This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# SpaceMind: Camera-Guided Modality Fusion for Spatial Reasoning in Vision-Language Models

Ruosen Zhao[1,2,*] Zhikang Zhang[1,*] Jialei Xu[1] Jiahao Chang[2]
Dong Chen[1,3] Lingyun Li[1] Weijian Sun[1] Zizhuang Wei[1,†]
[1] Huawei [2] The Chinese University of Hong Kong, Shenzhen
[3] The University of Hong Kong
*Equal contribution. †Corresponding author.

## Abstract

*Large vision-language models (VLMs) show strong multimodal understanding but still struggle with 3D spatial reasoning, such as distance estimation, size comparison, and cross-view consistency. Existing 3D-aware methods either depend on auxiliary 3D information or enhance RGB-only VLMs with geometry encoders through shallow feature fusion. We propose SpaceMind, a multimodal large language model trained on SpaceMind-700K for spatial reasoning from RGB inputs. The model adopts a dual-encoder architecture integrating VGGT as a spatial understanding encoder and InternViT as a 2D visual encoder. The key idea is to treat the camera representation as an active guiding modality rather than passive metadata. Specifically, SpaceMind introduces a lightweight Camera-Guided Modality Fusion module before the language model to replace shallow fusion. It applies camera-conditioned biasing to spatial tokens, assigns query-independent weights reflecting their geometric importance, and uses the camera embedding to gate the fused representation. Empirically, SpaceMind establishes new **state-of-the-art** results on **VSI-Bench**, **SQA3D** and **SPBench**, surpassing both open and proprietary systems on VSI-Bench and SPBench by large margins and achieving state-of-the-art performance on SQA3D. These results demonstrate that camera-guided modality fusion is an effective and practical inductive bias for equipping VLMs with genuinely spatially grounded intelligence. We will release code and model checkpoints to support future research.*

## 1. Introduction

Humans perceive space not merely by seeing, but by understanding *from where* they see. This coupling between visual observation and viewpoint underlies our ability to estimate distances, compare sizes, infer connectivity, and navigate unfamiliar environments. Endowing multimodal models with such **spatial intelligence** [68] is crucial for a wide range of applications. However, despite remarkable progress in large language models (LLMs) [6, 7, 28, 45, 48, 49, 56, 57, 63, 76] and multimodal large language models (MLLMs) [1–3, 9, 19, 20, 26, 34, 39] on open-ended reasoning, instruction following, and long-context understanding, recent studies consistently show that they struggle with explicitly 3D-aware tasks, such as metric comparison, layout inference, and multi-view consistency.

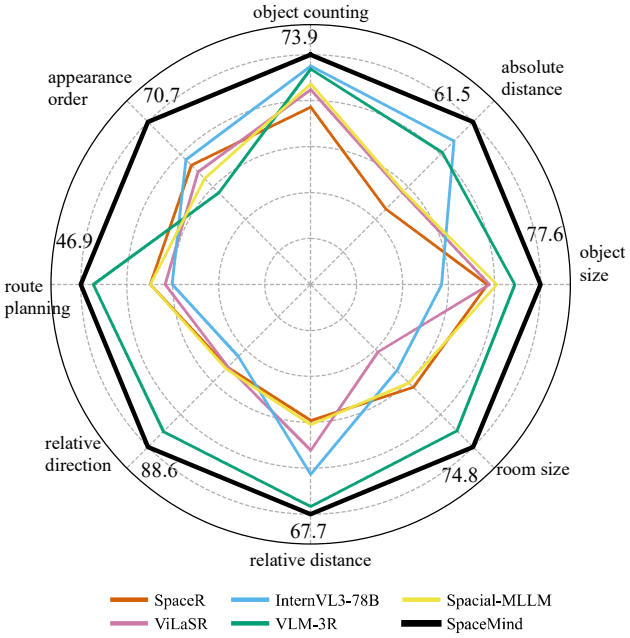



Figure 1. Performance on VSI-Bench across different spatial reasoning categories. SpaceMind achieves consistently strong visuospatial intelligence compared to existing systems.

