A. Annotation Tool

We set up an interactive annotation tool for data collection based on SQA3D [14]. We present a visualization of the user interface (UI) in Fig. A.1, including a 3D scene viewer (left), an annotation editor (middle), and object information (right). There are three Grounding-Chains (G-Chains) and three Grounding-QA-Chains (GQA-Chains) to be annotated in the annotation editor for each target object.

Two panels on the right exhibit details of each annotation:

- For the grounding task, the human annotator is supposed to fill the referential text with precise and natural language, and then select the involved knowledge types and a list of objects that match the referential text.
- For the question answering (QA) task, the human annotator first generates a QA pair based on the "grounding text", which lists three *primary grounding texts* from the G-Chains. Then, the annotator labels the knowledge type and the flag of *extra knowledge*, *e.g.*, "no" if the answer is covered by the "grounding text".

B. Baselines

ViL3DRel [4]. This is a 3D vision-language (3D-VL) specialist model for grounding, trained in a single-task scheme. We use the official checkpoint trained on ScanRefer [3].

3D-VisTA [23]. While 3D-VisTA adopts task-specific finetuning for downstream tasks by default, we perform multi-task training by aggregating the datasets it uses. The datasets for grounding include ScanRefer, Nr3D [1], Sr3D [1], and Multi3DRefer [21]. The datasets for QA include ScanQA [2] and SQA3D [14].

PQ3D [24]. PQ3D is a 3D-VL generalist model that supports both grounding and QA tasks. We directly use the checkpoint after pretraining and multi-task training. The training datasets include Scan2Cap [5] in addition to the datasets for 3D-VisTA.

SceneVerse [12]. SceneVerse is a 3D-VL model pretrained on large-scale grounding datasets. To make it a generalist model for grounding and QA, we finetune a QA head while freezing the pretrained backbone weights to preserve its grounding ability. The datasets for fine-tuning include ScanQA and SQA3D.

GPT-4o [16]. As a state-of-the-art large language model (LLM), GPT-4o is selected as a specialist model for QA to probe the upper bound of LLMs. We adopted the evaluation pipeline outlined in [13] to assess GPT-4o's performance. In our evaluation, we prompt GPT-4o to answer the questions based on a collection of objects, which comprises the category, location, size, and attributes of each object. The object attributes are extracted with GPT-4V [16].

LEO-multi. To address the lack of grounding capability in LEO [11], we design a grounding loss alongside the original autoregressive language modeling loss. The grounding loss resembles contrastive learning (CLIP [17]) on the alignment between object tokens (the input to LLM) and text embeddings. With the multi-task objective, we train LEO-multi by combining grounding (ScanRefer and Nr3D) with instruction-tuning tasks (ScanQA, SQA3D, 3RScan-QA [11], 3RScan-Plan [11], and 3RScan-Dialog [11]).

LEO-curricular. Similar to LEO-multi, LEO-curricular incorporates the contrastive grounding loss but learns grounding and QA in a curricular strategy. We first train the 3D encoder of LEO-curricular with grounding loss on ScanRefer and Nr3D. We then freeze the 3D encoder and finetune the LLM with LoRA [9] on instruction-tuning datasets.

PQ3D-LLM. This is a model variant based on PQ3D, substituting the original T5-Small [18] with Vicuna-7B [6], which is finetuned with LoRA. The training setting is identical to PQ3D.

Chat-Scene [10]. Chat-Scene is designed to be a 3D-VL generalist model, using object identifiers and LLM to perform grounding. The training datasets include ScanRefer, Multi3DRefer, Scan2Cap, ScanQA, and SQA3D. We directly use its released checkpoint for evaluation.

C. Additional Analyses

C.1. Outliers and Prospective Questions

We observe several outliers in our evaluation results. Below, we address these outliers and answer prospective questions:

Poor grounding for LEO-multi and LEO-curricular. The grounding performance of these two models falls significantly below that of others. We attribute this to our implementation of the grounding task learning, which employs contrastive learning between object tokens and text embeddings of pretrained LLM (*e.g.*, Vicuna). We receive two lessons from this: (1) contrastive learning demands large-scale data while the scarce 3D-VL data proves insufficient; and (2) unlike CLIP, the text embeddings of pretrained LLM may not be suitable for contrastive learning.

