GAP-MLLM: Geometry-Aligned Pre-training for Activating 3D Spatial Perception in Multimodal Large Language Models

Jiaxin Zhang[1#], Junjun Jiang[1†], Haijie Li[2], Youyu Chen[1], Kui Jiang[1], and Dave Zhenyu Chen[3]

[1] Harbin Institute of Technology, China
[2] School of Electronic and Computer Engineering, Peking University, China
[3] Huawei, China
[#]Intern at Huawei. [†]Corresponding Author.
https://gapmllm.github.io/

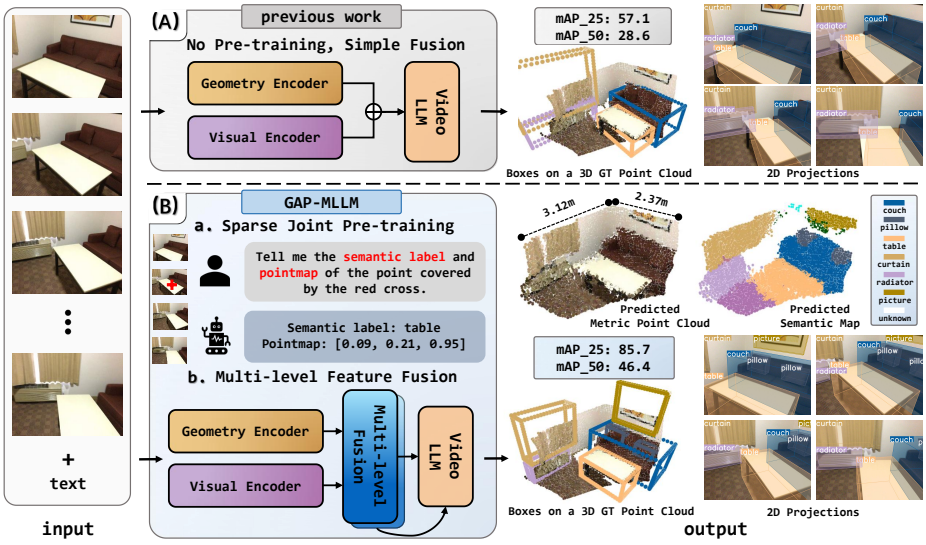

Fig. 1: Geometry-aligned pre-training significantly improves 3D perception in MLLMs. (A) Naive fusion without geometry-aware pre-training leads to limited geometric utilization and inaccurate 3D detection. (B) Our sparse geometry–semantics joint pre-training and multi-level fusion module progressively integrate geometric priors, activating structural perception and yielding substantial gains in 3D detection mAP.

Abstract. Multimodal Large Language Models (MLLMs) demonstrate exceptional semantic reasoning but struggle with 3D spatial perception when restricted to pure RGB inputs. Despite leveraging implicit geometric priors from 3D reconstruction models, image-based methods still exhibit a notable performance gap compared to methods using explicit 3D data. We argue that this gap does not arise from insufficient geometric priors, but from a misalignment in the training paradigm: text-dominated fine-tuning fails to activate geometric representations within

MLLMs. Existing approaches typically resort to naive feature concatenation and optimize directly for downstream tasks without geometry-specific supervision, leading to suboptimal structural utilization. To address this limitation, we propose **GAP-MLLM**, a **G**eometry-**A**ligned **P**re-training paradigm that explicitly activates structural perception before downstream adaptation. Specifically, we introduce a visual-prompted joint task that compels the MLLMs to predict sparse pointmaps alongside semantic labels, thereby enforcing geometric awareness. Furthermore, we design a multi-level progressive fusion module with a token-level gating mechanism, enabling adaptive integration of geometric priors without suppressing semantic reasoning. Extensive experiments demonstrate that GAP-MLLM significantly enhances geometric feature fusion and consistently enhances performance across 3D visual grounding, 3D dense captioning, and 3D video object detection tasks.

Keywords: Multimodal Large Language Models · 3D Spatial Perception · Geometry-Aligned Pre-training

1 Introduction

Multimodal large language models (MLLMs), including vision-language models (VLMs) [2, 25, 29] and vision-language-action models [27, 40, 62], have demonstrated strong cross-modal reasoning and semantic understanding capabilities. Extending these models to 3D spatial perception is a crucial step toward grounding MLLMs in the physical world [13, 60], enabling structured scene understanding and object localization under natural language supervision.

Recent efforts [18, 58] enhance 3D spatial perception under pure RGB input by incorporating geometric priors derived from feed-forward reconstruction models [47, 48, 51]. These implicit geometric priors provide pixel-aligned structural cues without requiring explicit 3D inputs such as point clouds [35, 59], offering scalability and compatibility with vision-language pipelines. However, despite their promise, implicit prior-based methods consistently underperform in spatial perception compared to approaches operating on explicit 3D representations.

We argue that this gap stems from a geometry–semantics imbalance in existing training paradigms rather than from insufficient geometric priors. As illustrated in Fig. 2, point cloud-based methods [35] possess strong geometric integrity but lack the high-level semantic reasoning of LLMs, whereas image-based MLLMs [58] enriched with geometric encoders preserve semantic prowess yet fail to fully exploit structural information for precise spatial inference. Most prior works simply fuse geometric and visual features and fine-tune on text-prioritized downstream tasks without dedicated geometry-aware pre-training. Under this optimization regime, semantic supervision dominates the learning process, and geometric representations remain weakly activated, limiting their effective contribution to spatial perception and accurate localization.

