ROSS3D: Reconstructive Visual Instruction Tuning with 3D-Awareness
– *Supplementary Material* –

Haochen Wang[1,2] Yucheng Zhao[3†] Tiancai Wang[3*] Haoqiang Fan[3]
Xiangyu Zhang[4,5] Zhaoxiang Zhang[1,2*]
[1]NLPR, MAIS, CASIA [2]UCAS [3]Dexmal [4]MEGVII Technology [5]StepFun
{wanghaochen2022, zhaoxiang.zhang}@ia.ac.cn wtc@dexmal.com

Project Page: https://haochen-wang409.github.io/ross3d

Supplementary Material

A. More Implementation Details

A.1. Position-Aware Video Representation

To inject 3D information into vanilla video frames, this paper utilizes the representation proposed by [33]. Specifically, it adopts sinusoidal position encoding on absolute 3D coordinates (x, y, z), where the coordinate of the pixel located at (i, j) is computed using depth maps $\boldsymbol{D} \in \mathbb{R}^{H \times W}$, the extrinsic matrix $\boldsymbol{T} \in \mathbb{R}^{4 \times 4}$, and a camera intrinsic matrix $\boldsymbol{K} \in \mathbb{R}^{3 \times 3}$

$$\begin{bmatrix} x & y & z & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{D}_{ij} \cdot \begin{bmatrix} j & i & 1 \end{bmatrix} \cdot (\boldsymbol{K}^{-1})^\top & 1 \end{bmatrix} \cdot \boldsymbol{T}^\top. \tag{S1}$$

The encoded positions are then added with the original video features extracted by the vision backbone, *e.g.*, CLIP [20].

A.2. Training Dataset

Our **ROSS3D** is a generalist model that handles multiple tasks within a single set of learned parameters. To achieve this, **ROSS3D** is trained on a combined dataset, including 3D question answering dataset [1, 17], 3D dense captioning dataset [7], and 3D visual grounding dataset [3, 30], in the multi-task manner similar to [33].

The statistics training set is illustrated in Table S1. All data have been converted to the format of LLaVA [16]. There are 223K training samples in total.

A.3. Training Objectives

For general 3D scene understanding tasks such as 3D question answering and 3D dense captioning, we use cross-entropy loss to supervise text outputs and our proposed denoising loss to supervise visual outputs. For 3D visual grounding, to locate more accurately, we only use 3D visual grounding loss introduced next.

We follow previous works [14, 26, 33, 35] and regard the visual grounding task as a classification problem for specific object proposals. Specifically, given a list of object proposals, we obtain object features for each object by aggregating visual embeddings. For each object with a bounding box b_i, we average the features of patches where more than 50% of their points lie within b_i. These object features are then added with the 3D position embedding of the center coordinate. InfoNCE [18, 22, 23, 25] is applied to optimize the similarity between the ground truth object feature and the hidden states of the special `<ground>` token.

Source	# samples	# scenes	Question Length	Answer Length
SQA3D [17]	79,445	518	37.8	1.1
ScanQA [1]	26,515	562	13.7	2.4
Scan2Cap [7]	36,665	562	13.0	17.9
ScanRefer [3]	36,665	562	24.9	–
Multi3DRefer [30]	43,838	562	34.8	–

Table S1. **Detailed statistics for training data.** Average lengths for questions and answers are obtained from [33].

A.4. Evaluation Details

For ScanRefer [3], we simply select the object proposal with the highest similarity as the prediction. For Multi3DRefer [30], we choose the objects with the highest probabilities until the cumulative probability of selecting these objects surpasses 25%. For Scan2Cap [7], we follow [13, 33] to evaluate the captioning performance by inserting special `<sos>` and `<eos>` tokens at the start and end of the prediction, respectively. Greedy sampling is utilized for both 3D dense captioning and 3D question answering tasks.

*Corresponding authors. † Project lead.

