

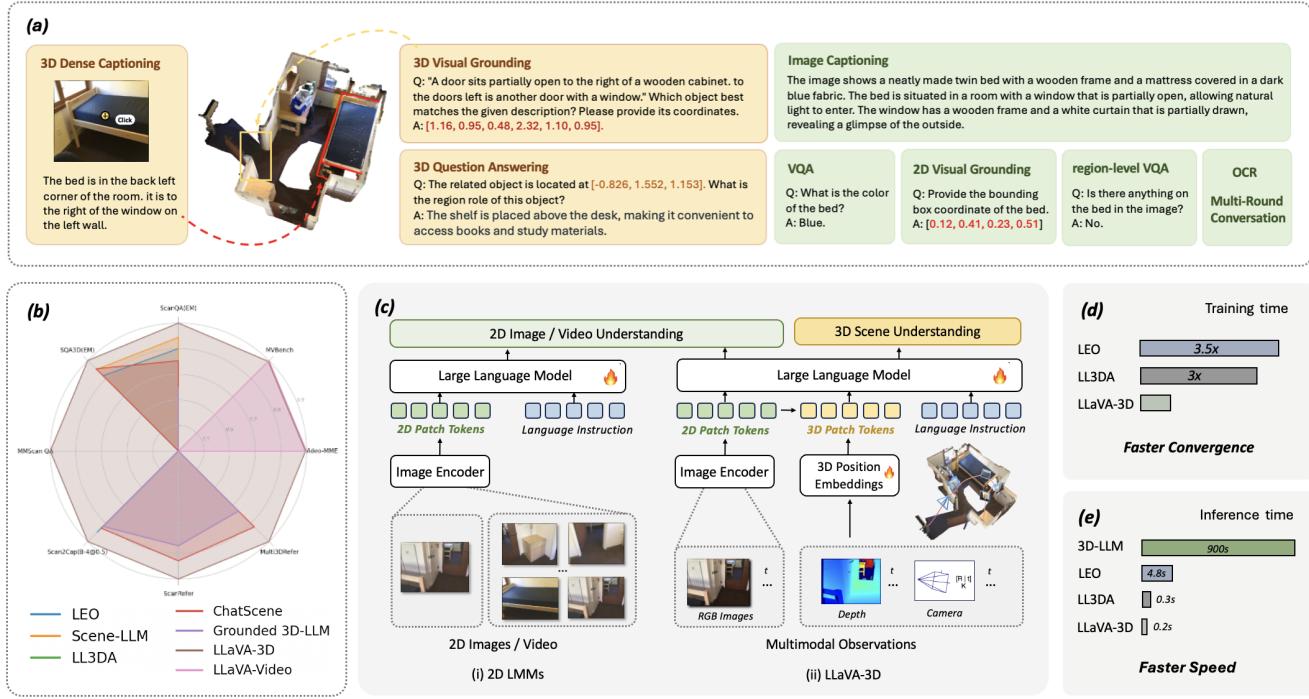
# LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D Capabilities

Chenming Zhu<sup>1,2</sup> Tai Wang<sup>2,†</sup> Wenwei Zhang<sup>2</sup> Jiangmiao Pang<sup>2</sup> Xihui Liu<sup>1,†</sup>

<sup>1</sup>The University of Hong Kong <sup>2</sup>Shanghai AI Laboratory

<https://zcmx.github.io/projects/LLaVA-3D>

† corresponding author



**Figure 1. Overview of LLaVA-3D.** The left block (b) shows that our LLaVA-3D achieves state-of-the-art performance across various 3D scene understanding benchmarks. Notably, LLaVA-3D maintains the comparable performance on 2D multimodal benchmarks compared to LLaVA-Video [53]. The middle block (c) demonstrates that LLaVA-3D leverages 3D patches to endow the 2D LMMs with 3D spatial awareness, enabling it to perform various 3D vision-and-language tasks in the physical world. The right block (d) and (e) highlight the significantly faster convergence and inference speeds of LLaVA-3D compared to existing 3D LMMs.

## Abstract

Recent advancements in Large Multimodal Models (LMMs) have greatly enhanced their proficiency in 2D visual understanding tasks, enabling them to effectively process and understand images and videos. However, the development of LMMs with 3D scene understanding capabilities has been hindered by the lack of large-scale 3D vision-language datasets and powerful 3D encoders. In this paper, we introduce a simple yet effective framework called **LLaVA-3D**.

Leveraging the strong 2D visual understanding priors from LLaVA, our LLaVA-3D efficiently adapts LLaVA for 3D scene understanding without compromising 2D understanding capabilities. To achieve this, we utilize the 3D position embeddings to enhance the 2D CLIP Patches with 3D spatial context information and construct 3D patches. By integrating the 3D position embeddings into 2D LMMs and employing joint 2D and 3D vision-language instruction tuning, we establish a unified architecture for both 2D visual understanding and 3D scene understanding. In contrast to previous 3D

*LMMs, LLaVA-3D supports decoding accurate 3D spatial perception outputs, e.g., 3D bounding boxes, directly from these 3D patches, without relying on the time-consuming off-the-shelf 3D segmentors. Experimental results show that LLaVA-3D converges 3.5× faster than existing 3D LMMs when trained on 3D vision-language datasets. Moreover, LLaVA-3D not only achieves state-of-the-art performance across various 3D tasks but also maintains comparable 2D visual understanding and vision-language conversation capabilities with LLaVA.*

## 1. Introduction

Recent advancements in Large Multimodal Models (LMMs) [2, 5, 27, 45] have significantly enhanced their ability to understand and reason over visual and language inputs, leading to remarkable performance in 2D visual tasks. Despite their advanced perceptual and reasoning capabilities, LMMs are primarily confined to virtual interactions through images or video, lacking the critical ability to interact with the physical world. To enable their deployment in real-world applications and to facilitate the emergence of new capabilities through physical interactions, it is imperative to equip LMMs with 3D spatial intelligence.

A key aspect of 3D spatial intelligence is the ability to perceive and understand the 3D world. Similar to how 2D LMMs align 2D visual features with language models using large-scale 2D vision-language datasets, a common approach to developing 3D LMMs [8, 16, 18] involves integrating 3D features encoded from point clouds into Large Language Models (LLMs) and training them on 3D point cloud-language datasets. However, in contrast to the abundance of large-scale 2D datasets, 3D datasets remain relatively scarce. Meanwhile, there are no powerful pre-trained 3D point cloud encoders, akin to CLIP ViT [43] in 2D, to provide strong and generalizable 3D features to LLMs.

