BROWN
Computer Science

Probing VideoLLMs for 3D Awareness

Bozheng Li

Brown University

{bozheng_li}@brown.edu

Abstract

Multimodal Large Language Models (MLLMs) now dominate general vision–language tasks, yet robust spatial reasoning remains challenging and is crucial for visual intelligence. To solve 3D question answering (3DQA), a basic task for spatial understanding, recent work has built 3D-aware Video Large Language Models (VideoLLMs) via 3D input enhancement; however, flawed benchmarks leave their true 3D awareness unclear. We therefore evaluate 3D-aware VideoLLMs on the high-quality Beacon3D benchmark and probe whether pure video–language training without 3D data induces intrinsic 3D awareness using feature probes and attention analyses. We reveal several key findings. (i) 3D-aware VideoLLMs show limited generalization, indicating that injected 3D information rarely yields robust 3D awareness, except under reconstruction-based vision supervision. (ii) Video–language training induces cross-frame, distributed spatial modeling that strengthens 3D awareness, and spatial mid-training can further boost performance by compensating for missing cross-view, richly annotated spatial supervision, reaching SOTA zero-shot results on Beacon3D. (iii) Spatial invariance, the ability to maintain robust spatial understanding regardless of view order, is a vital but overlooked facet of 3D awareness in VideoLLMs. Architectural and training biases make VideoLLMs sensitive to frame order, and simple frame-shuffle augmentation during training only partially mitigates this.

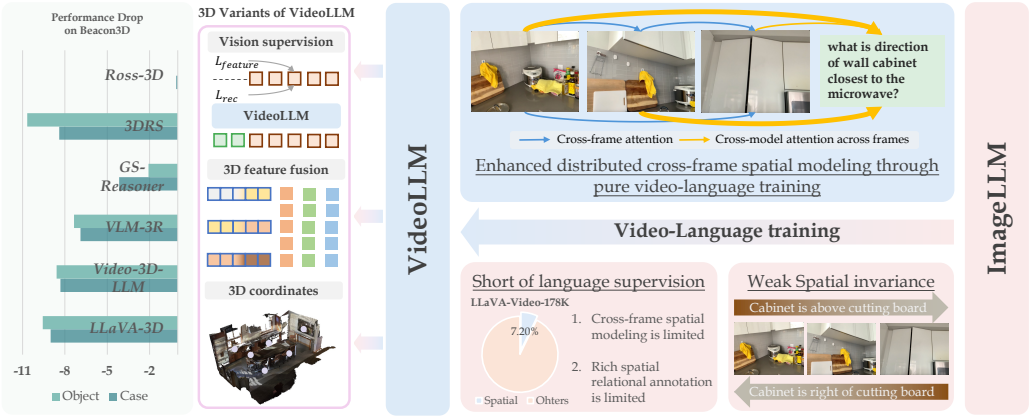


Figure 1 | **Left.** Most 3D-augmented VideoLLMs fail to generalize on the high-quality Beacon3D benchmark Huang et al. (2025a). **Right.** video–language training strengthens distributed cross-frame modeling, yielding 3D awareness without explicit 3D data.

1. Introduction

Spatial understanding Azuma et al. (2022); Chen et al. (2020b); Huang et al. (2025a); Ma et al. (2022); Majumdar et al. (2024); Yang et al. (2025b) is a foundational competency for intelligent systems operating in the physical world, yet reliably bridging spatial perception with language-based reasoning remains challenging. A basic task of spatial understanding is 3D question answering (3D-QA): the model must understand a 3D scene and answer relational questions in natural language Azuma et al. (2022); Huang et al. (2025a); Ma et al. (2022). Earlier 3D-QA systems were limited by their language backbones Huang et al. (2025a); Jia et al. (2024); Zhu et al. (2023, 2024b). Later 3D-LLM approaches fuse 3D features with stronger LLMs Hong et al. (2023); Huang et al. (2024, 2023), but the cost of well-curated 3D data makes scaling and alignment to language backbones harder than in 2D Xue et al. (2023).

Against this backdrop, Video-LLMs Zhang et al. (2024a) enable spatial understanding without explicit 3D inputs. Recent work inject 3D awareness to VideoLLM via 3D positional encodings Zheng et al. (2025b); Zhu et al. (2024a), geometry-aware feature fusion Chen et al. (2025b); Fan et al. (2025), or 3D-aware vision supervision Wang et al. (2025a), reporting steady gains. However, commonly used benchmarks contain substantial flawed content Huang et al. (2025a), which undermines the validity of reported improvements, leading to the first question: *To what extent do the reported gains reflect robust spatial intelligence of 3D-aware VideoLLM?*

To answer the question, we select Beacon3D Huang et al. (2025a), an open-ended spatial understanding benchmark with verified and unique answers, to probe the 3D awareness of 3D-aware VideoLLM in a zero-shot manner. To the best of our knowledge, we present the first systematic evaluation of 3D-aware Video-LLMs on Beacon3D, which is originally designed for 3D-LLMs with explicit 3D inputs. As shown in Figure 1, zero-shot probing reveals that all 3D-aware VideoLLMs lag behind LLaVA-Video Zhang et al. (2024a) on in-domain scenes with novel QA and on out-of-domain scene splits, with up to a 10% overall performance drop on Beacon3D. Meanwhile, we also find that reconstruction-based vision supervision effectively regularizes training and improves the generalization of spatial understanding.

Despite extensive efforts, explicitly integrating 3D awareness into VideoLLMs has not yielded robust 3D understanding and still depends heavily on explicit 3D inputs Chen et al. (2025b); Fan et al. (2025); Wang et al. (2025a); Zheng et al. (2025b); Zhu et al. (2024a). This motivates our second question: *Is it possible for VideoLLM to directly acquire better spatial modeling and 3D awareness through pure video training?* Holding the architecture and the pre-training stage fixed, we compare two representative systems, LLaVA-OneVision-SI Li et al. (2024) and LLaVA-Video Zhang et al. (2024a), where LLaVA-OneVision-SI is trained with purely image–language supervision, and LLaVA-Video is initialized from it and further trained on video data. We probe the emergence of 3D awareness along three axes: (i) Vision encoder: apply established probing protocols El Banani et al. (2024) to measure 3D awareness before and after video training. (ii) LLM reasoning: use attention knockout and weight analyses to reveal cross-frame spatial modeling inside the LLM. (iii) Data supervision: analyze composition of video-language training data to identify the sources of spatial supervision. Our analysis unveil that video training preserves a broadly similar vision feature space, but develops more distributed cross-view modeling in early LLM layers, aggregating information more uniformly across frames and performing critical global spatial composition within the first 6 layers. From a data perspective, we find a non-trivial yet insufficient proportion of spatially supervised samples, which helps explain the large out-of-domain drop when SFT relies on simple QA data, suggesting that spatial understanding, a foundational capability to be acquired during pretraining, remains limited due to absence of dense spatial supervision.