Poor QA for PQ3D and PQ3D-LLM. Despite the strong performance in grounding for these two models, their performance in QA is notably weak. We attribute this to the choice

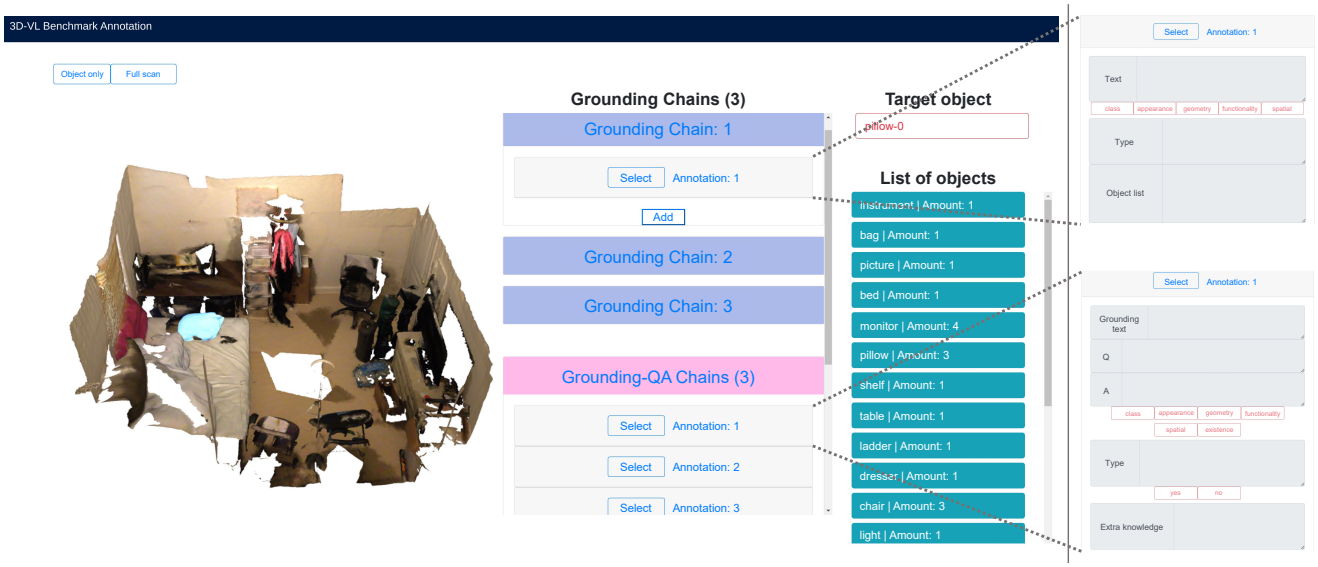

Figure A.1. **Overview of our annotation tool.** The interface includes a 3D viewer (left), an annotation editor (middle), and object information (right). Two panels on the right exhibit details of each annotation for the grounding and QA task, respectively.

of language encoder. Compared to 3D-VisTA, PQ3D adopts a similar overall architecture but differs in language encoder: 3D-VisTA uses BERT [8], whereas PQ3D uses CLIP. The reasonable QA performance of 3D-VisTA indicates that the CLIP language encoder is suboptimal for QA task, despite being adequate for grounding. This further underscores the linguistic gap between grounding and QA tasks: grounding texts encompass descriptive language while questions involve diverse querying patterns. It reveals the limitations of the CLIP language encoder in addressing this disparity.

Why is PQ3D-LLM worse than PQ3D in grounding? While the LLM incorporated by PQ3D-LLM is only used for QA, it introduces a significant number of extra parameters for optimization, which may hinder the learning of grounding during multi-task learning and consequently weaken the grounding performance.

Why is PQ3D-LLM not better than PQ3D in QA? In PQ3D, the input to the QA head (*e.g.*, LLM) only comprises object tokens, which can be regarded as foreign language for LLM. The challenge of utilizing these tokens for QA cannot be alleviated by incorporating LLM, despite its strength in language processing. Additionally, incorporating LLM for QA is prone to overfitting given the scarcity of 3D QA data.

Strong performance of GPT-4o in QA. We observe that GPT-4o significantly outperforms 3D-VL models in QA, especially in questions related to appearance (App.) and existence (Exi.). This showcases the upper bound of using explicit textual information (*e.g.*, object lists with attributes), which bypasses 3D perception. The considerable gap between GPT-4o and 3D-VL models further suggests that 3D perception remains a key bottleneck in 3D-VL models.

