Scale3DUR: scaling data generation for 3D understanding and reasoning

APPENDIX

A. Data Distribution

Figure.1 illustrates the analysis of the first four question types based on an expanded dataset, revealing four distinct aspects: inquiries about local objects (e.g., "What color is it?"), questions regarding the global context (e.g., "How many are there?"), complex relational questions (e.g., "What is to the right of the trash can?"), and those addressing directions or positions (e.g., "Where is it?"). Meanwhile, Figure 2 presents the distribution of answers by highlighting the top 20 most frequently occurring words in responses, where the X-axis indicates the answers, the Y-axis reflects their counts, and different colors distinguish the words.

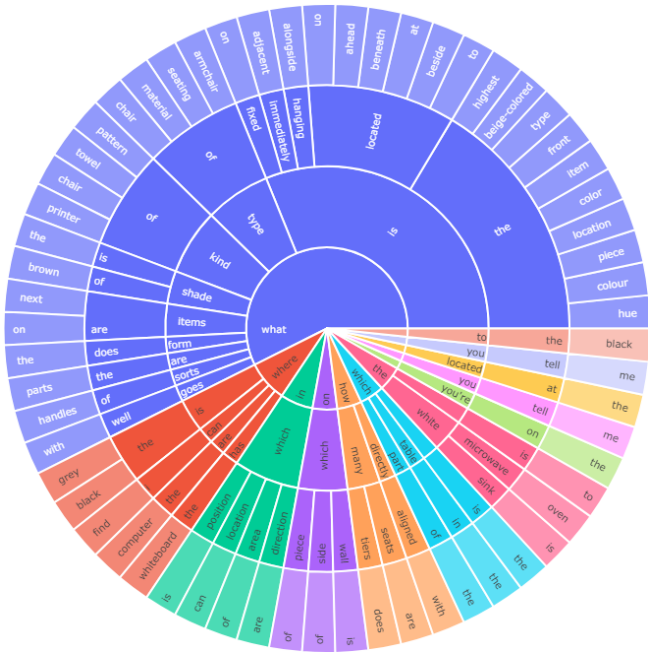

Fig. 1. we analyzed the distribution of the first four types of questions based on the expanded dataset. The results indicate that these questions cover several aspects: Firstly, there are questions concerning local objects, such as "What color is it?" and "What type of thing is it?"; Secondly, there are questions concerning the global context, such as "How many are there?" and "Are there more than how many tables?"; Thirdly, there are questions involving complex relationships between multiple objects, such as "What is to the right of the trash can?" and "What is in the middle?"; Finally, there are questions concerning directions or positions, such as "Where is it?" and "Which way is it facing?".

B. More Qualitative Results

As shown in Figure.3, In the 3D Dense Captioning and 3D Question Answering tasks, Vision Prompts and Instance

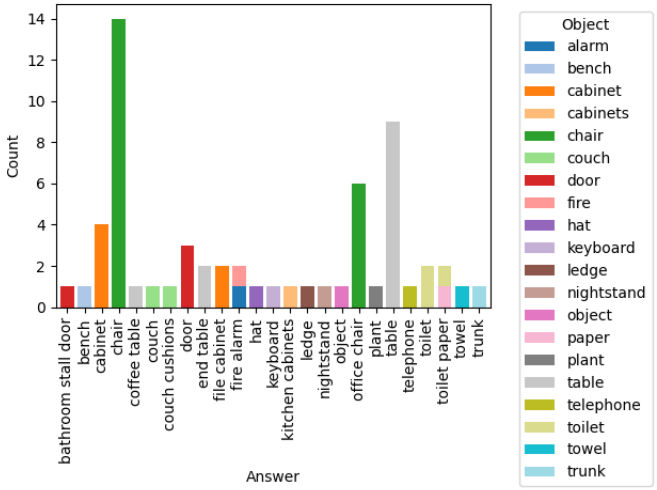

Fig. 2. We analyzed the top 20 words that appeared most frequently in the responses based on an expanded dataset, showing the distribution of answers to different types of questions. The X-axis of the chart represents the answers, the Y-axis represents the count, and different colors represent different words.".

Prompts play crucial roles in guiding the model's understanding and generation of outputs. Vision Prompts help in the spatial localization of objects by framing the specific region or object in question, ensuring that the model focuses on the correct target within the 3D environment. This is particularly important in complex scenes where multiple objects may overlap or clutter the view. On the other hand, Instance Prompts assign unique identifiers to the objects being described, facilitating the model's ability to distinguish between multiple instances of similar objects. This instance-level annotation helps improve the accuracy of the model's responses by allowing it to produce more structured and specific descriptions, crucial for both QA and dense captioning tasks in rich 3D environments.