Since real-world embodied agents typically rely on egocentric, multi-view images as raw observations, we aim to build a 3D foundation model based on such inputs rather than 3D point clouds. There have been attempts [14, 15] to leverage the 2D foundation models, like CLIP, alongside LLMs to advance this goal. These methods resort to 2D object segmentation results [24] to extract and aggregate CLIP features from object-centric image patches, constructing pixel-aligned 3D scene features [22]. However, this pipeline is inherently complex and computationally intensive. In contrast, 2D LMMs [2, 5, 26, 27, 33, 34, 53] directly leverage CLIP’s image patch features with richer, fine-grained information for effective 2D understanding and reasoning. This naturally leads to the question: *Can we directly build a 3D LMM upon the strong 2D priors from 2D LMMs, bypassing the obstacles in 3D data scale and 3D encoders?*

In light of recent progress in 2D LMMs, we propose a

simple yet effective framework, **LLaVA-3D**, which extends the well-established LLaVA model to efficiently comprehend the 3D world while preserving its robust 2D multimodal perception and reasoning capabilities. Inspired by ODIN [21], which leverages the positional encodings for unified 2D and 3D segmentation, our LLaVA-3D encodes the 3D spatial coordinates into 3D position embeddings, and incorporate them into the 2D CLIP patches in LLaVA to construct 3D patches. These 3D patches then undergo adaptive token compression via efficient pooling strategies before LLM processing. To further adapt LLaVA for tasks involving accurate 3D information in input or output, we devise an effective 3D-aware position encoding and decoding approach based on the 3D patches, without relying on the off-the-shelf 3D segmentors used in previous works [16–18]. Fine-tuned on the existing 3D vision-language datasets, our model converges rapidly and acquires 3D spatial understanding and grounding capabilities. Furthermore, the unified model architecture allows LLaVA-3D to retain the strong 2D understanding and reasoning abilities of LLaVA through joint instruction-tuning on 2D vision-language datasets.

Although LLaVA-3D is first built upon the LLaVA-family LMMs, it is a general extension method with simple yet effective designs for equipping any 2D LMM with 3D capabilities. With preliminary attempts based on different 2D LMMs [12, 34, 53] as the foundation model, we empirically observe that LLaVA-3D can benefit from stronger perception and understanding capabilities from large-scale 2D multi-modal pretraining, especially the video-based pre-training given the intrinsic connection between video and multi-view 3D scene representations. As illustrated in Fig. 1, experimental results demonstrate that LLaVA-3D achieves state-of-the-art performance on a wide range of 3D scene understanding benchmarks [3, 10, 38–40, 56], covering tasks such as 3D dense captioning and 3D question answering. Notably, thanks to the minimal 3D designs and powerful pre-training, LLaVA-3D attains these results with significantly less training time and fewer epochs than existing 3D LMMs, without large-scale data and pretraining costs for alignment. It could also achieve promising 3D visual grounding results without relying on the time-consuming offline 3D object preprocessing [16–18]. Furthermore, LLaVA-3D maintains capabilities comparable to state-of-the-art 2D LMMs in 2D visual understanding, reasoning, and conversation through joint tuning on 2D and 3D vision-language instructions.

## 2. Related Work

**2D LMMs.** Building on the success of recent LLMs, numerous studies [2, 5, 27, 31, 33, 34] explored LMMs that can jointly process visual and linguistic information. For example, LLaVA [33, 34] aligned 2D images with language models through an image encoder and a projection layer,

while BLIP2 [27] employed a sophisticated Q-Former architecture to guide the compression of visual features using textual cues. However, most early 2D LMMs were trained on single-image datasets, limiting their ability to tackle multi-image understanding. Recently, there has been increasing interest in expanding LMMs to handle multi-image inputs, addressing the demands of real-world scenarios. For video LMMs [26, 28, 30, 50], multi-image input forms the basis for capturing temporal or action-related dynamics across sequences of video frames. On the other hand, multi-view images of the 3D scene can implicitly reveal 3D spatial relationships and other abstract relations in the environment. Recent works [32, 40] explored whether 2D LMMs [1, 46] can leverage multi-view images to perform spatial understanding. However, these methods primarily relied on implicit learning from the data, without directly modeling the 3D world. In contrast, our LLaVA-3D explicitly models the 3D world from multi-view images, enabling advanced 3D spatial understanding and grounding capabilities.

**Injecting 3D into LLMs.** As 2D LMMs achieved substantial progress in visual perception, similar efforts have been made in the 3D domain. For 3D scene-level understanding, recent works explored ways to integrate 3D inputs such as point clouds [8, 16–18] or multi-view images [14, 15, 42, 55] into LLMs to enable advanced 3D scene understanding and reasoning. An important distinction among these methods is how they construct the 3D scene representation. LL3DA [8] directly used a scene-level 3D point cloud encoder to extract the 3D scene representation. LEO [18] and Chat-Scene [17] first segmented 3D objects from the scene point cloud using the off-the-shelf 3D instance segmentation model and then independently extracted 3D object features with object-level 3D encoders to represent the 3D scene. On the other hand, starting from multi-view images, 3D-LLM [15] and Scene-LLM [14] resorted to manually crafted 2D object segmentation to extract and aggregate CLIP features from object-centric image patches, constructing pixel-aligned 3D point representation. Unlike these approaches, our LLaVA-3D directly builds on the well-trained 2D LMM with multi-view images as input. Utilizing the 3D position embeddings, it brings the 2D patches within a 3D spatial context to construct 3D Patches. This 3D scene representation enables quick adaption of LLaVA for 3D scene understanding while preserving its strong 2D image understanding ability.

**Joint Modeling of 2D and 3D.** Recent works [21, 35, 36] explored leveraging existing 2D foundation models to enhance 3D perception for detection and segmentation tasks. These methods extract 2D features from multi-view images using 2D foundation models, and then construct 3D position-aware features by incorporating 3D position embeddings for improved 3D detection and segmentation. ODIN [21] utilized the posed RGB-D images as input for 3D instance segmentation. It leverages the powerful 2D pre-trained backbone and

differentiates between 2D and 3D features by using distinct learnable position encodings, with 2D features represented by pixel coordinates and 3D features represented by 3D coordinates. This unified architecture facilitated joint training on both 2D and 3D datasets, further enhancing the performance of 3D segmentation. Our LLaVA-3D first integrates the 3D position-aware features into 2D LMMs, enabling 2D LMMs to achieve 3D understanding. Analogous to ODIN [21], this modeling approach enables joint 2D-3D training, allowing the model to process and reason about both 2D and 3D tasks in a unified framework.

### 3. Method

Previous 2D LMMs typically consist of a visual encoder to extract 2D image features, which are then aligned with the LLM via the projection layer for joint visual and language reasoning tasks. In this section, we introduce how to bridge the 2D image features within 3D spatial context to construct 3D patches (Sec. 3.1, 3.2), and then demonstrate the 3D-aware pooling strategies to compress the 3D patches (Sec. 3.2) and finally present the 3D-aware position encoding and decoding process (Sec. 3.4), as illustrated in Fig. 2.