γ	SQA3D	ScanQA	ScanRefer	Multi3DRefer
0.125	62.0	105.6	60.2	59.1
0.25	**63.0**	**107.0**	**61.1**	**59.6**
0.5	61.8	105.3	60.8	**59.6**
0.75	61.2	104.9	60.8	59.0

Table S2. **Ablations on the masking ratio γ.** A relatively small masking ratio performs slightly better, but overall, **ROSS3D** is robust against γ.

Δt	SQA3D	ScanQA	ScanRefer	Multi3DRefer
4	**63.0**	**107.0**	61.1	**59.6**
2	62.6	105.4	60.9	59.2
1	61.8	104.8	**61.2**	59.5

Table S3. **Ablations on the interval Δt.** We implement our $\mathcal{L}_{3D}^{corss}$ and $\mathcal{L}_{3D}^{global}$ every Δt steps.

BEV res.	filter	SQA3D	ScanQA	ScanRefer
256×256	✓	62.3	106.5	60.9
432×432	–	61.8	104.6	60.2
432×432	✓	**63.0**	**107.0**	61.1
1024×1024	✓	62.7	106.5	**61.4**

Table S4. **Ablations on global-view reconstruction.** "Filter" indicates whether filtering out black spaces or not.

α	SQA3D	ScanQA	Scan2Cap	ScanRefer	Multi3DRef
0.5	62.3	30.9	**83.4**	60.8	59.3
1	**63.0**	30.8	81.3	**61.1**	**59.6**
5	61.9	**31.0**	81.1	60.9	59.2

Table S5. **Ablation of the denoising loss weight α**, where our **ROSS3D** is quite robust against different values of α.

B. More Experiments

B.1. More Ablation Studies

Design Choices for Cross-View Reconstruction. We ablate the masking ratio γ and the interval Δt in Table S2 and Table S3, respectively. These designs alleviate the discrepancy between training and testing. Empirically, a relatively *small masking ratio*, *i.e.*, 25%, together with an appropriate interval, *i.e.*, 4, performs the best among others. But overall, **ROSS3D** is robust against these designs.

Design Choices for Global-View Reconstruction. We ablate the BEV resolution and the filtering technique in Table S4. **ROSS3D** is quite robust against these designs.

Denoising Loss Weight α. The denoising loss is around 0.2, while the cross-entropy loss is around 1. Therefore, we simply add these two terms. We study different weights α for the denoising loss in Table S5. **ROSS3D** is robust against α.

	Method	Avg.	What	Is	How	Can	Which	Others
1	Video-3D-LLM	41.5	39.4	49.4	42.4	45.8	32.9	38.2
2	① + vanilla	41.7	38.0	49.6	43.2	44.7	35.0	40.3
3	① + cross-view	45.6	41.6	53.4	47.2	48.7	41.5	43.5
4	① + global-view	47.6	45.1	54.5	50.2	48.4	43.5	42.4
5	① + ③ + ④	**51.5**	52.0	56.0	53.1	47.6	47.1	48.3

Table S6. **Ablations on the multiple-choice version of SQA3D** [17], where we leverage Qwen2.5-72B-Instruct [27] to generate candidate options.

	Method	Avg.	*Count*	*A.Dis.*	*Object*	*Room*	*R.Dis.*	*R.Dir.*	*Route*	*Order*
1	Video-3D-LLM	27.0	36.4	9.1	25.1	10.2	43.2	44.6	29.7	16.9
2	① + vanilla	27.9	58.0	11.5	29.8	13.9	34.8	27.8	38.6	8.6
3	① + cross-view	30.6	57.5	24.1	22.8	17.2	41.6	34.9	30.8	15.7
4	① + global-view	31.1	60.7	19.1	20.1	15.5	45.3	44.2	25.1	19.1
5	① + ③ + ④	**34.7**	65.6	24.4	32.5	15.8	46.7	43.2	29.7	19.2

Table S7. **Ablations on VSI-Bench** [28] on the ScanNet [8] subset, where depth images and camera poses are incorporated.