C. More Evaluations

In this section, we evaluate the effectiveness of the Scaling Data Paradigm by applying the newly extended dataset to multiple models, particularly focusing on the 3DLLM's performance in 3D Question Answering tasks(As shown in Table.I). The goal is to investigate whether scaling the dataset improves model performance across key evaluation metrics, including BLEU, CIDEr, ROUGE-L, and METEOR. As the table shows, the model trained on 100 epochs with the extended dataset demonstrates notable improvements, particularly in BLEU-4 (+1.13) and CIDEr (+3.28) scores,

TABLE I

EVALUATION OF SCALING DATA PARADIGM ON MODEL PERFORMANCE BOLD: THIS INDICATOR OUTPERFORMS THE OTHER TASKS BEING COMPARED, WITH CIDER SERVING AS THE PRIMARY EVALUATION METRIC.

Dataset	Evaluation Metrics						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L	METEOR
3D-LLM [1](BLIP2-flant5)	**39.30**	**25.20**	**18.40**	12.00	69.40	**35.70**	14.50
Custom Dataset (50 epochs)	35.37	22.07	15.38	10.71	69.63	35.37	14.31
Custom Dataset (100 epochs)	37.93	25.10	18.14	**13.13**	**72.68**	35.42	**15.14**

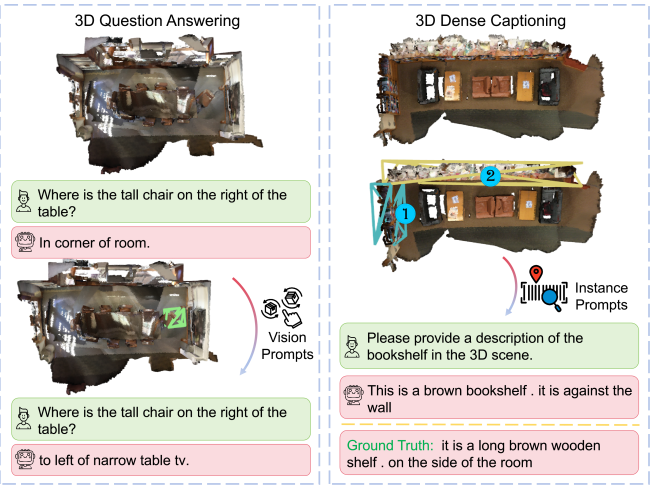

Fig. 3. Vision and Instance Prompts improve object localization and distinction, enhancing model accuracy in 3D captioning and question answering.

TABLE II

EVALUATION OF SCALING DATA PARADIGM ON DIFFERENT BACKBONE VERSIONS SCALE3DUR: USING THE SCALING DATA PARADIGM WE PROPOSED. BOLD: THIS INDICATOR OUTPERFORMS THE OTHER TASKS BEING COMPARED, WITH CIDER SERVING AS THE PRIMARY EVALUATION METRIC. NOTE: THE MODEL USED IN THIS TABLE IS THE LL3DA[2] WITH DIFFERENT BACKBONE VERSIONS.

Dataset	Evaluation Metrics			
	BLEU-4	CIDEr	ROUGE-L	METEOR
opt-1.3b	13.53	76.79	37.31	15.88
opt-1.3b(Scale3DUR)	**14.17**	78.94	37.89	**16.07**
Qwen1.5-7B	13.61	78.78	38.09	15.95
Qwen1.5-7B(Scale3DUR)	13.98	**79.22**	**38.41**	16.00

indicating enhanced capability in generating coherent and contextually relevant answers in complex 3D scenes. This suggests that the Scaling Data Paradigm is effective in improving model performance as the data volume increases, leading to better understanding and reasoning within 3D environments.

D. More backbone version

As shown in Table.II, We evaluate the impact of the extended dataset generated using the Scaling Data Paradigm on different backbone versions of the LL3DA model: 'facebook/opt-1.3b'[3] and 'Qwen/Qwen1.5-7B'[4]. As seen in the table, applying the expanded dataset consistently improves the performance across both backbones, particularly in CIDEr and BLEU-4 metrics. For instance, in the case of 'facebook/opt-1.3b', the BLEU-4 score increases from 13.53 to 14.17, and CIDEr improves from 76.79 to 78.94. Similarly, the Qwen backbone shows an increase in BLEU-4 from 13.61 to 13.98, and CIDEr from 78.78 to 79.22. These improvements demonstrate the effectiveness of the Scaling Data Paradigm in enhancing the model's ability to comprehend and generate accurate responses in the 3D environment, with both larger and smaller backbone versions benefiting from increased data diversity.

REFERENCES

[1] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 20 482–20 494, 2023.

[2] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, "Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 428–26 438.

[3] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[4] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.