Table 1 | Summary of 3D-aware Video-LLM selection, configurations, and comparison. Here ScanNet-223K refer to the combination of 5 ScanNet-based training set from ScanQA Azuma et al. (2022), SQA3D Ma et al. (2022), Scan2Cap Chen et al. (2020b), ScanRefer Chen et al. (2020a) and Mutli3DRefer Zhang et al. (2023b). †All list models are initialized from LLaVA-Video, except for LLaVA-3D, which only open-sourced checkpoint based on LLaVA-1.5 Liu et al. (2024a)

Model	3D Input	Training Data	Trainable Parameters
†LLaVA-3D Zhu et al. (2024a)	3D embedding	LLaVA-3D-Instruct-1M	3D embed., VL-proj., LLM
Video-3D-LLM Zheng et al. (2025b)	3D coordinates	ScanNet-223K	Vision enc., VL-proj., LLM
Ross-3D Wang et al. (2025a)	3D coordinates	ScanNet-223K	Projector, LLM
VLM-3R Fan et al. (2025)	CUT3R Wang et al. (2025c) feat. fusion	VSTI-138K	Spatial fusion mod., VL-proj., LLM
3D-RS Huang et al. (2025b)	VGGT Wang et al. (2025b) feat. superv.	ScanNet-223K	Vision enc., VL-proj., LLM
GS-Reasoner Chen et al. (2025b)	PTv3 Wu et al. (2024) feat. fusion	GCoT-156K, ScanNet-223K	Spatial fusion mod., VL-proj., LLM

Guided by the above observation, we apply a simple yet effective training strategy: continue mid-training of VideoLLMs on spatially enriched captions. Concretely, we curate Scan2Cap-Video-36K from Scan2Cap Chen et al. (2020b), transforming visual grounding samples into spatially enriched mid-training caption data through visual prompting, projecting 3D boxes into videos. With 36K mid-training samples, our model improves zero-shot performance on Beacon3D by 1.25% on case accuracy and 3.06% on object accuracy, achieving state-of-the-art performance through pure video training, meanwhile reaching higher object-consistency.

While mid-training strengthens spatial reasoning, we uncover a complementary weakness: spatial variance. For a fixed scene and question, response of VideoLLM should be invariant to permutations of frame order Biederman (1987); DiCarlo and Cox (2007); Marr (2010). We formalize an invariance metric Correct Difference Ratio (CDR) to quantify the spatial invariance, and observe degrading spatial invariance after video-language training. We ease the weaken spatial invariance brought by fixed order training by apply a minimal augmentation: randomly shuffling frame order during training for spatial understanding samples irrelevant to temporal order. This reduces variance and yields VideoLLM with stronger spatial invariance.

Our contributions are summarized as follows:

- We diagnose via Beacon3D that current 3D-aware VideoLLMs overfit flawed benchmarks and generalize poorly in spatial understanding outside biased training data, lacking robust spatial understanding capability.
- We show that pure video training induces 3D awareness via distributed cross-view modeling under spatially relevant language supervision, from which we advocate mid-training with spatially enriched captions based on introduced Scan2Cap-Video-36K, achieving state-of-the-art results on Beacon3D without 3D input.
- We identify the spatial variance issue in video training and propose a lightweight data augmentation through random frame shuffling for spatial invariance examples.

2. Related Work

Spatial Understanding Benchmark. Initial work on 3D vision–language problems spanning visual grounding Achlioptas et al. (2020); Chen et al. (2020a); Zhang et al. (2023b), captioning Chen et al. (2020b), and question answering Azuma et al. (2022); Hong et al. (2023); Ma et al. (2022); Ye et al. (2022) laid the groundwork for evaluation by casting them as 3D counterparts to well developed 2D vision–language tasks Antol et al. (2015); Hudson and Manning (2019); Mao et al. (2016); Marino et al. (2019). As large language models advanced rapidly Achiam et al.

(2023), a line of studies Chen et al. (2024b); Chu et al. (2024); Fu et al. (2024b); Hong et al. (2023); Huang et al. (2024, 2023); Kang et al. (2025); Thai et al. (2025); Yang et al. (2024); Zhang et al. (2024b) combined 3D encoders with LLMs to push progress on 3D vision–language tasks. In contrast to the swift expansion of 2D benchmarks that has kept pace with large vision–language models Chen et al. (2024a); Fu et al. (2024a); Li et al. (2023); Liu et al. (2024b); Rahmanzadehgervi et al. (2024); Tong et al. (2024); Yue et al. (2024a); Yuksekgonul et al. (2022), the maturation of stronger 3D benchmarks remains behind. Several efforts Achlioptas et al. (2020); Azuma et al. (2022); Chen et al. (2020a); Ma et al. (2022) extend weaknesses of 3D vision–language evaluation in existing datasets, including hallucination, brittleness under distribution shifts, and language biases Deng et al. (2024); Huang et al. (2025a); Kang et al. (2025); Yang et al. (2025a). In parallel, additional benchmarks Yang et al. (2025b) have been proposed to assess spatial understanding from a video perspective. However, most recent spatially aware multimodal LLMs still report on benchmarks such as ScanQA Azuma et al. (2022) and SQA3D Ma et al. (2022), which have been shown to contain flawed content Huang et al. (2025a), casting doubt on the reported improvements in spatial capability.

VideoLLM for Spatial Understanding. With the rapid progress of VideoLLMs Bai et al. (2025); Li et al. (2024); Lin et al. (2024); Zhang et al. (2023a, 2024a), recent work treats video as a variety of multi-view observations of a scene Yang et al. (2025b) and builds 3D-aware VideoLLMs that inherit generalization from large-scale video–language training. Video-3D-LLM Zheng et al. (2025b) and LLaVA-3D Zhu et al. (2024a) inject explicit 3D coordinates via positional encodings or coordinate projection into embedding space. ROSS3D Wang et al. (2025a) adds reconstruction-based supervision for global spatial modeling. VLM-3R Fan et al. (2025), GS-reasoner Chen et al. (2025b), spatial-MLLM Wu et al. (2025), and VG-LLM Zheng et al. (2025a) fuse features from pretrained 3D encoders with video features, and 3DRS Huang et al. (2025b) supervises using geometry encoder features Wang et al. (2025b). Despite these advances, weaknesses in common benchmarks cast doubt on whether the reported gains reflect robust spatial reasoning, and most methods still depend on explicit 3D inputs, which limits scalability when only pure video data are available.

