ROSS3D: Reconstructive Visual Instruction Tuning with 3D-Awareness

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

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

Abstract

*The rapid development of Large Multimodal Models (LMMs) for 2D images and videos has spurred efforts to adapt these models for interpreting 3D scenes. However, the absence of large-scale 3D vision-language datasets has posed a significant obstacle. To address this issue, typical approaches focus on injecting 3D awareness into 2D LMMs by designing 3D input-level scene representations. This work provides a new perspective. We introduce **reconstructive visual instruction tuning with 3D-awareness** (**ROSS3D**), which integrates 3D aware visual supervision into the training procedure. Specifically, it incorporates cross-view and global-view reconstruction. The former requires reconstructing masked views by aggregating overlapping information from other views. The latter aims to aggregate information from all available views to recover Bird's-Eye-View images, contributing to a comprehensive overview of the entire scene. Empirically,* **ROSS3D** *achieves state-of-the-art performance across various 3D scene understanding benchmarks. More importantly, our semi-supervised experiments demonstrate significant potential in leveraging large amounts of unlabeled 3D vision-only data.*

1. Introduction

Embodied Artificial Intelligence systems are designed to effectively interact with physical environments [7, 8, 21, 47, 62], offering transformative potential applications. Central to these systems is the ability to understand 3D scenes comprehensively. This involves both modeling spatial relationships between objects [10, 17, 35] and comprehending the overall layout [5, 51, 79], which is critical for enabling embodied agents to navigate, manipulate objects, and perform complex tasks in diverse environments.

The remarkable success of Large Multimodal Modals (LMMs) in handling images [4, 18, 27, 44, 45, 53, 54, 69–72, 76] and videos [72, 83, 86] has motivated researchers to adapt these models to interpret 3D scenes [24, 32, 58, 84, 88, 89]. Similar to 2D LMMs, a straightforward approach is to develop 3D LMMs by projecting 3D point cloud features into the feature space of Large Language Models (LLMs) using point cloud-text pairs [13, 33, 73, 78]. However, unlike the abundance of large-scale 2D image-text pairs, 3D datasets remain extremely limited. Moreover, there are no powerful pre-trained 3D point cloud encoders, such as CLIP [59] in 2D, to provide strong language-aligned 3D features, leading to unsatisfactory performance shown in Figure 1. Therefore, researchers [58, 88, 89] begin to focus on building a 3D LMM based on the strong 2D priors from 2D LMMs. In such a setting, incorporating 3D-awareness into LMMs originally trained on 2D data becomes a significant challenge. To address this issue, previous attempts preliminary focus on *crafting 3D-aware input representations* as illustrated in Figure 2, including fusing 3D point cloud features with 2D image features [24, 32, 84] in Figure 2a,

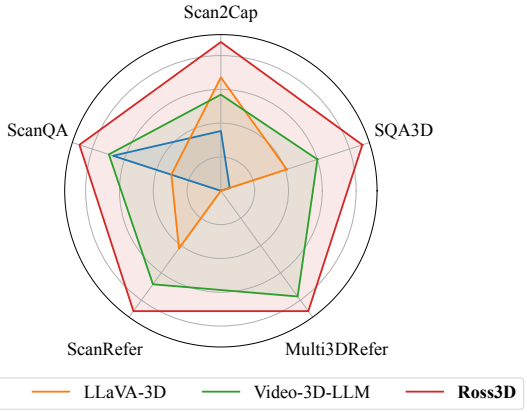
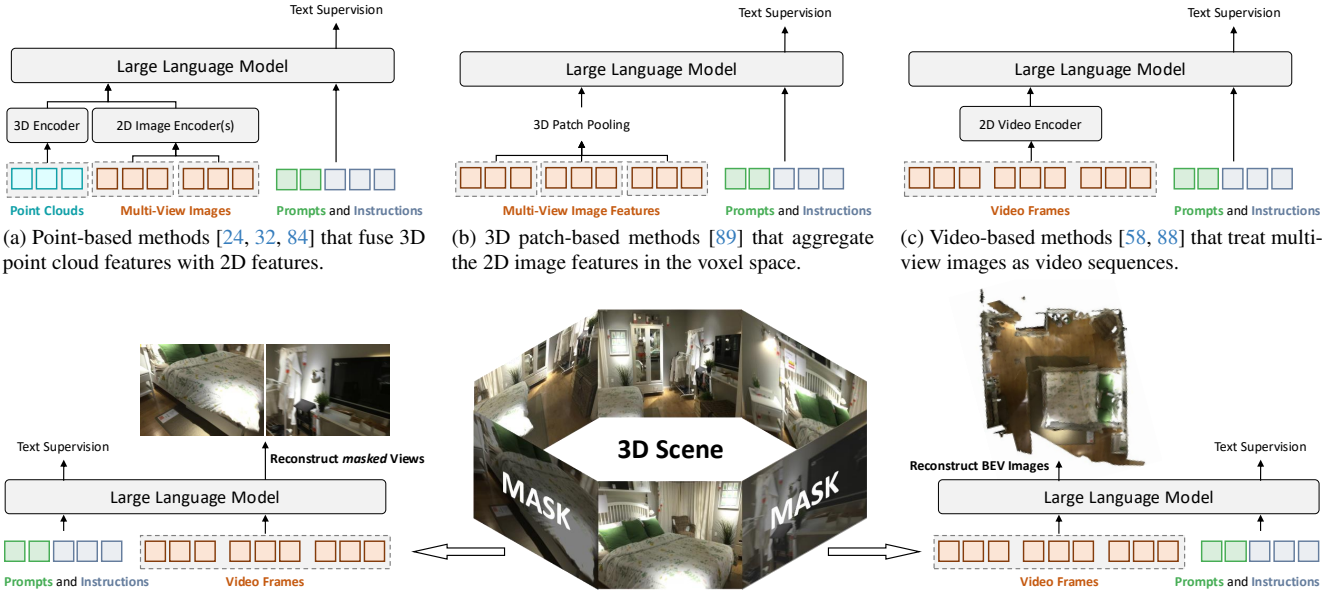


Figure 1. **Performance of ROSS3D** compared with state-of-the-art alternatives. We report EM on SQA3D [51], CIDEr on ScanQA [5], ROUGE on Scan2Cap [17], Acc@0.25 on ScanRefer [10], and F1@0.25 on Multi3DRefer [85]. With 3D-aware visual supervision, **ROSS3D** significantly outperforms other approaches across various benchmarks.

*Corresponding authors. † Project lead.



(a) Point-based methods [24, 32, 84] that fuse 3D point cloud features with 2D features.