Existing approaches to spatially grounded multimodal reasoning can be roughly grouped into two families. The first family augments language or vision-language models with *explicit 3D inputs*, such as point clouds, depth maps, reconstructed meshes, or BEV/voxel features [13, 25, 41, 43, 61, 71, 77, 78, 81, 82]. These methods leverage depth sensors or multi-view reconstruction pipelines to build 3D scene representations, which are then encoded and aligned with language. They can capture rich geometric details, but inherit several structural limitations: reliance on specialized hardware or pre-scanned environments, heavy multi-stage processing, sensitivity to reconstruction failures and scale ambiguity, and difficulty scaling to unconstrained in-the-wild video. The second family, exemplified by recent 3D-aware MLLMs, targets *visual-based spatial intelligence* from monocular or multi-view RGB [15, 65]. They typically combine a general-purpose visual encoder [3, 72] with a feed-forward geometry encoder [58, 59] and then fuse the resulting features via simple mechanisms such as MLP projection or one-stage cross-attention. These designs avoid explicit 3D supervision and are more scalable, but the way they integrate spatial cues remains largely ad hoc.

A closer look at these two lines of work reveals a common limitation. Explicit-3D methods over-commit to precomputed geometry and do not address how viewpoint should interact with language-native reasoning. Visual-based methods, on the other hand, tend to treat all geometric signals—image features, spatial features, and camera information—as if they belonged to a single homogeneous feature space. In practice, camera-related signals are often appended as auxiliary embeddings or implicitly blended into fusion layers. This conflation ignores a principle long emphasized in 3D vision [27, 35, 51, 58, 74]: *camera* (ego/viewpoint) and *scene* (allocentric content) play fundamentally different roles.

Motivated by this insight, we present **SpaceMind**, a large vision-language model explicitly designed for 3D spatial reasoning from RGB inputs. SpaceMind treats the camera representation as a *dedicated modality* that guides how spatial information is injected into the visual stream, rather than as a passive conditioning vector. Concretely, SpaceMind uses a visual encoder [9] and a spatial encoder [58] to obtain semantic and geometry-aware tokens, and introduces a *Camera-Guided Modality Fusion* (CGMF) module before the language model. CGMF applies camera-conditioned biasing over spatial tokens, a query-independent spatial importance weighting, and camera-conditioned gating on fused features, making the role of the camera explicit in the fusion process while preserving the interface of standard VLMs. This design is intentionally simple—built from standard components—but aligns viewpoint, spatial cues, and semantics within a unified multimodal framework. Empirically, SpaceMind achieves a new state of the art on **VSI-Bench** [68], with a clear margin over existing open and proprietary systems, and delivers comparable or superior performance on other spatial reasoning benchmarks such as SQA3D [40] and SPBench [55], indicating strong generalization across datasets and task formats. In summary, our contributions are four-fold:

- We propose **SpaceMind**, a camera-guided multimodal large language model tailored for 3D spatial reasoning from visual observations.
- We identify a key conceptual gap in prior work: existing fusion paradigms often conflate camera and scene features, whereas explicitly treating the camera representation as a guiding modality leads to more coherent and robust spatial reasoning.
- We introduce a simple yet effective Camera-Guided Modality Fusion (CGMF) module that integrates camera, spatial, and visual tokens within a unified framework.
- We show that SpaceMind achieves new **state-of-the-art** performance across diverse spatial reasoning datasets and tasks, consistently outperforming all prior methods.

## 2. Related Work

### 2.1. MLLMs for Image and Video Understanding

Multimodal large language models extend large language models to vision and other modalities by aligning visual and textual tokens in a shared embedding space. CLIP [50] and ALIGN [1] learn joint image-text representations via contrastive pretraining, enabling strong zero-shot transfer. Flamingo [1] and BLIP-2 [34] introduce token-level or feature-level fusion modules (e.g., Q-Former, cross-attention bridges) to couple frozen LLM backbones with vision encoders, while instruction-tuned systems such as the LLaVA family [31, 32, 37, 38, 75], MiniGPT-4 [79], Qwen-VL [3], and InternVL [9] further enhance open-ended visual QA. Recent video MLLMs [36, 42, 54, 64, 73] extend these designs to temporal inputs through spatiotemporal pooling, causal attention over frame tokens, or long-context architectures, and achieve strong performance on video understanding benchmarks.

Despite their success, these models are mostly optimized for semantic and temporal understanding rather than geometric reasoning. They typically treat videos as sequences of 2D observations, with little explicit modeling of camera motion, global scene layout, or cross-view consistency. As a result, they perform well on recognition-style tasks but remain unreliable for metric judgments or layout reconstruction, motivating architectures with stronger inductive bias toward 3D spatial understanding.