Interpretability of Vision and Language Model. Several studies propose probing protocols to assess the 3D awareness of vision encoders Chen et al. (2025c); Danier et al. (2025); El Banani et al. (2024); Man et al. (2024); You et al. (2024); Yue et al. (2024b); Zhan et al. (2024). As MLLM models develop, analysis methods such as attention knockout Geva et al. (2023) and attention mechanism analysis Bi et al. (2025); Chen et al. (2025a) from the NLP community are used to study how MLLM processes information. Most of these analysis targets image-based MLLMs Kaduri et al. (2025) and remains limited for VideoLLMs. Recent work on video language modeling Gou et al. (2025); Zhang et al. (2025b) makes progress on general capabilities but lacking a focus on spatial awareness from a video perspective.

3. Probing 3D-aware VideoLLM

3.1. Preliminaries

For 3D QA with VideoLLM, let V denote a video of a static scene and $\mathcal{F} = \{f_i\}_{i=1}^{N}$ be N uniformly sampled RGB frames from V. VideoLLM π_{vid} comprises a vision encoder π_v, a projector π_p, a tokenizer π_t, and a language model π_ℓ. Given a spatial-relevant question Q, the model encodes video tokens $f_v = \pi_p(\pi_v(\mathcal{F}))$ and text tokens $f_q = \pi_t(Q)$, then generates the answer $A = \pi_\ell(f_v, f_q)$.

Table 2 | The performance of 3D-aware VideoLLM on Beacon3d Huang et al. (2025a) benchmark. The base model LLaVA-Video Zhang et al. (2024a) beat all the 3D-aware variant in overall performance except Ross3D Wang et al. (2025a), indicating the weakness on spatial modeling for 3D aware VideoLLM.

Model	ScanNet		MultiScan		3RScan		Overall	
	case	obj.	case	obj.	case	obj.	case	obj.
LLaVA-Video Zhang et al. (2024a)	67.19	29.87	53.88	13.21	59.28	18.75	61.96	22.92
LLaVA-3D Zhu et al. (2024a)	57.55	17.66	41.98	9.43	52.63	9.56	52.88$_{\downarrow 9.1}$	13.36$_{\downarrow 9.6}$
Video-3D-LLM Zheng et al. (2025b)	59.76	19.48	40.51	6.29	52.45	11.76	53.57$_{\downarrow 8.4}$	14.34$_{\downarrow 8.6}$
VLM-3R Fan et al. (2025)	62.49	24.42	41.14	5.66	52.48	9.19	54.99$_{\downarrow 7.0}$	15.69$_{\downarrow 7.2}$
GS-Reasoner Chen et al. (2025b)	68.46	32.99	40.88	5.66	52.36	12.87	57.72$_{\downarrow 4.2}$	20.96$_{\downarrow 2.0}$
3DRS Huang et al. (2025b)	60.84	19.48	39.73	5.00	51.10	6.60	53.48$_{\downarrow 8.5}$	12.38$_{\downarrow 10.5}$
ROSS3D Wang et al. (2025a)	67.38	31.43	54.04	15.09	58.36	16.18	61.77$_{\downarrow 0.2}$	23.16$_{\uparrow 0.2}$

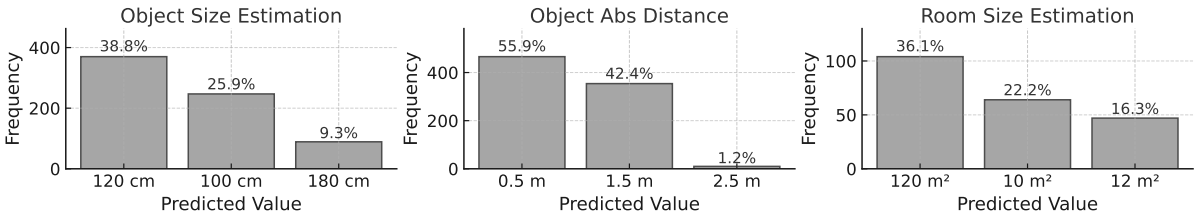


Figure 2 | Biased output distribution of LLaVA-Video on measurement estimation task from VSI-Bench Yang et al. (2025b).

3.2. Model and Task Selection

3D-aware VideoLLM. To inject 3D awareness into VideoLLM, two strategies are commonly adopted. **(i) Explicit 3D inputs**: A 3D encoder π_{3D} consumes geometric signals $\mathcal{S}_{geo}$, such as 3D coordinates Zhu et al. (2024a), coordinate-based positional encodings Zheng et al. (2025b), or features from a geometry encoder Chen et al. (2025b); Fan et al. (2025), and outputs $f_{3D} = \pi_{3D}(\mathcal{S}_{geo})$. A fusion module π_f forms $\tilde{f}_v = \pi_f(f_v, f_{3D})$, and the language model then generates the final answer from the enhanced representation: $A = \pi_\ell(\tilde{f}_v, f_q)$. **(ii) 3D aware vision supervision**: the inference architecture remains unchanged, while during training, geometry-informed auxiliary training objectives are obtained through either self-construction loss Wang et al. (2025a) or feature alignment Huang et al. (2025b). To conduct a fair and thorough comparison of the 3D-aware VideoLLM, we select LLaVA-Video-7B Zhang et al. (2024a) as the base VideoLLM since it serves as the initial model for most existing 3D-aware VideoLLMs. For 3D-aware variants, LLaVA-3D Zhu et al. (2024a), Video-3D-LLM Zheng et al. (2025b), VLM-3R Fan et al. (2025) and GS-reasoner Chen et al. (2025b) mainly utilize the first strategy, while ROSS3D Wang et al. (2025a) and 3DRS Huang et al. (2025b) belong to the second category. We exclude models like spatial-MLLM Wu et al. (2025) and VG-LLM Zheng et al. (2025a) due to limited open-source and different base models. Detailed configuration of models are listed in Table 1.

Task Selection. For the probing task type, we prioritize relative spatial relations over measurement estimation.Measurement estimation requires fine-grained spatial understanding, which sits at odds with the high-level, abstract language space, making it a particularly challenging spatial task for current VideoLLMs. We can tell from the figure 2 that LLaVA-Video Zhang et al. (2024a) is generating language-biased answer instead of a diverse and fine-grained abso-

Table 3 | The performance gap between MCQA and open-ended format question on relative distance and relative direction subset on VSI-bench Yang et al. (2025b). Δ denotes the Direct Difference, **CDR** denotes the Correct Difference Ratio.