(b) 3D patch-based methods [89] that aggregate the 2D image features in the voxel space.

(c) Video-based methods [58, 88] that treat multi-view images as video sequences.

(d) **The high-level idea of ROSS3D.** We inject vision-centric 3D-aware visual *supervision signals* into 2D LMMs by incorporating 3D-aware pretext tasks, including cross-view (left) and global-view (right) reconstruction.

Figure 2. **Conceptual comparison of our ROSS3D with popular paradigms.** Unlike previous methods that preliminarily focus on *input-level* modifications to craft 3D-aware input representations, we incorporate 3D-aware visual pretext tasks.

aggregating 2D features into 3D voxel spaces [89] with 3D patch pooling in Figure 2b, and treating multi-view images as video sequences [58, 88] in Figure 2c.

Despite these advancements, these kinds of input-level modifications *alone* are insufficient to learn genuine 3D awareness. This is because the inherent inductive bias toward 2D data in LMMs impedes the effective integration of 3D information. As a result, even with enhanced inputs, LMMs struggle to produce optimal 3D scene representations, leading to suboptimal performance. Therefore, we argue that incorporating 3D-aware *visual pretext tasks* is crucial. This allows models *to be guided* to better understand both spatial relationships and comprehensive layouts.

To address this issue, we propose **r**ec**o**nstructive vi**s**ual in**s**truction tuning with **3D**-awareness (**ROSS3D**), which introduces *3D-aware vision-centric supervision signals* by adding a variety of vision-centric 3D-aware pretext tasks into the training procedure. The high-level idea is presented in Figure 2d, where we leverage input video frames to *supervise those visual outputs* of LMMs directly. To effectively inject 3D awareness, we consider two types of 3D-aware reconstructive visual pretext tasks, including cross-view reconstruction and global-view reconstruction.

(1) Cross-view reconstruction (the left part of Figure 2d) is responsible for the detailed modeling of relationships between different views. Specifically, it requires reconstruction on *masked* views by analyzing overlapping information from other views. This process is crucial for tasks requiring fine-grained perception and precise alignment across various viewpoints such as 3D visual grounding [10, 85].

(2) Global-view reconstruction (the right part of Figure 2d) contributes to the comprehensive understanding of the whole scene since it requires integrating information from all available perspectives to recover comprehensive Bird's-Eye View (BEV) images. It synthesizes a complete and coherent overview of the entire scene, effectively capturing the full context and layout of the environment. This approach is particularly useful for applications needing a comprehensive understanding such as 3D question-answering [5, 51].

With the aid of both (1) and (2), **ROSS3D** is equipped with a more accurate and effective 3D representation learning procedure. Technically, a small denoising network [57] is responsible for reconstructing specific targets, *i.e.*, masked views and BEV images for cross-view reconstruction and global-view reconstruction, respectively, conditioned on visual outputs from LMMs [70].

Empirically, as shown in Figure 1, **ROSS3D** outperforms previous approaches by a large margin across various evaluation protocols. For instance, it achieves 63.0 EM on SQA3D [51], 107.0 CIDEr on ScanQA [5], 66.9 ROUGE on Scan2Cap [17], 61.1 Acc@0.25 on ScanRefer [10], and 59.6 F1@0.25 on Multi3DRefer [85], outperforming the previous state-of-the-art Video-3D-LLM [88] by +4.4 EM, +4.9 CIDEr, +5.2 ROUGE, +3.0 Acc@0.25, and +1.6 F1@0.25, respectively. More importantly, towards the scarcity of high-quality 3D vision-language datasets, we leverage **ROSS3D** for semi-supervised learning by training on 50% text-labeled data and applying the proposed 3D-aware visual objective to another 50% unlabeled 3D vision-

only data. This approach even surpasses the 100% text-supervised baseline in certain settings, demonstrating the significant potential of leveraging large amounts of unlabeled 3D data. To summarize, our contributions are:

- We introduce **ROSS3D**, a novel approach that enhances the ability of LMMs to understand 3D scenes on both spatial relationships and comprehensive layouts.
- We propose two distinct types of 3D-aware reconstructive visual pretext tasks: (1) cross-view reconstruction that focuses on modeling detailed relationships between different views, and (2) global-view reconstruction that provides a comprehensive understanding of the whole scene layout and context.
- **ROSS3D** brings significant improvements over previous state-of-the-art methods across multiple benchmarks, and effectively demonstrates the potential of leveraging large-scale unlabeled 3D visual data.

We hope this work will inspire future work in designing appropriate 3D-aware supervision signals for 3D LMMs.

2. Related Works

3D Scene Understanding. As a fundamental requirement for embodied agents, 3D scene understanding has emerged as a focal point of research, witnessing numerous significant advancements over the years [7, 8, 21, 47, 58, 62, 88, 89], which have empowered embodied agents to accurately identify object positions, discern structures, and understand the relationships between objects within environments. These advancements have been built upon foundational 3D perception tasks such as 3D visual grounding [2, 10, 11, 35], 3D dense captioning [12, 14, 17], and 3D question answering [5, 51, 79], each demanding a comprehensive understanding of spatial positions and object relationships. In contrast to conventional approaches [11, 28, 35, 36, 50, 75, 81, 82, 87] that typically focus on specific tasks, our goal is to develop a generalist model, which is expected to address multiple aspects of 3D scene comprehension simultaneously interacting with humans using natural language.

