

Vision-Language Memory for Spatial Reasoning

Zuntao Liu¹, Yi Du¹, Taimeng Fu¹, Shaoshu Su¹, Cherie Ho², and
Chen Wang^{1*}

¹ Spatial AI & Robotics (SAIR) Lab, University at Buffalo, USA

² Stanford University, USA

lzuntao327@gmail.com, chenw@sairlab.org

<https://sairlab.org/vlm2>

Abstract. Spatial reasoning is a critical capability for intelligent robots, yet current vision-language models (VLMs) still fall short of human-level performance in video-based spatial reasoning. This gap mainly stems from two challenges: a semantic-geometric misalignment that prevents consistent 3D understanding, and the absence of persistent memory to retain 3D representation and understanding across frames. To address these limitations, we present VLM², a Vision-Language Model with persistent Memory for spatial reasoning with a view-consistent, 3D-aware representation purely from 2D videos. Specifically, we incorporate a dual-memory module consisting of a working memory that operates as a sliding window to focus on immediate context, and an episodic memory that consolidates and stores critical information across frames. This design enables bounded and efficient spatial reasoning under a fixed computational cost. Extensive experiments on multiple benchmarks show that VLM² achieves state-of-the-art performance among video-based models, significantly advancing the frontier of visual-spatial intelligence.

Keywords: Spatial Reasoning · Vision-Language · Memory

1 Introduction

Spatial reasoning is a fundamental capability of intelligent robots, enabling them to perceive, localize, and reason about spatial relationships in the physical world. Recent advances [8, 15, 27, 38, 51, 75] have sought to enhance this ability in vision-language models (VLMs) [3, 31, 34, 37, 73], for instance by incorporating 3D information such as point clouds and depth maps to enhance spatial awareness. Despite these efforts, current VLMs still fall short of human-level spatial reasoning [7, 64]. As in Fig. 1, existing models often fail to answer simple questions like “*How many chairs are there in this room?*”. One reason is that such questions require consistent feature alignment across multiple views (recognizing the same chairs across viewpoints) and memory across frames (maintaining an accurate count as views change). However, existing VLMs rarely address the two issues simultaneously. In practice, they often exhibit (1) semantic-geometric feature misalignment and (2) the absence of persistent memory.

* Corresponding author.

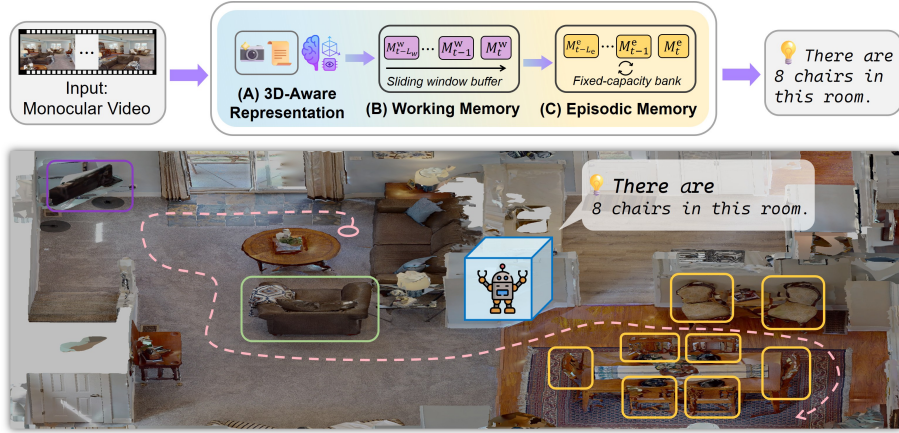


Fig. 1: VLM² is a Vision-Language Model with Memory for spatial reasoning that constructs view-consistent 3D-aware representations from monocular video and supports consistent reasoning across viewpoints. Such capabilities are critical for questions like “How many chairs are in this room?”, which require consistent cross-view alignment and reasoning over previously observed content.

First, while existing methods that leverage semantic features from 2D encoders provide strong categorical understanding, they lack metric grounding and positional awareness [16, 30, 69, 75], making it challenging to maintain spatial coherence across viewpoints. For instance, when moving from the living room to the kitchen, a model may incorrectly associate distinct chairs at different locations as the same instance, causing spatial inconsistency. In contrast, geometric features from 3D visual geometry models offer reliable structural cues; however, naïvely fusing them with semantic representations often produces view-dependent instabilities [6, 35, 36, 43], causing global spatial inconsistencies under camera motion.

Moreover, directly using predicted 3D point clouds as explicit inputs does not inherently establish semantic-geometric correspondence across views. Although they provide structural cues, point-level predictions are not naturally aligned with semantic tokens, thus requiring additional alignment mechanisms.

The second challenge lies in the absence of persistent memory, although memory mechanisms have been explored in computer vision for decades [22]. Mobile robots experience viewpoint changes and occlusions, where objects may temporarily disappear and reappear as the viewpoint shifts [33, 61]. Existing VLMs [3, 34, 42, 73] rely on token-based context windows that struggle to preserve coherent spatial representations across frames [25, 66]. As new visual inputs arrive, previously observed information may be overwritten, preventing consistent spatial reasoning across the video sequence. For example, when the viewpoint shifts and later revisits a previously observed area, a chair initially seen at the entrance may no longer be recalled, leading to failures in tasks such as object

counting. This highlights the need for an explicit memory mechanism capable of preserving and updating 3D-aware representations across frames and viewpoints.

To solve these challenges, existing approaches have explored several directions. Early works [8, 15, 27, 38, 51, 75] fine-tuned VLMs with additional 3D data, yet progress was constrained by limited dataset scale and diversity. More recent methods [17, 58, 76] leverage geometric priors from 3D visual geometry foundation models [53–55], but their reliance on simple feature fusion neither explicitly addresses semantic-geometric misalignment nor provides a dedicated mechanism to preserve representations across frames. As a result, achieving coherent 3D-aware representations across views remains challenging.

In this paper, we present VLM², a **V**ision-**L**anguage **M**odel with persistent attention-based **M**emory for spatial reasoning. First, we construct a 3D-aware representation from visual observations by enforcing the alignment between the semantic and geometric features before fusion. To this end, we introduce a viewpoint-aware geometry alignment module to align geometry tokens with their view tokens, ensuring geometric features from different viewpoints are distinct even if their geometry is similar. Furthermore, we inject the predicted 3D coordinates into their corresponding visual tokens to ensure geometric awareness. To avoid potential spatial inconsistency, we introduce a learnable gate to decide which of these 3D points are useful, allowing the model to be “adaptive.” These designs unlock the model’s ability to perform spatial reasoning, yielding view-consistent 3D-aware representations under camera motion.