To address this limitation, we propose **GAP-MLLM**, a **G**eometry-**A**ligned **P**re-training paradigm that explicitly activates structural perception before task-specific learning. Rather than directly fine-tuning fusion modules on target tasks,

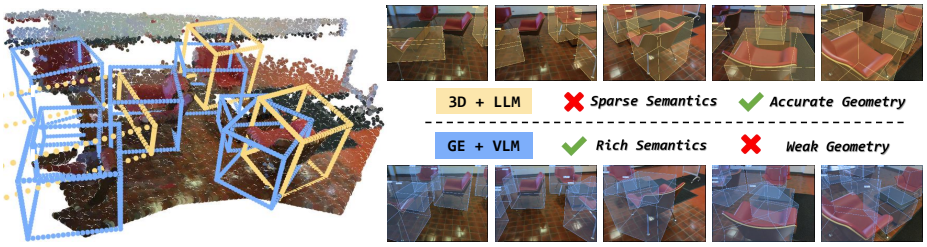

Fig. 2: Geometry–semantics imbalance in existing 3D perception paradigms. Point cloud-based methods ensure geometric accuracy but lack semantics, whereas image-based geometric encoders (GE) with VLMs retain semantics yet under-utilize geometry. As a result, both paradigms exhibit suboptimal performance in 3D perception.

we introduce a joint reconstruction-perception objective that aligns geometric priors with the inherent image-text representations of MLLMs across multiple feature levels. Specifically, we design a vision-prompt-based joint task requiring MLLMs to predict both sparse pointmaps and semantic labels (see Fig. 1). This objective enforces structural reasoning as a primary capability rather than an auxiliary feature, preventing geometry from being suppressed by dominant language supervision. Furthermore, inspired by the hierarchical attention patterns of geometric encoders, we develop a multi-level progressive fusion module with a token-level gating mechanism. This design enables adaptive geometric integration from abstract global contexts to fine-grained local details, ensuring that geometry and semantics interact synergistically without mutual suppression. Extensive experiments demonstrate that our geometry-aligned pre-training substantially enhances geometric feature utilization under sparse 3D supervision and consistently improves downstream performance on 3D visual grounding, 3D dense captioning, and 3D video object detection tasks.

Our contribution can be summarized as follows:

1. We propose a geometry-aligned pre-training paradigm that activates structural perception in image-only multimodal large language models with implicit geometric priors, moving beyond purely task-driven fine-tuning.
2. We develop a sparse geometry–semantics joint pre-training objective that aligns geometric priors with semantic representations through simultaneous prediction of 3D pointmaps and semantic labels.
3. We design a multi-level progressive fusion architecture with token-level gating, enabling hierarchical and adaptive integration of geometric priors from abstract to fine-grained representations across attention layers.
4. We show through extensive experiments that GAP-MLLM effectively activates the spatial structure perception capabilities even with a small amount of sparse data, and as pre-trained weights, consistently improves performance across a diverse range of downstream 3D perception tasks.

2 Related Works

Multimodal Large Language Models (MLLMs) Multimodal Large Language Models (MLLMs) have demonstrated strong semantic reasoning and visual–linguistic alignment capabilities across diverse 2D perception and reasoning tasks [2, 3, 25, 29, 31, 32, 41, 45]. Recent efforts [13, 17, 19, 20, 34, 52–54] extend these models to 3D spatial tasks such as grounding [1, 9, 10, 15], captioning [10,14,15], and spatial object detection [5–8,37,38], aiming to bridge language understanding with real-world spatial structure. Some approaches incorporate explicit geometric inputs, including depth maps, point clouds, or BEV representations [35, 39, 61], which improve spatial perception but rely on costly or scarce 3D annotations. Under pure RGB video settings, obtaining reliable 3D inputs and explicit geometry remains challenging, motivating the use of implicit geometric priors derived directly from image sequences.

Feed-Forward 3D Reconstruction Feed-forward 3D reconstruction provides an efficient mechanism for extracting geometric structure directly from RGB inputs. Unlike traditional iterative pipelines [21, 42], recent approaches [26, 33, 47–49, 51] predict pixel-aligned 3D geometry end-to-end. For example, DUSt3R [49] estimates pointmaps without known camera parameters, while VGGT [47] and MapAnything [26] extend to multi-view and metric-scale reconstruction. These models have also been applied to scene semantic segmentation [28, 30, 44], efficient reconstruction [43, 56], and dynamic scene reconstruction [46], demonstrating strong geometric expressiveness under pure RGB settings. More importantly, feed-forward reconstruction provides scalable implicit geometric priors for multimodal systems. However, integrating reconstruction-derived geometry into MLLMs while preserving semantic reasoning remains challenging.

Image-based 3D Spatial Perception with MLLMs MLLM-powered image-based 3D perception methods aim to combine geometric understanding from image sequences with structured language reasoning. [18, 35, 58, 59]. Existing approaches can be broadly divided into explicit-input and implicit-prior paradigms. Explicit-input methods encode geometric representations into language-aligned tokens. SpatialLM [35] reconstructs point clouds from images and converts them into structured tokens for spatial description, and Video-3D-LLM [59] injects explicit 3D cues into visual tokens via positional encoding. While effective, these methods rely on high-quality 3D annotations. Implicit-prior methods instead extract geometric cues from feed-forward reconstruction models and fuse them with visual features. VG-LLM [58] embeds a reconstruction encoder to inject geometric priors from video sequences, and VLM-3R [18] introduces spatial–visual–view fusion for spatial reasoning tasks. Although these approaches preserve strong semantic reasoning, most directly optimize fused representations on downstream tasks without geometry-aware pre-training. Under text-dominated supervision, geometric priors are often weakly activated, limiting structural perception and localization. In contrast, our work focuses on a geometry-aligned pre-training paradigm that activates geometric priors before downstream adaptation, enabling more effective geometric utilization within MLLMs.

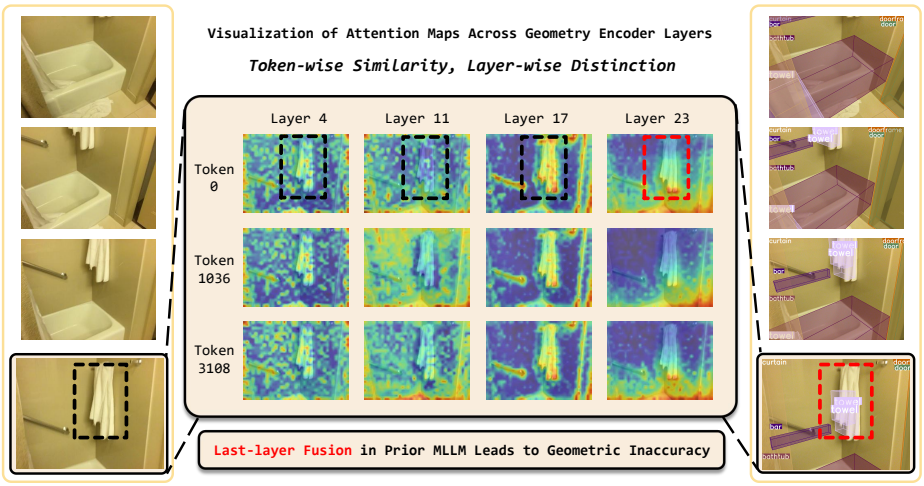

Fig. 3: Failure Analysis. Global attention maps of the geometric encoder are visualized using three representative tokens across layers. The same token exhibits distinct attention patterns at different layers. Prior method that relies on last-layer tokens as implicit geometric priors [58] leads to inaccurate 3D bounding box prediction.