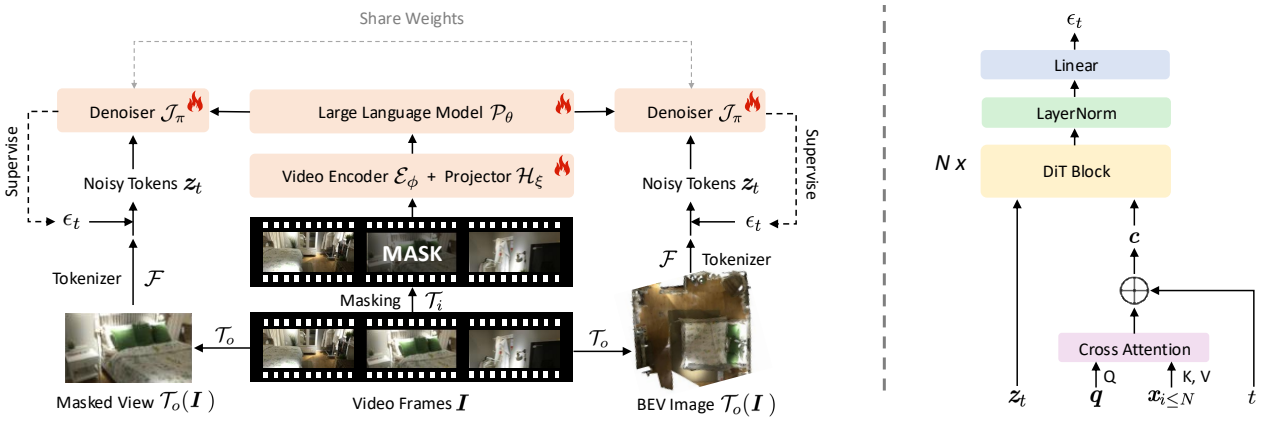
Large Multimodal Models for Scene Understanding. The impressive generalization capabilities of state-of-the-art 2D large multimodal models (LMMs) [18, 45, 70, 72, 76, 83, 86] has motivated researchers to adapt these models to understand 3D scenes [52, 58, 88, 89]. A critical challenge in this adaptation is designing appropriate 3D scene representations that align well with the original 2D LMMs. Prior attempts focus on utilizing features derived from 3D point clouds. For instance, 3D-LLM [32] aggregates features from off-the-shelf 3D reconstruction backbones [37, 38]. PointLLM [78] utilizes pre-trained 3D point cloud encoders, and LL3DA [13] further leverages an extra Q-former [46] to extract useful information. Several research [16, 33, 73, 84] further enhance these methods by incorporating 3D detectors to provide object-centric representations. However, these vision-only features are *not* aligned with the feature space of LMMs. To bridge this gap, LLaVA-3D [89] aggregates 2D-patch features from CLIP [59] in the voxel space. Video-3D-LLM [88] treats multi-view images as video sequences and incorporates 3D information into video LMMs. GPT4Scene [58] improves it by introducing an extra BEV image. This paper, on the basis of [88], regards multi-view images as videos and utilizes them as 3D scene representations, as it leverages the full potential of pre-trained 2D video LMMs and aligns more closely with human perception, where humans actually understand 3D scenes without explicit 3D point clouds. Instead of input-level modifications, we aim to explore 3D-aware vision-centric designs for enhanced 3D spatial understanding capabilities.

Vision-Centric Designs in LMMs. Typical LMMs based on visual instruction tuning [18, 45, 49, 72, 76] adopt a plug-in architecture, where pre-trained vision-language foundation models project images into visual tokens and subsequently serve as prefix tokens for multimodal comprehension. This type of design is preliminary LLM-centric, as supervision solely comes from text tokens [70]. To address this limitation, [70] pioneered the exploration of vision-centric supervision for 2D image LMMs. This paper aims to extend this idea to 3D scene understanding. However, this extension is non-trivial because it requires developing pretext tasks specific to 3D scenes rather than simply reconstructing original input images like [70]. Our exploration focuses on designing appropriate vision-centric pretext tasks to enhance 3D understanding within LMMs.

Visual Self-Supervised Learning. Visual self-supervised learning approaches rely on appropriate pretext tasks to extract scalable representations from large-scale data without any human annotations, along with representative studies in both images [15, 29, 30, 67, 68], videos [22, 64, 66, 74], and 3D scenarios [1, 23, 25, 40, 80]. In the context of 3D perception, most previous methods have primarily concentrated on designing pretext tasks for point clouds [34, 55, 60, 61, 77]. In contrast, this work represents 3D scenes with multi-view images similar to [3, 26], which offers greater scalability compared to point clouds and is naturally aligned with existing 2D foundation models.

3. Preliminaries

Visual Instruction Tuning. Autoregressive LLMs model the canonical causal distribution of a text sentence $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^{T}$ as $p_\theta(\boldsymbol{x}) = p_\theta(\boldsymbol{x}_i|\boldsymbol{x}_{<i})$, where θ indicates trainable parameters and T is the sequence length. To comprehend visual signals $\boldsymbol{I}$, typical visual instruction tuning-based methods [49] regard a sequence of visual features $\boldsymbol{v} = \mathcal{H}_\xi \circ \mathcal{E}_\phi(\boldsymbol{I})$ as prefix tokens, where $\mathcal{E}_\phi$ is a ϕ-



(a) The pipeline of **ROSS3D**.

(b) The architecture of the denoiser.

Figure 3. **Illustration of (a) ROSS3D and (b) the detailed architecture of the denoiser $\mathcal{J}_\pi$.** (a) Given raw video frames $\boldsymbol{I}$ for a 3D scene, we apply transformations to obtain inputs $\mathcal{T}_i(\boldsymbol{I})$ and targets $\mathcal{T}_o(\boldsymbol{I})$, respectively, and subsequently encourage LMMs to recover clean latent tokens $\boldsymbol{z}_0 = \mathcal{F} \circ \mathcal{T}_o(\boldsymbol{I})$ using noisy tokens $\boldsymbol{z}_t$ and visual outputs $\boldsymbol{x}_{i\leq N}$. (b) The denoiser is based on DiT [57]. Condition $\boldsymbol{c}$ is computed by a set of learnable queries $\boldsymbol{q}$, visual outputs $\boldsymbol{x}_{i\leq N}$, and timesteps t.

parameterized visual encoder such as CLIP [59] and $\mathcal{H}_\xi$ is a ξ-parameterized multimodal projector. Therefore, the canonical causal distribution for multimodal inputs becomes $p_\Theta = \prod_{i=1}^{T} p_\Theta(\boldsymbol{x}_i|\boldsymbol{x}_{<i}, v)$, where $\Theta = \{\theta, \phi, \xi\}$ indicates all parameters and $\boldsymbol{v} \in \mathbb{R}^{N\times D}$. N denotes the number of vision tokens and D is the feature dimension. The training objective is the standard cross-entropy, maximizing the log-likelihood of text outputs:

$$\mathcal{L}_{\text{text}}(\boldsymbol{x}, \boldsymbol{I}; \Theta) = -\frac{1}{T-N}\sum_{i=N+1}^{T} \log p_\Theta(\boldsymbol{x}_i|\boldsymbol{x}_{<i}, \boldsymbol{v}), \quad (1)$$

where *only text outputs $\boldsymbol{x}_{i>N}$ are supervised.*

Reconstructive Visual Instruction Tuning. [70] introduced a simple yet effective reconstructive objective into the training procedure with enhanced fine-grained comprehension capabilities. Specifically, it introduces an small π-parameterized denoising network $\mathcal{J}_\pi$ that is responsible to recover fine-grained tokens $\boldsymbol{z} = \mathcal{F}(\mathbf{I})$ conditioned on visual outputs $\boldsymbol{x}_{i\leq N}$, where $\mathcal{F}$ is the teacher tokenizer such as VAE [41]. This type of 2D objective is

$$\mathcal{L}_{\text{2D}}(\boldsymbol{x}, \boldsymbol{I}; \Theta) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[||\mathcal{J}_\pi(\boldsymbol{z}_t|\boldsymbol{x}_{i\leq N}, t) - \boldsymbol{\epsilon}||^2\right]. \quad (2)$$

While this objective is empirically effective in 2D understanding, it does not introduce any 3D awareness.

