LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D Capabilities
Supplementary Materials

Chenming Zhu[1,2] Tai Wang[2,†] Wenwei Zhang[2] Jiangmiao Pang[2] Xihui Liu[1,†]
[1]The University of Hong Kong [2]Shanghai AI Laboratory
[†] corresponding author

1. Implementation Details

LLaVA-3D is built upon the LLaVA-Video-7B [9], utilizing their pre-trained weights from the HuggingFace library, and follows a two-stage training process. Each subsequent stage builds upon the weights learned in the previous stage. The number of views V is set to 32. When adapting our method to LLaVA-1.5 [5], due to the LLM context length limitation, we use the voxelization pooling to compress the 3D patch token numbers, and the maximum number of 3D patch tokens after 3D pooling is set to 3096. Our grounding decoder consists of $L = 4$ decoder layers, as illustrated in Fig. 1. For query initialization, we employ farthest point sampling to select $N = 512$ instance queries from the 3D patches. All experiments are conducted on 16 × 80G A100 GPUs.

Settings of Stage 1. We use the Adam optimizer to train our model for one epoch with a total batch size of 16 and a warmup ratio of 0.03. During the warmup phase, the learning rates peak at 1e-5 for the LLM, 3D position encoding layer and grounding decoder, and 2e-6 for the vision encoder. The training objectives consist of the auto-regressive language modeling loss and the grounding decoder training loss.

Settings of Stage 2. In stage 2, we freeze all the components except for the grounding decoder. The model undergoes 40 training epochs on 16 A100 GPUs with a peak learning rate of 1e-4.

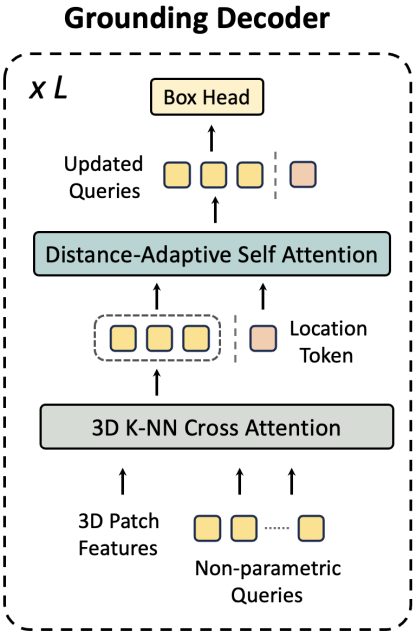


Figure 1. **Grounding Decoder Architecture.**

2. More Training Details

2.1. Training Convergence Speed

To further validate the effectiveness of 2D LMM-based Architecture and ensure fairness as much as possible, we choose LLaVA-1.5 as the base model and replace the LLaVA-3D-Instruct-86K dataset in stage 1 with the MMScan QA [6] training data. We record and evaluate the performance of LLaVA-3D under different training data ratios. Besides, we further fine-tune LEO [2] on full MMScan QA training data based on the officially released model checkpoint. Both models utilize Vicuna-7B as the LLM, ensuring comparable parameter counts. As illustrated in Fig. 2, LLaVA-3D surpasses LEO's full-step performance even when trained on less than 300 steps, indicating better data efficiency and 3.5x faster training convergence speed.

2.2. Training Objective

After matching the instance queries with ground truth 3D bounding boxes, for each match between a proposal and a ground truth object, we compute the DIOU loss [1] between predicted and ground truth boxes. We utilize InfoNCE loss [7] to optimize the similarity between the matched queries and the location token.

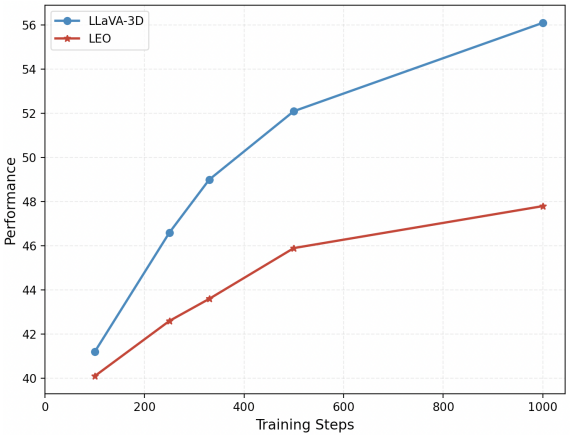


Figure 2. **Training convergence comparison.** LLaVA-3D achieves higher data efficiency and faster convergence speed during the instruction tuning stage compared with existing 3D LMM: LEO.

