Vi3D-LLaMA: Observe and Understand the 3D Scene with A Video Sequence

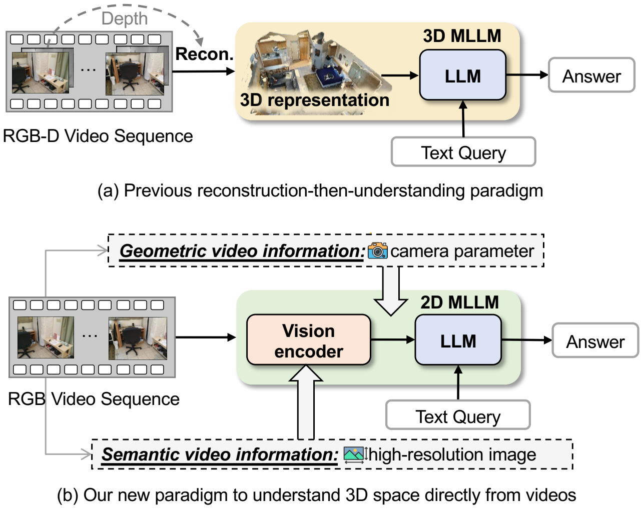
Yingjie Wang[1] , **Jiajun Deng**[2*] , **Yao Li**[1] , **Houqiang Li**[1] and **Yanyong Zhang**[1*]
[1]University of Science and Technology of China, China
[2]The University of Adelaide, Australia
{yingjiewang, zkdly}@mail.ustc.edu.cn, jiajun.deng@adelaide.edu.au, {lihq, yanyongz}@ustc.edu.cn

Abstract

Current 3D Multimodal Large Language Models (3D MLLMs) leverage explicit 3D input, *e.g.*, point clouds, to understand the 3D world and enable spatial reasoning. These explicit 3D data are usually obtained through reconstruction or additional depth sensors, affecting the model's scalability and deployment. In this work, we take a different stance and introduce **Vi3D-LLaMA**, a powerful MLLM operating without point cloud or depth data. Particularly, the proposed Vi3D-LLaMA directly performs 3D spatial reasoning with RGB video sequences. The core idea of this work is to empower the video MLLM with the capability of 3D understanding from two aspects: (1) 3D-Aware Geometric Encoding: Camera parameters and a frustum-based 3D position encoder are used to transform video representations into 3D-aware tokens, enabling implicit modeling of 3D structures with RGB frames. (2) Fine-Grained Semantic Enhancement: High-resolution (HR) images are progressively incorporated into the video representation through a lightweight HR adapter, facilitating video tokens with semantic details. We conduct extensive experiments and demonstrate that Vi3D-LLaMA, using only RGB data, can achieve comparable results with state-of-the-art 3D-MLLM-based methods. Additionally, we benchmark our method on the new VSI-Bench, showing consistent improvement over the video MLLM baseline.

1 Introduction

Recent advances in large language models (LLMs) [Achiam *et al.*, 2023; Touvron *et al.*, 2023] have established the paradigm of levering language as a universal interface for building versatile intelligent assistants. This milestone has paved the way for the emergence of Multimodal LLMs (MLLMs), which extend these capabilities to tackle a wide range of multi-modal tasks. While substantial progress has been achieved in 2D MLLMs [Liu *et al.*, 2023; Alayrac *et al.*, 2022; Li *et al.*, 2023], 3D scene understanding remains a critical frontier. Developing 3D MLLMs is essential for enabling systems to perceive and reason about the physical world in three dimensions, a capability crucial for applications such as autonomous driving [Ding *et al.*, 2024] and intelligent robotics [Li *et al.*, 2024].



(a) Previous reconstruction-then-understanding paradigm

(b) Our new paradigm to understand 3D space directly from videos

Figure 1: An illustration of the comparison between (a) previous reconstruction-then-understanding paradigm and (b) our new paradigm to understand 3D space directly from videos. Given an RGB video sequence and a text query, Vi3D-LLaMA proposes a novel paradigm for 3D scene understanding by injecting both geometry-aware and semantic video information into a 2D multimodal large language model (MLLM).

Current 3D MLLMs [Hong *et al.*, 2023; Chen *et al.*, 2024b; Zhu *et al.*, 2024a; Deng *et al.*, 2025] commonly rely on 3D geometric information, presented as point cloud or depth data, to comprehend the 3D world and perform spatial reasoning. This paradigm, referred to as "reconstruction-then-understanding" in Figure 1 (a), requires an additional depth sensor or reconstruction preprocessing to obtain explicit 3D data, which affects both the model's scalability and deployment. On the other hand, due to the abundance of image and video data on the Internet, MLLMs with RGB data have made significant advances [Zhang *et al.*, 2023; Liu *et al.*, 2023;

*Corresponding Author: Jiajun Deng and Yanyong Zhang.

Cheng *et al.*, 2024], resulting in a variety of powerful open-source RGB-based MLLMs. However, these well-trained models cannot directly operate on the point cloud data and RGB-D images, so the previous 3D MLLM paradigm cannot take advantage of internet-scale pre-trained MLLMs [Yang *et al.*, 2023; Wang *et al.*, 2024]. These two facts motivate us to explore an alternative way to design 3D MLLMs.

Particularly, in this work, we take a different stance on directly understanding 3D scenes with video sequences without 3D reconstruction and explicit depth information, as inspired by human beings are able to imagine the 3D space when watching videos. However, it is non-trivial to apply video MLLMs [Zhang *et al.*, 2023; Cheng *et al.*, 2024; Maaz *et al.*, 2024] for 3D understanding, especially for the following two challenges: (1) **Inadequate 3D geometric information.** Although the 3D structure can be reconstructed from video sequences [Schonberger and Frahm, 2016], the video MLLMs are not optimized to recover the 3D geometric information, but to capture the temporal coherence. The concurrent work, such as Video 3D-LLM [Zheng *et al.*, 2024], alleviates this issue by additionally involving depth input, leaving the method still relying on the extra depth sensor. (2) **Resolution gap.** Video MLLMs typically operate on low-resolution video frames (*e.g.*, 224 × 224) to understand actions [Ding *et al.*, 2024]. However, it is supposed that the frames have higher resolutions to keep semantic details for spatial modeling, especially for the small objects in the 3D environment. Simply increasing the input resolution can better preserve spatial information, but it significantly raises computational costs, which is not appreciated considering that the LLM already costs a large computation overhead.