3 Method

We present **GAP-MLLM**, a geometry-aligned pre-training framework designed to activate structural perception in multimodal large language models under pure RGB input. The framework consists of a multi-level geometric–visual fusion architecture and a two-stage 3D perception training paradigm that jointly enhance metric-aware spatial perception. We first describe the architecture in Sec. 3.1, followed by the geometry-aligned training strategy in Sec. 3.2.

3.1 Architecture

Visualization Analysis. Existing methods [18, 58] enhance the spatial perception capabilities of MLLMs by extracting implicit geometric priors via geometric encoders with feed-forward reconstruction, but typically utilize only the last-layer features of geometric encoders as implicit priors. Recent works [43, 56] have attempted to eliminate useless tokens by analyzing information redundancy in the intermediate feature space of feed-forward reconstruction. Different from these works that focus on token efficiency, we investigate how geometric representations at different layers affect downstream 3D perception when used as implicit priors. As shown in Fig. 3, we visualize the global attention maps of the geometric encoder across layers using representative tokens. The attention pattern of the same token varies significantly across layers, indicating that geometric cues are distributed hierarchically rather than concentrated in the last layer. When only the last-layer geometric tokens are injected to guide 3D bounding box prediction, the model exhibits inaccurate geometric structures caused by biased geometric attention. This observation suggests that intermediate-layer

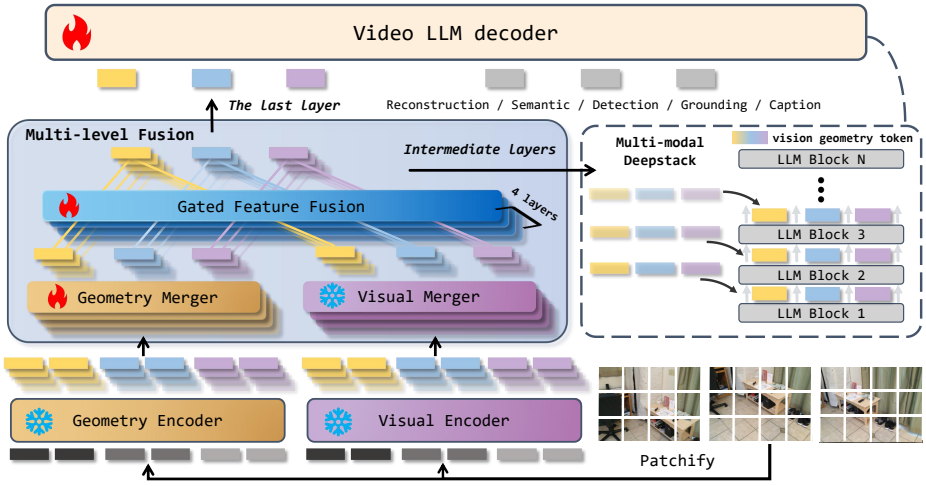

Fig. 4: Network Architecture. An image sequence is processed by parallel geometric and visual branches to extract multi-level structural and semantic tokens. After gated multi-level fusion, the final-layer fused tokens are aligned with task-related textual representations in the video LLM decoder, while selected intermediate-layer tokens are injected into early decoder blocks to preserve hierarchical geometric information.

geometric representations are not sufficiently activated in existing pipelines, motivating the necessity of multi-level geometric feature fusion in GAP-MLLM.

Overview. Fig. 4 illustrates the overall architecture of GAP-MLLM. During both training and inference, the model takes an image sequence $\{I_i\}_{i=1}^{n}$ and a task-related natural language query Q as input. The image sequence is processed by two parallel branches: a visual branch responsible for extracting semantic features and a geometric branch responsible for extracting structural representations derived from feed-forward reconstruction. Specifically, for each image $I_i \in \mathbb{R}^{h\times w\times 3}$, the visual and geometric branches respectively generate multi-level tokens $\{\mathcal{T}_{i,j}^{V}\}_{j=1}^{L_V}$ and $\{\mathcal{T}_{i,j}^{G}\}_{j=1}^{L_G}$, where

$$\mathcal{T}_{i,j}^{V} \in \mathbb{R}^{\lfloor h/p_V \rfloor \times \lfloor w/p_V \rfloor \times c}, \quad \mathcal{T}_{i,j}^{G} \in \mathbb{R}^{\lfloor h/p_G \rfloor \times \lfloor w/p_G \rfloor \times c}. \tag{1}$$

Here, p_V and p_G denote the patch sizes of the two branches, and L_V and L_G represent the total number of layers. For simplicity and alignment between modalities, we set $p_V = p_G = p$ and $L_V = L_G = L$ in subsequent formulations.

These multi-level tokens are then passed into a gated fusion module, where geometric and visual representations are hierarchically integrated across layers. The fused tokens are subsequently aligned with textual representations in the MLLM decoder. The final-layer fused tokens serve as the primary input to the decoder, while selected intermediate-layer tokens are progressively injected during decoding to preserve multi-level structural information. In this work, we adopt Qwen3-VL [2] as the MLLM backbone and visual branch, and VGGT [47] as the geometric branch to extract implicit geometric priors.

Multi-Level Feature Fusion. Existing implicit prior-based methods [58] typically fuse geometric and visual features via simple element-wise addition, relying only on last-layer geometric representations [18]. While effective, such static and single-layer fusion overlooks hierarchical geometric cues and may leave structural information insufficiently activated under dominant semantic supervision. To address this limitation, we introduce independent gating mechanisms across multiple layers for adaptive and hierarchical geometric integration.

The Multi-Level Fusion module takes multi-level visual tokens $\left\{\mathcal{T}_{i,j}^{V}\right\}_{j=1}^{L}$ and geometric tokens $\left\{\mathcal{T}_{i,j}^{G}\right\}_{j=1}^{L}$ as inputs. Note that the Qwen-VL series [2,3] compresses image tokens via token merging to reduce computational cost. To ensure spatial alignment between modalities, the geometric branch adopts an identical architectural design to the visual branch. Concretely, both branches group spatially adjacent 2×2 patches and pass them through a two-layer MLP to produce a single feature, yielding reduced visual tokens $\mathcal{T}_{i,j}^{V'} \in \mathbb{R}^{\lfloor \frac{h}{2p} \rfloor \times \lfloor \frac{w}{2p} \rfloor \times c}$ and geometric tokens $\mathcal{T}_{i,j}^{G'} \in \mathbb{R}^{\lfloor \frac{h}{2p} \rfloor \times \lfloor \frac{w}{2p} \rfloor \times c}$.