Additionally, some models like [75] require both 2D videos and accurate depth labels (or 3D point clouds) as inputs. This requirement limits their applicability in real-world scenarios, where reliable depth measurements are often unavailable. Although depth maps can be estimated using monocular depth estimation models, our empirical observations show that directly using such noisy depth predictions can significantly degrade model performance. In contrast, our experiments demonstrate that our proposed view-consistent 3D-aware representation effectively addresses these issues without relying on accurate depth inputs.

To equip the model with persistent memory, we introduce an attention-based dual-memory to preserve and update 3D-aware representations across frames and viewpoints. It comprises (1) a working memory that functions as a sliding window over recent frames to capture immediate context, and (2) an episodic memory that serves as a fixed-capacity bank for recall across frames. We design gated fusion and similarity-based update mechanisms to retain informative observations while preventing redundant accumulation, enabling coherent spatial reasoning across frames under bounded computation and storage. We note that 3DLLM-Mem [25] augments VLMs with memory mechanisms. In contrast to memory architectures that grow stored representations over time, our dual-memory design integrates a sliding-window working memory for recent context with a fixed-capacity episodic memory that performs gated fusion. This design ensures bounded memory usage while preserving reliable spatial reasoning.

We evaluate VLM² on multiple spatial reasoning and 3D understanding benchmarks, including VSI-Bench [64], VSTI-Bench [17], ScanQA [1], and SQA3D

[41]. VLM² achieves superior performance among video-based models and surpasses open-source VLMs and spatial-enhanced models [3, 17, 58, 74, 76], advancing the frontier of visual-spatial intelligence. Our main contributions include:

- We introduce VLM², a vision-language model, integrating a semantic-geometric consistent representation with a persistent dual-memory, enabling spatial reasoning from video. VLM² achieves state-of-the-art performance on multiple benchmarks, including VSI-Bench, VSTI-Bench, ScanQA, and SQA3D.
- We develop a 3D-aware representation for spatial reasoning that resolves semantic-geometric misalignment by grounding visual features into 3D space and enforcing cross-view consistency for coherent understanding. We also design a dual-memory module that couples a sliding-window working memory with a fixed-capacity episodic memory, achieving efficient spatial reasoning.

2 Related Work

2.1 3D Large Language Models

Recent efforts have focused on enabling 3D large language models to understand 3D scenes by introducing explicit 3D modalities [8, 9, 23, 26, 28, 51, 57, 75, 77]. Early work LL3DA [8] employs a point cloud encoder for scene-level representations, while Chat-3D [57], LEO [28], and Chat-Scene [26] segment objects and aggregate object-level 3D features for scenes. 3D-LLM [23] recovers 3D cues from rendered multi-view images via a 3D feature extractor, while LLaVA-3D [77] injects 3D position to form 3D-aware patch aggregation. More recently, Video-3D LLM [75] encodes 3D position into the visual representation. Ross3D [51] leverages reconstruction-based supervision to learn 3D-aware features. GS-Reasoner [9] constructs a 3D scene representation by integrating geometric features extracted from a point cloud encoder. These approaches demonstrate the effectiveness of incorporating explicit 3D information for 3D understanding. In contrast, our framework focuses on learning 3D-aware representations from 2D inputs, requiring no additional 3D data or supervision.

2.2 Spatial Reasoning in VLMs

Equipping VLMs with spatial reasoning capabilities has recently emerged as a significant area of research [5, 7, 12, 40, 44, 63, 65]. However, previous work has mainly focused on spatial understanding from 2D static images, leaving video-based spatial reasoning as an underexplored domain. To address this gap, VSI-Bench [64] introduces a video-based benchmark to evaluate how effectively VLMs can understand spatial relationships. Recent studies [17, 30, 58, 76] have sought to enhance VLM spatial reasoning by incorporating geometric priors from 3D Visual Geometry Foundation Models (VGFM). For instance, Spatial-MLLM [58] and VG-LLM [76] leverage VGFM (e.g. VGGT [53]) as spatial encoders to extract geometric features. Similarly, 3DRS [30] introduces supervision signals from pretrained VGFM, although this approach relies on additional 3D data to compute position information. VLM-3R [17] incorporates implicit 3D tokens from pretrained VGFM (e.g., CUT3R [54]) and further proposes the VSTI-Bench

to evaluate the comprehension of spatial relationships evolving over time. Despite these advances, a common limitation of these methods is their tendency to simply fuse geometric features with semantic cues. Without explicitly addressing the potential semantic-geometric misalignment, their performance improvements remain constrained. In contrast, our work introduces an explicit alignment mechanism to learn a view-consistent 3D-aware representation, which more effectively integrates semantic and geometric cues for spatial reasoning.

2.3 Memory for Spatial Reasoning

Memory mechanisms have been adopted in vision tasks that depend on long-range temporal or spatial information. Their applications include video understanding [20, 45, 71], video generation [29, 59, 68], 3D reconstruction [11, 52, 60], and robotic memory [48–50, 62]. This paradigm has also been explored for lifelong navigation in embodied agents [32], where memory representations are critical. Within this embodied context, several works focus on building spatial memories. For instance, MTU3D [79] maintains a dynamic spatial memory bank for grounding and exploration, while 3D-Mem [66] constructs a 3D scene memory from multi-view images for exploration and reasoning. More recently, 3DLLM-Mem [25] equips 3D LLMs with long-term memory to support diverse embodied tasks. However, this approach has limitations: its unpruned memory accumulates redundant entries and incurs high computational cost on long videos. In contrast, our work introduces a bounded dual-memory module to address these challenges. This module couples a sliding-window working memory with a fixed-capacity episodic memory, preserving salient information for efficient spatial reasoning.

3 Method

Overview We introduce VLM², a vision-language model that takes a monocular video $\mathcal{V} = \{I_t\}_{t=1}^N$ and a language instruction Q as input, and generates answers by using a memory mechanism built directly upon view-consistent 3D-aware representations. An overview of our architecture is shown in Fig. 2. Our method is built on two technical key innovations. First, to address semantic-geometric misalignment, we explicitly align visual, geometry, and view tokens into a view-consistent 3D-aware representation (Sec. 3.1) that maintains spatial consistency under camera motion. Second, building on this consistent 3D-aware representation, we introduce a dual-memory module (Sec. 3.2) that enables reasoning over both immediate and earlier observations. This allows the model to reason about spatial layouts and objects no longer in view, supporting temporally aware and consistent spatial reasoning. We implement VLM² on top of LLaVA-Video [74]. The 3D-aware, memory-enhanced representations are concatenated with the instruction embeddings and fed into the language backbone to generate answers.