C.2. Discussion on the Effect of LLM

LLM hinders grounding. This conclusion is drawn from the consideration of two categories of models:

- *LLM directly used for grounding.* Models that perform grounding based on LLM (*e.g.*, Chat-Scene) exhibit less robust performance compared to models without LLM. Specifically, despite the close performances on ScanRefer, Chat-Scene lags behind PQ3D and SceneVerse on BEACON3D, which implies the potential risk of overfitting for LLM-based grounding. However, LLM may be beneficial in more complex grounding tasks that require high-level reasoning or planning, *e.g.*, sequential grounding [22]. This suggests that the effect of LLM-based grounding varies according to task complexity.
- *LLM not directly used for grounding.* In models that do not rely on LLM for grounding (*e.g.*, PQ3D-LLM), we observe a weaker performance in grounding after incorporating LLM. This shows the negative effect of LLM's parameters on the learning of grounding during multi-task learning. A practical solution is to decompose multi-task learning into curricular learning, which disregards LLM's parameters during the learning of grounding.

LLM does not truly improve QA. We elaborate on this conclusion from three aspects: clarification on how we draw the conclusion, explanation on why per-case metrics do not matter, and analysis on why LLM may not help 3D QA.

Table A.1. **Evaluation results of grounding on BEACON3D (3RScan).** The settings and metrics follow the main paper. ** denotes models that have never been trained in 3RScan. * denotes models that have been trained in 3RScan but not on grounding. ‡ denotes only point feature is available.

	Knowledge type				Overall	
	Class	App.	Geo.	Spa.	Case	Obj.
w/o LLM						
ViL3DRel** [4]	41.5	44.9	37.4	37.3	41.5	18.4
3D-VisTA** [23]	45.6	38.3	37.4	40.9	45.6	21.7
PQ3D**‡ [24]	38.3	28.0	36.4	35.3	38.3	13.6
SceneVerse [12]	**61.8**	**51.4**	**53.3**	**57.3**	**61.8**	**37.5**
LLM-based						
LEO-multi*	10.1	9.9	9.7	8.8	10.1	0.4
LEO-curricular*	15.3	17.7	11.8	9.3	15.3	1.1
PQ3D-LLM**‡	30.3	27.6	24.6	25.5	30.3	8.5

Table A.2. **Evaluation results of QA on BEACON3D (3RScan).** † indicates text input (*i.e.*, object locations and attributes) instead of 3D point cloud. ** denotes models that have never trained in 3RScan. * denotes models that have been trained in 3RScan but not on QA. ‡ denotes only point feature is available.

	Knowledge type					Overall	
	Class	App.	Geo.	Spa.	Exi.	Case	Obj.
w/o LLM							
3D-VisTA** [23]	15.2	24.1	28.2	25.3	28.9	25.7	3.3
PQ3D**‡ [24]	6.5	19.6	13.6	16.6	52.6	25.7	0.7
SceneVerse* [12]	28.3	32.3	34.6	38.9	44.6	37.4	0.4
LLM-based							
GPT-4o† [16]	34.8	38.2	40.0	45.4	**60.7**	**46.1**	**11.0**
LEO-multi	**37.0**	35.0	**51.8**	**48.5**	46.5	44.1	1.8
LEO-curricular	19.6	**41.8**	48.2	**48.5**	50.7	45.6	7.4
PQ3D-LLM**‡	13.0	21.4	17.3	21.4	33.2	23.4	1.8

- *How we draw the conclusion.* The evidence mainly comes from two observations: (1) the results of LLM-based models are comparable to those without LLM under object-centric metrics; and (2) fragile grounding-QA coherence.
- *Why per-case metrics do not matter.* While LLM-based models show slightly better results in per-case metrics, these metrics do not reliably indicate true 3D QA capability. As demonstrated in the main paper, per-case metrics are not robust enough due to their vulnerability to shortcuts. Moreover, the advantage of LLM-based models in per-case metrics is marginal, which is intuitive given LLM's strength in general QA. We believe the marginal gap in per-case metrics cannot evidence a gap in the true capability of 3D QA.
- *Why LLM may not help 3D QA.* We conjecture the bottleneck in 3D QA lies in the alignment between 3D features and QA modules, rather than language generation, where the primary strength of LLM resides. Prior works [12, 23, 24] have shown that simple QA heads (*e.g.*, T5-Small or MCAN [20]) perform well in 3D QA, as the task demands only a basic level of language generation. This explains the minimal contribution of LLM to 3D QA.