To address the challenges, we introduce Vi3D-LLaMA, a novel MLLM that performs 3D scene understanding and spatial reasoning by intelligently extracting and utilizing the geometric and semantic knowledge embedded in video sequences (Figure 1 (b)). In contrast to prior approaches relying on cumbersome 3D data manipulation techniques [Hong *et al.*, 2023; Chen *et al.*, 2024b; Deng *et al.*, 2025; Huang *et al.*, 2023], Vi3D-LLaMA *bridges 2D video understanding and 3D scene comprehension by transforming 2D visual tokens into 3D-aware representations, therefore harnessing the powerful capabilities of existing 2D MLLM.*

Specifically, we propose a frame-wise 3D spatial tokenizer that generates structured scene description tokens from video input, which are optimized by integrating geometry-aware spatial modeling with fine-grained semantic feature enrichment, enabling the frozen large language model to achieve holistic 3D spatial understanding. The spatial modeling component establishes 3D geometric understanding through camera parameter-guided projection and frustum-based positional encoding, enabling precise spatial awareness without explicit depth estimation. Building upon this spatial representation, we strategically incorporate high-resolution visual features through an efficient HR adapter. The module selectively incorporates fine-grained HR features from pre-trained networks into transformer blocks, achieving detailed visual representation while maintaining computational efficiency. This dual-focus design allows our tokenizer to capture both accurate spatial structure and rich semantic details.

We conduct extensive experiments on ScanQA [Azuma *et al.*, 2022], SQA3D [Ma *et al.*, 2023], Scan2Cap [Chen *et al.*, 2021] and Nr3D [Achlioptas *et al.*, 2020] datasets to validate the capacities of Vi3D-LLaMA in understanding complex and diverse 3D environments. Vi3D-LLaMA achieves comparable performance with the state-of-the-art 3D MLLMs with point clouds or depth information. Notably, our method shows emerging spatial reasoning capability through zero-shot evaluation on the newly published visual-spatial intelligence benchmark (VSI-Bench) [Yang *et al.*, 2024].

In summary, we make three-fold contributions:

- We pioneer the design of Vi3D-LLaMA, a novel MLLM that directly leverages information embedded in video sequences to address a wide range of language-involved 3D understanding tasks.
- We develop a frame-wise 3D spatial tokenizer that produces structured scene-descriptive tokens optimized for the frozen LLM, by fusing geometry-aware spatial modeling with fine-grained semantic feature enrichment.
- Vi3D-LLaMA demonstrates superior performance across a range of tasks, outperforming existing methodologies in experiments involving the 3D Dense Captioning, 3D Question Answering and VSI tasks.

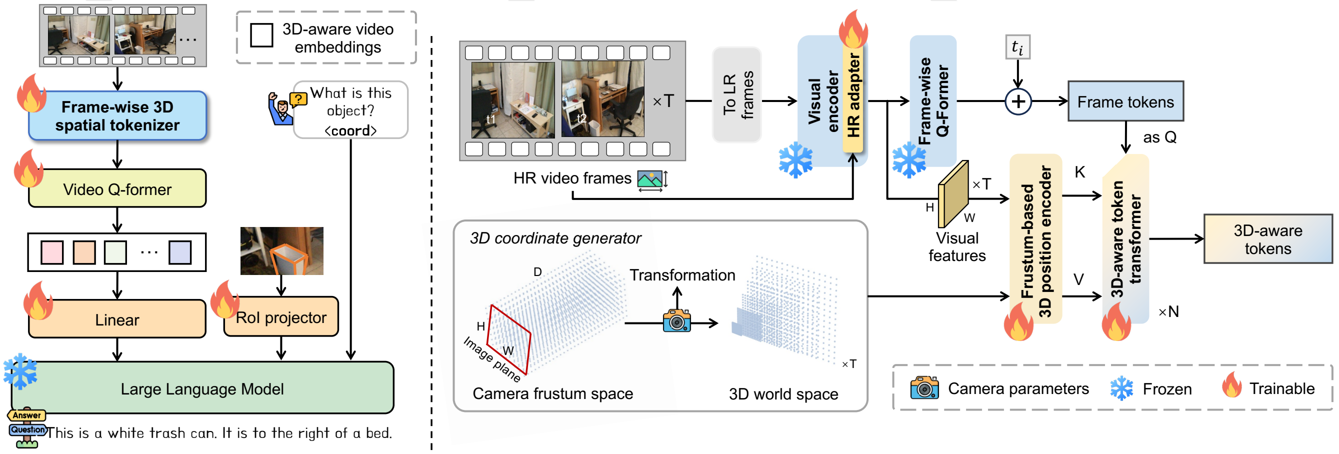
2 Related Work

2.1 3D Scene-language Understanding

3D Scene-Language Understanding has emerged as a crucial research direction, aiming to enable models to understand and reason about complex 3D environments through natural language instructions. Early approaches to 3D scene understanding leveraged explicit 3D data, such as point clouds, and combined them with large language models (LLMs) to interpret and act on spatial information [Huang *et al.*, 2023; Deng *et al.*, 2025; Chen *et al.*, 2024b]. One notable example is LL3DA [Chen *et al.*, 2024b], which directly extracts features from the 3D point cloud scene and can handle both visual prompts and textual instructions, offering diverse interaction capabilities in 3D environments. Meanwhile, some works have explored RGB-D based solutions, which use both RGB images and depth data to extract 3D spatial representations. For example, LLaVA-3D [Zhu *et al.*, 2024a] leverages 3D patches to enhance the spatial understanding of the 2D visual features by using additional depth data, bridging the gap between 2D features and 3D space. This approach allows the model to handle 3D reasoning tasks while maintaining the 2D image-based capabilities for semantic understanding. Despite their success, these 3D and RGB-D-based methods still face challenges regarding scalability and data accessibility. Depth sensors and 3D reconstruction often require specialized hardware and can be costly and time-consuming to generate. To address these issues, we directly converts video representations into 3D-aware representations, eliminating the need for explicit 3D representation.