Model	Relative Distance				Relative Direction			
	MCQA	OEQA	Δ	CDR	MCQA	OEQA	Δ	CDR
LLaVA-Video Zhang et al. (2024a)	41.54	31.26	10.28	34.50	41.43	44.81	3.38	40.80
ROSS3D Wang et al. (2025a)	41.05	38.71	2.34	31.83	34.82	36.88	2.06	28.62
Video-3D-LLM Zheng et al. (2025b)	27.85	29.17	1.32	16.48	34.27	25.47	8.80	27.07

lute measurement. Measurement estimation requires large-scale training to acquire as a base capability Cai et al. (2025). We therefore focus on relative spatial questions.

For the task format, we adopt open-ended questions rather than multi-choice. The multiple-choice setting exhibits answer format bias. We verify this by converting the Multi-Choice QA into open ended format, define Correct Difference Ratio (CDR) to quantify the performance varaiance among two inference settings: CDR = $\frac{n_\Delta}{n_{\text{correct}}}$, where n_{correct} is the number of questions answered correctly at least once and n_Δ is the number of questions whose answers differ across all n_{correct} answers. As shown in Table 3, CDR values are large across all models, indicating the spatial understanding capability is likely to be misdirected by the option-based language bias Wang et al. (2023); Zhang et al. (2025a). We provide detailed analysis in the Appendix.

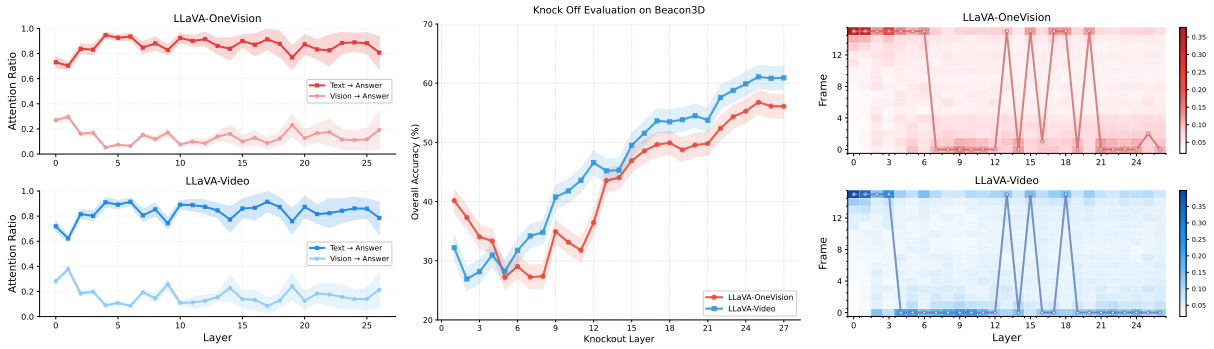
For open-ended relational understanding, existing benchmarks such as ScanQA Azuma et al. (2022) and SQA3D Ma et al. (2022) contain unanswerable samples as reported by recent study Huang et al. (2025a), but majority 3D-aware VideoLLMs still report them as primary results Huang et al. (2025b); Wang et al. (2025a); Zheng et al. (2025b), which raises concerns about reliability. Beacon3D Huang et al. (2025a) conducts a detailed analysis of these existing problems and constructs a new benchmark with verified and answerable questions across three scene distributions. Therefore, we select Beacon3D for evaluation.

Experiment Settings. Beacon3D composed of spatial QA derived from three subsets: ScanNet Dai et al. (2017), 3RScan Wald et al. (2019), and MultiScan Mao et al. (2022). The benchmark reports two metrics: Case accuracy and Object accuracy. Case accuracy is computed per QA pair, and Object accuracy is computed per object, requiring all three object-associated questions to be correct. During evaluation, $N = 32$ frames are uniformly sampled per video as input to the VideoLLM. For models that require explicit 3D input, aligned 3D coordinates are obtained from official resources Mao et al. (2022); Wang et al. (2024). Answers are scored by GPT5 Achiam et al. (2023).

3.3. Result and Analysis

Finding 1: 3D-aware VideoLLMs are overfitting benchmarks instead of acquiring 3D awareness.

As shown in Table 2, compared with the base model LLaVA-Video Zhang et al. (2024a), the majority of 3D-aware VideoLLMs exhibit significant performance drops on Beacon3D. With large-scale training on spatial QA built on ScanNet Azuma et al. (2022); Chen et al. (2020a,b); Ma et al. (2022); Zhang et al. (2023b), most models perform worse on the ScanNet subset and show severe declines in object accuracy compared to LLaVA-Video. Note that all these 3D-aware VideoLLMs are initialized from LLaVA-Video, indicating overfitting to language patterns and failure to acquire robust spatial understanding capability through injected 3D information. In



((a)) Cross Modal Attention Fraction Across Layers

((b)) Attention Knockout Across Layers

((c)) Frame-wise Attention Fraction

Figure 3 | Spatial modeling information flow inside LLM comparison between LLaVA-Video Zhang et al. (2024a) and LLaVA-OneVision Li et al. (2024) on Beacon3D Huang et al. (2025a). (a) answer generation attention contribution ratio between video and language tokens. (b) Attention knockout experiment result (c) frame-wise attention distribution from video to text tokens. Compared to LLaVA-OneVision, LLaVA-Video begins to show stable performance gains around layer 6, earlier than LLaVA-OneVision.