Table 1. **Comparison on different pooling strategies**.

Pooling Strategy	Voxel Size	Token Number	ScanQA	SQA3D
Voxelization	0.4	Dynamic	24.1	53.2
Voxelization	0.3	Dynamic	25.9	54.8
Voxelization	0.2	Dynamic	27.0	55.6
FPS	-	576	25.7	54.9
FPS	-	1024	26.3	55.2

3. More Components Analysis

To better understand the impact of different components and the generalizability of our LLaVA-3D, we conduct a thorough ablation study on the ScanQA and SQA3D benchmarks based on LLaVA-1.5 [5].

Impact of Pooling Strategy. Here we conduct various experiments to evaluate the effects of the different pooling strategies. For voxelization pooling, we adopt the simple voxelization approach from ODIN [3]. As shown in Tab. 1, the voxelization pooling strategy outperforms the FPS pooling method on 3D QA benchmarks. Model performance can be improved by either decreasing voxel size in voxelization pooling or increasing the number of 3D patch tokens in FPS pooling.

Multi-View Images Sampling Strategy. To balance computational efficiency with visual coverage, we sample V views from the egocentric images of each 3D scene. We investigate two sampling strategies during inference: *Uniform Sampling*, which evenly samples images across the scene, and *Text-Guided Sampling*, which selects frames based on CLIP image-text similarity scores to the input instruction. Since our experiments show a similar performance, we adopt uniform sampling for its simplicity.

Number of Views. An intuitive assumption is that sampling more views from the 3D scene will preserve more information about the 3D scene. We conduct a comparative experiment varying the number of views sampled from 3D scenes. Tab. 4 presents the Exact Match (EM) scores on ScanQA and SQA3D across different settings, revealing that the increase in EM score is marginal as the number of views increases. Additionally, the experimental results indicate that exceeding a certain number of views can degrade the model's performance.

Ablations on Multi-scale 3D k-NN Attention. Here we conduct further analysis and compare it with other cross-attention configurations: 1) *Full cross-attention:* instance queries attend to all 3D patches in each decoder layer, 2) *Single-scale 3D k-NN attention:* instance queries attend only to the features of the same k nearest 3D patch neighbors in each layer. Experimental results in Tab. 2 demonstrate that multi-scale 3D k-NN attention achieves the best performance on 3D VG benchmarks.

Table 2. **Ablations on multi-scale 3D k-NN attention.**

Setting	ScanRefer	Multi3DRefer
full cross-attention	41.9	39.7
$k = 16$	46.1	45.7
$k = 32$	46.3	45.2
$k = 64$	46.9	45.8
$k = 128$	44.3	43.2
$k = \{16, 32, 64, 128\}$	50.1	49.8

Ablations on Different Visual Encoders. Here we replace the CLIP with DINOv2, and report the performance on various 3D tasks. Experiment results in Tab. 3 show that 1) ***3D positional embeddings continue to provide improvements when replacing CLIP with DINOv2.*** 2) Replacing CLIP with DINOv2 improves 3D visual grounding performance that requires precise 3D perception. However, it leads to a notable decrease in 3D question answering and 3D dense captioning performance, likely due to the lack of language alignment in DINOv2's features.

Table 3. **Ablations on different visual encoders.**

Patch Type	Visual Encoder	ScanQA	SQA3D	Scan2Cap	ScanRefer
2D	CLIP	29.4	59.8	29.7	47.7
3D	CLIP	29.8 (+0.4)	60.1 (+0.3)	84.1 (+54.4)	50.1(+2.4)
2D	DINOv2	26.3	55.4	28.6	50.1
3D	DINOv2	26.9 (+0.6)	55.9 (+0.5)	80.7 (+52.1)	53.4(+3.3)

4. More Qualitative Results

3D Scene Understanding. We evaluate LLaVA-3D on various 3D scene understanding tasks and display more visualization results from Fig. 3 to Fig. 5. These examples demonstrate LLaVA-3D's robust 3D understanding abilities: comprehensive 3D scene understanding, accurate object recognition, and precise object localization in the 3D world.