#### 3.1. Preliminary

We choose LLaVA-Video [53] as the base model to build the 3D LMM. For each frame image, LLaVA-Video uses the pre-trained CLIP encoder to split the image  $X \in \mathbb{R}^{3 \times W \times H}$  into patches at the patch size  $P$  and extract the 2D patch features  $X' \in \mathbb{R}^{c \times w \times h}$ , where  $h = \lceil \frac{H}{P} \rceil$ ,  $w = \lceil \frac{W}{P} \rceil$ , and then align the 2D patch features  $X_v$  into with LLM space with the projection layer. For the 3D scene understanding, the multi-view image patch features  $X'_v \in \mathbb{R}^{V \times c \times w \times h}$  are sequentially sent into LLM. To empower LLaVA-Video with 3D capabilities, we incorporate the 3D position embeddings into 2D patches to obtain the 3D patches.

#### 3.2. 3D Patch

Our 3D patch representations are built upon the 2D patch features  $X'_v$  extracted from multi-view images with CLIP visual encoder to leverage the strong visual-semantic alignment. To construct the 3D patches, we inject the 3D position information into the 2D patches so that the 3D patches can explicitly model 3D spatial information while preserving the semantic information from 2D patches. As illustrated in left block of Fig. 2, given the multi-view 2D patch features  $X'_p \in \mathbb{R}^{V \times d \times w \times h}$  after the projection layer, we obtain their 3D positions  $P \in \mathbb{R}^{V \times 3 \times w \times h}$  in the 3D world, using nearest neighbor depth and known camera intrinsic and extrinsic parameters, following ODIN [21]. The 3D positions  $P$  are then encoded into 3D position embeddings  $P' \in \mathbb{R}^{V \times d \times w \times h}$  through the *3D Position Encoding Layer* which consists of a learnable two-layer MLP. The 3D position embeddings are

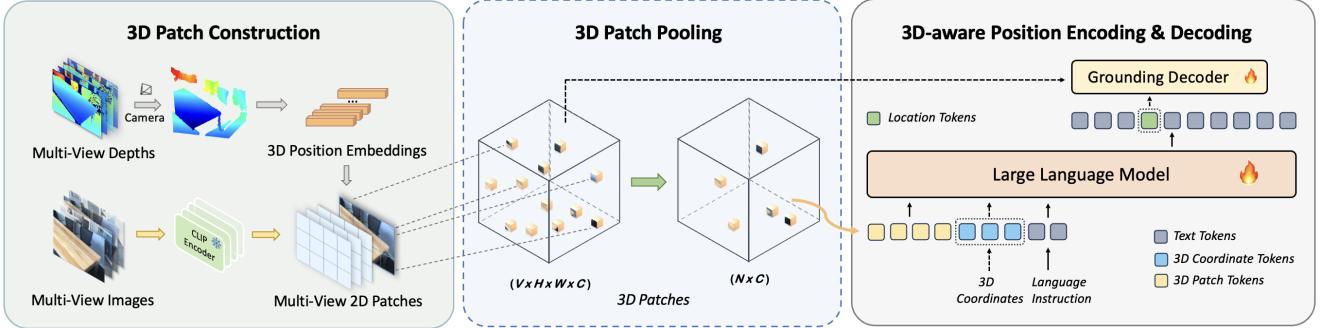


Figure 2. **LLaVA-3D Architecture.** Based on LLaVA-Video, we directly add the corresponding 3D position embeddings to 2D patch visual tokens of multi-view images to construct the 3D patches. Considering the context length support of the base model, we skip the 3D Patch Pooling stage and directly send the 3D patches into the 3D-aware Position Encoding and Decoding process to perform various 3D understanding tasks.

subsequently added to the 2D patch visual tokens, resulting in the 3D patches  $X'_{3D} \in \mathbb{R}^{V \times d \times w \times h}$ :

$$X'_{3D} = X'_p + P' \quad (1)$$

### 3.3. 3D Patch Pooling

While 3D patches enhance 2D patches with spatial information, they increase linearly with the number of input images and may exceed the context length of LLM. To address this, we introduce a 3D-aware pooling mechanism to reduce the number of 3D patches when token compression is needed, as illustrated in the middle block of Fig. 2.

In the 2D image or video domain, pooling is commonly applied along the 2D spatial or temporal dimensions to compress the number of tokens and extract essential semantic information. However, for 3D scene understanding, we pool the 3D patches based on their 3D locations to ensure these features can cover and preserve the entire scene’s structure as completely as possible. We explore two parameter-free pooling strategies to achieve this:

**Voxelization Pooling.** Voxelization discretizes the 3D space into a volumetric grid, with 3D patches undergoing average pooling within each occupied voxel, resulting in updated voxel visual tokens. Only the visual tokens from the occupied voxels are passed to the LLM, and the number of tokens varies across different 3D scenes. While the number of 3D patches scales with the number of images, the number of voxel tokens depends solely on the partitioning of the voxel grid. By adjusting the voxel size, we can effectively balance the trade-off between the number of visual tokens and the preservation of fine-grained scene features.

**FPS Pooling.** Farthest Point Sampling (FPS) is a widely used sampling strategy [20, 37] to select a representative subset of points from a larger set of points cloud. We apply FPS to sample 3D patches from multi-view images to a fixed number of tokens, ensuring that the sampled tokens represent

the entire scene structure. While fixing the number of tokens helps the LLM efficiently process visual information, it may also result in loss of scene information.

### 3.4. 3D-aware Position Encoding & Decoding

In the previous sections, we detailed the construction of the 3D scene representation from multi-view images, establishing the foundation for further interaction with the 3D scene. Building on this, the LLM could process multi-modal inputs such as the 3D scene, language instructions, and 3D coordinate cues to generate outputs such as language responses and 3D bounding boxes, as illustrated in the right block of Fig. 2. In this section, we introduce how the model is equipped to interpret accurate 3D coordinate information from inputs and subsequently output precise 3D bounding boxes when specific location-related task requirements are needed.

**Encoding of 3D Coordinate Input.** In scenarios such as 3D dense object captioning or object-centric question answering, the language instruction contains 3D coordinates. To handle such tasks, we introduce the *3D Coordinate Token* to allow the model to integrate the provided coordinates as context into its reasoning processes. Specifically, we obtain the 3D coordinate token by feeding the 3D coordinates through the 3D position encoding layer. The 3D coordinate tokens are fed into LLM together with 3D patch tokens and text tokens, enabling 3D coordinate-aware perception and reasoning.