Harnessing LLM for 3D-VL tasks. We first identify a critical problem in current 3D large vision-language models (LVLMs) and then propose an effective solution to harness LLM for 3D-VL tasks.

- *Problem.* Our investigation in the main paper reveals that overfitting to text is a critical problem in current 3D LVLMs. This implies a significant imbalance between 3D encoder and LLM, that is, LLM often overshadows 3D encoder during training. This issue is less pronounced in 2D LVLMs owing to the robust 2D features learned through extensive pretraining, which is infeasible for 3D encoders.
- *Solution.* We propose curricular learning, progressing from grounding to QA, as an effective solution to mitigate this issue by shielding 3D features from LLM interference. The effectiveness is evidenced by the advantages of SceneVerse and LEO-curricular.

C.3. Limitations and Future Work

First, our benchmark prioritizes focused and systematic analysis, which involves trade-offs in task scope and complexity. Our object-centric evaluation excludes more advanced tasks, such as multi-object grounding and complex reasoning. Extending this evaluation framework to include more complex tasks will be a key direction for future work. Second, our baselines may not cover the wide range of existing 3D-VL models. We will evaluate and analyze more models in the future. Third, we consider the performance of the grounding task as a proxy for the grounding implicitly performed in the QA task. This may be unfair to models whose grounding performance is locked due to issues like improper implementation (*e.g.*, LEO-multi and LEO-curricular). Nonetheless, we believe our approach remains practical for assessing grounding-QA coherence in most 3D-VL generalist models.

D. Domain Transfer

We follow the setting outlined in the main paper to evaluate the baselines in two novel domains: 3RScan [19] and MultiScan [15]. This evaluation is referred to as *domain transfer* since most baselines are only trained on ScanNet [7]. Notably, as Chat-Scene only provides model features for ScanNet, its evaluation on 3RScan and MultiScan is not feasible. We distinguish between two types of domain transfer:

- **: the model has never been trained in the target domain.
- *: the model has been trained in the target domain but on tasks other than the specific one.

Table A.3. **Evaluation results of grounding on BEACON3D (MultiScan).** The settings and metrics follow the main paper. ** denotes models that have never been trained in MultiScan. Only SceneVerse has been trained in MultiScan.

	Knowledge type				Overall	
	Class	App.	Geo.	Spa.	Case	Obj.
w/o LLM						
ViL3DRel** [4]	33.2	34.4	25.0	32.0	33.2	13.2
3D-VisTA** [23]	40.8	30.5	28.1	38.0	40.8	18.9
PQ3D** [24]	56.3	53.9	37.5	52.8	56.3	34.0
SceneVerse [12]	**59.5**	**54.6**	**53.1**	**56.6**	**59.5**	**35.9**
LLM-based						
LEO-multi**	9.0	9.1	9.4	9.0	9.0	1.3
LEO-curricular**	11.7	11.0	6.3	9.0	11.7	0
PQ3D-LLM**	51.0	46.8	37.5	49.0	51.0	25.8

Table A.4. **Evaluation results of QA on BEACON3D (MultiScan).** † indicates text input (*i.e.*, object locations and attributes) instead of 3D point cloud. ** denotes models that have never been trained in MultiScan. * denotes models that have been trained in MultiScan but not on QA.

	Knowledge type					Overall	
	Class	App.	Geo.	Spa.	Exi.	Case	Obj.
w/o LLM							
3D-VisTA** [23]	6.5	22.6	16.7	13.2	28.8	19.1	0
PQ3D** [24]	21.0	16.8	16.7	9.6	39.0	20.8	0.6
SceneVerse* [12]	16.2	32.1	12.5	26.5	38.1	28.9	3.1
LLM-based							
GPT-4o† [16]	**29.0**	**41.6**	33.3	25.7	**59.3**	**39.4**	**7.6**
LEO-multi**	12.9	24.1	41.7	24.3	32.2	25.6	2.5
LEO-curricular**	8.1	27.0	**50.0**	**28.7**	41.5	29.8	3.8
PQ3D-LLM**	6.5	21.9	8.3	11.0	25.4	17.0	0.6