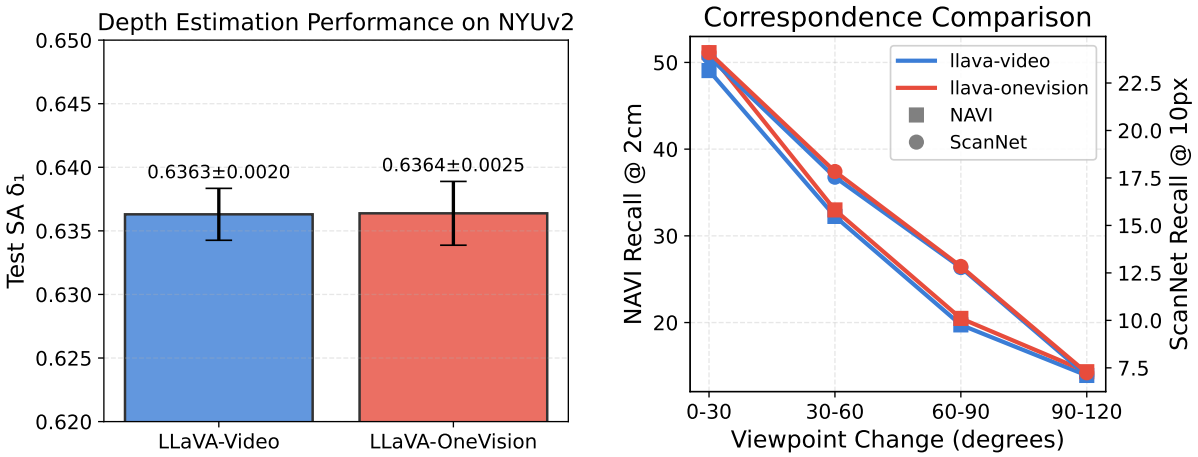
the novel scene subsets 3RScan and MultiScan, the gaps widen and object consistency drops further. Trained on 138K QA samples, VLM-3R reaches only half the object accuracy of LLaVA-Video on 3RScan. We attribute these gaps to two reasons. First, training on simple QA datasets with supervised fine-tuning tends to amplify language bias instead of inducing stable and generalizable spatial modeling inside the language model. GS-reasoner trained on CoT dataset GCoT-156k Chen et al. (2025b) exhibits higher performance on the ScanNet subset, benefiting from more diverse and general language supervision. Second, large-scale training on a fixed scene distribution creates strong visual token bias, with visual tokens accounting for a large percentage of the input token, the model overfits to the seen distribution and struggles with novel scenes.

Finding2: Self-supervised visual regularization can ease the overfitting and generalize well.

Using self-supervised vision regularization, ROSS-3D Wang et al. (2025a) eases the limitation brought by training on fixed scenes and simple QA. The random masking process during the training not only eases the model's overfitting of scene patterns but also adds strong regularization of the overfitting trend on simple QA questions, while still exhibiting a certain level of performance drop compared to the base model, it successfully reaches competitive object consistency compared to the base model. Considering the dataset ROSS3D was trained on, it exhibits the effectiveness of self-visual supervision on easing language bias in the mid-quality training corpus. Meanwhile, compared to 3DRS Huang et al. (2025b), trained on the same data corpus with shared inference architecture, visual supervision from explicit geometry encoder Wang et al. (2025b) failed to generalize to a novel dataset, suggesting the vision supervision steers the general spatial understanding model from the intrinsic cross-view representation better.

4. 3D awareness of VideoLLM

The 3D-aware VideoLLM still struggles for robust spatial understanding with aligned 3D input, which motivates us to explore a more straightforward setting: *Does pure video-language training bring a VideoLLM with stronger 3D awareness?* We conduct empirical analysis between



((a)) Monocular Depth Estimation

((b)) MultiView Correspondance

Figure 4 | Probing performance on vision encoder from LLaVA-Onevision Li et al. (2024) and LLaVA-Video Zhang et al. (2024a) on (a) monocular depth estimation and (b) multiview correspondence tasks.

LLaVA-Video Zhang et al. (2024a) and LLaVA-OneVision-SingleImage Li et al. (2024)(LLaVA-OV-SI). LLaVA-OV-SI is trained on a pure image-language corpus, while LLaVA-Video is trained on it using video-language data, serving as a suitable target for understanding the effect of video-language training on spatial understanding. As shown in Table 4, LLaVA-Video exhibits significant performance gain compared to LLaVA-OV-SI on Beacon 3D. To understand how the video-language training empowers VideoLLM with stronger 3D awareness without 3D data supervision, we probe its 3D awareness along three dimensions. (i) Vision encoder: whether video instruction tuning brings a more 3D-aware vision encoder. (ii) LLM: feature modeling pattern inside the LLM where cross-frame-cross-modality features interact. (iii) Data: We analyze the data distribution within the training corpus, showing the effect of spatially relevant labels and the insufficiency of spatial supervision.

4.1. Vision Encoder

Firstly, to probe the 3D awareness change of the vision encoder after video-language training, we adopt the Probe3D protocol El Banani et al. (2024). The vision encoder is decoupled from the VideoLLM and evaluated on two downstream tasks: probing through DPT head Ranftl et al. (2021) on monocular depth estimation Bhat et al. (2021) and evaluated on multi-view correspondence Sarlin et al. (2020). As shown in Fig. 4, the vision encoder before and after video–language training exhibits similar performance on depth estimation task and shows nearly overlapping multi-view feature correspondence. Though the vision encoder is unfrozen during video-language training, the 3D awareness of the encoder exhibits no significant change. Detailed settings of the vision encoder probing are in Appendix.

4.2. 3D Awareness Inside LLM

For VideoLLM architectures, cross-frame spatial modeling occurs inside the language model through causal self-attention. Our analysis, therefore, focuses on attention modules, which govern the flow of information across frames and the interaction between video and text.

Table 4 | **Spatial mid-training improves cross-view grounding.** Numbers from our evaluation. Object-level gains are most pronounced after Scan2Cap-based mid-training, surpassing 3D-aware methods. LLaVA-Video$_{mt}$ indicates LLaVA-Video after spatial mid-training on Scan2Cap-Video-36K. The best performance is highlighted in bold.

Model	ScanNet		MultiScan		3RScan		Overall	
	case	obj.	case	obj.	case	obj.	case	obj.
LLaVA-OV-SI Li et al. (2024)	61.30	23.64	50.26	9.43	57.05	20.22	57.73	19.73
LLaVA-Video Zhang et al. (2024a)	67.19	29.87	53.88	13.21	59.28	18.75	61.96	22.92
ROSS3D Wang et al. (2025a)	67.38	31.43	54.04	15.09	58.36	16.18	61.77	23.16
LLaVA-Video$_{mt}$	**67.97**	**32.73**	**57.08**	**16.35**	**60.08**	**22.06**	**63.21**	**25.98**

Cross Modality Attention Distribution. Let $\mathcal{I}$ be the index set of visual tokens in f_v, $\mathcal{T}$ the index set of text tokens in f_q, and $\mathcal{G} \subseteq \mathcal{T}$ denote the index set of generated answer tokens. Let $A^{(\ell,h)}$ be the post-softmax attention matrix at layer ℓ and head h, with entries $A^{(\ell,h)}_{q,k}$ from query q to key k. We use attention mass as a proxy for token contribution:

$$a^{(\ell)}_m = \frac{1}{H\,|\mathcal{G}|} \sum_{h=1}^{H} \sum_{q\in\mathcal{G}} \sum_{k\in S_m} A^{(\ell,h)}_{q,k}, \quad m \in \{\text{vis}, \text{txt}\},$$

where $S_{\text{vis}} = \mathcal{I}$ and $S_{\text{txt}} = \mathcal{T}$, and $a^{(\ell)}_{\text{vis}} + a^{(\ell)}_{\text{txt}} = 1$.