3.1 View-Consistent 3D-Aware Representation

Given a video, the goal is to produce a view-consistent, globally coherent 3D-aware representation built from varying viewpoints, in which local semantic and geometric information is aligned for improved spatial reasoning. First, we

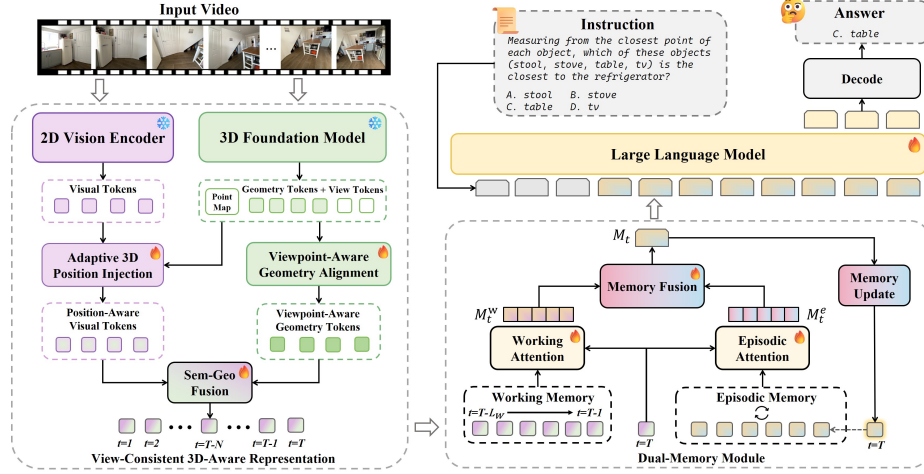


Fig. 2: Overview of the VLM² Architecture. Our model constructs a view-consistent 3D-aware representation via adaptive 3D position injection, viewpoint-aware geometry alignment and semantic-geometric fusion. A dual-memory module with a sliding-window working memory and a fixed-capacity episodic memory maintains these representations over time, supporting temporally aware spatial reasoning.

uniformly sample N frames $\{I_t\}_{t=1}^N$ from the input video, where $I_t \in \mathbb{R}^{H \times W \times 3}$. From each frame I_t , we extract visual tokens $F_t \in \mathbb{R}^{h \times w \times c}$ capturing semantic information from a pretrained vision encoder, and geometric priors from a 3D foundation model (π^3 [56]): geometry tokens $G_t \in \mathbb{R}^{h \times w \times c_g}$, view tokens $Z_t \in \mathbb{R}^{h \times w \times c_v}$, and per-pixel point maps $X_t \in \mathbb{R}^{H \times W \times 3}$. However, directly fusing these features suffers from a core semantic-geometric misalignment: the visual tokens F_t lack spatial grounding, while the geometry tokens G_t are viewpoint-ambiguous. We introduce three core modules to address this issue: (1) *Adaptive 3D Position Injection* injects predicted 3D coordinates into visual tokens while handling noisy predictions; (2) *Viewpoint-Aware Geometry Alignment* aligns geometry tokens with their corresponding view tokens to resolve view ambiguity; and (3) *Semantic-Geometric Fusion* fuses these aligned features into a 3D-aware representation that enables spatial reasoning consistent across camera motion.

Adaptive 3D Position Injection. Recent works such as Video-3D LLM [75] enhance spatial awareness by injecting ground-truth 3D points (obtained from ground-truth depth maps and camera parameters) into the model. This requirement limits their applicability in real-world scenarios, where reliable depth labels are often unavailable. Moreover, we find that directly injecting predicted 3D coordinates (from a monocular depth estimation model) may introduce noise and inaccuracies, potentially affecting performance. To address this, we introduce an adaptive gating mechanism that selectively incorporates reliable and useful 3D position cues (e.g. chairs instead of walls) while filtering out noisy or irrelevant predictions. We first obtain per-patch 3D coordinates $C_t \in \mathbb{R}^{h \times w \times 3}$ by pooling

the predicted point maps X_t from a 3D foundation model, then encode them into 3D position embeddings $C'_t \in \mathbb{R}^{h \times w \times c}$ via a sinusoidal position encoding and a two-layer MLP $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}^c$. The core of our adaptive design is a learnable gate α_t that modulates the injection:

$$F_t^{\text{pa}} = F_t + \alpha_t \odot \phi(C_t), \quad \alpha_t \in [0, 1]^{h \times w \times 1}, \quad (1)$$

where $\odot$ denotes element-wise multiplication. This produces position-aware visual tokens F_t^{pa} grounded in 3D, while mitigating the impact of prediction noise.

Viewpoint-Aware Geometry Alignment. Direct fusing visual and geometric features presents a crucial challenge: the geometric features alone are viewpoint-ambiguous (e.g., a chair’s front and back legs produce similar features). To make the geometric feature viewpoint-aware, we infuse geometry tokens with both patch-level and frame-level perspective cues. First, to disambiguate local geometric patterns, we enrich geometry tokens with the corresponding view tokens, projected via a linear layer $\psi_v : \mathbb{R}^{c_v} \rightarrow \mathbb{R}^{c_g}$ to produce $\hat{Z}_t$. This fusion produces viewpoint-aware features G_t^{va} that make patch-level geometry less ambiguous.

$$G_t^{\text{va}} = \text{MLP}(\text{Concat}[G_t; \hat{Z}_t]). \quad (2)$$

Second, to provide global viewpoint context, we append a global view descriptor $\bar{Z}_t \in \mathbb{R}^{1 \times 1 \times c_v}$, obtained by pooling view tokens Z_t and projecting via a linear layer $\psi_g : \mathbb{R}^{c_v} \rightarrow \mathbb{R}^{c_g}$. This provides a frame-level signal of the camera viewpoint.

$$G_t^{\text{vc}} = \text{Concat}[G_t^{\text{va}}; Z_t^g] \in \mathbb{R}^{(hw+1) \times c_g}. \quad (3)$$

Infused with both patch-specific and frame-level viewpoint information, the resulting geometric tokens G_t^{vc} are now viewpoint-aware across different viewpoints.

Semantic-Geometric Fusion. To form the final 3D-aware representation for each frame I_t , we fuse the position-aware visual tokens and viewpoint-aware geometry tokens through cross-attention. We use the position-aware visual tokens F_t^{pa} as queries and the viewpoint-aware geometry tokens G_t^{vc} as keys and values:

$$H_t = \text{Attn}(Q(F_t^{\text{pa}}), K(G_t^{\text{vc}}), V(G_t^{\text{vc}})). \quad (4)$$

The output H_t is a sequence of powerful 3D-aware tokens for frame I_t that bind visual semantics to consistent geometric structure, resolving the semantic-geometric misalignment and producing a coherent 3D-aware representation that supports reliable spatial reasoning across views.