**Decoding of 3D Bounding Box Output.** The integration of the 3D coordinate token enables the model to process 3D coordinate information from instructions effectively. However, experiments reveal that directly outputting 3D object location coordinates is non-trivial for the LLM, resulting in notably poor performance in 3D visual grounding. To handle this, previous 3D LMMs [16, 17, 55] tend to rely on the offline time-consuming 3D object extraction method and convert the task to a 3D object selection task, which is not suitable for real world application. In the section, we

introduce the efficient *Grounding Decoder*: The process begins with a set of instance queries sampled via Farthest Point Sampling from the 3D patches. The 3D patches contain rich language-aligned semantic information but lack detailed geometric structure. To handle the grounding task, the grounding decoder guides instance queries to capture the geometry awareness from the 3D patches and aggregate information from the LLM. Specifically, in each decoder layer, we conduct cross-attention between the instance queries and the 3D patch features and then concatenate the updated queries and location token for distance-adaptive self-attention [13] to capture the relative relationship. To model the local object geometry information at different scales and reduce the computational complexity, we apply *multi-scale 3D k-NN attention* with relative 3D positional embeddings when cross-attending to 3D patch features. The updated instance queries are sent to the grounding head to predict corresponding 3D bounding boxes, and the matching score is calculated based on the similarity of the queries and location token. More details are provided in our supplementary material.

## 4. Training

To achieve strong 3D understanding and grounding capabilities without compromising 2D understanding, we conduct the two-stage training strategy, inspired by [49]. The first stage equips the model with various 2D and 3D task instruction following capabilities, and the second stage further enhances the 3D grounding capability.

**Stage 1: Multi-Task Instruction Tuning.** During the instruction-tuning stage, LLaVA-3D is optimized to respond to complex 3D V&L tasks while maintaining its original 2D reasoning and instruction-following capabilities. To facilitate this capability, we fine-tune the model on the joint 2D and 3D data. Specifically, the 2D data is sampled from the LLaVA-Video training dataset, ensuring the preservation of 2D video comprehension and conversation abilities. For the 3D data, we collect the **LLaVA-3D-Instruct-86K** dataset, a hybrid collection of 3D data specifically tailored for instruction tuning covering various 3D tasks. The overall distribution of the dataset collection is shown in Fig 3. For the 3D tasks, the 3D position encoding layer will be added to jointly train with the other modules. Additionally, for tasks which require 3D bounding box outputs, the grounding decoder will be trained together. This training setup ensures that LLaVA-3D can effectively process both 2D and 3D visual tokens and is adaptive to various tasks.

**Stage 2: Decoder-only Fine-tuning.** Since the grounding decoder fails to converge within a single epoch during the first stage of training, we further train the grounding decoder for additional epochs using the 3D visual grounding data while keeping all other components frozen. During this stage, the location token is further trained with the grounding

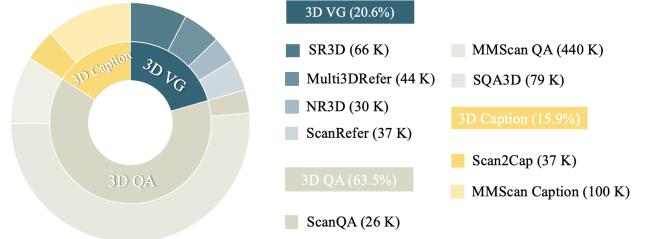


Figure 3. **LLaVA-3D-Instruct-86K.** The 3D Dataset Collection. Left: Distribution of data across categories, with the inner circle representing all categories and the outer circle illustrating data subset distribution. Right: Detailed dataset quantities.

Table 1. **Quantitative comparison with SOTA models on various 3D QA tasks.** “C” stands for “CIDEr”, “B-4” for “BLEU-4”, “M” for “METEOR”, “R” for “ROUGE”, and “EM@1” for top-1 exact match. Gray indicates evaluation results with refined exact-match protocol.

	ScanQA (val)					SQA3D (test)
	C	B-4	M	R	EM@1	EM@1
<i>Task-specific models</i>						
Scan2Cap [10]	-	-	-	-	-	41.0 <sup>†</sup>
ScanRefer+MCAN [51]	55.4	7.9	11.5	30.0	18.6	-
ClipBERT [25]	-	-	-	-	-	43.3
ScanQA [3]	64.9	10.1	13.1	33.3	21.1	47.2
3D-VisTA [57]	69.6	10.4	13.9	35.7	22.4	48.5
<i>3D LMMs</i>						
3D-LLM (FlanT5) [15]	69.4	12.0	14.5	35.7	20.5	
LL3DA [33]	76.8	13.5	15.9	37.3	-	
Chat-3D v2 [16]	87.6	14.0	-	-	-	54.7
LEO [18]	101.4	13.2	20.0	49.2	24.5	50.0
Scene-LLM [14]	80	12.0	16.6	40.0	27.2	54.2
ChatScene [17]	87.7	14.3	18.0	41.6	21.6	54.6
<i>Zero-shot 2D LMMs</i>						
VideoChat2 [29]	49.2	9.6	9.5	28.2	19.2	37.3
LLaVA-NeXT-Video [26]	46.2	9.8	9.1	27.8	18.7	34.2
LLaVA-Video [53]	88.7	-	-	-	-	48.5
GPT-4V	59.6	-	13.5	33.4	-	-
Gemini	68.3	-	11.3	35.4	-	-
Claude	57.7	-	10.0	29.3	-	-
LLaVA-3D	<b>103.1</b>	<b>16.4</b>	<b>20.8</b>	<b>49.6</b>	<b>30.6</b>	<b>60.1</b>

decoder for better 3D visual grounding performance, without influencing the strong 2D and 3D scene understanding capabilities.

## 5. Experiments

In this section, we conduct extensive evaluations to examine the capabilities of LLaVA-3D, based on LLaVA-Video. We compare our model’s 3D scene understanding (Sec. 5.1, 5.2, 5.3) and 2D video understanding (Sec. 5.4) capability with previous methods. Then we thoroughly analyze the effectiveness of the components and designs of LLaVA-3D and demonstrate the our LLaVA-3D can be integrated with any 2D LMMs to equip 3D capabilities (Sec. 5.5).

Table 2. **Quantitative comparison on MMScan QA benchmark.** “ST” stands for Single-target, “attr” for attribute, “OO” for Object-Object, and “OR” for Object Region.“S.-BERT”, “B-1”, “B-4”, “R.-L.”, “MET.” represents “Sentence-BERT”, “BLEU-1”, “BLEU-4”, “ROUGE-L”, “METEOR”, respectively. Here, we report the top-1 exact match with (the refined exact-match protocol results) for “EM@1”.