Results. We present the domain transfer results for 3RScan in Tabs. A.1 and A.2, and MultiScan in Tabs. A.3 and A.4. The overall trends are consistent with those reported in the main paper for ScanNet. For example, while models without LLM (*e.g.*, SceneVerse) excel in grounding, LLM-based models (*e.g.*, LEO-curricular) perform better under per-case metrics but struggle with object-centric metrics in QA. In particular, we report several specific findings regarding the domain transfer results:

- ***Challenge of domain transfer.*** All models exhibit notable performance declines, emphasizing the challenge of domain transfer (ScanNet → 3RScan; MultiScan). SceneVerse surpasses PQ3D owing to its comprehensive pretraining across diverse scene domains. Moreover, training on 3RScan-QA improves QA performance on 3RScan (LEO-multi and LEO-curricular). These findings highlight the inevitable domain gap and the benefits of cross-domain pretraining.
- ***Limitations of feature-dependent models.*** PQ3D and PQ3D-LLM experience considerable performance drops on 3RScan due to a lack of image and voxel features. While this issue results in only a marginal drop on ScanNet, as reported in the original paper [24], the considerable drop on 3RScan indicates the heightened challenges of transferring to novel domains for feature-dependent models such as PQ3D and Chat-Scene.
- ***More challenging 3D perception in MultiScan.*** Performance on MultiScan is consistently lower than on 3RScan, reflecting the increased difficulty of 3D perception in the domain of MultiScan. SceneVerse, despite using a simple QA head [20], outperforms LEO-multi and matches LEO-curricular. This suggests that the bottleneck in QA lies in 3D perception, suppressing the contribution of LLM. It further underscores the need for more powerful 3D encoders to address this bottleneck.
- ***Performance degradation of GPT-4o.*** GPT-4o exhibits noticeably lower performance on 3RScan and MultiScan compared to ScanNet, with the results on 3RScan approached by LEO-curricular. We attribute this degradation to incomplete object attributes stemming from insufficient multi-view images, which limits the object attribute extraction by GPT-4V. This reveals that, despite their strengths in 3D QA, LLMs and 2D LVLM are constrained by the availability of high-quality multi-view images.

E. Illustration of Data and Evaluation

We present a video demo to illustrate the process of data collection and evaluation (see attachment). Here we show the static overview in Fig. A.2 and A.3.