### 2.2. 3D Visual-Based Spatial Intelligence

A growing body of work aims to endow multimodal models with 3D awareness. One line adopts explicit 3D

inputs [13, 25, 41, 43, 61, 71, 77, 78, 81, 82]: models take point clouds, meshes, RGB-D scans, or voxel-style features as input and align them with language using Q-Formers, 3D detectors, or volumetric aggregation. These approaches demonstrate strong 3D grounding when high-quality geometry is available, but rely on depth sensors or reconstruction pipelines, making them sensitive to reconstruction errors and less scalable to unconstrained Alongside these sensor-dependent designs, a second line of work including SpaceR [47] and VILASR [66] attempts to elicit 3D reasoning directly from existing VLMs *without* introducing additional geometry encoders. Operating in the same RGB-only regime as our setting, these methods keep the native visual backbone mostly frozen and instead rely on carefully engineered data and training strategies, but improvements on rigorous spatial benchmarks remain modest.

monocular video.

Complementary work such as VLM-3R [15] and Spatial-MLLM [65] augments VLMs with geometry-aware encoders that infer spatial tokens and camera tokens from images or short videos, followed by lightweight fusion layers. These approaches demonstrate clear improvements on spatial benchmarks [40, 68], validating the value of injecting geometric priors. Yet, their fusion mechanisms typically mix camera and scene tokens within a shared space via simple concatenation or single-stage cross-attention, without explicitly modeling their distinct roles. Our work follows the same RGB-only, geometry-augmented paradigm but introduces a fusion design where the camera representation serves as an explicit guiding modality for spatial tokens.

### 2.3. 3D Reconstruction from Images

Progress in 3D reconstruction [29, 44, 62, 74] provides the geometric priors that many 3D-aware multimodal systems exploit. Classical pipelines based on Structure-from-Motion [52] and Multi-View Stereo [18] recover camera poses and dense depth through feature matching, bundle adjustment, and volumetric or patch-based aggregation, but require multi-stage optimization and often assume controlled capture conditions. Learnable MVS approaches, such as MVSNet and its variants [23, 69], predict per-view depth or cost volumes under known cameras, improving efficiency and accuracy while still depending on calibrated multi-view input.

Feed-forward visual-geometry models such as DUSt3R [60], MASt3R [30], and VGGT [58] directly infer pixel-wise 3D correspondences, point maps, and camera parameters from image pairs or short sequences, often without requiring explicit calibration. By casting reconstruction as a single forward pass of a transformer-style backbone, these models relax assumptions on camera poses, scale to in-the-wild data, and unify tasks such as depth estimation, camera pose recovery, and dense point cloud reconstruction. Such advances disentangle camera (viewpoint) and scene (geometry) representations and make it practical to obtain geometry-aware spatial tokens and camera tokens from ordinary RGB sequences, motivating architectures like SpaceMind that focus on fusing these signals to support reliable spatial reasoning in multimodal language models.

## 3. Method

We propose **SpaceMind**, a multimodal large language model for spatial reasoning from RGB inputs. As illustrated in Figure 2, SpaceMind treats language, vision, and camera viewpoint as three complementary modalities, each with distinct roles in spatial reasoning. Unlike prior work that conflates these signals, SpaceMind explicitly separates them: language provides semantic context, vision captures appearance, and camera encodes the geometric perspective from which the scene is observed. Rather than folding camera information into the visual or spatial stream as auxiliary metadata, SpaceMind exposes it as a first-class modality that actively shapes how geometric evidence is injected into the language backbone. We describe the overall architecture in Sec. 3.1 and the proposed Camera-Guided Modality Fusion (CGMF) module in Sec. 3.2.

### 3.1. SpaceMind Architecture

Given a text prompt $T$ and an image sequence $\{I_i\}_{i=1}^{N}$, SpaceMind routes inputs through three modality-specific streams before language generation.

**Visual stream.** A ViT-based visual encoder $e_v$ (InternViT-300M [80]) extracts per-frame semantic tokens:

$$f_v = e_v(\{I_i\}_{i=1}^{N}), \quad f_v \in \mathbb{R}^{N\times M_v\times d_v}. \tag{1}$$

**Spatial and camera streams.** A feed-forward spatial encoder $e_s$ (VGGT [58]) simultaneously produces geometry-aware patch tokens and a per-frame camera token encoding the viewpoint:

$$f_s,\ f_c = e_s(\{I_i\}_{i=1}^{N}) \tag{2}$$

Here $f_s \in \mathbb{R}^{N\times M_s\times d_s}$ carries allocentric scene geometry, while $f_c \in \mathbb{R}^{N\times 1\times d_s}$ encodes the egocentric viewpoint of each frame—two signals that 3D reconstruction models [58, 60, 74] have shown benefit from explicit disentanglement.