Methods	Setting	Overall	Single-target		Inter-target			Advanced	Data-driven Metrics		Traditional Metrics				
			ST-attr	ST-space	OO-attr	OO-space	OR		SimCSE	S.-BERT	B-1.	B-4.	R.-L	MET.	EM@1
3D-LLM [15]	Zero-Shot	28.6	37.8	18.8	13.7	26.3	15.4	20.8	40.4	40.3	13.4	1.5	17.3	6.0	6.2 (19.6)
Chat3D-v2 [16]		27.9	38.1	18.3	9.3	22.4	13.5	25.4	45.4	46.3	18.0	3.0	22.9	7.5	10.2 (19.6)
LL3DA [8]		15.8	15.5	14.7	14.2	25.2	4.3	6.4	40.7	43.6	5.4	2.1	16.4	4.4	8.3 (19.4)
LEO [18]		22.2	28.9	17.6	18.1	20.4	15.0	16.3	40.4	41.0	11.0	0.7	17.1	4.9	9.6 (18.7)
LL3DA [8]	Fine-tuning	38.5	40.4	46.2	14.7	47.1	26.4	7.1	65.3	67.0	26.4	8.5	44.3	14.7	30.2 (37.6)
LEO [18]		47.8	55.5	49.5	36.1	45.6	32.1	38.4	71.2	72.2	32.0	12.5	52.1	17.7	36.6 (44.5)
LLaVA-3D	Generalist	<b>55.4</b>	<b>63.2</b>	<b>57.1</b>	34.1	<b>63.2</b>	47.5	<b>44.9</b>	<b>76.2</b>	<b>78.3</b>	<b>39.2</b>	<b>13.9</b>	<b>57.5</b>	<b>20.3</b>	<b>50.1 (54.9)</b>

Table 3. **Quantitative comparison with SOTA models on OpenEQA benchmark.**

Models	Frame	Accuracy
LLaMA2 [47]	0	28.3
GPT-4 [1]	0	33.5
Claude3	20	36.3
Gemini-Pro [46]	15	44.9
GPT-4V [1]	15	54.6
GPT-4V [1]	50	55.3
Human	Full	86.8
LLaVA-3D	32	53.2

## 5.1. Evaluation on 3D Question Answering

3D Question Answering requires a model to generate responses to the natural language questions in a 3D scene. In this section, we validate LLaVA-3D performance on various 3D question answering benchmarks: ScanQA [3], SQA3D [39], MMScan QA [38], and OpenEQA [40].

**Spatial Understanding with ScanQA and SQA3D.** ScanQA and SQA3D are both built on the ScanNet dataset. The ScanQA dataset consists of 41363 questions about 800 scenes, including 32337 unique questions. SQA3D comprises 20.4k descriptions of 6.8k unique situations collected from 650 ScanNet scenes and 33.4k questions about these situations. Questions in ScanQA require basic recognition and 3D reasoning capabilities, and SQA3D further incorporates situation understanding and situated reasoning into embodied 3D scene understanding. As shown in Tab. 1, LLaVA-Video could achieve promising performance on these benchmarks in a zero-shot manner, even surpassing the task-specific methods. This phenomenon may suggest that these benchmarks do not truly assess the model’s 3D spatial understanding ability. Notably, our model could achieve the SOTA performance on these benchmarks.

### Coordinate Spatial Understanding with MMScan QA.

MMScan QA includes 5.2k scans from ScanNet, 3RScan, and Matterport3D, along with 116k training questions and 29k validation questions. These questions span existential inquiries, attribute understanding, and more advanced queries. Unlike ScanQA and SQA3D, some MMScan QA questions require 3D reasoning based on object coordinates rather than

Table 4. **Quantitative Comparisons with SOTA models for 3D Dense Captioning on Scan2Cap.** The n-gram metrics for Scan2Cap are governed by IoU@0.5.

	Scan2Cap (Val)			
	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap [10]	39.1	23.3	22.0	44.8
3D-VLP [23]	55.0	32.3	24.8	51.5
3D-VisTA [57]	61.6	34.1	26.8	55.0
Vote2Cap-DETR [7]	61.8	34.5	26.2	54.4
LL3DA [8]	65.2	36.8	26.0	55.0
LEO [18]	68.4	36.9	27.7	57.8
ChatScene [17]	77.2	36.3	28.0	58.1
LLaVA-3D	<b>84.1</b>	<b>42.6</b>	<b>29.0</b>	<b>63.4</b>

relying solely on text descriptions, demanding the model capable of understanding 3D coordinates information. We present the results under GPT-4 evaluation, data-driven metrics, and traditional metrics respectively in Tab. 2. Our LLaVA-3D achieves significantly better performance compared to LL3DA and LEO which are further fine-tuned on full MMScan QA. The results highlight the training efficiency of LLaVA-3D and its strong 3D understanding ability to serve as the generalist model.

**Embodied Question Answering with OpenEQA.** OpenEQA is the first open-vocabulary benchmark designed for spatial understanding and embodied reasoning in embodied question answering. It features an automated evaluation protocol powered by LLMs, which shows strong alignment with human judgment. Our evaluations are conducted using the EM-EQA data split of OpenEQA, which includes over 1,600 high-quality, human-generated questions from diverse real-world environments. The results in Tab. 3 demonstrate that LLaVA-3D surpasses Claude3 and Gemini-Pro, and achieves comparable performance with powerful GPT-4V on this benchmark with significantly fewer model parameters.

## 5.2. Evaluation on 3D Dense Captioning

3D dense captioning requires the model to localize all the objects in a 3D scene and then generate a descriptive sentence for each object. To evaluate our model on the 3D dense captioning tasks, we utilize the off-the-shelf 3D instance seg-

Table 5. Quantitative comparison with SOTA models on the MMScan Captioning benchmark.

model	Evaluator	Type	Color	Shape	Position	Function	Design	Overall
LL3DA [8]	GPT	10.0	26.3	40.6	38.9	67.5	21.7	33.6
LEO [18]	GPT	34.9	29.7	63.0	63.7	75.0	42.7	51.3
LLaVA-3D	GPT	<b>39.9</b>	<b>79.2</b>	<b>89.1</b>	<b>82.2</b>	<b>94.1</b>	<b>88.0</b>	<b>78.8</b>

mentation model [44] to generate object proposals. Then we further construct the 3D coordinate tokens based on the 3D object center coordinates to guide the model to handle the task. We report the performance of various methods on two 3D dense captioning benchmarks:

**Scan2Cap.** Scan2Cap requires the model to describe the object’s appearance and the spatial relations with nearby objects and output the corresponding 3D bounding box. As illustrated in Tab. 4, our method consistently outperforms the existing method on the Scan2Cap benchmark.

**MMScan Captioning.** MMScan Captioning focuses on identifying common aspects of 3D objects such as Object Type, Color, Shape, Position, Function, and Design. We benchmark various methods on MMScan Captioning benchmark in Tab. 5. The results show that our method surpasses existing approaches across all metrics by a substantial margin, especially achieving a 49.5% and 43.3% improvement in the Color score and the Design score respectively. The strong performance further demonstrates the advantages of architectures based on 2D LMMs.

Uniquely, LLaVA-3D takes multi-view images as inputs, enabling a user-friendly feature where users can simply click on the selected images to generate both 3D object captions and 3D bounding boxes, as illustrated in Fig. 4.