After token merging, we introduce an independent gating mechanism for each layer j (where $j = 1, 2, \ldots, L$) to dynamically balance geometric and visual contributions. The gating coefficient for layer j is defined as:

$$g_{i,j} = \sigma\left(\mathrm{MLP}\left(\left[\mathcal{T}_{i,j}^{V'}, \mathcal{T}_{i,j}^{G'}\right]\right)\right), \tag{2}$$

where σ denotes the sigmoid activation function mapping values to $[0, 1]$, MLP represents a two-layer perceptron, and $[\cdot, \cdot]$ indicates concatenation. The fused tokens $\left\{\mathcal{T}_{i,j}^{S}\right\}_{j=1}^{L}$ are then computed as:

$$\mathcal{T}_{i,j}^{S} = g_{i,j} \odot \mathcal{T}_{i,j}^{V'} + (1 - g_{i,j}) \odot \mathcal{T}_{i,j}^{G'}, \tag{3}$$

where $\odot$ denotes element-wise multiplication. This per-layer token-level gating allows geometry–semantics weighting to vary across representation layers and spatial locations, ensuring that structural information remains sufficiently activated rather than uniformly blended.

To further enhance hierarchical structural preservation, we draw inspiration from DeepStack [36] and Qwen3-VL [2] to design a multi-modal DeepStack strategy. Specifically, from $\left\{\mathcal{T}_{i,j}^{S}\right\}_{j=1}^{L}$, the final-layer tokens $\mathcal{T}_{i,L}^{S}$ are concatenated with textual embeddings as the primary input to the video LLM decoder. Meanwhile, selected intermediate-layer tokens $\mathcal{T}_{i,m}^{S}$ with $m \in \{L_1, L_2, L_3\}$ are injected into early decoder layers by direct addition to the corresponding hidden states. This hierarchical injection maintains geometric signals throughout decoding, preventing structural cues from being suppressed by dominant semantic representations and promoting sustained geometric activation.

3.2 Training

Recent works have proposed multiple object-level datasets for RGB-only 3D scene perception [55,58], covering tasks such as 3D visual grounding, dense captioning, and video object detection. However, directly fine-tuning MLLMs on

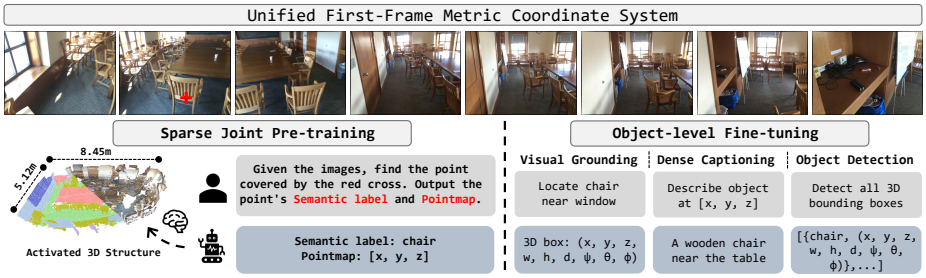

Fig. 5: Training Strategy. Sparse joint pre-training (left) activates 3D representations under a unified first-frame metric coordinate system. The learned representations are then transferred to object-level fine-tuning (right) for downstream 3D perception.

object-level supervision remains insufficient to activate implicit geometric priors, as optimization is dominated by language-driven objectives. We argue that an explicit geometry-aligned pre-training stage is necessary to strengthen structural perception before downstream adaptation. Accordingly, we design a two-stage training paradigm, as illustrated in Fig. 5, consisting of sparse pixel-level geometry–semantics pre-training followed by object-level perception fine-tuning.

Consistent Metric Coordinate System. RGB-based 3D perception under implicit geometric priors lacks explicit constraints on the output space. Without a unified coordinate definition, predictions across tasks become inconsistent and undermine structural supervision. To enforce geometry-aligned learning, we adopt a consistent metric coordinate system for all tasks. Following VGGT [47], we use the first-frame coordinate system as the shared reference frame. All numerical outputs are represented in metric space and rounded to two decimal places, ensuring structural consistency across tasks.

Sparse Geometry-Semantics Joint Pre-training. Inspired by DepthLM [4], we observe that pure vision models still struggle with metric spatial perception. Even when equipped with implicit geometric priors, multi-view structural prediction remains weak under language-dominated supervision. Therefore, we introduce a sparse geometry–semantics joint pre-training objective to explicitly activate geometric representations under cross-modal guidance.

Given image sequences $\{I_1, \ldots, I_n\}$, visual prompts, and a natural language query Q, the model predicts both the 3D coordinate (x, y, z) of the prompted pixel and its semantic label c under the unified first-frame coordinate system, as shown in Fig. 5. By jointly supervising structural coordinates and semantic categories, this objective aligns geometric priors with language representations and explicitly activates 3D structural perception during early optimization.

Object-level 3D Perception Fine-tuning. After sparse geometry-semantics joint pre-training, we evaluate whether the activated structural representations transfer to complex object-level 3D perception tasks.

3D Visual Grounding: Following previous works [57], we formulate 3D visual grounding as a 3D video grounding problem that predicts the frame index where

the target object appears and its 3D bounding box in that coordinate system. Unlike VG-LLM [58], which performs grounding in a single feed-forward pass, we decompose it into a two-stage procedure. In the first stage, given an image sequence and a language query, the model predicts the target frame index I_t. In the second stage, we reorganize the sequence with I_t as the first frame and predict the 3D bounding box $(x, y, z, w, h, d, \psi, \theta, \phi)$, where (x, y, z) denotes the center coordinate, (w, h, d) the object size, and (ψ, θ, ϕ) the rotation angles.

3D Dense Captioning: Similar to visual grounding, we adopt a two-stage dense captioning approach [61]. Given a frame sequence and the spatial coordinate (x, y, z) of an object center under the unified first-frame coordinate system, the model generates a descriptive caption conditioned on the corresponding 3D location. This formulation preserves coordinate consistency across tasks.