4. Method

We propose a 3D generalist model for indoor scene understanding, namely **ROSS3D**. In contrast to previous approaches that focus on injecting 3D information into 2D LMMs through input-level modifications, **ROSS3D** is equipped with novel 3D-aware vision-centric supervision signals. In the following, we elaborate on our **ROSS3D** step by step. First, we provide a comprehensive overview of our method in Section 4.1. Subsequently, implementation details of two proposed 3D-aware pretext tasks are provided in Section 4.2 and Section 4.3. Finally, detailed formulations of training objectives are introduced in Section 4.4.

4.1. Overview

An overview of our **ROSS3D** is presented in Figure 3a, which contains a video encoder $\mathcal{E}_\phi$, a large language model $\mathcal{P}_\theta$, and a denoiser $\mathcal{J}_\pi$. Different from conventional methods [32, 33, 52, 58, 84, 88, 89] that solely supervise text outputs $\boldsymbol{x}_{i>N}$, we design a series of 3D-aware vision-centric supervision $\mathcal{L}_{\text{3D}}$ for visual outputs $\boldsymbol{x}_{i\leq N}$.

$$\mathcal{L}_{\text{3D}}(\boldsymbol{x}, \boldsymbol{I}; \Theta) = \mathcal{D}(\mathcal{J}_\pi(\boldsymbol{x}_{i\leq N}), \mathcal{F} \circ \mathcal{T}_o(\boldsymbol{I})), \quad (3)$$

where $\mathcal{D}$ is a specific distance metric, and this paper takes the diffusion denoising process by default.

Transformations at the input-level $\mathcal{T}_i$ and the output-level $\mathcal{T}_o$ are applied to video frames $\boldsymbol{I}$ to obtain input videos and reconstruction targets, respectively. LMMs are required to recover $\mathcal{T}_o(\boldsymbol{I})$ while taking $\mathcal{T}_i(\boldsymbol{I})$ as inputs. In particular, Equation (3) degenerates into "vanilla reconstruction", *i.e.*, Equation (2), *without* 3D-awareness when both $\mathcal{T}_i$ and $\mathcal{T}_o$ are identity functions, *i.e.*, reconstruction targets are identical to inputs. Therefore, to inject 3D awareness, design choices of transformations are crucial.

In the following, we introduce how we choose appropriate transformations to conduct *3D-aware* self-supervised pretext tasks, including cross-view reconstruction in Section 4.2 and global-view reconstruction in Section 4.3.

Please note that the actual inputs for **ROSS3D** include video frames alongside a corresponding depth map for each frame to produce position-aware video representations discussed in the *Supplementary Material*. We omit the depth

inputs in this section as this simplification helps to focus on the essential aspects and does not influence our motivation or the overall pipeline.

4.2. Cross-View Reconstruction

Cross-view reconstruction enables reconstructing masked views based on other views, contributing to enhanced modeling of relationships between different views, which is crucial for tasks requiring fine-grained perception and precise alignment across various viewpoints such as 3D visual grounding [10, 85].

In general, input transformation $\mathcal{T}_i$ indicates randomly masking a subset of views, and output transformation $\mathcal{T}_o$ is obtaining those masked views. Formulaly, given multi-view images $\boldsymbol{I} \in \mathbb{R}^{M\times H\times W\times 3}$, where M indicates the number of views and (H, W) is the spatial resolution, we first generate a *view-aware* binary mask $\boldsymbol{M} \in \{0, 1\}^M$ with a mask ratio γ, *i.e.*, $\sum_{i=j}^{M} \boldsymbol{M}_j = (1-\gamma)M$, where 1 means unmasked views while 0 indicates the opposite. Features of those masked views are subsequently replaced with learnable mask tokens $\boldsymbol{m} \in \mathbb{R}^D$. Specifically, the encoded visual feature $\boldsymbol{v}_i$ for each frame i becomes:

$$\boldsymbol{v}_j = \begin{cases} \mathcal{H}_\xi \circ \mathcal{E}_\phi(\boldsymbol{I}_j), & \text{if } \boldsymbol{M}_j = 1, \\ \boldsymbol{m}_j, & \text{otherwise.} \end{cases} \tag{4}$$

Similarly, reconstruction targets $\boldsymbol{z}_0$ becomes the latent tokens provided by the teacher $\mathcal{F}$ for those *masked* views:

$$\boldsymbol{z}_0 = \{\mathcal{F}(\boldsymbol{I}_j) \mid \boldsymbol{M}_j = 0\}. \tag{5}$$

Therefore, the formulation of cross-view reconstruction can be obtained through rewriting Equation (3) by

$$\mathcal{L}_{\text{3D}}^{\text{cross}} = \frac{1}{\gamma M} \sum_{j=1}^{M} (1 - \boldsymbol{M}_j) \cdot \mathcal{D}(\mathcal{J}_\pi \circ \mathcal{P}_\theta(\boldsymbol{v}), \mathcal{F}(\boldsymbol{I}_j)), \tag{6}$$

where $\boldsymbol{v}$ indicates visual features defined in Equation (4). Here, language tokens are omitted as they are always behind visual tokens. Thus, they do not influence the forward motion of visual parts due to their causal nature.

Discussion. As masking may lead to discrepancies between training and testing, we apply this objective every Δt steps and we set $\Delta t = 4$ by default. For the same reason, a relatively small mask ratio, *e.g.*, 25%, is also important.

4.3. Global-View Reconstruction

Global-view reconstruction enables reconstructing the BEV image of the whole scene, resulting in an improved understanding of the environment, which is crucial for tasks requiring comprehensive comprehension such as 3D question-answering [5, 51].

Under this case, $\mathcal{T}_i$ is the same as cross-view reconstruction, while $\mathcal{T}_o$ is responsible for converting inputs into a BEV image $\boldsymbol{I}_{\text{BEV}}$ using both egocentric video, extrinsic parameters for each frame, and the corresponding camera intrinsic matrix. We use 3D reconstruction techniques to generate 3D meshes and point clouds, and render a BEV image from the top-down view.