### 5.3. Evaluation on 3D Visual Grounding

3D visual grounding aims to localize the target object in the 3D scene using the natural language description. In this section, we initially report the performance on the ScanRefer [4] and Multi3DRefer [52] benchmarks in Tab. 6. Previous task-specific methods [20, 56, 57] typically rely on point clouds extracted from 3D reconstructed meshes as input, which are not easily accessible in real-world applications. Meanwhile, current 3D LMMs [9, 16, 17] tend to decouple the grounding task into two stages: first, extract the 3D objects from the 3D scene using the off-the-shelf 3D segmentor and then convert the task into a 3D object selection task. Without relying on the constructed point cloud, our method could directly decode the accurate 3D bounding boxes from 3D patches and achieve the SOTA performance (49.8 Acc@0.25 on Multi3DRefer) in the single-stage manner.

### 5.4. Evaluation on 2D benchmarks

Since our model is trained on the joint 2D and 3D datasets, we evaluate it on the 2D video benchmarks to ensure it retains the 2D understanding capabilities of the original

Table 6. Quantitative comparison with SOTA models on various 3D VG tasks. † represents that we apply the current two stage 3D visual grounding approach to LLaVA-3D.

	ScanRefer		Multi3DRefer	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
<i>Task-specific models</i>				
ScanRefer [4]	37.3	24.3	-	-
MVT [19]	40.8	33.3	-	-
3DVG-Trans [54]	45.9	34.5	-	-
VIL3DRel [6]	47.9	37.7	-	-
BUTD-DETR [20]	52.2	39.8	-	-
ReGround3D [56]	53.1	41.1	-	-
M3DRef-CLIP [56]	51.0	44.7	42.8	38.4
<i>Two-Stage 3D LMMs</i>				
Chat-3D v2 [16]	35.9	30.4	-	-
Grounded 3D-LLM [9]	47.9	44.1	45.2	40.6
Chat-Scena [17]	55.5	50.2	57.1	52.4
LLaVA-3D†	63.9	58.6	68.1	62.9
<i>Single-Stage 3D LMMs</i>				
3D-LLM [15]	30.3	-	-	-
LLaVA-3D	50.1	42.7	<b>49.8</b>	<b>43.6</b>

Table 7. Quantitative Comparisons on 2D video benchmarks.

Method	MVBench	VideoMME
LLaVA-Video [53]	58.6	63.3
LLaVA-3D	58.1	62.8

LLaVA-Video. As demonstrated in Tab. 7, LLaVA-3D achieves performance comparable to LLaVA-Video across various 2D video understanding and conversation benchmarks, which the current existing 3D LMMs do not possess. This performance highlights the architectural advantages of our model over other 3D LMMs.

### 5.5. More Analysis

In this section, we first delve deeper into the architectural benefits and efficacy of adapting the 2D LMM to 3D, as opposed to developing a 3D LMM solely from LLMs. Then, we analyze the effectiveness of the components and validate the generalizability of our method on various 2D LMMs.

**Developing 3D LMM from LLM or 2D LMM.** As shown in Tab. 8, (a), (b), and (c) develop 3D LMM from the LLM with different 3D scene feature and 3D-language connectors, while (d) builds 3D LMM from well-trained 2D LMM: LLaVA-1.5 [34]. Due to the context length limitation of the LLM in (b), (c), and (d), we utilize 3D voxelization pooling to compress the 3D patch tokens. The results demonstrate: 1) the Q-Former (a) and Pooling + MLP (b) share a similar performance on 3D V&L benchmarks. 2) using CLIP (c) alone instead of SAM + CLIP (b) like 3D-LLM [15] achieves better performance and significantly reduces 3D scene feature construction time from 900s to 0.2s. 3) Developing from 2D LMM (d) instead of LLM (c) could greatly improve the 3D scene understanding performance, achieving improvements of 4.1% Acc@0.25 on ScanRefer.

**Effectiveness of 3D Patch.** To further ascertain the effec-

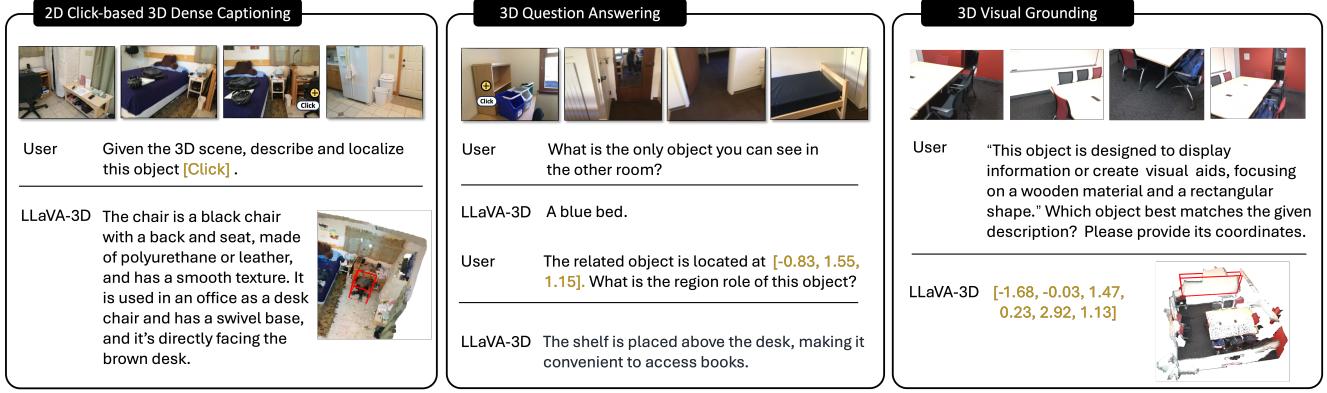


Figure 4. LLaVA-3D enables the user-friendly interaction with the 3D scene across various 3D understanding and reasoning tasks. It allows the users to just click on the 2D images or the video frame to simply conduct the interactive 3D question answering and 3D dense captioning.

Table 8. Analysis of model architecture and generalization on various 3D V&L Benchmark.

3D Feature		Connector	LLM / LMM	ScanQA	SQA3D	ScanRefer	Inference time
(a)	(SAM + CLIP) w / PE	Q-Former	Vicuna-7B	21.9	49.3	-	900s
(b)	(SAM + CLIP) w / PE	Pooling + MLP	Vicuna-7B	22.1	49.2	-	900s
(c)	CLIP w / PE	Pooling + MLP	Vicuna-7B	23.4	51.2	43.8	0.2s
(d)	CLIP w / PE	Pooling + MLP	LLaVA-1.5-7B [34]	27.0	55.6	47.9	0.2s
(e)	CLIP w / PE	MLP	InternVL2.5-7B [12]	29.1	58.8	49.3	0.3s
(f)	CLIP w / PE	MLP	LLaVA-Video-7B [53]	<b>30.6</b>	<b>60.1</b>	<b>50.1</b>	0.2s

Table 9. Effectiveness of 3D Patch Representation

Patch Type	ScanQA	SQA3D	MMScan QA	Scan2Cap
2D	29.4	59.8	42.1	29.7
3D	29.8 (+0.4)	60.1 (+0.3)	55.4 (+13.3)	84.1 (+54.4)

tiveness of our proposed 3D Patch, we conduct additional experiments across a variety of 3D question answering and 3D dense captioning benchmarks. As shown in Tab. 9, injecting 3D position information into 2D patches only brings minor improvements on ScanQA and SQA3D benchmarks. However, 3D patches prove instrumental in tasks requiring objects position information and more difficult 3D spatial understanding, yielding substantial performance gains of 13.3% on the MMScan QA benchmark and a remarkable 54.4% improvement on the Scan2Cap benchmark.