As shown in Fig. 3(a), visual tokens contribute more in the early few layers, while text tokens gradually dominate after layer 6, consistent with prior findings on image-based MLLMs Kaduri et al. (2025). Guided by this pattern, we further study cross-frame information flow through the attention-knockout Geva et al. (2023) experiment and frame-wise attention distributions analysis.

Attention knockout. We add a knockout mask to the attention logits from layer ℓ onward; earlier layers use only the causal mask. For any layer r and any head h, define

$$M^{\text{ko},(r)}_{q,k} = \begin{cases} -\infty, & r \geq \ell,\ q \in \mathcal{T},\ k \in \mathcal{I}, \\ 0, & \text{otherwise.} \end{cases}$$

Here $\mathcal{I}$ and $\mathcal{T}$ are the index sets of visual and text tokens. The mask is added to the standard causal mask, blocking video→text attention while keeping pathways unchanged.

As shown in Fig. 3(b), compared with LLaVA OneVision, LLaVA Video accumulates sufficient visual evidence earlier, between layers four and six, after which performance gains appear. This highlights the importance of early global visual composition Yu and Lee (2025).

Frame-wise attention distribution. We compute a per-frame attention distribution from question tokens to each frame vision tokens:

$$a^{(\ell)}_i = \frac{1}{H\,|\mathcal{Q}|} \sum_{h=1}^{H} \sum_{q\in\mathcal{Q}} \frac{1}{|\mathcal{I}_i|} \sum_{k\in\mathcal{I}_i} A^{(\ell,h)}_{q,k}.$$

Here $\mathcal{Q} \subseteq \mathcal{T}$ is the index set of question tokens and $\mathcal{I}_i \subseteq \mathcal{I}$ the visual tokens of frame i.

As shown in Fig. 3(c), LLaVA style models display a strong last frame bias in the earliest layer. LLaVA-Video spreads attention more evenly across frames and shifts the attention peak

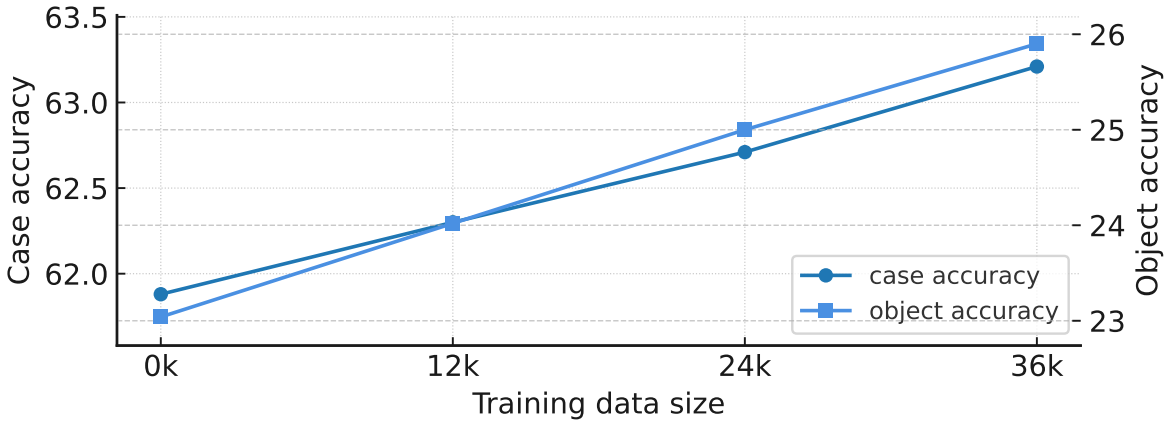


Figure 5 | Performance gain on Beacon3D as training data scale.

from the last layer earlier than LLaVA-OneVisio-SI, after the layer in which text tokens begin to dominate. As layers stack, LLaVA-Video consistently outperforms LLaVA-OneVision-SI on spatial understanding, indicating that video training fosters intrinsic cross-frame spatial modeling. Although a sequential bias persists and contributes to the spatial invariance issues discussed in Section 5, video training reduces the last frame bias, promotes more distributed spatial reasoning, and yields a clear performance gain on Beacon3D when moving from LLaVA-OneVision-SI to LLaVA-Video.

4.3. Data Analysis

We look into LLaVA-Video-178K, which is used to transform LLaVA-OneVision-SI into LLaVA-Video. With around 94k samples are spatial-understanding QAs drawn from diverse videos, included to equip the model with the ability to perceive spatial relationships, providing the primary supervision signal that encourages VideoLLMs to learn intrinsic cross-frame modeling for spatial understanding. However, 50% of these data are limited to simple multiple-choice QAs, and the open-ended questions are mostly answerable from a single frame with only simple spatial relations and short answers, while the spatial supervision present in detailed captions is sparse and mixed with general descriptions, offering weak signals for fine-grained spatial modeling require cross-frame 3D awareness. Therefore, higher-quality spatial video–language data are still needed for mid-training. More static analyses of the dataset composition are provided in the Appendix.

4.4. Spatial Mid-training

Based on these findings, we apply a simple strategy: continue mid-training on LLaVA-Video using a detailed spatial caption dataset. With no large-scale high-quality training corpus available, we transform the Scan2Cap Chen et al. (2020b) dataset into a pure video-based dataset. Instead of forcing the model to interpret out of the pretraining distribution 3D coordinates, we project the 3D coordinates back onto the video frames and prompt the model to generate detailed, spatially enriched captions for objects constrained within projected 3D bounding boxes, with visual prompting has been proven beneficial for stronger relational 3D awareness Cai et al. (2025). Through this simple strategy, we construct Scan2Cap-Video-36K, which contains 36k video–caption pairs with non-overlapping to the Beacon3D Huang et al. (2025a).