2.2 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) [Liu *et al.*, 2023; Zhu *et al.*, 2024b; Chen *et al.*, 2024a]have gained significant attention for their ability to process and understand



(a) Vi3D-LLaMA pipeline

(b) Design of frame-wise 3D spatial tokenizer

Figure 2: **Overview of the proposed approach.** (a) The overall pipeline of our Vi3D-LLaMA, with video as input, begins with a frame-wise 3D spatial tokenizer that transforms video frame representations into 3D-aware tokens by leveraging off-the-shelf visual knowledge. These 3D-aware tokens are then processed by a Video Q-Former, which aggregates them into fixed-length 3D-aware video embeddings. Finally, these video embeddings are projected as a prefix to textual instructions and fed into a frozen large language model (LLM) for downstream tasks. (b) The detailed design of the frame-wise 3D spatial tokenizer. The video frames are input into a Vision Transformer (ViT) to extract 2D visual tokens, enhanced with fine-grained HR features via an HR-adapter. A frame-wise Q-Former and position embedding are applied to obtain frame tokens. he discretized 3D meshgrid within the frustum space is transformed into 3D coordinates using camera parameters. These 3D coordinates, along with the 2D visual features, are passed through a 3D position encoder to generate 3D-aware features. The frame tokens, treated as queries, are updated through interaction with the 3D-aware features and output as 3D-aware tokens.

both text and visual inputs, allowing models to perform a wide range of tasks such as image captioning, visual question answering, and even multimodal dialogues. Building on the success of LLMs like GPT series [Achiam *et al.*, 2023], LLaMA [Touvron *et al.*, 2023] and etc. The integration of temporal information has prompted the development of video-based MLLMs [Zhang *et al.*, 2023; Maaz *et al.*, 2024; Cheng *et al.*, 2024; Lin *et al.*, 2024], which extend the capabilities of traditional MLLMs to handle video inputs. Video-LLaMA [Zhang *et al.*, 2023] is a notable example, which extends the LLaMA architecture to process video sequences and perform video-based tasks. By leveraging large-scale video-text datasets, these models enable tasks such as video captioning and action recognition. However, despite their success, these models still face limitations when it comes to understanding the 3D structure of scenes, as they rely on 2D representations without leveraging the spatial depth inherent in 3D environments. To address these limitations, we propose a novel approach that aims to directly convert video representations into 3D-aware representations, allowing us to better leverage pre-trained RGB-based foundation models. This approach eliminates the need for depth or point cloud information, enabling MLLMs to reason about and understand 3D scenes more effectively through the use of video inputs alone.

3 Method

This section introduces Vi3D-LLaMA, a method for learning 3D-aware and semantically enriched video representations from video inputs, which empowers MLLM with reasoning and understanding capabilities in 3D space. In this section, we begin with an overview of the Vi3D-LLaMA architecture (Section 3.1) and delve into the details of its core component—the frame-wise 3D spatial tokenizer (Section 3.2). Then, in Section 3.3, we elaborate on the details of each module in our Vi3D-LLaMA pipeline (Section 3.3).

3.1 Overall Framework

As shown in Figure 2 (a), the pipeline of our Vi3D-LLaMA begins with a frame-wise 3D spatial tokenizer, which converts each RGB frame into 3D-aware tokens by modeling the 3D geometric prior and incorporating the semantic details. These tokens are then processed by a Video Q-Former. The Video Q-Former aggregates frame-wise 3D-aware tokens into fixed-number feature embeddings. After that, the produced embeddings are projected as a prefix to textual instructions and input into a frozen LLM to perform downstream tasks. The key component, the frame-wise 3D spatial tokenizer (Figure 2 (b)), operates in several stages. First, video frames are encoded into 2D visual tokens using a Vision Transformer (ViT). These tokens are further enhanced into fine-grained, high-resolution (HR) features through an HR-adapter. Frame-wise Q-Former and position embeddings are then applied to generate frame tokens. Simultaneously, the shared frustum space across views is discretized into a 3D meshgrid, and its coordinates are transformed into 3D world space via camera parameters. The 2D visual features (reshaped from the frame tokens) and the transformed 3D coordinates are processed by a 3D position encoder to produce 3D position-aware features. Finally, the frame tokens, acting as queries, interact with these 3D position-aware features via a 3D-aware token transformer, producing the final 3D-aware tokens.

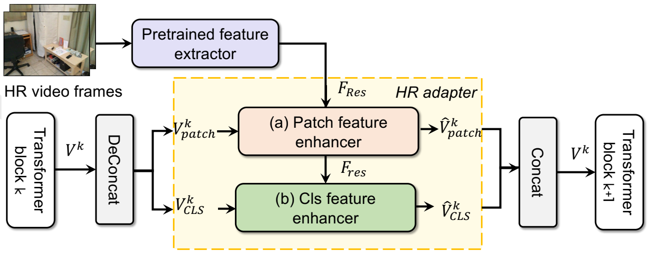


Figure 3: The HR-adapter between the transformer blocks in ViT. The feature $\mathbf{V}^k$ from the k-th transformer and the HR features F_{Res} extracted from HR frames using a pre-trained HR feature extractor (ResNet-101) are processed by the HR adapter. The adapter outputs $\hat{\mathbf{V}}^k$, which is then passed to the $(k+1)$-th transformer.

3.2 Frame-wise 3D Spatial Tokenizer

In this section, we detail the design of the core component, *i.e.*, Frame-wise 3D Spatial Tokenzier, in our Vi3D-LLaMA. This special tokenizer includes five components: a 3D Coordinates Generator, a Frustum-based 3D Position Encoder, a 3D-Aware Token Transformer, a ViT with an HR-Adapter, and a Frame-wise Q-Former.