**Modality fusion and language generation.** The three streams are combined by the CGMF module $F$, whose output preserves the shape of the visual tokens so that the pre-trained LLM backbone $g$ (InternVL3-8B [80]) requires no architectural changes:

$$f_{\text{fused}} = F(f_v,\ f_s,\ f_c), \quad f_{\text{fused}} \in \mathbb{R}^{N\times M_v\times d_v}, \tag{3}$$

$$R = g(f_{\text{fused}},\ T). \tag{4}$$

**Design principle.** Prior geometry-augmented VLMs [15, 65] concatenate camera tokens with spatial tokens and pass

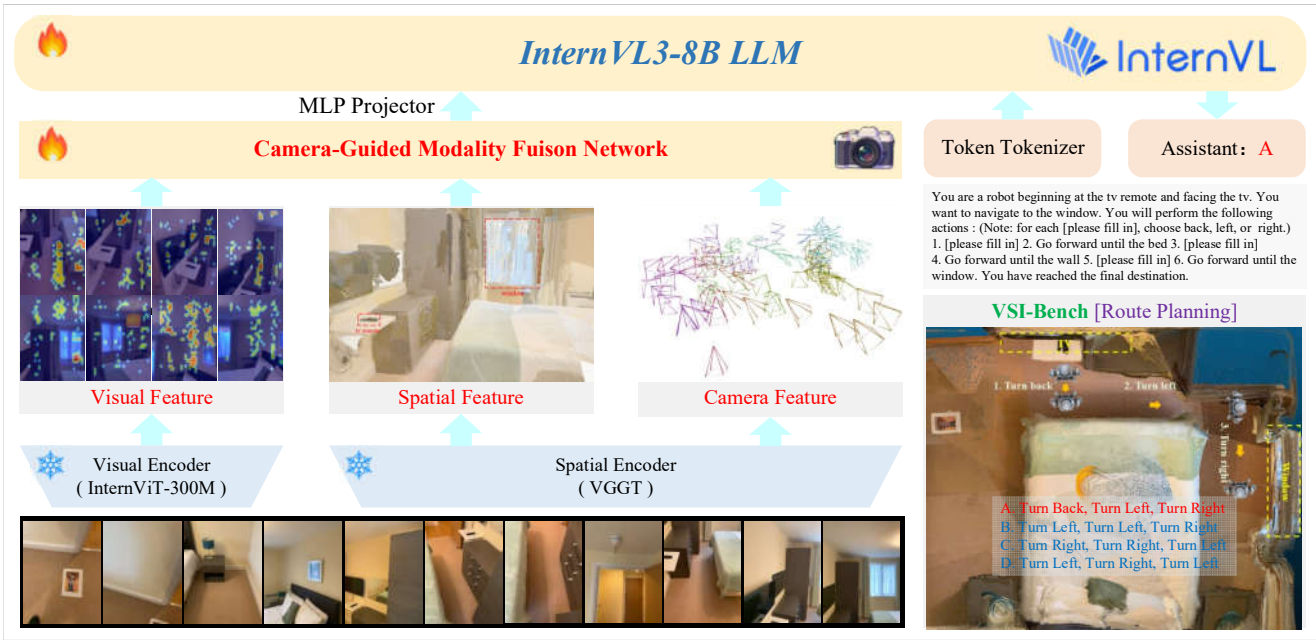



Figure 2. Overall pipeline of **SpaceMind**. Given a text prompt and an image sequence, a visual encoder produces semantic visual tokens, while a spatial encoder produces geometry-aware tokens together with per-frame camera tokens that summarize viewpoint information. The proposed Camera-Guided Modality Fusion (CGMF) module takes these three streams as input: it uses camera tokens to modulate spatial tokens, estimates their relative importance, and injects the resulting spatial cues into the visual tokens. The fused, view-aware visual tokens preserve the original token shape expected by the multimodal LLM, enabling SpaceMind to be trained end-to-end on RGB-only data while remaining compatible with standard VLM architectures and enhancing their 3D spatial reasoning ability.

the result through a single fusion stage, collapsing viewpoint and scene into one homogeneous sequence. CGMF instead exposes the camera as a *dedicated control modality*: $f_c$ first conditions how spatial tokens represent the scene from the current viewpoint, then gates how strongly the enriched spatial evidence influences the visual backbone—a role analogous to how language governs attention in instruction-following VLMs, but operating over geometry.