Figure S1. **Qualitative comparison with Video-3D-LLM** [33]. Thanks to the proposed two 3D-aware visual pretext tasks, **ROSS3D** has a stronger ability to interpret the overall 3D scene.

SQA3D-MCQ. In addition to conventional LM metrics, we introduce a more precise evaluation based on *LLM-generated multiple-choice QA* for SQA3D [17]. Under this new evaluation protocol demonstrated in Table S6, our results consistently demonstrate that 3D-aware visual pretext tasks are crucial.

VSI-Bench. Furthermore, in Table S7, we evaluate on VSI-Bench [28] on the ScanNet [8] subset, where depth images and camera poses are incorporated. The proposed two 3D-aware visual pre-text tasks are also effective on this advanced benchmark.

B.2. Qualitative Results

Qualitative Comparison with Video-3D-LLM [33]. We provided qualitative comparisons in Figure S1, where **ROSS3D** has a stronger ability to interpret the overall 3D scene thanks to the proposed two 3D-aware pretext tasks.

Failure Cases. We incorporate some failure cases on SQA3D [17] in Figure S2. **(1)** *Mismatched perspectives (left):* The user describes the clothing rack as "behind me" but the video shows it in front of the table. **(2)** *Subtle linguistic cues (right):* "Twiddling my thumbs together of boredom" implies there is no computer in front of the user.

General Video Understanding.

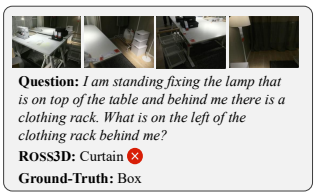

Figure S2. **Failure cases on SQA3D** [33]. It struggles with mismatched perspectives (left) and subtle linguistic cues (right).

We evaluate **ROSS3D** on Video-MME [10] using VLMEvalKit [9], *without* depth images and camera poses as inputs, where **ROSS3D** surpasses GPT4Scene [19] and Video-3D-LLM [33].

Method	Video-MME
GPT4Scene$_{64f}$	58.4
Video-3D-LLM$_{64f}$	60.1
ROSS3D$_{64f}$	**60.7**

B.3. Full Comparison

We present full comparisons with previous approaches with the complete metrics for all benchmarks. Specifically, we provide Table S8 for SQA3D [17], Table S9 for ScanQA [1], Table S10 for ScanRefer [3], and Table S11 for Multi3DRefer [30], respectively. Our **ROSS3D** significantly outperforms across all benchmarks, highlighting the effectiveness of 3D-aware visual supervision for 3D LMMs.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, 2022. 1, 3, 4

[2] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16464–16473, 2022. 5

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, pages 202–221. Springer, 2020. 1, 3, 5

[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)*, 35: 20522–20535, 2022. 5

[5] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao 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 (CVPR)*, pages 26428–26438, 2024. 4

[6] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024. 4, 5

[7] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2021. 1

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 2

[9] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 3

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118, 2025. 3

[11] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 4

[12] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 20482–20494, 2023. 4, 5

[13] 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. *arXiv preprint arXiv:2311.12871*, 2023. 1, 4