3D coordinates generator. To establish the correspondence between 2D image views and 3D space, we start by discretizing the view frustum space into a meshgrid with dimensions (W, H, D), inspired by [Liu *et al.*, 2022; Hu *et al.*, 2019]. This meshgrid serves as a shared sampling space across multiple views, each point in the frustum space is defined as:

$$f^m_j = (u_j \cdot d_j, v_j \cdot d_j, d_j, 1)^T, \tag{1}$$

where (u_j, v_j) represents the pixel coordinates in the image plane, and d_j indicates the depth along the axis orthogonal to the image plane. We can then project each frustum point f^m_j of i-th view into 3D space as follows via the transformation matrix $K_i \in \mathbb{R}^{4\times 4}$:

$$K_i = C_i^\top \cdot C_p, \tag{2}$$

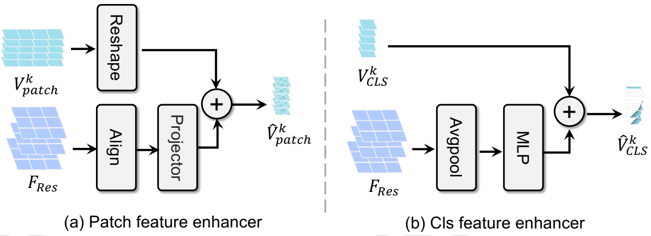
$$\mathbf{P}^{3d} = \{p^{3d}_{i,j} \mid p^{3d}_{i,j} = K_i \cdot f^m_j\}, \tag{3}$$

where $C_i^\top$ denotes the known transpose of the camera intrinsics, C_p denotes the camera poses. This process establishes a one-to-one correspondence between the 2D image data and the 3D spatial coordinates, enabling accurate representation in the 3D perception coordinate system.

Frustum-based 3D position encoder. The purpose of the 3D position encoder is to obtain 3D-aware features $\mathbf{F}^{3d}$, by associating 2D visual features $\mathbf{F}^{2d}$ with 3D position information $\mathbf{P}^{3d}$. The process can be summarized as follows:

$$\mathbf{F}^{3d} = \mathbf{F}^{2d} \diamond \psi(\mathbf{P}^{3d}), \tag{4}$$

The 2D features $\mathbf{F}^{2d} \in \mathbb{R}^{T\times W\times H\times D}$ are obtained by reshaping the output tokens of the ViT. Specifically, given the 2D features $\mathbf{F}^{2d}$ and 3D coordinates $\mathbf{P}^{3d}$, the $\mathbf{P}^{3d}$ is first fed into a multi-layer perception (MLP) network $\psi(\cdot)$ and transformed to the 3D position embedding (PE) via $PE^{3d} = \psi(\mathbf{P}^{3d})$. The 3D PE PE^{3d} is added with 2D features to obtain the key value for the following transformer decoder. The flattened 2D features are used as the value component for the transformer decoder.



(a) Patch feature enhancer

(b) Cls feature enhancer

Figure 4: Illustration of the detailed architecture for patch feature enhancer in (a) and cls feature enhancer in (b).

3D-aware token transformer. The 3D-aware token transformer is designed to convert frame representations into 3D-aware video tokens. In this process, frame tokens generated from the frame-based branch are treated as object queries. At each decoder layer, these object queries interact with the 3D position-aware features through a multi-head attention mechanism and a feed-forward network. Formally, the transformation at the l-th layer can be expressed as:

$$\mathbf{Q}^{l+1} = \text{MHA}(\mathbf{Q}^l, \mathbf{F}^{3d}, \mathbf{F}^{2d}) + \mathbf{Q}^l, \tag{5}$$

$$\mathbf{Q}^{l+1} = \text{FFN}(\mathbf{Q}^{l+1}) + \mathbf{Q}^{l+1}, \tag{6}$$

where $\mathbf{Q}^l$ represents the object queries at the l-th layer, $\mathbf{F}^{3d}$ denotes the 3D position-aware features, MHA$(\cdot)$ refers to the multi-head attention mechanism, and FFN$(\cdot)$ is the feed-forward network. This process is iteratively performed across the N transformer layers, allowing the frame tokens to fuse effectively with the 3D position-aware features. The final output is the frame-wise token that encompasses both geometric and semantic information.

ViT with HR-adapter. Given a video with L frames, each frame is processed by the ViT to produce visual features at the k-th transformer layer, denoted as $\mathbf{V}^k = \{v^k_i\}^L_{i=1}$, where $v^k_i \in \mathbb{R}^{N_v\times C_v}$. Here, N_v is the patch number and C_v is the dimension. To address the semantic ambiguity inherent in low-resolution visual features in MLLM's input, we enhance the video representations by incorporating high-resolution (HR) features extracted from the raw video frames using an HR-adapter module. As shown in Figure 3, for the k-th transformer block, we first split visual tokens $\mathbf{V}^k$ into V^k_{cls} and V^k_{patch}, which are then fed into the HR adapter module for separate processing. This module consists of two components: the *PatchFeatureEnhancer* for refining spatial details in patch tokens and the *ClsFeatureEnhancer* for enriching the global semantics of the cls token. Specifically, ViT patch features V^k_{patch} are reshaped into $(H_{\text{ViT}}, W_{\text{ViT}})$, where $N_v = H_{\text{ViT}} \times W_{\text{ViT}}$. The HR feature maps F_{Res}, extracted from T frames by pre-trained ResNet-101 [He *et al.*, 2015], are adjusted via bilinear interpolation to match the spatial dimensions. After that, we apply the projection on F_{Res} to align the channel dimensions, shown in Figure 4 (a). The resized HR features are fused with V^k_{patch} to enhance spatial details. The above process is summarized as follows:

$$\hat{V}^k_{\text{patch}} = V^k_{\text{patch}} + \text{Projector}(\text{Alignment}(F_{\text{Res}})). \tag{7}$$

Here, Alignment denotes bilinear interpolation, Projector aligns dimensions using 1×1 Conv2D, and $\hat{V}^k_{\text{patch}}$ denotes the enhanced patch tokens. For the cls token in Figure 4 (b), global semantic information is extracted from F_{Res} using global pooling, linearly projected to the same dimensionality as the ViT cls token, and fused with V^k_{cls} :

$$\hat{V}^k_{\text{cls}} = V^k_{\text{cls}} + \text{MLP}(\text{GlobalPool}(F_{\text{Res}})). \tag{8}$$

Here, GlobalPool extracts global features, MLP maps the global features to V^k_{patch}'s dimension, and $\hat{V}^k_{\text{cls}}$ represents the enhanced cls token.