**Effectiveness of Grounding Decoder.** An ideal scenario for the 3D LMM to perform 3D visual grounding would be to directly output 3D bounding boxes in text format, just like the 2D LMM. However, while the 3D LMM can accurately output the object dimensions represented by a given 3D coordinate, it struggles to directly output the 3D position of the target object. We attempt to train the LLaVA-3D model to directly output the coordinates of the target object in the 3D visual grounding task in text format or using special tokens [15], resulting in 7.8 Acc@0.25 and 8.2 Acc@0.25 on

the ScanRefer benchmark. However, utilizing our grounding decoder could achieve a performance of 50.1 Acc@0.25.

**Generalization to various 2D LMMs.** In Tab. 8, we adapt our LLaVA-3D to other 2D LMMs, such as LLaVA-1.5 [34] (d) and InternVL2.5 [12] (e). Experiments demonstrate that our method could be a general extension to equip 2D LMMs with 3D capabilities due to the simple yet effective designs. Comparing (d) and (e), we can observe that LLaVA-3D can benefit from stronger 2D understanding capabilities of the 2D LMM base model. Besides, due to the intrinsic consistency between multi-view image 3D scene representation and video, the results in (e, f) demonstrate that our method can enjoy the significant benefits from video LMMs.

## 6. Conclusion

We propose LLaVA-3D, a simple yet effective framework built upon the well-established LLaVA. LLaVA-3D extends LLaVA’s capabilities to perform 3D scene understanding and grounding by leveraging the 3D patches and grounding decoder while efficiently preserving the 2D visual understanding and reasoning capability. Experimental results show that our method achieves state-of-the-art performance on various 3D tasks and benchmarks. We hope that our model will inspire new ideas for building 3D LMMs, and we plan to explore the application of LLaVA-3D in more downstream scenarios, such as robot manipulation and navigation.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [3](#), [6](#)
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. [2](#)
- [3] 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*, pages 19129–19139, 2022. [2](#), [5](#), [6](#)
- [4] 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*, pages 202–221. Springer, 2020. [7](#)
- [5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. [2](#)
- [6] 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*, 35:20522–20535, 2022. [7](#)
- [7] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11124–11133, 2023. [6](#)
- [8] 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*, pages 26428–26438, 2024. [2](#), [3](#), [6](#), [7](#)
- [9] 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. [7](#)
- [10] 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*, pages 3193–3203, 2021. [2](#), [5](#), [6](#)
- [11] 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. [2](#)
- [12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. [2](#), [8](#)
- [13] Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d llms with omni superpoint transformer. *arXiv preprint arXiv:2501.01163*, 2025. [5](#), [2](#)
- [14] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhao Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. [2](#), [3](#), [5](#)
- [15] 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. *arXiv preprint arXiv:2307.12981*, 2023. [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [16] Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023. [2](#), [3](#), [4](#), [5](#), [6](#)
- [17] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [3](#), [4](#), [5](#), [6](#), [7](#)
- [18] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puha Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiang Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. [2](#), [3](#), [5](#), [6](#), [7](#), [1](#)
- [19] 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*, pages 15524–15533, 2022. [7](#)
- [20] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 417–433. Springer, 2022. [4](#), [7](#)
- [21] 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](#), [3](#)
- [22] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Osama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. [2](#)
- [23] 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*, pages 10984–10994, 2023. 6
- [24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2
- [25] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021. 5
- [26] 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. 2, 3, 5
- [27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3
- [28] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3
- [29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 5
- [30] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 3
- [31] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 2
- [32] Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*, 2024. 3
- [33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 5
- [34] 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. 2, 7, 8, 1
- [35] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 3
- [36] Yingfei Liu, Junjie Yan, Fan Jia, Shuaolin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 3
- [37] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 4
- [38] 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. 2, 6, 1
- [39] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 6
- [40] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul McAvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498, 2024. 2, 3, 6
- [41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2
- [42] 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
- [43] 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*, pages 8748–8763. PMLR, 2021. 2
- [44] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 7
- [45] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023. 2
- [46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3, 6
- [47] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6
- [48] 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
- [49] Jannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhui Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multi-modal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2025. 5
- [50] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 3
- [51] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 5
- [52] 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*, pages 15225–15236, 2023. 7
- [53] 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, 2, 3, 5, 7, 8
- [54] 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*, pages 2928–2937, 2021. 7
- [55] 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. 3, 4
- [56] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Empowering 3d visual grounding with reasoning capabilities. *arXiv preprint arXiv:2407.01525*, 2024. 2, 7
- [57] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. *arXiv preprint arXiv:2308.04352*, 2023. 5, 6, 7

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

## Supplementary Material

### A. Implementation Details

LLaVA-3D is built upon the LLaVA-Video-7B [53], 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 [34], 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. All experiments are conducted on  $16 \times 80G$  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  $1e-4$ .

### B. 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 MM-Scan QA [38] training data. We record and evaluate the performance of LLaVA-3D under different training data ratios. Besides, we further fine-tune LEO [18] on full MM-Scan 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. 5, 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.

### C. More Architecture Details

In this section, we provide more details about how to connect LLM with the grounding decoder via special token embeddings, which enables the end-to-end optimization of the entire model and the grounding decoder details.

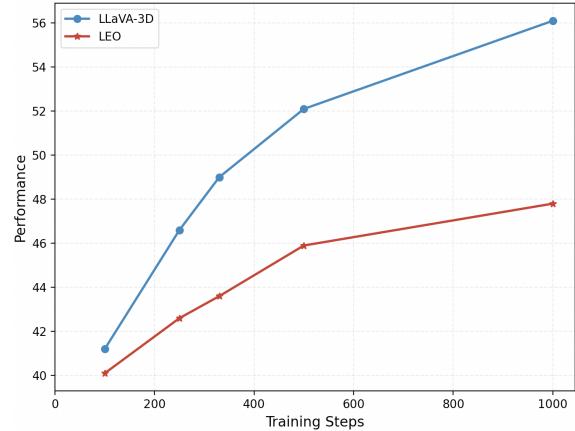


Figure 5. **Training convergence comparison.** LLaVA-3D achieves higher data efficiency and faster convergence speed during the instruction tuning stage compared with existing 3D LMM: LEO.