[14] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15524–15533, 2022. 1, 5

[15] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10984–10994, 2023. 4, 5

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023. 1

[17] 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, 2, 3, 4

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

Method	Question Type						Avg. (EM)	EM-R
	What	Is	How	Can	Which	Others		
Expert Models								
SQA3D [17]	31.6	63.8	46.0	69.5	43.9	45.3	46.6	–
3D-VisTA [35]	34.8	63.3	45.4	69.8	47.2	48.1	48.5	–
2D LLMs								
InternVL2-8B [21]	30.5	53.8	5.5	47.3	25.8	36.3	33.0	45.3
Qwen2-VL-7B [24]	29.0	59.2	33.4	50.5	44.2	43.2	40.7	46.7
LLaVA-Video-7B [31]	42.7	56.3	47.5	55.3	50.1	47.2	48.5	–
3D LMMs								
LEO [13]	–	–	–	–	–	–	50.0	52.4
Scene-LLM [11]	40.9	69.1	45.0	**70.8**	47.2	52.3	54.2	–
ChatScene [29]	45.4	67.0	52.0	69.5	49.9	55.0	54.6	57.5
LLaVA-3D [34]	–	–	–	–	–	–	55.6	–
Video-3D-LLM [33]	51.1	72.4	55.5	69.8	51.3	56.0	58.6	–
GPT4Scene-HDM‡ [19]	55.9	69.9	50.8	68.7	53.3	60.4	59.4	62.4
ROSS3D	**56.0**	**79.8**	**60.6**	70.4	**55.3**	**60.1**	**63.0**	**65.7**

Table S8. **Full comparison of 3D question answering** on SQA3D [17] test set. "‡" indicates this result is achieved by adopting a larger input resolution (512×490) and incorporating extra BEV inputs.

Method		BLEU-n Metrics				Language Generation Metrics		
	EM	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
Expert Models								
ScanQA [1]	21.1	30.2	20.4	15.1	10.1	33.3	13.1	64.9
3D-VLP [15]	21.7	30.5	21.3	16.7	11.2	34.5	13.5	67.0
3D-VisTA [35]	–	–	–	–	13.9	35.7	–	–
2D LLMs								
InternVL2-8B [21]	16.9	20.0	9.8	5.2	2.7	32.6	14.5	55.3
Qwen2-VL-7B [24]	19.0	27.8	13.6	6.3	3.0	34.2	11.4	53.9
LLaVA-Video-7B [31]	–	39.7	26.6	9.3	3.1	44.6	17.7	88.7
3D LMMs								
3D-LLM [12]	20.5	39.3	25.2	18.4	12.0	35.7	14.5	69.4
Chat-3D [26]	–	29.1	–	–	6.4	28.5	11.9	53.2
LL3DA [5]	–	–	–	–	13.5	37.3	15.9	76.8
LEO [13]	24.5	–	–	–	11.5	39.3	16.2	80.0
Scene-LLM [11]	27.2	43.6	26.8	19.1	12.0	40.0	16.6	80.0
ChatScene [29]	21.6	43.2	29.1	20.6	14.3	41.6	18.0	87.7
Grounded 3D-LLM [6]	–	–	–	–	13.4	–	–	72.7
LLaVA-3D [34]	27.0	–	–	–	14.5	50.1	20.7	91.7
Video-3D-LLM [33]	30.1	47.1	31.7	22.8	16.2	49.0	19.8	102.1
GPT4Scene-HDM‡ [19]	28.2	44.4	30.3	22.3	15.5	46.5	18.9	96.3
ROSS3D	**30.8**	**49.2**	**33.7**	**24.9**	**17.9**	**50.7**	**20.9**	**107.0**

Table S9. **Full comparison of 3D question answering** on ScanQA [17] validation set. "‡" indicates this result is achieved by adopting a larger input resolution (512×490) and incorporating extra BEV inputs.

[19] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025. 3, 4, 5

[20] 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)*, pages 8748–8763. PmLR, 2021. 1

[21] OpenGVLab Team. InternVL2: Better than the Best—Expanding Performance Boundaries of Open-Source