Frame-wise Q-Former. For the T frames of vision tokens output by the ViT with HR-adapter, a *Frame-wise Q-Former* is utilized to aggregate vision tokens from each individual image frame. Temporal information is subsequently injected into the aggregated tokens across different frames to capture cross-frame dependencies. Finally, frame-level representations, referred to as frame tokens, are obtained. These frame tokens are then fed into the 3D-aware Token Transformer to generate the final 3D-aware frame tokens.

3.3 Details in Pipeline

Video Q-Former. The Video Q-Former, sharing the same architecture as the Q-Former in BLIP-2 [Li *et al.*, 2023], takes 3D-aware tokens generated by a frame-wise 3D spatial tokenizer as input. By introducing a fixed number of learnable query vectors (learnable queries) to extract key information from the frame features, it ultimately produces 3D-aware video embeddings. This module empowers the model to capture both spatial and semantic features across frames.

Linear and RoI projection. After obtaining the video embeddings, we apply projection layers to project the output video embeddings and the visual prompt into the same dimension as the text embeddings of LLMs, respectively. For the dense captioning task, we leverage pretrained Mask3D [Schult *et al.*, 2023] to extract the corresponding Region of Interest (ROI) features from the video frames. These ROI features provide the model with a focused view of the objects that need to be described, acting as a visual prompt. The visual prompt helps identify the specific object or region of interest within the video that requires captioning.

Large Language Model (LLM). Given the visual tokens, we leverage captions of multiple tasks to fine-tune the pre-trained video LLM to understand the 3D world space. The input to the LLM is the concatenated muli-modal tokens including visual RoI tokens and the text embeddings, tokenized from text prompts.Then, the pre-trained LLM receives the muli-modal tokens to generate language in an auto-regressive way.

4 Experiments

4.1 Datasets and Metrics

Datasets. We train our Vi3D-LLaMA on the video frames provided by the training set of ScanNet [Dai *et al.*, 2017], which includes 1,201 and 312 diverse and complex indoor 3D scenes for training and validation. The language annotations used in this study are sourced from Scan2Cap [Chen *et al.*, 2021], Nr3D [Achlioptas *et al.*, 2020], ScanQA [Azuma *et al.*, 2022], SQA3D [Ma *et al.*, 2023], and the ScanNet subset of 3D-LLM. This combination covers a variety of tasks, including instance and scene descriptions, conversations and question answering. We also evaluate our approach on a novel video-based visual-spatial intelligence benchmark (VSI-Bench) [Yang *et al.*, 2024], consisting of over 5,000 question-answer pairs, to validate the method's spatial intelligence. Note that we only utilize the Scannetv2 subset of VSI-Bench for zero-shot evaluation.

Metrics. We adopt CiDEr (C), BLEU-4 (B-4), METEOR (M), and Rouge-L (R) to evaluate the quality of the generated text response on ScanQA [Azuma *et al.*, 2022]. Different from the setting of ScanQA, there is a definite answer to situated question answering dataset SQA3D, therefore we leverage extract match accuracy (EM) as well as the refined version (EM-R) as the metric. For 3D Dense Captioning, we use the $m@k$IoU metric. Here, $m \in$ (C, B-4, M, R), and the m score of a caption is set to 0 if the IoU between the predicted box and the object is less than the given threshold k. For VSI-Bench, we use accuracy (ACC), based on exact matching (with possible fuzzy matching), as the primary metric.

4.2 Implementation Details

We develop our Vi3D-LLaMA on Video-LLaMA-7B [Zhang *et al.*, 2023] and finetune the model with task-specific data. Our proposed method is implemented in PyTorch trained using a single machine with 8 NVIDIA A100 GPUs. The input video frames are resized and cropped to the spatial size of 224 × 224 for MLLM input and 640 × 480 for HR adapter input. We uniformly sample T = 16 frames from the entire video. We sample 16 points along the depth axis following the linear-increasing discretization (LID) [Reading *et al.*, 2021]. We set the region to [-5.0m, 5.0m] for the X and Y axis, and [0m, 3.0m] for Z axis. The 3D coordinates in 3D world space are normalized to [0, 1]. We initialized the MLLM with their official pretrained weights, freezing these weights during training and only training the parameters of ST-Adapters and our additional modules (3D position encoder, 3D token transformer, and HR-adapater module). The 3D token transformer contains N = 6 layers. We use AdamW [Loshchilov and Hutter, 2019] as the optimizer and cosine annealing scheduler [Loshchilov and Hutter, 2016] as the learning rate scheduler with an initial learning rate of 1e-4. We train all models in a total of 20 epochs.

4.3 Comparison with SoTA Models

We compare the proposed Vi3D-LLaMA with other models and present the results in Table 1. The models compared in this table are divided into three groups: specialist models, generalist models with 3D input, and generalist models without 3D input. The specialist model is designed to address a single kind of task. All of the specialist models in this table are without LLMs. Generalist models with 3D input are trained to perform multiple language-related 3D tasks based on 3D representations. The last kind, generalist models without 3D input, these models take RGB inputs and leverage existing LLMs or MLLMs for further training and adaptation.