### C.1. Connecting with Grounding Decoder

For 3D understanding tasks like 3D visual grounding, we employ the grounding decoder to localize objects according to the user query. Specifically, we introduce a special localization token  $\langle \text{LOC} \rangle$  into the LLM vocabulary, and the LLM is trained to predict the special location token to represent the 3D bounding boxes prediction when the task necessitates 3D bounding box outputs. The last layer embedding of this location token is then sent to the 3D grounding decoder as a condition. Grounding Decoder receives both the 3D patch features and the obtained location token embeddings as inputs and predicts the 3D visual grounding results.

### C.2. Grounding Decoder Details

Here illustrate more architectural and training objective details about the grounding decoder. Our grounding decoder consists of  $L = 4$  decoder layers, as illustrated in Fig. 6. For query initialization, we employ farthest point sampling to select  $N = 512$  instance queries from the 3D patches. We initialize the value of queries to zeros and only the 3D coordinates of the sampled queries are used to set the corresponding learnable positional encoding.

**Multi-Scale 3D k-NN Cross Attention.** During the cross-attention between the queries and 3D patch features, the instance queries can only attend to the features of  $k$  nearest 3D patch neighbors to accelerate training convergence speed and reduce the memory usage. To capture the object

geometry information at different scales, the  $k$  is set to be  $\{16, 32, 64, 128\}$  in different decoder layers. Inspired by ODIN [21], we encode the relative position between the query and its neighbor into the position encoding with a learnable MLP. The position encoding is incorporated into the cross-attention computation by adding it to the instance queries and their corresponding 3D patch neighbors.

**Distance-Adaptive Self Attention.** After attending to the 3D patch features, we utilize the distance-adaptive self-attention layer [13] to model the relative spatial relationship among the queries and achieve the visual-language feature communication. Distance-adaptive self-attention introduces a bias term based on the distances between instance queries. The pairwise attention between the  $i$ -th instance query and the  $j$ -th instance query is computed as:

$$Attn(Q_i, K_j, V_j) = \text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{C}} - \sigma \cdot D\right) V_j, \quad (2)$$

where  $Q, K, V$  is the query, key, and value of the attention module,  $C$  is the channel of the embedding,  $\sigma$  is a learnable parameter based on the query, and  $D$  indicates the Euler distance between the position of these two instance queries. For the location token that does not have the 3D coordinate information, the bias term between this location token and the instance queries is set to zero.

**Box Head.** The box head consists of a two-layer MLP, which takes the updated instance queries in each decoder layer as input and predicts the corresponding 3D bounding boxes.

### C.3. 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 [11] between predicted and ground truth boxes. We utilize InfoNCE loss [41] to optimize the similarity between the matched queries and the location token.

## D. 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 [34].

**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 [21]. As shown in Tab. 10, the voxelization pooling strategy outperforms the FPS pooling method on 3D QA benchmarks. Model performance can

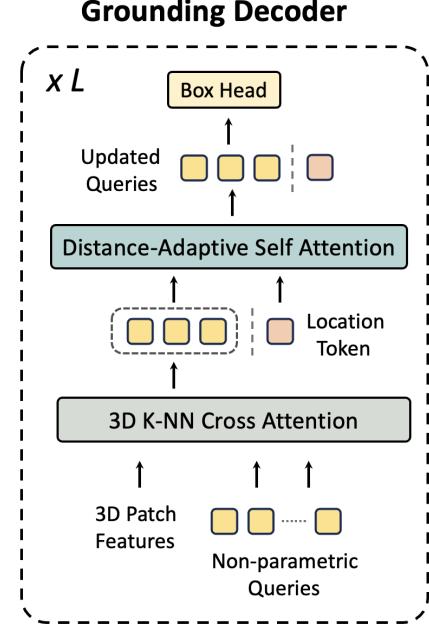


Figure 6. **Grounding Decoder Architecture.**

Table 10. **Comparsion 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

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. 11 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

Table 11. 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

model’s performance.

## E. More Qualitative Results

**3D Scene Understanding.** We evaluate LLaVA-3D on various 3D scene understanding tasks and display more visualization results from Fig. 7 to Fig. 9. 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. Besides, our model enables the users to more easily interact with the 3D scene through the 2D images.

## F. 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 [48], 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 [26] in Fig. 10, 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



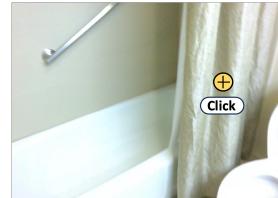
User      The related object is located at [Click]. Describe the object in detail and localize the object.

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 7. 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.

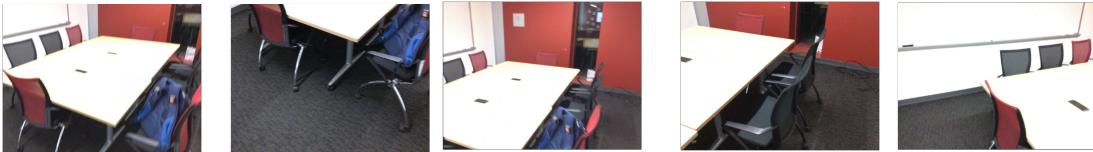


User      The related object is located at [Click]. Looking at the front of the object, select the cabinet that is to the left of it. What is the state of the cabinet?

LLaVA-3D The cabinet is full of ovens.

Figure 8. 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]



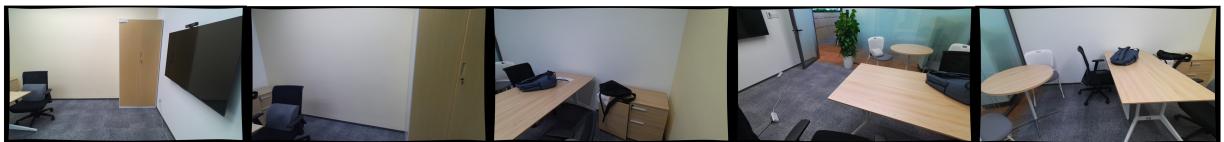
User     " This is a padded chair with no arms and is checkerboard color blue and light blue or white. it belongs to the second table from the front of the class on the side with the windows and is the second chair closest to the middle window." Which object best matches the given description? Please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

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



Figure 9. LLaVA-3D exhibits powerful 3D visual grounding capability, enabling accurate 3D bounding boxes output.

### Video Demo



User If I'm standing and facing the front of TV, which side should I turn to go out?

LLaVA-OneVision-Qwen2-72B-OV-chat Turn to left side to go out. X LLaVA-3D-7B Right. ✓

User How many white chairs are in the room?

LLaVA-OneVision-Qwen2-72B-OV-chat There are 2 white chairs in the room. ✓ LLaVA-3D-7B There are two white chairs in the room. ✓

User What color is the object between the round table and the largest rectangular table?

LLaVA-OneVision-Qwen2-72B-OV-chat White. X LLaVA-3D-7B Black. ✓

Figure 10. LLaVA-3D achieves superior performance in 3D spatial reasoning and relationship understanding between objects with significantly fewer parameters compared with powerful LLaVA-OneVision 72B.