3D Video Object Detection: Following [58], our 3D video object detection requires the model to output all objects appearing in a multi-frame sequence under the same first-frame coordinate system. The output is represented as $\{(b_i, c_i)\}$, where b_i follows the same bounding box parameterization as in 3D visual grounding and c_i denotes the object category.

4 Experiments

We evaluate **GAP-MLLM** on RGB-only 3D scene perception to validate our claim that geometry-aligned pre-training paradigm activates structural perception and improves implicit-prior utilization under sparse supervision. Sec. 4.1 describes datasets, baselines, metrics, and implementation details. Sec. 4.2 reports results on 3D visual grounding, dense captioning, and video object detection, and includes metric 3D reconstruction as a probe of metric-aware perception after sparse prompt-based pre-training. Sec. 4.3 analyzes our sparse geometry-semantics pre-training and multi-level gated fusion.

4.1 Setting

Datasets and Baselines. We first conduct sparse pixel-level geometry semantics pre-training, and then perform multi-task fine-tuning on object-level RGB-only 3D perception datasets.

1. ***Sparse Joint Pre-training***: We use EmbodiedScan [50] and ScanNet [16]. To improve semantic label quality, we adopt frame-aligned 3D annotations from EmbodiedScan. Specifically, object centers are projected to 2D and then back-projected to metric 3D coordinates in the first-frame coordinate system using ScanNet depth maps and camera poses. Semantic labels are inherited from EmbodiedScan. In total, the pre-training set contains approximately 500K samples; however, since supervision is provided only at sparse prompted pixels, the effective pixel-level supervision is roughly equivalent to two 680×480 images. Each training sample contains four consecutive frames sampled at 1 FPS, where a red cross marks the prompted pixel location.

2. ***3D Visual Grounding***: We use ScanRefer [9] and formulate it as 3D video grounding: given a text query, the model predicts the target frame index

and a 3D bounding box under that frame's camera coordinates. Following VG-LLM [58], EmbodiedScan [50] annotations are used as ground-truth boxes.

3. ***3D Dense Captioning***: We adopt Scan2Cap [14]. Mask3D proposals extracted by LEO [24] are used as input, and the model generates captions conditioned on object center coordinates under the first-frame coordinate system.

4. ***3D Video Object Detection***: We use EmbodiedScan [50] with frame-aligned dense 3D bounding boxes. Four consecutive frames are sampled at 1 FPS, and all annotations are transformed to the first-frame coordinate system [58]. The dataset is split into 958 training and 243 evaluation scenes.

Baselines. We compare against both *explicit 3D-input* methods and *RGB-only* methods that rely on implicit geometric priors. For 3D visual grounding and captioning, we include task-specific models ScanRefer [9] and Scan2Cap [14], as well as generalist 3D/LLM models Chat-3D v2 [23], Grounded 3D-LLM [12], LL3DA [11], LLaVA-3D [61], Video-3D LLM [59], and VG-LLM [58]. For 3D video object detection, we compare with SpatialLM [35] and VG-LLM [58]. For metric 3D reconstruction, we compare with CUT3R [48] and MapAnything [26].

Metrics. ScanRefer: Acc@0.25/0.5. Scan2Cap: CIDEr, BLEU-4, ROUGE on IoU ≥ 0.5. 3D Video Detection: Precision/Recall/F1 at IoU 0.25 over 20 classes. Metric 3D Reconstruction: Accuracy/Completeness/Overall on 10 ScanNet [16] validation scenes under aligned and metric evaluation.

Implementation Details. Our main instantiation of GAP-MLLM is built upon Qwen3-VL-2B [2], with VGGT-1B [47] as the geometric encoder. To demonstrate architectural scalability, we also apply the same training paradigm and fusion design to the VG-LLM setting [58]. Both sparse joint pre-training and mixed-task fine-tuning are conducted for one epoch using Adam, with a warm-up ratio of 0.03 and a peak learning rate of 1e-5. The batch size is 32 for joint pre-training and 16 for mixed-task fine-tuning. The visual and geometric encoders each contain $L = 24$ layers. For multi-level fusion, we select intermediate layers $L_1 = 5$, $L_2 = 11$, and $L_3 = 17$, while the final layer L is used as the primary decoder input. We freeze the visual encoder of the MLLM and the geometric encoder to provide stable multimodal token representations, and optimize the MLLM backbone (LLM) together with the multi-level fusion module.

4.2 Evaluations

Results on 3D Visual Grounding. As shown in Tab. 1, GAP-MLLM significantly outperforms prior RGB-only methods that rely on implicit geometric priors. Compared with VG-LLM-4B [58], our 3B model improves Acc@0.25 and Acc@0.5 by over 11 points while using fewer parameters, achieving comparable Acc@0.25 to 3D-input methods. These results indicate that geometry-aligned pre-training effectively activates structural perception within MLLMs. Fig. 6 further illustrates that our predictions exhibit tighter geometric alignment and more accurate box structures. Even when both models identify the same object, our bounding boxes better conform to the underlying spatial geometry.

Table 1: ScanRefer Results. Ours surpasses prior RGB-only methods and approaches explicit 3D models.

Model	3D Input	Acc@0.25	Acc@0.5
ScanRefer [9]	✓	37.3	24.3
3D-LLM [22]	✓	30.3	-
Chat-3D v2 [23]	✓	35.9	30.4
Grounded 3D-LLM [12]	✓	47.9	44.1
ChatScene [23]	✓	55.5	50.2
LLaVA-3D [61]	✓	54.1	42.4
Video-3D LLM [59]	✓	**58.1**	**51.7**
SPAR [57]	✗	31.9	12.4
VG-LLM-4B [58]	✗	36.4	11.8
VG-LLM-4B (w/ GAP)	✗	49.7	23.2
GAP-MLLM-3B (Ours)	✗	**53.1**	**26.0**

Table 2: Scan2Cap Results. GAP-MLLM achieves the best overall performance among both 3D-input and RGB-only methods.