3D Question Answering is the task that asks the model to observe the visual information of the 3D scene and give a

Method	Modality	ScanQA (val)				SQA3D (test)		Scan2Cap (val)				Nr3D (val)			
		C↑	B-4↑	M↑	R↑	EM↑	EM-R↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Specialist Models:															
ScanQA[Azuma *et al.*, 2022]	PC	64.9	10.1	13.1	33.3	46.6	-	-	-	-	-				
3D-VLP[Jin *et al.*, 2023]	PC	67.0	11.2	13.5	34.5	-	-	54.9	32.3	24.8	51.5	-	-	-	-
3D-VisTA[Zhu *et al.*, 2023]	PC	69.6	10.4	13.9	45.7	48.5	-	61.6	34.1	26.8	55.0	-	-	-	-
Scan2Cap[Chen *et al.*, 2021]	PC	-	-	-	-	41.0	-	39.1	23.3	22.0	44.8	27.5	17.2	21.8	49.1
Vote2Cap-DETR [Chen *et al.*, 2023]	PC	-	-	-	-	-	-	61.8	34.5	26.2	54.4	43.8	26.7	25.4	54.4
Generalist Models with 3D Input:															
Grounded 3D-LLM [Chen *et al.*, 2024c]	PC	72.7	13.4	-	-	-	-	70.6	35.5	-	-	-	-	-	-
LL3DA* [Chen *et al.*, 2024b]	PC	76.8	13.5	15.9	37.3	-	-	65.2	36.8	26.0	55.1	**51.2**	28.8	25.9	56.6
3D-LLM[Hong *et al.*, 2023]	PC+RGB	69.4	12.0	14.5	35.7	-	-	-	-	-	-	-	-	-	-
Scene-LLM [Fu *et al.*, 2024]*	PC+RGB	80.0	12.0	16.8	40.0	54.2	-	-	-	-	-	-	-	-	-
LEO [Huang *et al.*, 2024]	PC+RGB	101.4	13.2	20.0	49.2	50.0	52.4	72.4	38.2	27.9	58.1	-	-	-	-
Video-3D LLM (MC) [Zheng *et al.*, 2024]	RGB+Depth	**100.5**	-	**29.5**	-	**57.7**	-	**80.0**	40.2	-	-	-	-	-	-
LLaVA-3D [Zhu *et al.*, 2024a]	RGB+Depth	91.7	14.5	20.7	**50.1**	55.6	-	79.2	**41.1**	30.2	**63.4**	-	-	-	-
Generalist Models w/o 3D Input:															
Uni3DR2-LLM [Chu *et al.*, 2024]	RGB	70.3	12.2	14.9	36.3	-	-	-	-	-	-	-	-	-	-
Vi3D-LLaMA (ours)	RGB	81.1	**14.7**	20.3	48.1	55.0	**55.9**	71.4	39.3	**31.6**	56.2	50.2	**29.9**	**29.8**	**59.5**

Table 1: **Performance comparison among state-of-the-art methods.** “Specialist Model” means this model can be utilized to perform 3D question answering and 3D dense captioning. “Generalist models with 3D input” indicates leveraging 3D representations to perform multiple language-related 3D tasks. We add a “*” to indicate further fine-tuned on each dataset before evaluation. ‘PC” means point cloud and “RGB” means multi-view images. “Depth” means extra depth input. Please note that LEO [Huang *et al.*, 2024]’s results on ScanQA is marked with a gray color and not compared to other methods, since it is in a different setting that accesses the ground truth object related to the question. The best result is highlighted in bold font, while the remaining top-2 entries for each metric are marked with an underline.

Method	Rel.Dist.	Rel.Dir.	Router Plan	Appr. Order
Random*	25.0	36.1	28.3	25.0
Freqency*	25.1	47.9	28.4	25.2
Video-LLaMA [Zhang *et al.*, 2023]	29.8	35.5	32.1	39.1
Vi3D-LLaMA	**33.2** (+4.4)	**40.1** (+4.6)	**33.0** (+0.9)	**44.2** (+3.1)

Table 2: **Evaluation on a subset of VSI-Bench for multiple-choice answers.** Rel.Dist. denotes relative distance, Rel.Dir. denotes relative direction, and Appr. denotes appearance. * indicates results reported in [Yang *et al.*, 2024].

precise response to the user’s question involving some part of the scene. We conduct the comparison between our Vi3D-LLaMA and other methods on both the conventional 3D question-answering dataset ScanQA [Azuma *et al.*, 2022] and the situated question-answering dataset SQA3D[Ma *et al.*, 2023]. As shown in Table 1, our method ranks among the top three in terms of CiDEr, BLEU-4, METEOR, and Rouge-L score. Remarkably, compared to LL3DA [Chen *et al.*, 2024b] which only uses point cloud as input and is finetuned, our Vi3D-LLaMA achieves **4.3%** CiDEr, **1.2%** BLEU-4, **4.4%** METEOR, and **10.8%** Rouge-L improvement. This indicates that learning 3D-aware representations from only RGB inputs can achieve or even surpass the performance of 3D inputs. When compared to the strongest competitor LLaVA-3D, which takes RGB and extra depth as input, our Vi3D-LLaMA achieves 0.2% BLEU-4 and comparable performance on other metrics on ScanQA. On SQA3D, our Vi3D-LLaMA reports comparable extract match accuracy as that of LLaVA-3D (55.0% V.S. 55.6%).

3D Dense Captioning demands the model to describe the object and its spatial relationship to the surrounding instances within the scene. In this experiment, we follow the common practice of using the predicted mask proposals of Mask3D [Schult *et al.*, 2023]. The proposal features are encoded as the visual prompt. Results in the table show that our Vi3D-LLaMA also achieves remarkable performance in generating instance-level descriptions on both the Scan2Cap and NR3D datasets. For example, for 3D input model LL3DA, our Vi3D-LLaMA achieves 6.2% CiDEr and 2.5% BLEU-4, 5.6% METEOR, and 1.1% Rouge-L improvements on Scan2Cap. For the strongest competitor, LLaVA-3D, Vi3D-LLaMA achieves a 1.4% improvement on METEOR, while performing slightly worse on other metrics, which can be attributed to the fact that we did not utilize additional depth information. This experiment further validates the effectiveness and scalability of the proposed Vi3D-LLaMA.