As shown in Table 4, when trained on Scan2Cap-Video-36K rather than large scale simple QA data, LLaVA-Video exhibits enhanced spatial awareness and achieves performance improvements on the Beacon3D benchmark under a zero shot setting, with 1.25% on case accuracy and 3.06% on object accuracy, showing enhanced spatial understanding in a more object-consistency way without introducing any explicit 3D input, demonstrating the effectiveness of spatial mid-training and the necessity of more high quality, detailed spatial supervision labels. Meanwhile, as the data scale used for mid-training growth, the model exhibits higher performance on the benchmark as depicted in Figure 5. Therefore, we expect that further scaling will lead to additional performance improvements, shed light on how to increase the effective training data in future work. We provide more analysis on the data mix in the Appendix.

5. Spatial Invariance

5.1. Invarance Issue

While video training enhances the intrinsic 3D awareness of VideoLLM, enhancing spatial understanding covering geometry, spatial relationship and object attributes covered in Beacon3D, we uncover **spatial invariance**, an important yet under-explored dimension of 3D awareness of VideoLLM. Spatial invariance for video-based spatial QA is formed as follows: Denoted by $A = \pi_{vid}(Q, \mathcal{F})$, the answer generated by the VideoLLM π_{vid} given a question Q and frames $\mathcal{F}$. We say that model is spatially invariant if the generated answer is invariant to any frame permutation:

$$\forall \sigma \in S_N : \quad \pi_{vid}(Q, \sigma(\mathcal{F})) = A = \pi_{vid}(Q, \mathcal{F}),$$

where $\sigma(\mathcal{F}) = \{f_{\sigma(i)}\}_{i=1}^{N}$ denotes the frames under a permutation σ of the frame indices.

While VideoLLMs rely on sequential order to achieve temporal understanding Li et al. (2025), the answer to a non-temporal spatial question should be identical regardless of the presentation order of the same views, maintaining robust spatial invariance as an essential spatial understanding capability.

The concept of spatial invariance is consistent with object constancy and viewpoint invariance in human cognition Biederman (1987); DiCarlo and Cox (2007), aligns with cross-view consistency in multi-view geometry and neural rendering Chan et al. (2022); Hartley and Zisserman (2003); Mildenhall et al. (2021), and connects to permutation-invariant set modeling and equivariant architectures in machine learning Cohen and Welling (2016); Fuchs et al. (2020); Lee et al. (2019); Qi et al. (2017); Satorras et al. (2021); Zaheer et al. (2017). In classical 3D reconstruction pipelines, where multi-view inputs are processed symmetrically, this invariance is guaranteed naturally Kerbl et al. (2023); Mildenhall et al. (2021). In contrast, the inherent order bias of autoregressive architectures can weaken the spatial invariance of VideoLLMs, making predictions unnecessarily sensitive to frame order in non-temporal queries.

5.2. Spatial Invariance Quantification

Experiment Setting. To evaluate the spatial invariance criterion, frame reversals are applied as a simplified version of arbitrary permutations. For each example (q, F, a) with ground-truth answer a, the VideoLLM π_{vid} are queried twice: once with the standard order F and once with the reversed order $\text{rev}(F)$, obtaining predictions: $\hat{a} = \pi_{vid}(q, F), \quad \tilde{a} = \pi_{vid}\big(q, \text{rev}(F)\big)$. We apply metric CDR mentioned in section 3 to measure the performance variance between different input orders. Lower CDR indicates stronger spatial invariance. For performance on Beacon3D, we report the average of two queries.

Table 5 | Overall spatial invariance level on Beacon3D Huang et al. (2025a). Large-scale video training and spatial mid-training with a fixed video input sequence lead to higher spatial variance. LLaVA-Video$_{mtsf}$ denotes mid-training with frame shuffle augmentation.

Model	Case	Object	Δ	CDR
LLaVA-OV-SI Li et al. (2024)	57.73	19.55	0.04	12.99
LLaVA-Video Zhang et al. (2024a)	61.41	22.55	0.94	15.82
LLaVA-Video$_{mt}$	63.58	26.17	0.57	15.30
LLaVA-Video$_{mtsf}$	62.85	25.50	**0.25**	**14.32**

Figure 6 | Qualitative results on Beacon3D. Same QA pair with forward (top) vs reversed (bottom) frames. While video–language training improves 3D awareness, it also reduces spatial invariance under fixed frame order; frame-shuffle augmentation eases it.

Experimental Results. As shown in Table 5, simply reversing the input frame order yields only a small change in overall accuracy, yet the CDR becomes substantial. Compared to the base LLaVA-OneVision, both LLaVA-Video and LLaVA-Video$_{mt}$ exhibit a higher CDR of 15.82% and 15.30%, respectively, indicating sequential bias introduced by video training that affects the spatial invariance of VideoLLMs. We attribute this to two aspects: First, LLaVA-style architectures require causal attention over both visual and textual tokens; although video training encourages the model to extract more information from early frames as mentioned in Sec. 4, the model tends to under-utilize middle and later frames; Second, during training, a scene video typically contains rich visual evidence that can align with many language QA pairs, but while the QA pairs change, the input sequence of the same scene video remains fixed. Consequently, spatial-aware training improves spatial modeling capacity but simultaneously induces pronounced spatial variance when the model overfits to frame order.

5.3. Spatial-invarance training

To reduce spatial variance in video training, we randomly shuffle the frames during spatial mid-training. Since the spatial QA pairs are order-invariant, we optimize

$$\min_{\theta} \mathbb{E}_{(Q,\mathcal{F},A)\sim\mathcal{D}}\left[\, \ell_{\mathrm{CE}}\big(\pi_{\mathrm{vid}}(Q, \mathrm{shuffle}(\mathcal{F})),\, A\big) \,\right]$$

where shuffle($\mathcal{F}$) denotes a uniformly random shuffle of the frames in $\mathcal{F}$, and ℓ_{CE} is the standard cross-entropy loss. As shown in Table 5, simply shuffling the input order leads to more stable performance across different input orders, yielding a 6% decrease of CDR compared to mid-training while still maintaining zero-shot performance growth. Meanwhile, the order sensitivity remains; we attribute it to the nature of causal attention–based architectures and leave further

architectural exploration to future work. Figure 6 provides qualitative examples, illustrating improved spatial invariance after spatial mid-training.