Table 4. **Comparison on performance on 3D QA tasks under different number of multi-view images**.

Number of Views	Number of Tokens	ScanQA	SQA3D
16	9216	26.2	55.1
20	11520	27.0	55.6
24	13824	27.0	55.4
40	23040	26.7	55.2

Besides, our model enables the users to more easily interact with the 3D scene through the 2D images.

5. Video Demo Comparision

To enhance real-world applicability, we design our framework to process 2D videos - a widely accessible data format that users can capture with standard mobile devices. Our pipeline processes these inputs by uniformly sampling 32 frames and leveraging DUST3R [8], an efficient offline MVS method, to obtain depth maps, camera parameters, and poses. Notably, DUST3R completes this process within one minute, enabling seamless conversion of conventional video inputs into our model's required format. To validate our approach, we conduct comprehensive evaluations against LLaVA-OneVision 72B [4] in Fig. 6, a state-of-the-art multimodal model that demonstrates strong capabilities across diverse 2D scenarios, including single-image understanding, multi-view reasoning, and video understanding. The qualitative results reveal that our method achieves superior performance in 3D spatial reasoning and relationship understanding between objects with significantly fewer parameters (7B), highlighting the effectiveness of our 3D-aware architecture.

2D Click-based 3D Dense Captioning


User The related object is located at [Click]. Describe the object in detail and localize the object.

LLaVA-3D This is a single pillow placed on the sofa. The pillow is rectangular in shape, with a white background and black and white zigzag pattern. It is made of soft fabric, likely cotton or a similar material. The pillow is of medium size and is in good condition, showing no obvious signs of wear. It stands upright on the sofa, leaning against the backrest. The design of the pillow features a zigzag pattern, adding a touch of modernity to the overall look of the sofa. The placement of the pillow on the sofa indicates that it is used for decoration rather than providing support or comfort.

Figure 3. LLaVA-3D could perform 2D Click-based 3D dense captioning, generating the corresponding object caption and 3D bounding box.

2D Click-based 3D Question Answering

User The related objects are located at [Click1] [Click2]. Are these two objects similar in size?

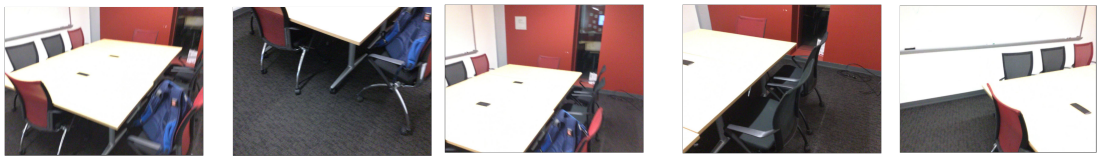
LLaVA-3D No.

User The related object is located at [Click]. Looking at the front of the object, select the cabinet that is to the left of it. What is the state of the cabinet?

LLaVA-3D The cabinet is full of ovens.

Figure 4. LLaVA-3D could perform 2D Click-based 3D question answering, now users could click on the 2D images and ask the question.

3D Visual Grounding



User "This object is designed to display information or create visual aids, focusing on a wooden material and a rectangular shape." Which object best matches the given description? Please provide its coordinates.

LLaVA-3D [-1.68, -0.025, 1.47, 0.23, 2.92, 1.13]

User " This is a padded chair with no arms and is checkerboard color blue and light blue or white. it belongs to the second table from the front of the class on the side with the windows and is the second chair closest to the middle window." Which object best matches the given description? Please provide its coordinates.

LLaVA-3D [-0.37, 0.76, 1.02, 0.94, 0.38, 2.11]

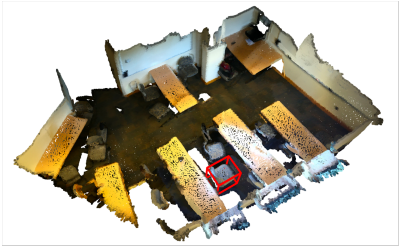


Figure 5. LLaVA-3D exhibits powerful 3D visual grounding capability, enabling accurate 3D bounding boxes output.