Formally, the objective of global-view reconstruction can be obtained through rewriting Equation (3) by

$$\mathcal{L}_{\text{3D}}^{\text{global}}(\boldsymbol{x}, \boldsymbol{I}; \Theta) = \mathcal{D}(J_\pi \circ \mathcal{P}_\theta(\boldsymbol{v}), \mathcal{F}(\boldsymbol{I}_{\text{BEV}})). \tag{7}$$

Discussion. Since BEV images are actually rendered from sparse scene point clouds, this process can result in numerous black blocks. Therefore, we simply filter out these blank spaces during reconstruction.

4.4. Training Objectives

Following [70], we leverage a simple denoising objective, as vanilla regression may suffer from heavy spatial redundancy of visual signals, and thus fail to produce meaningful supervision for LMMs. Technically, as demonstrated in Figure 3a, our **ROSS3D** regards high-level visual outputs $\boldsymbol{x}_{i\leq N}$ as conditions to recover clean latent tokens $\boldsymbol{z}_0$ from noisy tokens $\boldsymbol{z}_t$. By default, we take a continuous VAE [41] regularized by Kullback–Leibler (KL) divergence provided by FLUX [43] since it is believed to capture sufficient image details. The training follows a diffusion process [31]:

$$\mathcal{D}(\mathcal{J}_\pi \circ \mathcal{P}_\theta(\boldsymbol{v}), \boldsymbol{z}_0) = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[||\mathcal{J}_\pi(\boldsymbol{z}_t | \mathcal{P}_\theta(\boldsymbol{v}), t) - \boldsymbol{\epsilon}||^2 \right], \tag{8}$$

where $\boldsymbol{z}_t$ is sampled from $\mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t \mathbf{1})$ and $\mathbf{1}$ here indicates the identity matrix. Following [31], $\boldsymbol{z}_t$ could be sampled directly from $\boldsymbol{z}_0$ by letting $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$

$$\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{1}). \tag{9}$$

Other objectives include standard cross-entropy loss introduced in Equation (1) and grounding loss described later in the *Supplementary Material*.

5. Experiments

Datasets. To evaluate the 3D scene understanding capabilities of our **ROSS3D**, we conduct experiments across five representative benchmarks, including SQA3D [51] for situated reasoning, ScanQA [5] for spatial understanding, Scan2Cap [17] for captioning specific objects, ScanRefer [10] and Multi3DRefer [85] for detecting objects in single-target and multiple-target scenarios. All these datasets are derived from ScanNet [20], which is an extensively annotated collection of RGB-D video data, encompassing 1,513 scans of 3D indoor scenes.

Metrics. Widely used evaluation metrics are utilized for each benchmark. For SQA3D [51], we evaluate the performance using exact match accuracy (EM) and the refined exact match protocol (EM-R) following [33, 89]. For

Method	Point Encoder	Vision Encoder	SQA3D$_{test}$		ScanQA$_{val}$				
			EM	EM-R	CIDEr	BLEU-4	METEOR	ROUGE	EM
Expert Models									
SQA3D [51]	✓	–	46.6	–	–	–	–	–	–
ScanQA [5]	✓	–	–	–	64.9	10.1	13.1	33.3	21.1
3D-VLP [39]	✓	–	–	–	–	11.2	13.5	34.5	21.7
3D-VisTA [90]	✓	–	–	–	–	–	13.9	35.7	22.4
2D LMMs									
InternVL2-8B [63]	–	✓	33.0	45.3	62.5	3.3	14.5	34.3	–
Qwen2-VL-7B [72]	–	✓	40.7	46.7	53.9	3.0	11.4	29.3	–
LLaVA-Video-7B [86]	–	✓	48.5	–	88.7	3.1	17.7	44.6	–
3D LMMs									
Chat-3D [73]	✓	–	–	–	53.2	6.4	11.9	28.5	–
3D-LLM [32]	✓	✓	–	–	69.4	12.0	14.5	35.7	20.5
Scene-LLM [24]	✓	✓	53.6	–	80.0	11.7	15.8	35.9	27.2
LL3DA [13]	✓	–	–	–	76.8	–	15.9	37.3	–
LEO [33]	✓	✓	50.0	52.4	80.0	11.5	16.2	39.3	21.5
ChatScene [84]	✓	✓	54.6	57.5	87.7	14.3	18.0	41.6	21.6
Grounded 3D-LLM [16]	✓	✓	–	–	72.7	13.4	–	–	–
LLaVA-3D [89]	–	✓	55.6	57.6	91.7	14.5	20.7	50.1	27.0
Video-3D-LLM [88]	–	✓	58.6	–	102.1	16.4	20.0	49.3	30.1
GPT4Scene-HDM‡ [58]	–	✓	59.4	62.4	96.3	15.5	18.9	46.5	–
ROSS3D	–	✓	**63.0**	**65.7**	**107.0**	**17.9**	**20.9**	**50.7**	**30.8**

Table 1. **Evaluation of 3D question-answering** on SQA3D [51] and ScanQA [5]. "Expert models" are customized for specific tasks with task-oriented decoders. General 2D LMMs [63, 72, 86] are evaluated in a zero-shot setting. "EM" stands for top-1 exact match and "EM-R" means the refined exact match following [33]. "–" indicates the number is not available for us. "‡" indicates this result is achieved by adopting a larger input resolution, *i.e.*, 512×490, and incorporating extra BEV inputs.

ScanQA [5], we use EM, BLEU-4 [56], METEOR [6], ROUGE [48], and CIDEr [65]. For Scan2Cap [17], we combine captioning metrics (CIDEr, BLEU-4, METEOR, and ROUGE) with an IoU threshold of 0.5 between predicted and reference bounding boxes. For ScanRefer [10], we report Acc@0.25 and Acc@0.5, where a prediction is considered correct only if the IoU exceeds 0.25 and 0.5, respectively. For Multi3DRefer [85], we combine F1 scores and IoU thresholds, *i.e.*, F1@0.25 and F1@0.5.

Implementation Details. We build our **ROSS3D** based on LLaVA-Video-7B [86], which is then fine-tuned on the combination of training sets of SQA3D [51], ScanQA [5], Scan2Cap [17], ScanRefer [10], and Multi3DRefer [85] for one epoch, using the AdamW optimizer with a global batch size of 256. The learning rates peak at 1e-5 for the LLM during the warmup phase and the vision encoder is kept frozen. All experiments are conducted with 8×A100-80G. Each scene is represented by 32 frames, with the resolution of each frame being 384×384. BEV images are rendered from point clouds with a resolution of 432×432.