Model	3D Input	C@0.5↑	B-4@0.5↑	R@0.5↑
Scan2Cap [14]	✓	39.1	23.3	44.8
LL3DA [11]	✓	65.2	36.8	55.0
Chat-3D-v2 [23]	✓	63.9	31.8	-
Grounded 3D-LLM [12]	✓	70.2	35.0	-
LEO [24]	✓	72.4	38.2	58.1
Chat-Scene [23]	✓	77.1	36.5	-
LLaVA-3D [61]	✓	79.2	**41.1**	**63.4**
Video-3D LLM [59]	✓	**80.0**	40.2	61.7
VG-LLM-4B [58]	✗	78.6	40.9	62.4
VG-LLM (w/ GAP)	✗	80.6	41.1	62.8
GAP-MLLM-3B (Ours)	✗	**84.7**	**42.1**	**63.1**

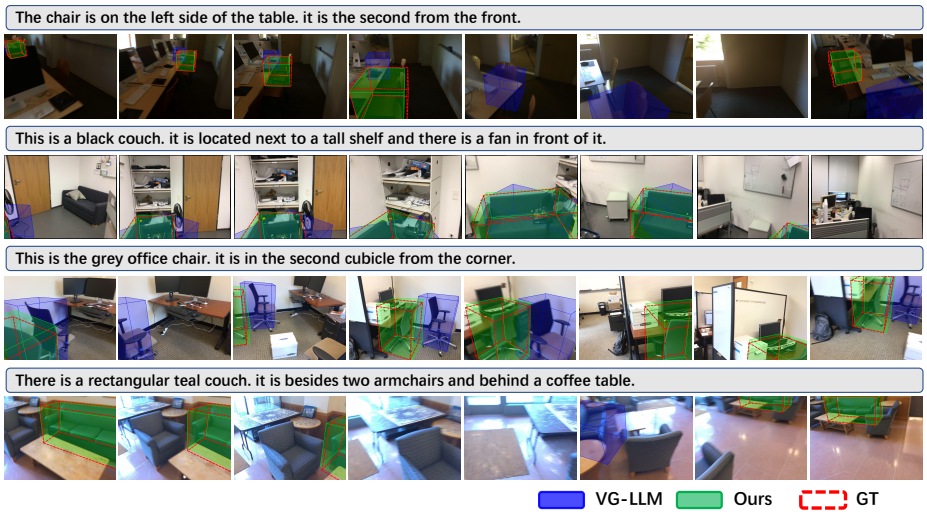

Fig. 6: Qualitative comparison on 3D visual grounding. We compare VG-LLM and GAP-MLLM against ground truth (GT). Our predictions exhibit tighter geometric alignment and more accurate box fitting under the same textual descriptions.

Results on 3D Dense Captioning. Tab. 2 shows that GAP-MLLM achieves the best performance among RGB-only approaches and surpasses several explicit 3D-input methods. Our model obtains 84.7 CIDEr and 42.1 BLEU-4 at IoU 0.5, demonstrating that geometry-aligned pre-training improves coordinate-conditioned grounding and description in image sequences.

Results on 3D Video Object Detection. As shown in Tab. 3 and Fig. 7, SpatialLM [35], which relies on explicit 3D input, achieves reasonable precision but low recall, indicating sensitivity to incomplete geometry. RGB-based VG-LLM [58] improves recall through stronger semantic perception but remains limited in overall F1. In contrast, GAP-MLLM consistently achieves the highest precision, recall, and F1 under both 4-frame and 6-frame settings. These results demonstrate that geometry-aligned pre-training and gated fusion enable more effective utilization of implicit geometric priors for accurate spatial localization.

Table 3: 3D Video Object Detection Results. GAP-MLLM achieves the highest precision, recall, and F1 under both 4-frame and 6-frame settings.

Model	3D Input	4-Frame Setting			6-Frame Setting		
		P_{25}	R_{25}	$F1_{25}$	P_{25}	R_{25}	$F1_{25}$
SpatialLM [35]	✓	40.8	23.7	29.1	42.1	26.1	31.1
VG-LLM-4B [58]	✗	41.7	35.7	38.2	39.7	34.0	36.4
VG-LLM-8B [58]	✗	43.4	39.6	41.2	43.5	38.7	40.8
VG-LLM-4B (w/ GAP)	✗	47.0	41.2	43.5	48.8	40.2	43.6
GAP-MLLM-3B (Ours)	✗	**54.2**	**48.1**	**50.6**	**52.9**	**45.4**	**48.5**

Fig. 7: Qualitative comparison on 3D video object detection. We compare SpatialLM [35], VG-LLM [58], and GAP-MLLM. Our method produces more accurate and spatially consistent 3D bounding boxes across multi-frame inputs.

Results on Pointmaps and Semantic Labels. Our sparse joint pre-training enables the MLLM to generate metric-consistent point and semantic maps. As shown in Fig. 8, the reconstructed geometry is structurally coherent and spatially accurate under metric supervision. The predicted semantic maps inherit the abstract semantic capacity of the MLLMs. Although boundaries are not sharp due to sparse supervision, the model reliably captures object extents and structure.

Tab. 4 reports aligned and metric evaluations on 10 random ScanNet [16] scenes. While CUT3R [48] achieves slightly better performance under Sim(3) aligned evaluation, GAP-MLLM performs best under metric evaluation, indicating stronger metric-aware reconstruction. Furthermore, semantic supervision yields improvements in metric consistency compared to pointmap only training, highlighting the role of semantic guidance in stabilizing metric structure.

Table 4: PointMap Estimation Results on ScanNet (meter). We report Accuracy, Completeness, and Overall under aligned evaluation (after Sim(3) alignment to GT) and metric evaluation (direct comparison with GT without alignment).

Model	Aligned Evaluation						Metric Evaluation					
	Acc. ↓		Comp. ↓		Overall. ↓		Acc. ↓		Comp. ↓		Overall. ↓	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
CUT3R [48]	0.0302	0.0227	0.0273	0.0207	0.0287	0.0217	0.1199	0.0998	0.1045	0.0871	0.1122	0.0937
Mapanything [26]	0.0688	0.0397	0.0610	0.0480	0.0649	0.0438	0.3409	0.2721	0.2498	0.2265	0.2953	0.2493
GAP-MLLM-3B (w/o semantic)	0.0424	0.0295	0.0326	0.0246	0.0375	0.0270	0.0691	0.0580	0.0570	0.0465	0.0630	0.0523
GAP-MLLM-3B (Ours)	0.0421	0.0293	0.0327	0.0247	0.0374	0.0270	0.0675	0.0553	0.0557	0.0441	0.0616	0.0497