3.2 Dual-Memory Module

Inspired by CUT3R [54], which maintains a continuously updating state for 3D reconstruction, we propose a dual-memory module that stores and updates 3D-aware representations, enabling temporally spatial reasoning across frames.

Working Memory for Immediate Retrieval. An agent’s immediate environment is dense with information, but not all recent observations are equally relevant. To focus on important short-term context, we design a sliding-window working memory $\mathcal{W}_t$ to store the most recent L_w representations. Our key insight

is that the model should retrieve only what is relevant rather than treating all information equally. We achieve this selective retrieval by using the current representation H_t as queries to attend over the working memory via cross-attention:

$$M_t^w = \text{Attn}(Q(H_t), K(\mathcal{W}_t), V(\mathcal{W}_t)). \quad (5)$$

The resulting feature M_t^w is an enhanced representation of the current state, enriched with relevant context from the immediate past and recent observations.

Episodic Memory for Recall Across Frames. While the working memory captures recent context, it cannot retain crucial information beyond the immediate window, limiting spatial reasoning that depends on earlier observations. For instance, the chair counting example in the introduction requires retrieving observations no longer stored in the sliding window. To overcome this, we introduce an episodic memory $\mathcal{E}_t = \{E_1, E_2, \dots, E_{L_e}\}_{i=1}^{L_e}$, a fixed-capacity bank that stores the most salient observations across frames. Through end-to-end training on QA tasks, the model implicitly learns what constitutes a “task-relevant” memory, guiding it to preserve representations useful for spatial reasoning. Similar to the working memory, we use the current representation H_t to query episodic memory via cross-attention, yielding an episodic-enhanced representation M_t^e .

Memory Fusion and Update. The working-enhanced representation M_t^w is fused with the episodic-enhanced representation M_t^e to produce the final memory-enhanced feature for the current step. We employ a learnable gate γ_t to control the combination of these two information streams:

$$\gamma_t = \sigma(\text{MLP}(\text{Concat}[M_t^w; M_t^e])), \quad (6a)$$

$$M_t = \gamma_t \odot M_t^w + (1 - \gamma_t) \odot M_t^e. \quad (6b)$$

This produces the final memory-enhanced representation M_t . To maintain a diverse and non-redundant episodic memory bank, we replace the most similar existing entry with current memory-enhanced representation M_t . Specifically, we identify the most similar entry $E_{i_t^*}$ in episodic memory and update it with M_t .

$$i_t^* = \text{argmax}_{i \in \{1, \dots, L_e\}} \cos(M_t, E_i). \quad (7)$$

This similarity-based update mechanism ensures episodic memory remains bounded and diverse, supporting temporally aware spatial reasoning across frames.

4 Experiments

Implementation Details. Our model is built on LLaVA-Video-7B [74], a video LLM based on Qwen2-7B [46], and uses π^3 [56] as the 3D foundation model. We train for one epoch on a mixed dataset using the same learning objective as VLM-3R [17]. We use the AdamW optimizer with a batch size of 128 and a peak learning rate of $1e-5$ for the LLM during the warmup phase. For fine-tuning, we apply Low-Rank Adaptation (LoRA [24]) with a rank of 128 and a scaling factor of 256. Both the vision encoder and 3D foundation model are kept fully frozen during training. All experiments are conducted on 8 NVIDIA H200 GPUs.

Table 1: Evaluations on VSI-Bench [64] for 3D spatial reasoning tasks. We compare against proprietary models, open-sourced VLMs, and spatial-enhanced models specifically designed for spatial reasoning. Bold indicates the best performance.

Methods	Rank Avg.		Numerical Qusetion				Multiple-Choice Question					
			Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr.	Order	
Proprietary Models (API)												
GPT-4o	4	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5		
GPT-5	1	55.0	53.3	34.5	73.3	47.5	63.7	48.7	50.3	68.9		
Gemini-2.5-Pro	3	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3		
Gemini-3-Pro	2	52.5	38.0	37.8	72.7	44.1	59.9	55.7	45.9	66.0		
Open-sourced VLMs												
LongVA-7B	13	29.2	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7		
InternVL2-8B	9	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6		
InternVL2-40B	7	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6		
LongVILA-8B	15	21.6	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5		
VILA-1.5-8B	14	28.9	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8		
VILA-1.5-40B	12	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9		
Qwen2.5VL-7B	10	33.0	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8		
Qwen2.5VL-72B	6	37.0	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9		
Qwen3-VL-2B	3	50.4	62.2	40.3	71.5	49.8	52.3	42.0	30.4	54.5		
Qwen3-VL-8B	1	57.9	67.6	47.0	76.3	61.9	58.0	51.0	35.1	66.3		
Qwen3.5-4B	2	53.6	56.5	36.5	67.5	53.8	60.3	57.5	34.0	62.3		
LLaVA-OneVision-7B	11	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4		
LLaVA-OneVision-72B	5	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6		
LLaVA-NeXT-Video-7B	8	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6		
LLaVA-NeXT-Video-72B	4	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6		
Spatial-Enhanced Models												
VG-LLM-4B	5	47.3	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4		
VG-LLM-8B	3	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2		
Spatial-MLLM-4B	4	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3		
VLM-3R-7B	2	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1		
VLM ² -7B (Ours)	1	68.8	72.5	59.6	70.8	69.9	69.0	87.8	52.6	68.3		
Improve ↑	-	+7.9	+2.3	+10.2	+1.6	+2.8	+3.6	+7.3	+7.2	+28.2		

4.1 Spatial Reasoning Benchmarks

Datasets and Metrics. We evaluate our model’s spatial reasoning performance on VSI-Bench [64], which contains over 5,000 QA pairs collected from egocentric videos in ScanNet [14], ScanNet++ [67], and ARKitScenes [4]. We further report results on VSTI-Bench [17], which is built on the same video sources and comprises approximately 6,000 QA pairs. This benchmark evaluates the model’s ability to interpret spatial layouts and reason about camera-object interactions under viewpoint changes in monocular videos. For Multiple-Choice Answer (MCA) tasks, we report standard *Accuracy* [18,21,70], and for Numerical Answer (NC) tasks, we report *Mean Relative Accuracy* [64].

Baselines. We compare our model with a diverse set of proprietary and open-source VLMs (e.g., GPT-4o [31], Gemini-2.5 Pro [13], LLaVA-NeXT-Video [73], Qwen2.5-VL [3], Qwen3-VL [2]). We also include recent spatial-enhanced models, including SPAR [72], VG-LLM [76], Spatial-MLLM [58], and VLM-3R [17].