5.1. Comparison with State-of-the-Arts

Comparison Alternatives. We include both expert models designed for specific tasks and LMM-based models with different input representations. *Expert models* include

Method	Scan2Cap$_{val}$ (IoU@0.5)			
	ROUGE	BLEU-4	METEOR	CIDEr
Expert Models				
Scan2Cap [17]	44.5	23.3	22.0	35.2
3DJCG [9]	50.8	31.0	24.2	49.5
3D-VLP [39]	51.5	32.3	24.8	54.9
3D-VisTA [90]	54.3	34.0	26.8	61.6
3D LMMs				
LL3DA [13]	55.1	36.8	26.0	65.2
LEO [33]	58.1	38.2	27.9	72.4
ChatScene [84]	58.1	36.3	–	77.1
LLaVA-3D [89]	63.4	41.1	30.2	79.2
Video-3D-LLM [88]	62.3	42.4	28.9	**83.8**
GPT4Scene-HDM‡ [58]	59.3	40.6	–	–
ROSS3D	**66.9**	**43.4**	**30.3**	81.3

Table 2. **Evaluation of 3D dense captioning** on Scan2Cap [17]. "‡" indicates this result is achieved by adopting a larger input resolution, *i.e.*, 512×490, and incorporating extra BEV inputs.

ScanQA [5], Scan2Cap [17], ScanRefer [10], MVT [35], 3DVG-Trans [87], ViL3DRel [11], M3DRef-CLIP [85], 3D-VLP [39], 3DJCG [9], and 3D-VisTA [90]. *2D LMMs* include InternVL2 [63], Qwen2-VL [72], and LLaVA-Video [86]. *3D LMMs* include 3D-LLM [32] that leverages 2D encoders pre-trained 3D tasks, Scene-LLM [24] and

LL3DA [13] that utilize point cloud features, Chat-3D [73], LEO [33], ChatScene [84], and Grounded 3D-LLM [16] that incoperate object representations, LLaVA-3D [89] that aggregates 2D features in the 3D voxel space, and Video-3D-LLM [88] and GPT4Scene [58] that treat multi-view images as video sequences.

3D Question Answering. We compare our **ROSS3D** with other methods on 3D question answering in Table 1. As demonstrated in the table, **ROSS3D** achieves 63.0 EM on SQA3D [51] and 107.0 CIDEr on ScanQA [5], outperforming previous state-of-the-art Video-3D-LLM [88] by +4.4 EM on SQA3D [51] and +4.9 CIDEr on ScanQA [5].

3D Dense Captioning. We compare our **ROSS3D** with other methods on 3D dense captioning in Table 2. As demonstrated in the table, **ROSS3D** achieves 66.9 ROUGE, 43.4 BLEU-4, and 30.3 METEOR on Scan2Cap [17], outperforming previous state-of-the-art Video-3D-LLM [88] by +4.6 ROUGE, +1.0 BLEU-4, and +1.4 METEOR.

3D Visual Grounding. We compare **ROSS3D** with other methods on 3D visual grounding in Table 3. As demonstrated in the table, **ROSS3D** achieves 61.1 Acc@0.25 and 54.4 Acc@0.5 on ScanRefer [10], and 59.6 F1@0.25 and 54.3 F1@0.5 on Multi3DRefer [85], outperforming previous state-of-the-art Video-3D-LLM [88] by +3.0 Acc@0.25 and +2.7 Acc@0.5 on ScanRefer [10], and +1.6 F1@0.25 and +1.6 F1@0.5 on Multi3DRefer [85], respectively.

Method	ScanRefer$_{val}$		Multi3DRefer$_{val}$	
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
Expert Models				
ScanRefer [10]	37.3	24.3	–	–
MVT [35]	40.8	33.3	–	–
3DVG-Trans [87]	47.6	34.7	–	25.5
ViL3DRel [11]	47.9	37.7	–	–
3DJCG [9]	49.6	37.3	–	26.6
M3DRef-CLIP [85]	51.9	44.7	42.8	38.4
3D LMMs				
3D-LLM [32]	30.3	–	–	–
Ground 3D-LLM [16]	47.9	44.1	45.2	40.6
ChatScene [84]	55.5	50.2	57.1	52.4
LLaVA-3D [89]	54.1	42.4	–	–
Video-3D-LLM [88]	58.1	51.7	58.0	52.7
GPT4Scene-HDM‡ [58]	62.6	57.0	64.5	59.8
ROSS3D	**61.1**	**54.4**	**59.6**	**54.3**

Table 3. **Evaluation of 3D visual grounding** on ScanRefer [10] and Multi3DRefer [85]. "‡" indicates this result is achieved by adopting a larger input resolution, *i.e.*, 512×490, ≈1.7× pixels than ours, and incorporating extra BEV inputs.

5.2. Ablation Studies

In this section, we report EM for SQA3D [51] , CIDEr for ScanQA [5], Acc@0.25 for ScanRefer [10], and F1@0.25 for Multi3DRefer [85] by default.

Effectiveness of Each 3D-Aware Pretext Task. We study the effectiveness of each 3D-aware pretext task using different input representations in Table 4. Specifically, we compare our proposed two *3D-aware* tasks, *i.e.*, cross-view reconstruction and global-view reconstruction, with vanilla reconstruction *without* 3D-awareness, and visual instruction tuning baselines. Input representations include scene-level 3D features provided by 3D-LLM [32] and position-aware video representations proposed by Video-3D-LLM [88]. According to [32], these 3D point cloud features are obtained by (1) extracting object masks using Mask2Former [19] and SAM [42], (2) extracting features of each object using BLIP-2 [46], and (3) reconstructing 3D features from extracted multi-view 2D features. Following the official implementation of 3D-LLM [32], we load the v2 pre-trained model and fine-tune on each task *separately* for 100 epochs. We fail to conduct experiments on ScanRefer [10] with 3D-LLM [32] as this part of fine-tuning code is unavailable.

As demonstrated in the table, we can draw the following three important conclusions. (1) Pretext tasks *with* 3D-awareness is crucial, as vanilla reconstruction brings marginal improvements. (2) Each 3D-aware pretext task is effective. (3) The proposed two pretext tasks promote each other, contributing to significant improvements *across different input representations*. Moreover, cross-view reconstruction is particularly effective for 3D visual grounding on ScanRefer [10] while global-view reconstruction is effective for 3D question answering on [5, 51].