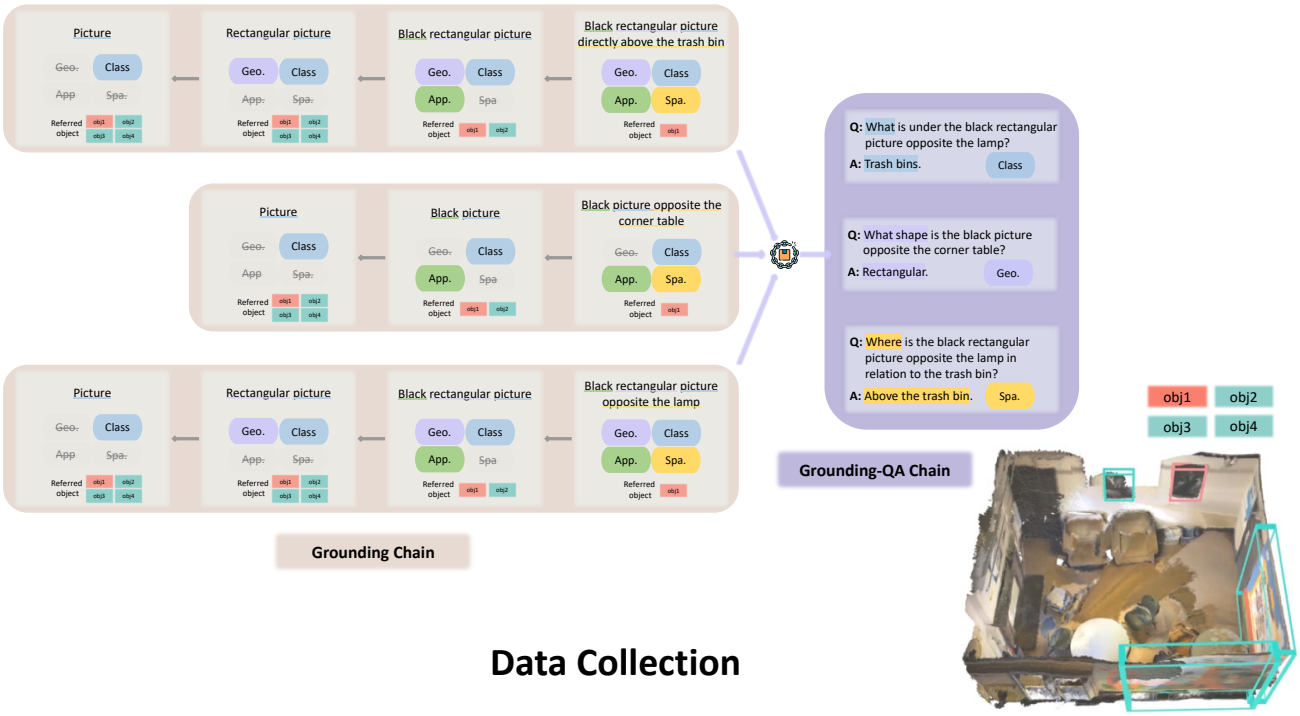

Figure A.2. **Static overview of data collection.** Check the dynamic process in our video demo in the attachment.

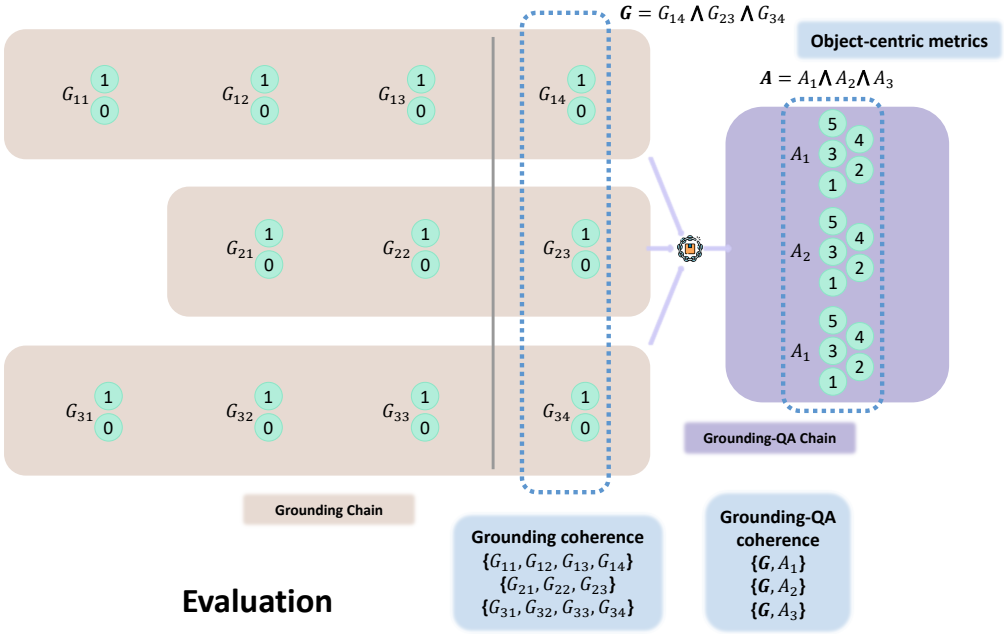

Figure A.3. **Static overview of evaluation.** Check the dynamic process in our video demo in the attachment.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natu-

ral language. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[4] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3, 4

[5] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna, 2023. 1

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 2

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 1

[10] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1

[11] 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 *International Conference on Machine Learning (ICML)*, 2024. 1

[12] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 3, 4

[13] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS Datasets and Benchmarks)*, 2024. 1

[14] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[15] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[16] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3, 4

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2020. 1

[19] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[20] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[21] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *International Conference on Computer Vision (ICCV)*, 2023. 1

[22] Zhuofan Zhang, Ziyu Zhu, Pengxiang Li, Tengyu Liu, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. Task-oriented sequential grounding in 3d scenes. *arXiv preprint arXiv:2408.04034*, 2024. 2

[23] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 3, 4

[24] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 3, 4