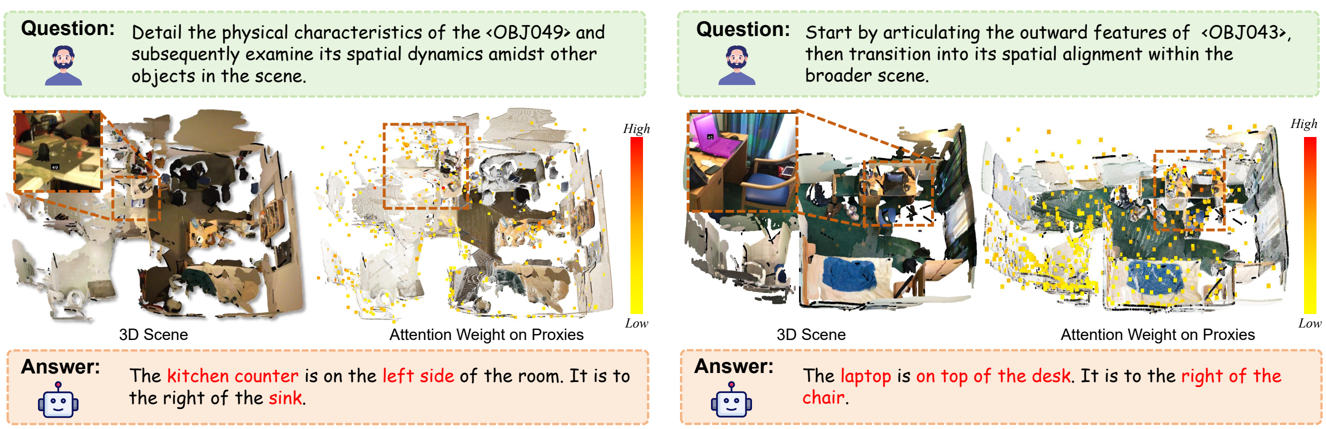

Figure 3. **Visualization of a normalized attention for our 3D proxy tokens.**

[4] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In *ICML*, 2024. 1

[5] Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang. Unveiling the mist over 3D vision-language understanding: Object-centric evaluation with chain-of-analysis. In *CVPR*, 2025. 1

[6] Jiangyong Huang, Xiaojian Ma, Xiongkun Linghu, Yue Fan, Junchao He, Wenxin Tan, Qing Li, Song-Chun Zhu, Yixin Chen, Baoxiong Jia, and Siyuan Huang. LEO-VL: Towards 3D vision-language generalists via data scaling with efficient representation. *arXiv:2506.09935*, 2025. 1

[7] Ruiyuan Lyu, Jingli Lin, Tai Wang, Shuai Yang, Xiaohan Mao, Yilun Chen, Runsen Xu, Haifeng Huang, Chenming Zhu, Dahua Lin, and Jiangmiao Pang. MMScan: A multi-modal 3D scene dataset with hierarchical grounded language annotations. In *NeurIPS*, 2024. 1

[8] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3d scenes. In *NeurIPS*, 2023. 1

[9] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing mllms in video spatial reasoning. *arXiv:2504.01805*, 2025. 1

[10] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan.

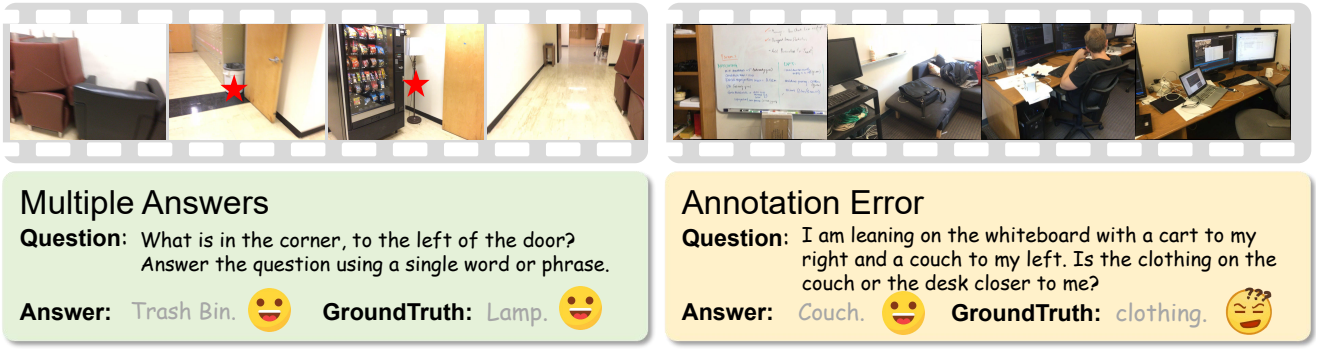

Figure 4. **Dataset-related failure cases in Scannet-based benchmarks.** These benchmarks demonstrate flaws in question design, with questions that has multiple candidate answers and annotation errors.

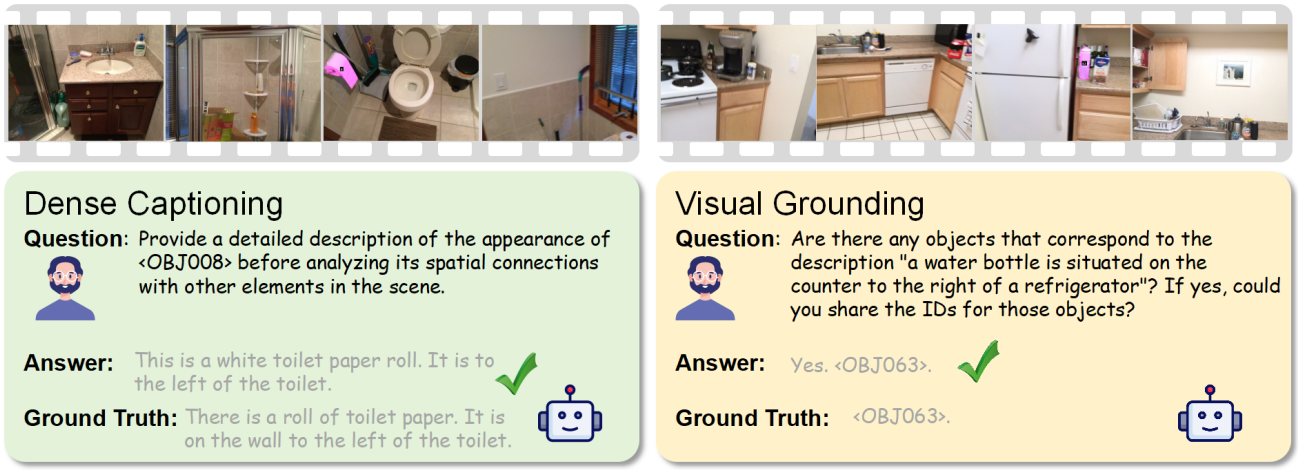

Figure 5. **More visual samples on the success of Proxy3D.** Proxy3D is not only able to accurately locate large objects, but also track small objects with few occurrences.

Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *NeurIPS*, 2025. 1

[11] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3DRefer: Grounding text description to multiple 3d objects. In *ICCV*, 2023. 1