	Method	SQA3D	ScanQA	ScanRefer
Point-based Representation				
1	3D-LLM‡ [32]	49.4	68.9	–
2	① + vanilla	48.9 ↓ 0.5	68.9 – 0.0	–
3	① + cross-view	51.0 ↑ 1.6	70.3 ↑ 0.4	–
4	① + global-view	54.3 ↑ 4.9	71.0 ↑ 1.1	–
5	① + ③ + ④	**55.0** ↑ **5.6**	**73.1** ↑ **2.2**	–
Video-based Representation				
6	Video-3D-LLM [88]	58.6	102.1	58.1
7	⑥ + vanilla	58.8 – 0.0	103.5 ↑ 1.4	58.2 ↑ 0.1
8	⑥ + cross-view	60.0 ↑ 1.4	103.6 ↑ 1.5	60.3 ↑ 2.1
9	⑥ + global-view	61.6 ↑ 3.0	105.6 ↑ 3.5	58.8 ↑ 0.7
10	⑥ + ⑧ + ⑨	**63.0** ↑ **4.4**	**107.0** ↑ **4.9**	**61.1** ↑ **3.0**

Table 4. **Ablations on 3D-aware pretext tasks** with different input representations, including point-based [32], and video-based [88]. "Vanilla" indicates directly reconstruction *without* 3D-awareness. Our default setting is highlighted in color. "‡" means our reproduction using the official code. "–" indicates this part of code is unavailable for us.

ROSS3D *v.s.* Adding 3D Features. In Table 5, we compare our output-level supervision solution with the input-level

	Method	SQA3D	ScanQA	ScanRefer
1	Video-3D-LLM [88]	58.6	102.1	58.1
2	① + 3D features	59.1 ↑ 0.5	102.4 ↑ 0.3	57.8 ↓ 0.3
3	**Ross3D**	**63.0** ↑ **4.4**	**107.0** ↑ **4.9**	**61.1** ↑ **3.0**

Table 5. **Ross3D *v.s.* Adding 3D Features.** Scene-level 3D features are provided by [32], which are aggregated with video features via cross attention.

	50% Data	+ 50% Data	SQA3D	ScanQA	ScanRefer
1	$\mathcal{L}_{\text{text}}$	–	55.1	100.3	57.0
2	$\mathcal{L}_{\text{text}}$ + $\mathcal{L}_{\text{3D}}$	–	56.4 ↑ 1.3	101.7 ↑ 0.6	57.4 ↑ 0.4
3	$\mathcal{L}_{\text{text}}$ + $\mathcal{L}_{\text{3D}}$	$\mathcal{L}_{\text{3D}}$	**57.7** ↑ **2.6**	**103.2** ↑ **2.9**	**57.9** ↑ **0.9**
4	$\mathcal{L}_{\text{text}}$	$\mathcal{L}_{\text{text}}$	58.6	102.1	58.1

Table 6. **Semi-supervised learning with Ross3D.** "✓" indicates applying the particular objective on this part of data, while "–" means the opposite. ④ is actually the standard Video-3D-LLM [88] baseline. *The proposed $\mathcal{L}_{3D}$ enables learning from raw visual signals effectively.*

aggregation alternative. Specifically, the second row in Table 5 indicates we aggregate scene-level 3D point cloud features provided by [32] with original position-aware video representations. This table demonstrates that our **Ross3D** is much more effective than simply adding 3D features.

Semi-Supervised Learning with Ross3D. As high-quality 3D vision-language data is quite limited, scaling up 3D LMMs poses a significant challenge. To address this issue by *leveraging knowledge directly from raw 3D visual data*, we conduct the following *semi-supervised* experiments in Table 6, as **Ross3D** naturally allows learning from 3D sequences *without* text annotations. Specifically, we split the training set into two non-overlapping parts and then run 4 settings: ① training with 50% text data with conventional $\mathcal{L}_{\text{text}}$, ② training with 50% text data with visual supervision provided by **Ross3D**, *i.e.*, with $\mathcal{L}_{\text{3D}}$ in the table, ③ training 50% with text data and applying **Ross3D** on the other 50% *without* text annotations, and ④ training with 100% text data. As demonstrated in the table, ③ significantly outperforms both ① and ②. It even surpasses the supervised upper bound represented by ④ on ScanQA (103.2 *v.s.* 102.1 on CIDEr). This result highlights the effectiveness of $\mathcal{L}_{\text{3D}}$ in learning *directly* from visual signals.

Training Costs. We analyze the extra computational costs brought by the proposed two 3D-aware visual pretext tasks in Table 7. Evaluations are conducted using 8 A100 GPUs with a global batch size of 256. Due to the limited GPU memory, we accumulate 32 gradient steps and the batch size per GPU is 1. The whole stage requires 871 training steps. GPU memories are averaged over 8 GPUs with DeepSpeed Zero 3. As shown in the table, the denoising process introduces a *negligible* increase.

$\mathcal{L}_{\text{3D}}^{\text{cross}}$	$\mathcal{L}_{\text{3D}}^{\text{global}}$	Speed (s/iter)	GPU Memory
–	–	111.6	57.2 G
✓	✓	125.2 (1.12×)	58.6 G (1.02×)

Table 7. **Training cost comparison.** All entries are based on LLaVA-Video-7B [86] with 32 frames as inputs. $\mathcal{L}_{\text{3D}}$ brings *marginal* extra computational costs.

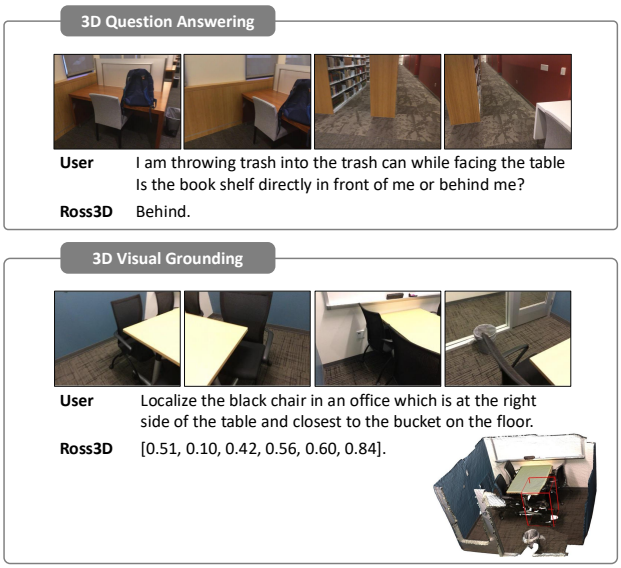


Figure 4. **Qualitative results.** Examples of 3D question answering and 3D visual grounding are sampled from $\text{ScanQA}_{\text{val}}$ [51] and $\text{ScanRefer}_{\text{val}}$ [10], respectively.