Method	Unique		Multiple		Overall	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
Expert Models						
ScanRefer [3]	76.3	53.5	32.7	21.1	41.2	27.4
3D-VLP [15]	84.2	64.6	43.5	33.4	51.4	39.5
3D-VisTA [35]	81.6	75.1	43.7	39.1	50.6	45.8
MVT [14]	77.7	66.5	31.9	25.3	40.8	33.3
3DVG-Trans [32]	81.9	60.6	39.3	28.4	47.6	34.7
ViL3DRel [4]	81.6	68.6	40.3	30.7	47.9	37.7
3DJCG [2]	83.4	64.3	41.4	30.8	49.6	37.3
M3DRef-CLIP [30]	85.3	77.2	43.8	36.8	51.9	44.7
3D LMMs						
3D-LLM [12]	–	–	–	–	30.3	–
Grounded 3D-LLM [6]	–	–	–	–	47.9	44.1
LLaVA-3D [34]	–	–	–	–	54.1	42.2
ChatScene [29]	**89.6**	**82.5**	47.8	42.9	55.5	50.2
Video-3D-LLM [33]	88.0	78.3	50.9	45.3	58.1	51.7
GPT4Scene-HDM‡ [19]	90.3	83.7	56.4	50.9	62.6	57.0
ROSS3D	87.2	77.4	**54.8**	**48.9**	**61.1**	**54.4**

Table S10. **Full comparison of 3D visual grouding** on ScanRefer [3] validation set. "‡" indicates this result is achieved by adopting a larger input resolution (512×490) and incorporating extra BEV inputs. "Unique" and "Multiple" depend on whether there are other objects of the same class as the target object.

Method	ZT w/o D	ZT w/ D	ST w/o D		ST w/ D		MT		ALL	
	F1	F1	F1@0.25	F1@0.5	F1@0.25	F1@0.5	F1@0.25	F1@0.5	F1@0.25	F1@0.5
Expert Models										
3DVG-Trans [32]	87.1	45.8	–	27.5	–	16.7	–	26.5	–	25.5
M3DRef-CLIP [30]	81.8	39.4	53.5	47.8	34.6	30.6	43.6	37.9	42.8	38.4
3DJCG [2]	94.1	66.9	–	26.0	–	16.7	–	26.2	–	26.6
3D LMMs										
ChatScene [29]	90.3	62.6	**82.9**	**75.9**	49.1	44.5	**45.7**	**41.1**	57.1	52.4
Video-3D-LLM [33]	**94.7**	**78.5**	82.6	73.4	52.1	47.2	40.8	35.7	58.0	52.7
GPT4Scene-HDM‡ [19]	97.4	84.4	85.0	77.7	59.9	55.1	48.6	44.6	64.5	59.8
ROSS3D	93.6	77.8	80.2	72.1	**54.7**	**49.6**	44.3	39.1	**59.6**	**54.3**

Table S11. **Full comparison of 3D visual grouding** on Multi3DRefer [30] validation set. "‡" indicates this result is achieved by adopting a larger input resolution (512×490) and incorporating extra BEV inputs. "ZT" means zero-target. "ST" denotes single-target and "MT" is multi-target. "D" indicates distractor.

Multimodal Models with the Progressive Scaling Strategy, 2024. 4

[22] Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Liwei Wu, Yuxi Wang, and Zhaoxiang Zhang. Pulling target to source: A new perspective on domain adaptive semantic segmentation. *International Journal of Computer Vision*, pages 1–24, 2024. 1

[23] Haochen Wang, Yuchao Wang, Yujun Shen, Junsong Fan, Yuxi Wang, and Zhaoxiang Zhang. Using unreliable pseudo-labels for label-efficient semantic segmentation. *International Journal of Computer Vision (IJCV)*, pages 1–23, 2024. 1

[24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4

[25] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4248–4257, 2022. 1

[26] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023. 1, 4

[27] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo

Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 2

[28] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024. 2

[29] Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15459–15469, 2024. 4, 5

[30] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15225–15236, 2023. 1, 3, 5

[31] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 4

[32] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2928–2937, 2021. 5

[33] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. *arXiv preprint arXiv:2412.00493*, 2024. 1, 2, 3, 4, 5

[34] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 4, 5

[35] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2911–2921, 2023. 1, 4, 5