### 3.2. Camera-Guided Modality Fusion (CGMF)

The CGMF module $F$ takes $f_v$, $f_s$, and $f_c$ as input and returns fused features of the same shape as $f_v$. The module realizes the camera-as-modality principle through three successive operations: (i) camera-conditioned spatial encoding, (ii) geometry-reliability weighting, and (iii) camera-conditioned output gating.

**Projection into a shared space.** All inputs are first projected into a shared attention space of width $d_a$:

$$Q = P_Q(\mathrm{LN}(f_v)), \quad K = P_K(\mathrm{LN}(f_s)),$$
$$V = P_V(\mathrm{LN}(f_s)), \quad C = P_C(f_c), \tag{5}$$

where $Q \in \mathbb{R}^{N\times M_v\times d_a}$, $K, V \in \mathbb{R}^{N\times M_s\times d_a}$, and $C \in \mathbb{R}^{N\times 1\times d_a}$.

**(i) Camera-conditioned spatial encoding.** Rather than treating spatial keys and values as view-agnostic features, we re-centre them with respect to the current camera. A small MLP jointly processes each spatial token with its frame-wise camera token to produce a geometric bias $B_g$, which is added to both $K$ and $V$:

$$B_g = \mathrm{MLP}([\mathrm{LN}(f_s),\, C_{\mathrm{expand}}]) \in \mathbb{R}^{N\times M_s\times d_a},$$
$$K \leftarrow K + B_g, \quad V \leftarrow V + B_g. \tag{6}$$

This lets the model encode viewpoint-dependent geometric structure directly into the attention memory.

To further make the camera modality visible across all frames, we build a cross-frame camera memory by tiling $C$ across the frame axis, adding sine positional encodings to index frame order, and reinforcing the diagonal (current-frame) entry with a learnable embedding $e_{\mathrm{cur}}$. The resulting $\tilde{C} \in \mathbb{R}^{N\times N\times d_a}$ is prepended to the key–value memory so that visual queries can directly attend to viewpoint information from any frame:

$$K' = [\tilde{C};\, K], \quad V' = [\tilde{C};\, V]. \tag{7}$$

**(ii) Geometry-reliability weighting.** Geometric predictions are not uniformly reliable across a scene: occluded regions, textureless surfaces, and depth discontinuities all yield noisier estimates. Motivated by the per-pixel confidence maps in DUSt3R [60] and VGGT [58], we predict a