6. Conclusion

In this work, we present the first empirical study of VideoLLMs as proxies for spatial understanding, covering both 3D-injected variants and the intrinsic 3D awareness learned from pure video–language training. We show that many 3D-enhanced VideoLLMs overfit flawed benchmarks, obscuring true spatial ability and generalizing poorly to new question types and unseen scenes. Analyzing cross-frame modeling from pure video–language training, we find that VideoLLMs offer a strong basis for spatial reasoning, but current datasets lack sufficient fine-grained spatial–language supervision. Mid-training on spatially enriched scene captions addresses this gap and achieves state-of-the-art zero-shot performance. Finally, we identify spatial invariance as an overlooked facet of 3D awareness: sequence-order sensitivity persists and is amplified by autoregressive training, but a simple frame-shuffling strategy mitigates it. We hope this work steers attention on high-quality spatial–language data and generalizable VideoLLMs for spatial reasoning.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

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, pages 422–440. Springer, 2020.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 19129–19139, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4009–4018, 2021.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 4135–4144, 2025.

Irving Biederman. Recognition-by-components: a theory of human image understanding. Psychological review, 94(2):115, 1987.

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. arXiv preprint arXiv:2509.25413, 2025.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16123–16133, 2022.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In European conference on computer vision, pages 202–221. Springer, 2020a.

Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in rgb-d scans, 2020b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems, 37:27056–27087, 2024a.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. arXiv preprint arXiv:2503.01773, 2025a.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26428–26438, 2024b.

Yiming Chen, Zekun Qi, Wenyao Zhang, Xin Jin, Li Zhang, and Peidong Liu. Reasoning in space via grounding in the world. arXiv preprint arXiv:2510.13800, 2025b.

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 6348–6361, 2025c.

Tao Chu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Qiong Liu, and Jiaqi Wang. Unified scene representation and reconstruction for 3d large language models. arXiv preprint arXiv:2404.13044, 2024.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In International conference on machine learning, pages 2990–2999. PMLR, 2016.

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 conference on computer vision and pattern recognition, pages 5828–5839, 2017.

Duolikun Danier, Mehmet Aygün, Changjian Li, Hakan Bilen, and Oisin Mac Aodha. Depthcues: Evaluating monocular depth perception in large vision models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 20049–20059, 2025.

Weipeng Deng, Jihan Yang, Runyu Ding, Jiahui Liu, Yijiang Li, Xiaojuan Qi, and Edith Ngai. Can 3d vision-language models truly understand natural language? arXiv preprint arXiv:2403.14760, 2024.

James J DiCarlo and David D Cox. Untangling invariant object recognition. Trends in cognitive sciences, 11(8):333–341, 2007.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21795–21806, 2024.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv. org/abs/2306.13394, 2(8), 2024a.

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, 2024b.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. Advances in neural information processing systems, 33:1970–1981, 2020.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. arXiv preprint arXiv:2304.14767, 2023.

Chenhui Gou, Ziyu Ma, Zicheng Duan, Haoyu He, Feng Chen, Akide Liu, Bohan Zhuang, Jianfei Cai, and Hamid Rezatofighi. An empirical study on how video-llms answer video questions. arXiv preprint arXiv:2508.15360, 2025.

Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003.

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, 36:20482–20494, 2023.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. Advances in Neural Information Processing Systems, 37:113991–114017, 2024.

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

Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang. Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 24570–24581, 2025a.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. Mllms need 3d-aware representation supervision for scene understanding. arXiv preprint arXiv:2506.01946, 2025b.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709, 2019.

Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In European Conference on Computer Vision, pages 289–310. Springer, 2024.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 14549–14558, 2025.

Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3905–3915, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph., 42(4):139–1, 2023.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In International conference on machine learning, pages 3744–3753. PMLR, 2019.

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.

Bozheng Li, Mushui Liu, Gaoang Wang, and Yunlong Yu. Frame order matters: A temporal sequence-aware model for few-shot action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 18218–18226, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In European conference on computer vision, pages 216–233. Springer, 2024b.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. arXiv preprint arXiv:2210.07474, 2022.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16488–16498, 2024.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. Advances in Neural Information Processing Systems, 37:76819–76847, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 11–20, 2016.

Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. Advances in neural information processing systems, 35:9058–9071, 2022.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 3195–3204, 2019.

David Marr. Vision: A computational investigation into the human representation and processing of visual information. MIT press, 2010.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1):99–106, 2021.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 652–660, 2017.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In Proceedings of the Asian Conference on Computer Vision, pages 18–34, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF international conference on computer vision, pages 12179–12188, 2021.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4938–4947, 2020.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In International conference on machine learning, pages 9323–9332. PMLR, 2021.

Anh Thai, Songyou Peng, Kyle Genova, Leonidas Guibas, and Thomas Funkhouser. Splattalk: 3d vqa with gaussian splatting. arXiv preprint arXiv:2503.06271, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9568–9578, 2024.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7658–7667, 2019.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. arXiv preprint arXiv:2504.01901, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 5294–5306, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 10510–10522, 2025c.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19757–19767, 2024.

Zhecan Wang, Long Chen, Haoxuan You, Keyang Xu, Yicheng He, Wenhao Li, Noel Codella, Kai-Wei Chang, and Shih-Fu Chang. Dataset bias mitigation in multiple-choice visual question answering and beyond. arXiv preprint arXiv:2310.14670, 2023.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747, 2025.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4840–4851, 2024.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1179–1189, 2023.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 7694–7701. IEEE, 2024.

Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 29501–29512, 2025a.

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. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 10632–10643, 2025b.

Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. IEEE Transactions on Visualization and Computer Graphics, 30(3):1772–1786, 2022.

Yang You, Yixin Li, Congyue Deng, Yue Wang, and Leonidas Guibas. Multiview equivariance improves 3d correspondence understanding with minimal feature finetuning. arXiv preprint arXiv:2411.19458, 2024.

Zhuoran Yu and Yong Jae Lee. How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. arXiv preprint arXiv:2508.20279, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, 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, pages 9556–9567, 2024a.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In European Conference on Computer Vision, pages 57–74. Springer, 2024b.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? arXiv preprint arXiv:2210.01936, 2022.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. Advances in neural information processing systems, 30, 2017.

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. Advances in Neural Information Processing Systems, 37:43468–43498, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023a.

Hao Zhang, Chen Li, and Basura Fernando. Mitigating easy option bias in multiple-choice question answering. arXiv preprint arXiv:2508.13428, 2025a.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture. arXiv preprint arXiv:2509.02359, 2025b.

Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15225–15236, 2023b.

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, 2024a.

Zhuofan Zhang, Ziyu Zhu, Pengxiang Li, Tianxu Wang, Tengyu Liu, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. Task-oriented sequential grounding in 3d scenes. 2024b.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. arXiv preprint arXiv:2505.24625, 2025a.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 8995–9006, 2025b.

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, 2024a.

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, pages 2911–2921, 2023.

Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In European Conference on Computer Vision, pages 188–206. Springer, 2024b.