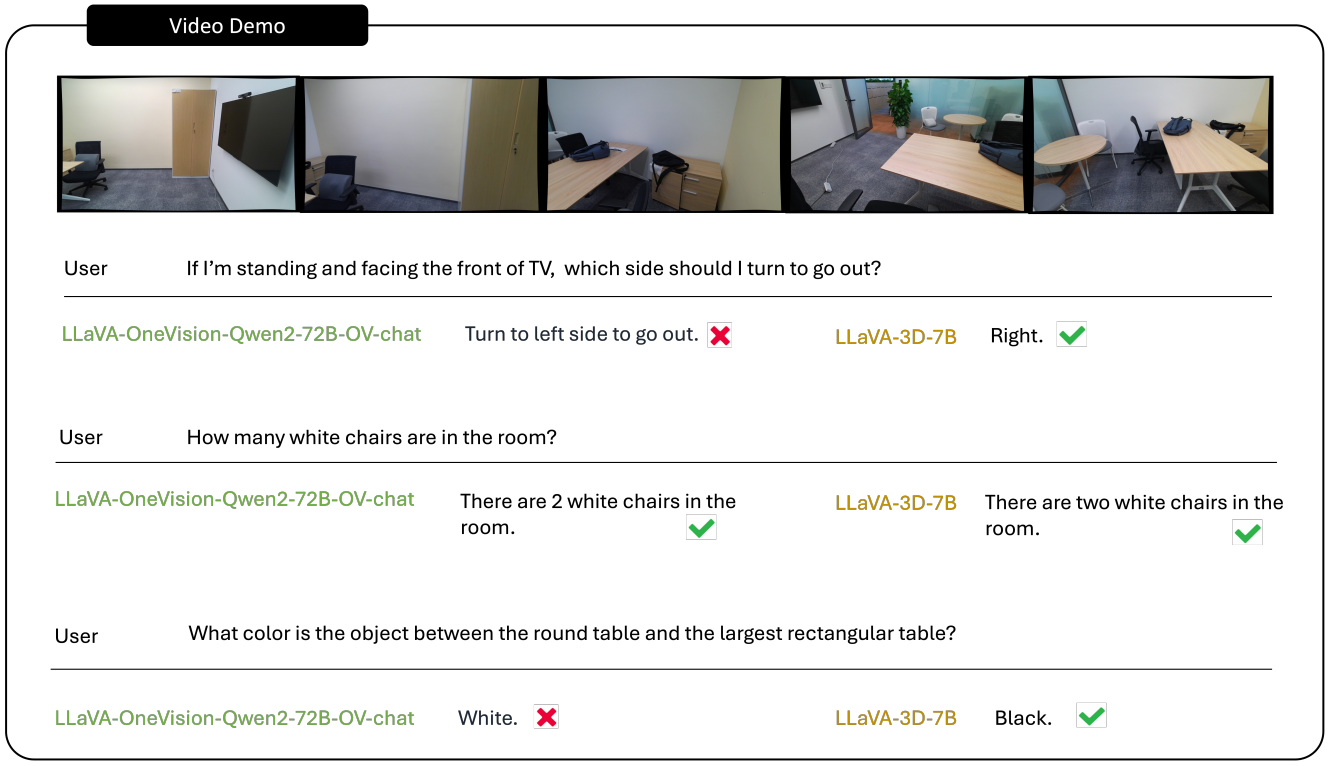


Figure 6. LLaVA-3D achieves superior performance in 3D spatial reasoning and relationship understanding between objects with significantly fewer parameters compared with powerful LLaVA-OneVision 72B.

References

[1] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18109–18119, 2023. 1

[2] 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

[3] Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W Harley, Gabriel Sarch, Kriti Aggarwal, Vishrav Chaudhary, and Katerina Fragkiadaki. Odin: A single model for 2d and 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3564–3574, 2024. 2

[4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2

[6] Ruiyuan Lyu, Tai Wang, Jingli Lin, Shuai Yang, Xiaohan Mao, Yilun Chen, Runsen Xu, Haifeng Huang, Chenming Zhu, Dahua Lin, et al. Mmscan: A multi-modal 3d scene dataset with hierarchical grounded language annotations. *arXiv preprint arXiv:2406.09401*, 2024. 1

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

[8] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[9] 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. 1