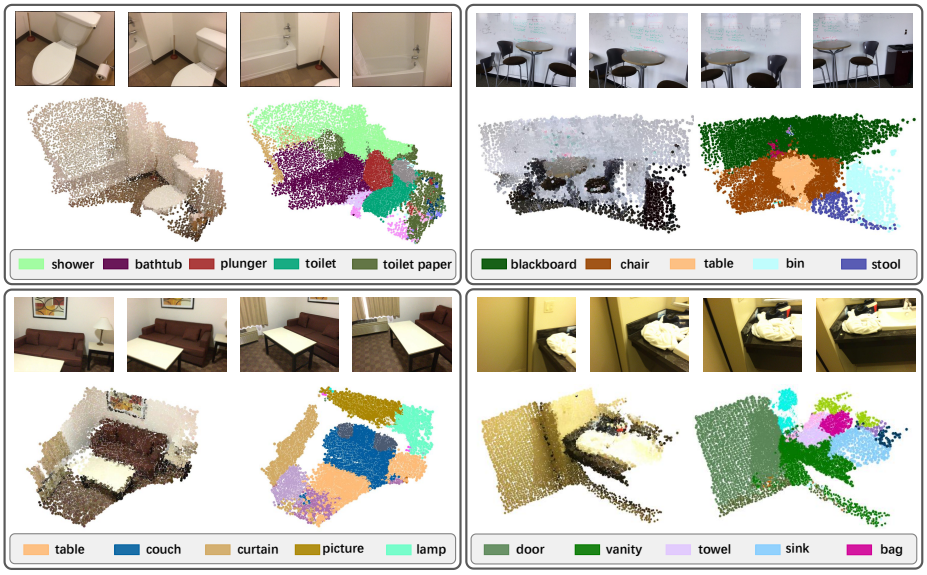

Fig. 8: Qualitative results on pointmap and semantic prediction. Despite being trained with sparse supervision, GAP-MLLM reconstructs metrically consistent 3D pointmaps and dense semantic regions. White points denote the *unknown* category.

4.3 Analysis

Ablation Study on Core Contribution. Tab. 5 shows that both sparse joint pre-training (A) and multi-level fusion (B) improve metrics in 3D video object detection. When combined, they achieve the best performance, demonstrating complementary effects. Applying GAP to VG-LLM [58] yields consistent gains, confirming that our geometry-aligned design generalizes across architectures.

Sparse Geometry-Semantics Joint Pre-training. To examine whether geometry aware pre-training enhances inherent spatial perception, we evaluate a pure Qwen3-VL [2] backbone without geometric modules. As shown in Tab. 6, sparse joint pre-training consistently improves all metrics, indicating strengthened geometric perception even without explicit 3D inputs or implicit priors.

Table 5: Ablation on 3D video object detection. Both sparse joint pre-training (A) and multi-level gated fusion (B) contribute complementary gains.

A	B	VG-LLM-4B [58] (Qwen2.5-VL + VGGT)			GAP-MLLM-3B (Qwen3-VL + VGGT)		
		P_{25}	R_{25}	$F1_{25}$	P_{25}	R_{25}	$F1_{25}$
		41.7	35.7	38.2	48.8	42.1	44.7
✓		44.1	39.6	41.4	52.7	45.9	48.7
	✓	43.3	36.9	39.4	51.8	44.6	47.5
✓	✓	**47.0**	**41.2**	**43.5**	**54.2**	**48.1**	**50.6**

Table 6: Effect of geometry-aware pre-training on pure VLM. Pre-training enhances geometric perception, even without geometric encoder.

Model	P_{25}	R_{25}	$F1_{25}$
Qwen3-VL	43.0	37.3	39.7
+ Joint Pre-training	**45.5**	**40.1**	**42.4**

Table 7: Effect of joint geometric-semantic pre-training. Joint supervision outperforms pointmap-only training in 3D perception.

Model	P_{25}	R_{25}	$F1_{25}$
VG-LLM [58]	41.7	35.7	38.2
+ Pointmap	43.1	38.3	40.1
+ Pointmap + Semantic	**44.1**	**39.6**	**41.4**

We further compare pointmap-only and joint geometric–semantic supervision on VG-LLM [58]. As shown in Tab. 7, joint supervision yields additional gains, suggesting that semantic guidance helps structure geometric representations. This trend aligns with the metric reconstruction results in Tab. 4.

Ablation on Fusion Strategies. Tab. 8 compares different fusion strategies. Simple addition or weighted fusion provides limited gains, while cross-attention improves interaction. Our gated fusion achieves the best results, highlighting the importance of adaptive token-level integration.

Table 8: Ablation on token-level fusion strategies. Gated fusion achieves the best performance with higher flexibility.

Fusion Type	Precision	Recall	F1
Add	51.7	45.3	47.9
Weighted	51.4	45.5	47.8
Cross-Attention	52.7	46.3	49.1
Gated (Ours)	**54.2**	**48.1**	**50.6**

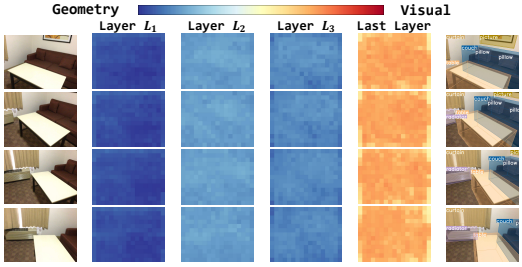

Fig. 9: Visualization of Layer-wise gating weights. Intermediate layers emphasize geometry, while later layers focus on semantics.

Fig. 9 visualizes layer-wise gating weights. Intermediate layers emphasize geometric priors, while the final layer assigns higher weight to semantic features. This layer-dependent behavior confirms that geometry and semantics play complementary roles across stages, and gated fusion enables flexible balancing.

5 Conclusion

In this work, we revisit RGB-only 3D perception from a training perspective and argue that the performance gap of implicit geometric prior-based methods does not stem from insufficient geometry, but from the lack of effective activation within text-dominated fine-tuning paradigms. To address this issue, we propose **GAP-MLLM**, a geometry-aligned pre-training paradigm that explicitly activates structural perception before downstream adaptation. Through sparse geometry–semantics joint supervision and multi-level gated fusion, GAP-MLLM strengthens metric-aware spatial perception while preserving semantic capability. Extensive experiments across 3D visual grounding, 3D dense captioning, 3D video object detection, and metric 3D reconstruction demonstrate that geometry-aligned pre-training consistently improves the utilization of implicit geometric priors and generalizes across different MLLM backbones. We hope this work provides a new perspective on training MLLMs for 3D spatial understanding and highlights the importance of activation-oriented pre-training for multimodal perception in the absence of explicit 3D inputs.