Results on VSI-Bench. Table 1 presents the quantitative results on VSI-Bench. Our method consistently outperforms proprietary and open-source VLMs across all task categories. The performance gains are particularly pronounced on *Absolute Distance* and *Relative Direction*, as these tasks demand strong spatial awareness and fine-grained 3D understanding. This suggests that our model

Table 2: Evaluations on VSTI-Bench [17] for 3D spatial-temporal reasoning tasks. *Abbr.:* CO-AbsD = Cam-Obj Abs. Dist.; CDisp = Cam. Displace.; CDir = Cam. Mov. Dir.; OO-RelP = Obj-Obj Rel. Pos.; CO-RelD = Cam-Obj Rel. Dist. Bold indicates best performance in each model category.

Methods	Avg.	Numerical Question		Multiple-Choice Question		
		CO-AbsD	CDisp	CMDir	OO-RelP	CO-RelD
Proprietary Models (API)						
GPT-4o	38.2	29.5	23.4	37.3	58.1	42.5
GPT-5	44.1	24.3	11.1	41.9	85.8	57.3
Gemini-2.5-Pro	42.4	29.2	8.0	30.0	84.0	60.7
Gemini-3-Pro	43.3	30.7	9.7	35.6	82.5	58.1
Open-sourced VLMs						
LongVA-7B	32.3	13.5	5.1	43.7	57.9	41.2
InternVL2-8B	43.5	32.9	13.5	48.0	68.0	55.0
InternVL2-40B	43.2	11.9	34.9	33.3	63.8	72.2
LongVILA-8B	30.5	20.0	11.6	35.4	52.3	33.4
VILA-1.5-8B	37.3	30.1	27.3	42.2	50.4	36.7
VILA-1.5-40B	38.2	28.2	15.7	28.8	65.4	53.0
Qwen2.5VL-7B	38.2	22.9	4.9	47.4	65.9	49.9
Qwen2.5VL-72B	40.3	18.0	10.0	41.0	74.2	58.4
Qwen3-VL-2B	45.4	30.2	35.0	37.9	65.0	59.0
Qwen3-VL-8B	47.0	26.1	31.8	37.1	79.7	60.1
Qwen3.5-4B	32.7	26.0	24.8	24.9	71.9	16.0
LLaVA-OneVision-7B	41.7	29.9	19.3	47.5	62.1	49.8
LLaVA-NeXT-Video-7B	40.0	28.2	1.8	49.8	64.7	55.6
LLaVA-NeXT-Video-72B	44.0	32.3	10.5	48.1	78.3	50.9
Spatial-Enhanced Models						
VLM-3R-7B	58.8	39.4	39.6	60.6	86.5	68.6
VLM ² -7B (Ours)	65.3	43.1	44.1	76.8	87.7	74.9
Improve ↑	+6.5	+3.7	+4.5	+16.2	+1.2	+6.3

learns view-consistent and spatially-aware representations that effectively capture geometric cues. Moreover, our method attains state-of-the-art results on *Route Plan* and *Appearance Order*. These tasks involve reasoning over changing viewpoints and maintaining consistent object relationships, highlighting the effectiveness of our representation and dual-memory module design.

Results on VSTI-Bench. As shown in Table 2, we evaluate our model on VSTI-Bench, which assesses spatial reasoning under temporally evolving dynamics in monocular videos. Our method achieves state-of-the-art performance across all tasks, with an overall accuracy of 65.3, a relative improvement of 11.1% over the previous best method VLM-3R (58.8), reflecting consistent gains on both spatial and temporal reasoning categories. We attribute these gains to our view-consistent 3D-aware representation and dual-memory design, which support reasoning about camera motion and object interactions across frames.

4.2 3D Scene Understanding Benchmarks

Datasets and Metrics. We further evaluate our approach on 3D question answering benchmarks. We evaluate on ScanQA [1] for 3D spatial reasoning and SQA3D [41] for situated reasoning, both built on the ScanNet [14]. We follow the

Table 3: Evaluation on ScanQA [1] and SQA3D [41] for 3D understanding tasks. “C” stands for “CIDEr”, “B-4” for “BLEU-4”, “M” for “METEOR”, “R” for “ROUGE”, and “EM-1” for top-1 exact match. **Bold** and underline denote the best-performing and second-best performing models in each category, respectively.

Methods	Video Input	ScanQA (val)					SQA3D (test)
		C	B-4	M	R	EM-1	EM-1
<i>Task-Specific Models</i>							
ScanQA [1]	✗	<u>64.9</u>	10.1	13.1	33.3	<u>21.1</u>	<u>47.2</u>
SQA3D [41]	✗	-	11.2	<u>13.5</u>	<u>34.5</u>	-	46.6
3D-VisTA [78]	✗	69.6	<u>10.4</u>	13.9	35.7	22.4	48.5
<i>3D/2.5D-Input Models</i>							
3D-LLM [23]	✗	69.4	12.0	14.5	35.7	20.5	-
Chat-3D v2 [27]	✗	87.6	14.0	-	-	-	54.7
LL3DA [8]	✗	76.8	13.5	15.9	37.3	-	-
ChatScene [26]	✗	87.7	14.3	18.0	41.6	21.6	54.6
LLaVA-3D [77]	✗	103.1	16.4	<u>20.8</u>	49.6	<u>30.6</u>	60.1
Video-3D LLM [75]	✗	102.1	16.4	20.0	49.3	30.1	58.6
3DRS [30]	✗	<u>104.8</u>	<u>17.2</u>	20.5	<u>49.8</u>	30.3	<u>60.6</u>
Ross3D [51]	✗	107.0	17.9	20.9	50.7	30.8	63.0
<i>Video-Input Models</i>							
InternVL2-8B [10]	✓	62.5	3.3	14.5	34.3	-	33.0
Qwen2-VL-7B [3]	✓	53.9	3.0	11.4	29.3	-	46.5
Qwen2-VL-72B [3]	✓	66.9	12.0	13.0	35.2	-	47.0
LLaVA-Video-7B [74]	✓	88.7	3.1	17.7	44.6	-	48.5
Oryx-34B [39]	✓	72.3	-	15.0	37.3	-	-
Spatial-MLLM-4B [58]	✓	<u>91.8</u>	<u>14.8</u>	<u>18.4</u>	<u>45.0</u>	-	<u>55.9</u>
VLM²-7B (Ours)	✓	105.5	17.7	20.5	50.3	30.7	60.4
Improve ↑		+13.7	+2.9	+2.1	+5.3	-	+4.5

standard evaluation settings for each benchmark and use widely adopted evaluation metrics to assess answer quality in a consistent manner. For ScanQA, we report exact match accuracy (EM), CIDEr, BLEU-4, METEOR, and ROUGE-L. For SQA3D, we evaluate performance using exact match accuracy (EM).