Qualitative Results. We provide qualitative results on both 3D question answering and 3D visual grounding in Figure 4. Our **Ross3D** is able to understand the whole scene comprehensively on 3D question answering, and perceive fine-grained details on 3D visual grounding.

6. Conclusion

We introduce **Ross3D** that significantly enhances the 3D scene understanding capabilities of LMMs. Unlike previous attempts that primarily focus on crafting 3D-aware input representations, we incorporate visual pretext tasks as 3D-aware supervision. These tasks, including cross-view and global-view reconstructions, enable accurate spatial relationship modeling and comprehensive scene layout comprehension, respectively. **Ross3D** demonstrates substantial improvements across various benchmarks compared to previous alternatives. More importantly, semi-supervised learning by training on 50% text-labeled data and applying the proposed 3D visual objectives on the other 50% vision-only data even surpasses the 100% text-supervised baseline in certain settings, demonstrating the significant potential of leveraging large amounts of unlabeled 3D data. We hope that our research draws the community's attention to the design of 3D-aware visual supervision signals for 3D LMMs.

Acknowledgements

This work was supported in part by the National Science and Technology Major Project (No. 2023ZD0121300), and the National Natural Science Foundation of China (No. U21B2042, No. 62320106010).

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning (ICML)*, pages 40–49. PMLR, 2018. 3

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, pages 422–440. Springer, 2020. 3

[3] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3773–3782, 2022. 3

[4] Anthropic. Claude 3.5 sonnet, 2024. 1

[5] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, 2022. 1, 2, 3, 5, 6, 7

[6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[7] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1, 3

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 1, 3

[9] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16464–16473, 2022. 6, 7

[10] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, pages 202–221. Springer, 2020. 1, 2, 3, 5, 6, 7, 8

[11] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 20522–20535, 2022. 3, 6, 7

[12] 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 (CVPR)*, pages 11124–11133, 2023. 3

[13] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26428–26438, 2024. 1, 3, 6, 7

[14] Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, YU Gang, Taihao Li, and Tao Chen. Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 3

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 3

[16] 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. 3, 6, 7

[17] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2021. 1, 2, 3, 5, 6, 7

[18] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 3

[19] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 7

[20] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 5

[21] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024. 1, 3

[22] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 35946–35958, 2022. 3

[23] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *European Conference on Computer Vision (ECCV)*, pages 228–244, 2022. 3

[24] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 1, 2, 6

[25] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 3

[26] Xiang Gao, Wei Hu, and Guo-Jun Qi. Self-supervised multi-view learning via auto-encoding 3d transformations. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(1):1–23, 2023. 3

[27] Google. Our next-generation model: Gemini 1.5, 2024. 1

[28] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15372–15383, 2023. 3

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 3

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 5

[32] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 20482–20494, 2023. 1, 2, 3, 4, 6, 7, 8

[33] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 1, 3, 4, 5, 6, 7

[34] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6535–6545, 2021. 3

[35] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15524–15533, 2022. 1, 3, 6, 7

[36] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision (ECCV)*, pages 417–433. Springer, 2022. 3

[37] Krishna Murthy Jatavallabhula, Soroush Saryazdi, Ganesh Iyer, and Liam Paull. gradslam: Automagically differentiable slam. *arXiv preprint arXiv:1910.10672*, 2019. 3

[38] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 3

[39] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10984–10994, 2023. 6

[40] Longlong Jing, Yucheng Chen, Ling Zhang, Mingyi He, and Yingli Tian. Self-supervised modal and view invariant feature learning. *arXiv preprint arXiv:2005.14169*, 2020. 3

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4, 5

[42] 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 (ICCV)*, pages 4015–4026, 2023. 7

[43] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 5

[44] Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1

[45] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 3

[46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742, 2023. 3, 7

[47] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023. 1, 3

[48] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023. 3

[50] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 16454–16463, 2022. 3

[51] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 3, 5, 6, 7, 8

[52] Guofeng Mei, Wei Lin, Luigi Riz, Yujiao Wu, Fabio Poiesi, and Yiming Wang. Perla: Perceptive 3d language assistant. *arXiv preprint arXiv:2411.19774*, 2024. 3, 4

[53] OpenAI. GPT-4V(ision) System Card, 2023. 1

[54] OpenAI. Hello GPT-4o, 2024. 1

[55] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision (ECCV)*, pages 604–621. Springer, 2022. 3

[56] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002. 6

[57] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 2, 4

[58] 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. 1, 2, 3, 4, 6, 7

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PmLR, 2021. 1, 3, 4

[60] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *European Conference on Computer Vision (ECCV)*, pages 626–642. Springer, 2020. 3

[61] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 3

[62] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023. 1, 3

[63] OpenGVLab Team. InternVL2: Better than the Best—Expanding Performance Boundaries of Open-Source Multimodal Models with the Progressive Scaling Strategy, 2024. 6

[64] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:10078–10093, 2022. 3

[65] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 6

[66] Haochen Wang, Junsong Fan, Yuxi Wang, Kaiyou Song, Tiancai Wang, Xiangyu Zhang, and Zhaoxiang Zhang. Bootstrap masked visual modeling via hard patches mining. *arXiv preprint arXiv:2312.13714*, 2023. 3

[67] Haochen Wang, Junsong Fan, Yuxi Wang, Kaiyou Song, Tong Wang, and Zhaoxiang Zhang. Droppos: Pre-training vision transformers by reconstructing dropped positions. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:46134–46151, 2023. 3

[68] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10375–10385, 2023. 3

[69] Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025. 1

[70] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Ge Zheng, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *International Conference on Learning Representations (ICLR)*, 2025. 2, 3, 4, 5

[71] Jiacong Wang, Zijiang Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025.

[72] 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. 1, 3, 6

[73] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023. 1, 3, 6, 7

[74] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14668–14678, 2022. 3

[75] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19231–19242, 2023. 3

[76] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 1, 3

[77] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision (ECCV)*, pages 574–591. Springer, 2020. 3

[78] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision (ECCV)*, pages 131–147. Springer, 2024. 1, 3

[79] 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. 1, 3

[80] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 3

[81] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1856–1866, 2021. 3

[82] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8563–8573, 2022. 3

[83] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1, 3

[84] Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15459–15469, 2024. 1, 2, 3, 4, 6, 7

[85] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15225–15236, 2023. 1, 2, 5, 6, 7

[86] 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, 3, 6, 8

[87] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2928–2937, 2021. 3, 6, 7

[88] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. *arXiv preprint arXiv:2412.00493*, 2024. 1, 2, 3, 4, 6, 7, 8

[89] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 1, 2, 3, 4, 5, 6, 7

[90] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2911–2921, 2023. 6