4.4 Analysis of Visual-Spatial Intelligence

To validate the model’s Visual-Spatial Intelligence capabilities, we conduct the zero-shot evaluation on a subset of VSI-Bench, focusing on multiple-choice answer (MCA) tasks such as Relative Distance, Relative Direction, Appearance Order, and Route Planning. As shown in the Table 2, we compare the performance of our method against ”Random Choose,” ”Frequency Choose,” and our strong baseline, Video-LLaMA. The most naive approach for selecting answers in multiple-choice questions is based on random guessing or frequency, as shown in the first two rows of the table. For Video-LLaMA, after fine-tuning on the ScanNet dataset, its overall performance surpasses the chance-level methods. Building upon this baseline, our method learns 3D-aware representations from video streams, resulting in significant improvements across all four Visual-Spatial Intelligence tasks. For example, we find that learning 3D-aware representation from video sequence improves the MLLM (Video-LLaMA)’s relative distance accuracy by 4.6%.

4.5 Ablation Study

Effect of each module in frame-wise 3D spatial tokenizer. Table 3 reports the effect of key modules in the frame-wise 3D spatial tokenizer. *Method (a)* represents the base-

	3D spatial tokenizer		ScanQA		Scan2Cap	
	Frustum-based 3D PE	HR-adapter	C↑	B-4↑	C↑	B-4↑
(a)			60.2	10.5	32.7	16.7
(b)	✓		74.2	14.5	64.8	30.7
(c)		✓	73.1	13.3	36.2	17.9
(d)	✓	✓	**81.1**	**14.7**	**67.3**	**39.3**

Table 3: **Ablation study of modules in the frame-wise 3D spatial tokenizer.** The models are compared in terms of CiDEr and BLEU-4 on ScanQA [Azuma *et al.*, 2022] and Scan2Cap [Chen *et al.*, 2021]. Our default setting is highlighted with light blue.

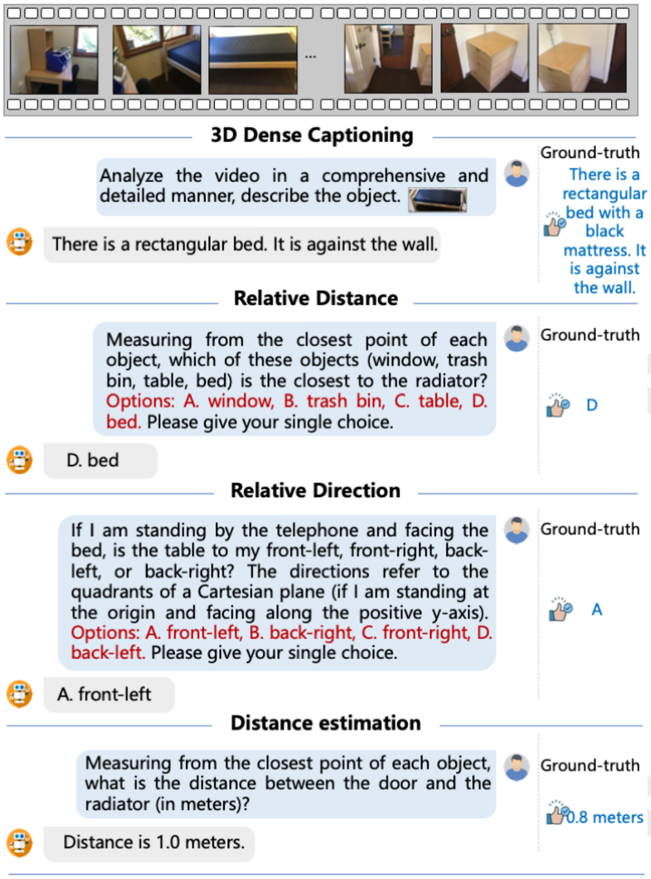
line model without the Frustum-based 3D Position Encoding (PE) and HR-adapter modules. *Method (b)* incorporates the Frustum-based 3D PE module into the baseline. This addition significantly improves performance, with gains of 14.0% in CiDEr and 4.2% in BLEU-4 on ScanQA, resulting in scores of 74.2% and 14.5%, respectively. For Scan2Cap, the module increases CiDEr by 32.1% and BLEU-4 by 14.0%, achieving scores of 64.8% and 30.7%. These results indicate that encoding 3D spatial relationships through Frustum-based 3D PE is highly effective in improving the model's understanding of 3D environments. *Method (c)* extends the baseline by incorporating the HR-adapter module, which progressively refines visual tokens using high-resolution images. This improves CiDEr by 13.1% and BLEU-4 by 2.8% on ScanQA, reaching scores of 73.1% and 13.3. On Scan2Cap, CiDEr improves by 3.5% and BLEU-4 by 1.2%. Although the HR-adapter offers a more modest improvement compared to Frustum-based 3D PE, it still contributes to better semantic understanding. *Method (d)* integrates both modules—Frustum-based 3D PE and HR-adapter—in a unified manner, representing our complete Vi3D-LLaMA framework. This configuration achieves the highest performance across all metrics. Compared to the baseline, the full model gains 20.9% in CiDEr and 4.2% in BLEU-4 on ScanQA, and 34.6% in CiDEr and 22.6% in BLEU-4 on Scan2Cap, demonstrating the complementary benefits of the two modules.

4.6 Qualitative Results

Figure 5 demonstrates Vi3D-LLaMA's capabilities across diverse 3D scene understanding tasks, showcasing its adaptability to handle both spatial reasoning and semantic understanding challenges. Through the 3D Dense Captioning task, Vi3D-LLaMA showcases its ability to understand the semantic meaning of the scene. It can accurately provide a detailed description of the rectangular bed against the wall. Additionally, the Relative Distance and Relative Direction tasks highlight Vi3D-LLaMA's spatial reasoning capabilities. It can precisely locate the bed as the closest object to the radiator, and determine the position of the table in relation to the user's perspective, demonstrating a nuanced understanding of the 3D layout.