Baselines. We compare against models covering different input settings and task specializations. Task-specific models including ScanQA [1], SQA3D [41], and 3D-VisTA [78] are designed for 3D question answering tasks. We also compare with methods that operate on explicit 3D or 2.5D inputs, such as point clouds or depth maps, including Video-3D LLM [75], 3DRS [30], and Ross3D [51]. In addition, we compare with video-based models such as SPAR [72] and VG-LLM [76], which process multi-view video observations.

Results. We present the quantitative results on the ScanQA and SQA3D benchmarks in Table 3. Our method achieves state-of-the-art performance among video-based models on both benchmarks and also outperforms task-specific models. Compared with approaches that leverage explicit 3D or 2.5D inputs, Ross3D achieves better performance, as it exploits additional point clouds to render BEV images for reconstructive supervision. Our approach surpasses other 3D/2.5D input models, such as Video-3D LLM and 3DRS. We attribute these gains to our model’s ability to explicitly address semantic-geometric misalignment when con-

Table 4: Zero-shot generalization on CV-Bench [47], SPAR-Bench [72], and BLINK [19]. We report accuracy without additional fine-tuning.

Methods	Avg.	CV-Bench			SPAR-Bench				BLINK
		2D	3D	Avg.	Low	Medium	High	Avg.	Avg.
GPT-4o [31]	57.3	69.4	81.3	75.4	29.3	24.9	45.1	36.4	60.0
Qwen2.5VL-3B [3]	47.3	69.1	72.2	70.6	15.3	26.4	32.2	24.6	46.6
Qwen2.5VL-7B [3]	55.0	75.0	83.1	79.0	17.5	29.5	41.8	30.2	55.9
VG-LLM-8B [76]	<u>62.4</u>	72.2	91.1	81.7	<u>39.6</u>	<u>54.0</u>	<u>71.7</u>	<u>54.0</u>	51.5
VLM²-7B (Ours)	71.1	<u>73.6</u>	<u>85.2</u>	<u>79.4</u>	58.6	71.2	78.8	66.8	67.2

structuring 3D-aware representations from video. Our viewpoint-aware geometry alignment and adaptive 3D position injection jointly enforce alignment between visual and geometry tokens, leading to view-consistent 3D-aware representations.

4.3 Analysis and Ablation Studies

In this section, we provide further analysis of VLM², including its zero-shot generalization, performance across video durations, and detailed ablation studies. We first evaluate it on additional benchmarks without fine-tuning, then analyze model performance across varying video durations, and finally examine the contribution of key components and memory configurations.

Zero-Shot Generalization. To evaluate zero-shot generalization, we assess VLM² on three additional benchmarks: CV-Bench [47], SPAR-Bench [72], and BLINK [19]. Importantly, we do not perform any additional fine-tuning on these benchmarks. As shown in Table 4, VLM² achieves competitive performance across all three benchmarks and attains the highest average results. These results suggest that the proposed design learns transferable 3D-aware representations that generalize effectively beyond specific spatial reasoning benchmarks.

Performance Across Video Durations. To analyze how performance varies with video duration, we partition the input videos in VSI-Bench [64] and VSTI-Bench [17] into three groups based on length: Short (< 1 min), Mid (1-2 min), and Long (> 2 min). Table 5 shows the comparisons. On VSI-Bench, VLM-3R’s performance degrades as video length increases, indicating challenges in maintaining consistent spatial reasoning under longer sequences, whereas our model shows more stable performance across durations. On VSTI-Bench, while VLM-3R’s performance remains relatively stable across lengths, our model achieves higher accuracy in all groups, with larger gains on long videos. Overall, VLM² improves average accuracy by +6.2 and +5.9 points on VSI-Bench and VSTI-Bench, respectively, with the larger gains on long videos. These results suggest that our view-consistent 3D-aware representation and dual-memory design contribute to reliable spatial reasoning across varying video durations.

Overall Component Analysis. Our component ablation studies are summarized in Table 6. We establish a strong baseline by fine-tuning LLaVA-NeXT-Video-7B [73] with the same language backbone, training schedule, and compute budget for all variants. Incorporating our 3D-aware representation into the

Table 5: Spatial reasoning performance on VSI-Bench [64] and VSTI-Bench [17] across varying video durations. Videos are grouped into Short (< 1 min), Mid (1–2 min), and Long (> 2 min).

Methods	VSI-Bench				VSTI-Bench			
	Avg.	Short	Mid	Long	Avg.	Short	Mid	Long
<i>Open-sourced VLMs</i>								
Qwen2.5VL-7B [3]	29.5	34.3	28.3	28.4	45.9	43.3	48.4	45.5
Qwen2.5VL-72B [3]	36.3	39.2	36.6	34.7	49.4	45.1	50.8	55.2
LLaVA-NeXT-Video-7B [73]	36.8	38.5	37.3	35.5	46.4	43.0	47.9	50.1
LLaVA-NeXT-Video-72B [73]	41.1	41.3	40.7	41.4	51.3	47.6	51.9	57.6
<i>Spatial-Enhanced Models</i>								
Spatial-MLLM-4B [58]	48.1	49.8	48.3	47.2	-	-	-	-
VG-LLM-4B [76]	47.4	49.1	48.7	45.2	-	-	-	-
VG-LLM-8B [76]	50.5	51.9	49.7	50.6	-	-	-	-
VLM-3R-7B [17]	<u>63.2</u>	<u>67.1</u>	<u>64.7</u>	<u>60.0</u>	<u>63.6</u>	<u>61.1</u>	<u>64.7</u>	<u>66.4</u>
VLM²-7B (Ours)	69.4	71.0	70.0	68.1	69.5	65.9	70.6	74.6
Improve $\uparrow$	+6.2	+3.9	+5.3	+8.1	+5.9	+4.8	+5.9	+8.2

Table 6: Ablation of model components. We evaluate the effect of 3D-aware representation, working memory, and episodic memory. Each module contributes complementary improvements, and combining all yields the best overall performance.

3D-Aware Rep.	Work. Mem.	Epis. Mem.	Avg.	Numerical Question		Multiple-Choice Question		
				Abs. Dist.	Room Size	Rel. Dist.	Rel. Dir.	Route Plan
$\times$	$\times$	$\times$	55.2	43.3	62.4	62.1	67.8	40.2
$\checkmark$	$\times$	$\times$	63.8	52.9	67.5	65.6	84.8	48.5
$\checkmark$	$\checkmark$	$\times$	65.9	56.3	68.3	67.3	86.2	51.3
$\checkmark$	$\times$	$\checkmark$	66.1	57.7	68.6	67.6	85.9	50.5
$\checkmark$	$\checkmark$	$\checkmark$	67.8	59.6	69.9	69.0	87.8	52.6

baseline yields an average of 8.6% accuracy gain. Building on this representation, introducing working memory or episodic memory brings additional gains of 2.1% and 2.3%, respectively. When both memory modules are combined, our full model achieves the best performance, improving by 4.0% over the 3D-aware representation and by 12.6% over the baseline. The benefit is particularly pronounced on tasks such as *Route Plan* (+ 12.4), which requires spatial structure awareness, highlighting the importance of an explicit memory mechanism. These findings suggest that the view-consistent 3D-aware representation and the dual-memory module are crucial components for advanced spatial reasoning.