Acknowledgments. The research was supported by the National Natural Science Foundation of China (U23B2009, 62471158).

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In: European conference on computer vision. pp. 422–440. Springer (2020)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: 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. Cai, Z., Yeh, C.F., Xu, H., Liu, Z., Meyer, G., Lei, X., Zhao, C., Li, S.W., Chandra, V., Shi, Y.: Depthlm: Metric depth from vision language models. arXiv preprint arXiv:2509.25413 (2025)
5. Cao, Y., Jv, Y., Xu, D.: 3dgs-det: Empower 3d gaussian splatting with boundary guidance and box-focused sampling for 3d object detection. arXiv preprint arXiv:2410.01647 (2024)
6. Cao, Y., Wu, F., Chen, D.Z., Zhong, Y., Hong, L., Xu, D.: Vggt-det: Mining vggt internal priors for sensor-geometry-free multi-view indoor 3d object detection. arXiv preprint arXiv:2603.00912 (2026)
7. Cao, Y., Yihan, Z., Xu, H., Xu, D.: Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. Advances in Neural Information Processing Systems **36**, 71862–71873 (2023)

8. Cao, Y., Zeng, Y., Xu, H., Xu, D.: Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3d object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
9. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in RGB-D scans using natural language. In: ECCV (2020)
10. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D 3 net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In: European Conference on Computer Vision. pp. 487–505. Springer (2022)
11. 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)
12. Chen, Y., Yang, S., Huang, H., Wang, T., Xu, R., Lyu, R., Lin, D., Pang, J.: Grounded 3d-llm with referent tokens. arXiv preprint arXiv:2405.10370 (2024)
13. 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)
14. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3193–3203 (2021)
15. Chen, Z., Hu, R., Chen, X., Nießner, M., Chang, A.X.: Unit3d: A unified transformer for 3d dense captioning and visual grounding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 18109–18119 (2023)
16. 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)
17. Dwedari, M.M., Niessner, M., Chen, D.Z.: Generating context-aware natural answers for questions in 3d scenes. arXiv preprint arXiv:2310.19516 (2023)
18. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279 (2025)
19. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
20. Gao, X., Zhang, Z., Chen, D.Z., Xu, S., Quan, L., Pérez-Pellitero, E., Jang, Y.: Map2thought: Explicit 3d spatial reasoning via metric cognitive maps. arXiv preprint arXiv:2601.11442 (2026)
21. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
22. 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)
23. 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)
24. 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)
25. 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)

26. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
27. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
28. Koch, S., Wald, J., Matsuki, H., Hermosilla, P., Ropinski, T., Tombari, F.: Unified semantic transformer for 3d scene understanding. arXiv preprint arXiv:2512.14364 (2025)
29. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
30. Li, H., Zou, Z., Liu, F., Zhang, X., Hong, F., Cao, Y., Lan, Y., Zhang, M., Yu, G., Zhang, D., et al.: Iggt: Instance-grounded geometry transformer for semantic 3d reconstruction. arXiv preprint arXiv:2510.22706 (2025)
31. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
32. Li, Q., Ye, Z., Feng, X., Zhong, W., Ma, W., Feng, X.: Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 31645–31653 (2026)
33. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
34. Ma, X., Smart, B., Bhalgat, Y., Chen, S., Li, X., Ding, J., Gu, J., Chen, D.Z., Peng, S., Bian, J.W., et al.: When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models. arXiv preprint arXiv:2405.10255 (2024)
35. Mao, Y., Zhong, J., Fang, C., Zheng, J., Tang, R., Zhu, H., Tan, P., Zhou, Z.: Spatiallm: Training large language models for structured indoor modeling. In: Advances in Neural Information Processing Systems (2025)
36. Meng, L., Yang, J., Tian, R., Dai, X., Wu, Z., Gao, J., Jiang, Y.G.: Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. Advances in Neural Information Processing Systems **37**, 23464–23487 (2024)
37. Qi, C.R., Chen, X., Litany, O., Guibas, L.J.: Imvotenet: Boosting 3d object detection in point clouds with image votes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4404–4413 (2020)
38. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019)
39. Qi, Z., Zhang, Z., Fang, Y., Wang, J., Zhao, H.: Gpt4scene: Understand 3d scenes from videos with vision-language models. arXiv preprint arXiv:2501.01428 (2025)
40. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al.: Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint arXiv:2501.15830 (2025)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

42. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
43. Shen, Y., Zhang, Z., Qu, Y., Zheng, X., Ji, J., Zhang, S., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025)
44. Sheng, Y., Deng, J., Zhang, X., Zhang, Y., Hua, B., Zhang, Y., Ji, J.: Spatialsplat: Efficient semantic 3d from sparse unposed images. arXiv preprint arXiv:2505.23044 (2025)
45. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
46. Wang, H., Zhou, H., Liu, H., Yan, L.: 4d-vggt: A general foundation model with spatiotemporal awareness for dynamic scene geometry estimation. arXiv preprint arXiv:2511.18416 (2025)
47. 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)
48. 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)
49. 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)
50. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., 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. pp. 19757–19767 (2024)
51. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
52. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
53. Xu, Y., Zhang, J., Huang, Z., Chen, Y., Zhou, Y., Chen, Z., Yuan, Y.J., Xia, P., Huang, G., Cai, X., et al.: Uniugg: Unified 3d understanding and generation via geometric-semantic encoding. arXiv preprint arXiv:2508.11952 (2025)
54. 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)
55. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025)
56. Yuan, S., Yang, Y., Yang, X., Zhang, X., Zhao, Z., Zhang, L., Zhang, Z.: Infinitevggt: Visual geometry grounded transformer for endless streams. arXiv preprint arXiv:2601.02281 (2026)
57. 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)
58. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. arXiv preprint arXiv:2505.24625 (2025)

59. 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)
60. Zheng, X., Dongfang, Z., Jiang, L., Zheng, B., Guo, Y., Zhang, Z., Albanese, G., Yang, R., Ma, M., Zhang, Z., et al.: Multimodal spatial reasoning in the large model era: A survey and benchmarks. arXiv preprint arXiv:2510.25760 (2025)
61. Zhu, C., Wang, T., Zhang, W., Pang, J., Liu, X.: Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. arXiv preprint arXiv:2409.18125 (2024)
62. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)