5 Conclusion

In this work, we present Vi3D-LLaMA, a novel video-based Multimodal Large Language Model (MLLM) designed to understand and reason about 3D scenes without relying on explicit 3D representations, such as point clouds or depth inputs. The core innovation of Vi3D-LLaMA lies in its 3D spatial tokenizer, which significantly enhances the spatial awareness and semantic expressiveness of the video MLLM by integrating geometric encoding and fine-grained semantic enrichment. Vi3D-LLaMA demonstrates outstanding performance in tasks such as 3D dense captioning and question answering while requiring only RGB video inputs. By addressing the limitations of the video MLLM in 3D spatial reasoning, Vi3D-LLaMA bridges the gap between 2D perception and 3D understanding, offering an efficient and versatile solution for advancing 3D scene comprehension.



Analyze the video in a comprehensive and detailed manner, describe the object.

There is a rectangular bed. It is against the wall.

Ground-truth
There is a rectangular bed with a black mattress. It is against the wall.

Measuring from the closest point of each object, which of these objects (window, trash bin, table, bed) is the closest to the radiator? Options: A. window, B. trash bin, C. table, D. bed. Please give your single choice.

D. bed

Ground-truth
D

If I am standing by the telephone and facing the bed, is the table to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis). Options: A. front-left, B. back-right, C. front-right, D. back-left. Please give your single choice.

A. front-left

Ground-truth
A

Measuring from the closest point of each object, what is the distance between the door and the radiator (in meters)?

Distance is 1.0 meters.

Ground-truth
0.8 meters

Figure 5: **Visualization of Vi3D-LLaMA's response on various tasks.** We provide several visualization results on various 3D vision and language tasks including 3D dense captioning on Scan2Cap and relative distance / direction on VSI-Bench. The video data is from ScanNetv2. Best to zoom in.

Acknowledgments

We sincerely thank all the reviewers for their insightful feedback and appreciate the area chair's effort in handling our paper. This work was supported by the National Natural Science Foundation of China (No. 62332016) and the Key Research Program of Frontier Sciences, CAS (No.ZDBS-LY-JSC001).

References

[Achiam *et al.*, 2023] 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.

[Achlioptas *et al.*, 2020] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV (1)*, volume 12346 of *Lecture Notes in Computer Science*, pages 422–440. Springer, 2020.

[Alayrac *et al.*, 2022] 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.

[Azuma *et al.*, 2022] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, pages 19107–19117. IEEE, 2022.

[Chen *et al.*, 2021] Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in RGB-D scans. In *CVPR*, pages 3193–3203. Computer Vision Foundation / IEEE, 2021.

[Chen *et al.*, 2023] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *CVPR*, pages 11124–11133. IEEE, 2023.

[Chen *et al.*, 2024a] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. LION : Empowering multimodal large language model with dual-level visual knowledge. In *CVPR*, pages 26530–26540. IEEE, 2024.

[Chen *et al.*, 2024b] 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 *CVPR*, pages 26418–26428. IEEE, 2024.

[Chen *et al.*, 2024c] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *CoRR*, abs/2405.10370, 2024.

[Cheng *et al.*, 2024] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024.

[Chu *et al.*, 2024] Tao Chu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Qiong Liu, and Jiaqi Wang. Unified scene representation and reconstruction for 3d large language models. *CoRR*, abs/2404.13044, 2024.

[Dai *et al.*, 2017] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 2432–2443. IEEE Computer Society, 2017.

[Deng *et al.*, 2025] Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. *arXiv preprint arXiv:2501.01163*, 2025.

[Ding *et al.*, 2024] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. *CoRR*, abs/2401.00988, 2024.

[Fu *et al.*, 2024] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *CoRR*, abs/2403.11401, 2024.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[Hong *et al.*, 2023] 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. In *NeurIPS*, 2023.

[Hu *et al.*, 2019] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *CVPR*, pages 1575–1584. Computer Vision Foundation / IEEE, 2019.

[Huang *et al.*, 2023] 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. *CoRR*, abs/2312.08168, 2023.

[Huang *et al.*, 2024] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[Jin *et al.*, 2023] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *CVPR*, pages 10984–10994. IEEE, 2023.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[Li *et al.*, 2024] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *CVPR*, pages 18061–18070. IEEE, 2024.

[Lin *et al.*, 2024] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pages 5971–5984. Association for Computational Linguistics, 2024.

[Liu *et al.*, 2022] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: position embedding transformation for multi-view 3d object detection. In *ECCV (27)*, volume 13687 of *Lecture Notes in Computer Science*, pages 531–548. Springer, 2022.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[Loshchilov and Hutter, 2016] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019.

[Ma *et al.*, 2023] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: situated question answering in 3d scenes. In *ICLR*. OpenReview.net, 2023.

[Maaz *et al.*, 2024] Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL (1)*, pages 12585–12602. Association for Computational Linguistics, 2024.

[Reading *et al.*, 2021] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, pages 8555–8564. Computer Vision Foundation / IEEE, 2021.

[Schonberger and Frahm, 2016] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[Schult *et al.*, 2023] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, pages 8216–8223. IEEE, 2023.

[Touvron *et al.*, 2023] 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.

[Wang *et al.*, 2024] 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.

[Yang *et al.*, 2023] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

[Yang *et al.*, 2024] 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.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP (Demos)*, pages 543–553. Association for Computational Linguistics, 2023.

[Zheng *et al.*, 2024] 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.

[Zhu *et al.*, 2023] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pretrained transformer for 3d vision and text alignment. In *ICCV*, pages 2899–2909. IEEE, 2023.

[Zhu *et al.*, 2024a] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *CoRR*, abs/2409.18125, 2024.

[Zhu *et al.*, 2024b] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*. OpenReview.net, 2024.