Effectiveness of 3D-Aware Representation. We ablate the choice of 3D backbone and fusion strategy on top of the aforementioned baseline, as shown in Table 7. To incorporate geometry tokens from 3D foundation models into visual tokens, we compare two fusion strategies. A simple *Concat-MLP* fusion, which concatenates 2D visual tokens with 3D geometry tokens followed by an MLP, yields lower performance (40.4–43.3), indicating that naïve concatenation leaves semantic-geometric misalignment unresolved. Replacing Concat-MLP

Table 7: Ablation of 3D-aware representations. We compare different 3D foundation models (VGGT [53], CUT3R [54], π^3 [56]) and fusion strategies (*Concat-MLP*, *Cross-Attn*). Built on π^3 and *Cross-Attn*, our viewpoint-aware geometry alignment (VAGA) and adaptive 3D position injection (A3PI) achieve the best results.

Methods	Avg.	Numerical Question		Multiple-Choice Question		
		Abs. Dist.	Room Size	Rel. Dist.	Rel. Dir.	Route Plan
<i>Baseline</i>						
LLaVA-NeXT-Video-SFT	55.2	43.3	62.4	62.1	67.8	40.2
<i>Sem-Geo Fusion (Concat-MLP)</i>						
Baseline + VGGT	41.4	38.6	45.6	38.7	47.6	36.6
Baseline + CUT3R	43.3	39.8	51.4	41.3	45.9	38.1
Baseline + π^3	40.4	38.3	42.5	41.0	45.2	35.1
<i>Sem-Geo Fusion (Cross-Attn)</i>						
Baseline + VGGT	60.2	51.0	63.9	61.4	81.8	42.7
Baseline + CUT3R	59.4	50.2	63.1	62.8	79.3	41.8
Baseline + π^3	61.0	51.9	64.7	63.1	81.9	43.3
<i>VLM² (Ours)</i> [3D Foundation: π^3 ; Sem-Geo Fusion: Cross-Attn]						
Baseline + A3PI (w/o adapt.)	58.9	50.3	64.5	60.5	77.2	41.9
Baseline + A3PI	61.6	52.3	65.3	63.7	82.1	44.7
Baseline + VAGA	62.9	52.5	66.4	64.9	84.0	46.6
Baseline + A3PI + VAGA	63.8	52.9	67.5	65.6	84.8	48.5

with a *Cross-Attn* fusion using 3D backbones (VGGT, CUT3R, π^3) raises performance to 59.4–61.0, showing that geometric cues help with a stronger fusion strategy but still fall short of our approach. Our viewpoint-aware geometry alignment (VAGA) with adaptive 3D position injection (A3PI) achieves the best accuracy, boosting performance from 61.0 to 63.8. Removing the adaptive mechanism from A3PI reduces the performance to 58.9, indicating that injecting position information into all visual tokens introduces noise in irrelevant regions and distorts feature distributions, highlighting the importance of adaptive injection.

Ablation of Dual-Memory Length. We study the effect of memory length by varying the working memory window L_w and episodic memory capacity L_e .

As reported in Table 8, increasing episodic memory capacity L_e improves performance up to $L_e = 32$, highlighting the role of episodic capacity. In contrast, a larger working memory window L_w is not always helpful. Working memory attends to recent frames, so a short window may miss spatial cues needed to disambiguate nearby objects, whereas a long window introduces irrelevant tokens that dilute attention. We find that $L_w = 8$, $L_e = 32$ achieves the best performance, indicating an effective balance between the two memory components.

5 Conclusion

In this paper, we propose VLM², a vision-language model for video-based spatial reasoning that addresses two critical challenges: semantic-geometric mis-

Table 8: Ablation of dual-memory length. We vary the length of working memory (L_w) and episodic memory (L_e) on VSI-Bench to study the effect of memory capacity. The setting $(L_w, L_e)=(8, 32)$ achieves the best overall performance.

Memory Size		VSI-Bench		
L_w	L_e	Numerical	Multiple-Choice	Avg.
4	8	64.1	58.5	61.3
8	8	65.6	66.9	66.3
8	16	67.1	67.4	67.3
8	32	68.2	69.4	68.8
16	8	65.2	65.7	65.5
16	16	67.3	67.9	67.6
16	32	67.9	68.6	68.3

alignment and the absence of persistent memory. VLM² constructs a view-consistent 3D-aware representation from video by adaptively grounding visual features into 3D space and enforcing cross-view consistency, enabling coherent 3D understanding. On top of this representation, we introduce a dual-memory module that combines a sliding-window working memory for immediate context with a fixed-capacity episodic memory for recall across frames, allowing efficient temporally aware spatial reasoning. Extensive experiments across multiple spatial reasoning benchmarks demonstrate that our method achieves state-of-the-art performance, advancing the frontier of visual-spatial intelligence.

References

1. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19129–19139 (2022)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. arXiv preprint arXiv:2111.08897 (2021)
5. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
6. Cen, J., Zhou, X., Fang, J., Wen, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Tackling view-dependent semantics in 3d language gaussian splatting. In: Forty-second International Conference on Machine Learning (2025)

7. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
8. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: 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. pp. 26428–26438 (2024)
9. Chen, Y., Qi, Z., Zhang, W., Jin, X., Zhang, L., Liu, P.: Reasoning in space via grounding in the world. arXiv preprint arXiv:2510.13800 (2025)
10. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
11. Chen, Z., Qin, M., Yuan, T., Liu, Z., Zhao, H.: Long3r: Long sequence streaming 3d reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5273–5284 (October 2025)
12. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems* **37**, 135062–135093 (2024)
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
14. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
15. Deng, J., He, T., Jiang, L., Wang, T., Dayoub, F., Reid, I.: 3d-llava: Towards generalist 3d llms with omni superpoint transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3772–3782 (2025)
16. Diomataris, M., Gkanatsios, N., Pitsikalis, V., Maragos, P.: Grounding consistency: Distilling spatial common sense for precise visual relationship detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15911–15920 (2021)
17. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Wang, P., Qu, H., Zhou, S., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 31054–31065 (2026)
18. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24108–24118 (2025)
19. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
20. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13504–13514 (2024)

21. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
23. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems* **36**, 20482–20494 (2023)
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *ICLR* **1**(2), 3 (2022)
25. Hu, W., Hong, Y., Wang, Y., Gao, L., Wei, Z., Yao, X., Peng, N., Bitton, Y., Szpektor, I., Chang, K.W.: 3d-llm-mem: Long-term spatial-temporal memory for embodied 3d large language model. *Advances in Neural Information Processing Systems* **38**, 67856–67884 (2026)
26. Huang, H., Chen, Y., Wang, Z., Huang, R., Xu, R., Wang, T., Liu, L., Cheng, X., Zhao, Y., Pang, J., et al.: Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems* **37**, 113991–114017 (2024)
27. Huang, H., Wang, Z., Huang, R., Liu, L., Cheng, X., Zhao, Y., Jin, T., Zhao, Z.: Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *CoRR* (2023)
28. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.C., Jia, B., Huang, S.: An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871* (2023)
29. Huang, J., Hu, X., Han, B., Shi, S., Tian, Z., He, T., Jiang, L.: Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198* (2025)
30. Huang, X., Wu, J., Xie, Q., Han, K.: 3drs: Mllms need 3d-aware representation supervision for scene understanding. In: *Conference on Neural Information Processing Systems* (2025)
31. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
32. Khanna, M., Ramrakhya, R., Chhablani, G., Yenamandra, S., Gervet, T., Chang, M., Kira, Z., Chaplot, D.S., Batra, D., Mottaghi, R.: Goat-bench: A benchmark for multi-modal lifelong navigation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16373–16383 (2024)
33. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* **62**(1), 1–62 (2022)
34. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-onevision: Easy visual task transfer. *TMLR* (2025)
35. Liao, G., Li, J., Bao, Z., Ye, X., Li, Q., Liu, K.: Clip-gs: Clip-informed gaussian splatting for view-consistent 3d indoor semantic understanding. *ACM Transactions on Multimedia Computing, Communications and Applications* **21**(8), 1–24 (2025)
36. Liao, G., Zhou, K., Bao, Z., Liu, K., Li, Q.: Ov-nerf: Open-vocabulary neural radiance fields with vision and language foundation models for 3d semantic understanding. *IEEE Transactions on Circuits and Systems for Video Technology* (2024)
37. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoybi, M., Han, S.: Vila: On pre-training for visual language models. In: *CVPR* (2024)

38. Liu, Y., Ma, M., Yu, X., Ding, P., Zhao, H., Sun, M., Huang, S., Wang, D.: Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *Advances in Neural Information Processing Systems* **38**, 123926–123958 (2026)
39. Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., Rao, Y.: Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. In: *International Conference on Learning Representations*. vol. 2025, pp. 85485–85507 (2025)
40. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6924–6934 (2025)
41. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. In: *International Conference on Learning Representations* (2023)
42. OpenAI: Gpt-4o (2024), <https://openai.com/index/hello-gpt-4o/>
43. Shi, J.C., Wang, M., Duan, H.B., Guan, S.H.: Language embedded 3d gaussians for open-vocabulary scene understanding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5333–5343 (2024)
44. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 15768–15780 (2025)
45. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18221–18232 (2024)
46. Team, Q., et al.: Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024)
47. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems* **37**, 87310–87356 (2024)
48. Wang, C., Ji, K., Geng, J., Ren, Z., Fu, T., Yang, F., Guo, Y., He, H., Chen, X., Zhan, Z., et al.: Imperative learning: A self-supervised neuro-symbolic learning framework for robot autonomy. *The International Journal of Robotics Research* p. 02783649251353181 (2024)
49. Wang, C., Qiu, Y., Wang, W., Hu, Y., Kim, S., Scherer, S.: Unsupervised online learning for robotic interestingness with visual memory. *IEEE Transactions on Robotics* **38**(4), 2446–2461 (2021)
50. Wang, C., Wang, W., Qiu, Y., Hu, Y., Scherer, S.: Visual memorability for robotic interestingness via unsupervised online learning. In: *European Conference on Computer Vision*. pp. 52–68. Springer (2020)
51. Wang, H., Zhao, Y., Wang, T., Fan, H., Zhang, X., Zhang, Z.: Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9275–9286 (2025)
52. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061* (2024)
53. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5294–5306 (2025)
54. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 10510–10522 (2025)

55. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
56. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3 : Scalable permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
57. Wang, Z., Huang, H., Zhao, Y., Zhang, Z., Zhao, Z.: Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. arXiv preprint arXiv:2308.08769 (2023)
58. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mlm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems* **38**, 13569–13597 (2026)
59. Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., Wetzstein, G.: Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems* **38**, 49371–49393 (2026)
60. Wu, Y., Zheng, W., Zhou, J., Lu, J.: Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. arXiv preprint arXiv:2507.02863 (2025)
61. Xu, Z., He, Z., Wu, J., Song, S.: Learning 3d dynamic scene representations for robot manipulation. In: Conference on Robot Learning (CoRL) (2020)
62. Yang, F., Frivik, P., Hoeller, D., Wang, C., Cadena, C., Hutter, M.: Spatially-enhanced recurrent memory for long-range mapless navigation via end-to-end reinforcement learning. *The International Journal of Robotics Research* p. 02783649251401926 (2025)
63. Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., Gu, Y., Cai, M., Ye, S., Jang, J., et al.: Magma: A foundation model for multimodal ai agents. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14203–14214 (2025)
64. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
65. Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., et al.: Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. arXiv preprint arXiv:2502.09560 (2025)
66. Yang, Y., Yang, H., Zhou, J., Chen, P., Zhang, H., Du, Y., Gan, C.: 3d-mem: 3d scene memory for embodied exploration and reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17294–17303 (2025)
67. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
68. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
69. Yuan, Z., Peng, Y., Ren, J., Liao, Y., Han, Y., Feng, C.M., Zhao, H., Li, G., Cui, S., Li, Z.: Empowering large language models with 3d situation awareness. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19435–19445 (2025)
70. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal

- understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
71. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Dai, J., Jin, X.: Flash-vstream: Memory-based real-time understanding for long video streams. arXiv preprint arXiv:2406.08085 (2024)
 72. Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. arXiv preprint arXiv:2503.22976 (2025)
 73. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (2024)
 74. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
 75. Zheng, D., Huang, S., Wang, L.: Video-3d llm: Learning position-aware video representation for 3d scene understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8995–9006 (2025)
 76. Zheng, D., Li, Y., Wang, L., et al.: Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems* **38**, 20560–20586 (2026)
 77. Zhu, C., Wang, T., Zhang, W., Pang, J., Liu, X.: Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4295–4305 (2025)
 78. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2911–2921 (2023)
 79. Zhu, Z., Wang, X., Li, Y., Zhang, Z., Ma, X., Chen, Y., Jia, B., Liang, W., Yu, Q., Deng, Z., et al.: Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8120–8132 (2025)