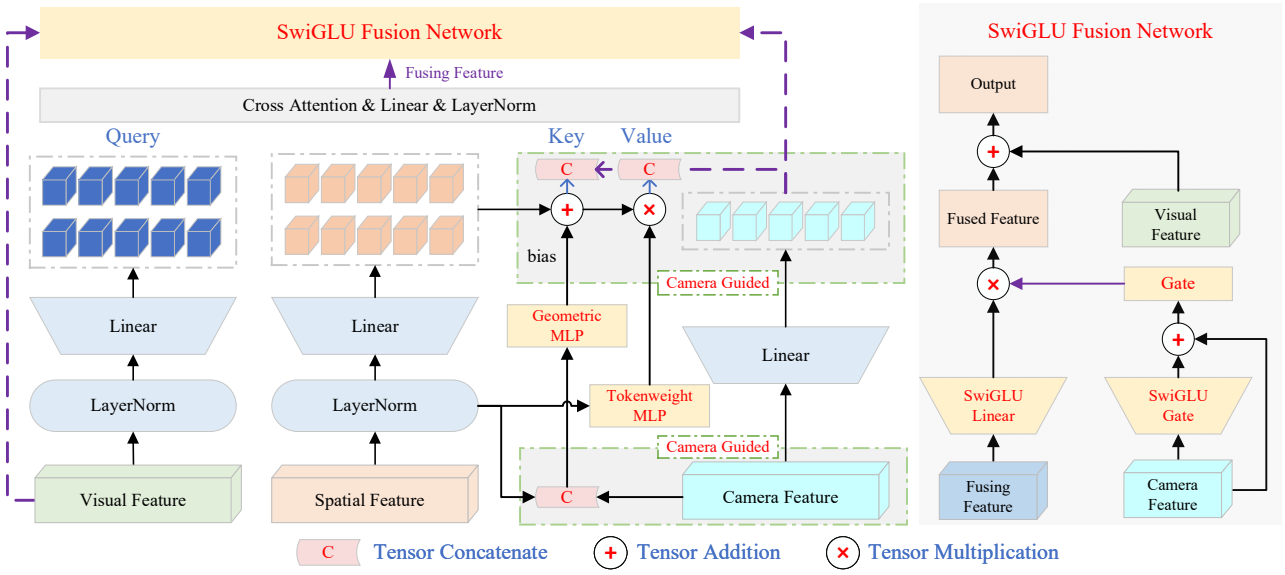



Figure 3. The architecture of the CGMF module. CGMF takes visual tokens $f_v$, spatial tokens $f_s$, and camera tokens $f_c$ as input, and outputs fused visual tokens with the same shape as $f_v$. The camera features are leveraged to refine the construction of geometric residuals, which further guide the cross-attention fusion between visual and spatial features. In addition, the SwiGLU Fusion Network follows the SwiGLU mechanism to achieve efficient multimodal feature fusion under the guidance of camera information.

scalar importance weight for each spatial token purely from the spatial branch, providing a query-independent prior that down-weights unreliable geometric evidence before attention is computed. Concretely, spatial tokens undergo a lightweight self-attention pass for context aggregation, and the result is fed through an MLP to yield per-token weights:

$$W_t = \sigma\big(\mathrm{MLP}(\mathrm{SelfAttn}(f_s) + f_s)\big) \in \mathbb{R}^{N\times M_s\times 1}. \quad (8)$$

Only the spatial slots of $V'$ are rescaled by $W_t$; the $N$ prepended camera memory slots are left unmodified, so that viewpoint information is never suppressed by the geometric reliability filter. Visual queries then attend over the conditioned memory:

$$\hat{f} = \mathrm{Attn}(Q,\, K',\, V') \in \mathbb{R}^{N\times M_v\times d_a}. \quad (9)$$

**(iii) Camera-conditioned output gating.** After projecting $\hat{f}$ back to the visual feature space, the camera modality makes one final contribution: it controls how strongly the spatially enriched features modify the visual backbone's representation. Inspired by SwiGLU [11, 53], we derive a per-frame gate $g$ from the camera summary $\bar{C} = \mathrm{squeeze}(C)$:

$$g = \mathrm{Swish}(P_{g,1}(\bar{C})) \odot P_{g,2}(\bar{C}) \in \mathbb{R}^{N\times d_v}, \quad (10)$$

and apply it as a residual modulation over the projected output:

$$f_{\text{fused}} = P_L(\mathrm{LN}(P_O(\hat{f}))) \odot g[:, \mathrm{None}, :] + f_v. \quad (11)$$

The gate $g$ is produced entirely from the camera modality, so its magnitude reflects how informative the current viewpoint is for spatial reasoning—naturally amplifying the spatial contribution when viewpoint provides strong geometric cues and suppressing it otherwise.

Overall, CGMF instantiates the camera-as-modality principle at three levels of the fusion pipeline: it shapes the geometric content of spatial tokens (i), filters their reliability (ii), and calibrates their influence on the visual backbone (iii). These three operations work synergistically: the camera first re-centers spatial geometry to the current viewpoint, then the reliability filter ensures only trustworthy geometric evidence is propagated, and finally the camera-conditioned gate controls how strongly this evidence modifies the visual representation. All operations are built from standard components—MLPs, self-attention, cross-attention, and SwiGLU gating—and require no architectural changes to the LLM backbone, preserving full compatibility with standard VLM training and inference pipelines.

## 4. Experiments

### 4.1. Implementation Details

**Backbone.** SpaceMind adopts InternVL3-8B [80] as the language model backbone and its paired image encoder, InternViT-300M [80], as the visual encoder. For geometric understanding, We adopt VGGT [58] to ensure a fair com-

Table 1. **Evaluation on VSI-Bench [68].** SpaceMind sets a new state-of-the-art, achieving the best average score and outperforming all prior models on every individual subtask, often by a large margin.

| Methods | Avg. | Numerical Question | | | | Multiple-Choice Question | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Obj. Cnt. | Abs. Dist. | Obj. Size | Room Size | Rel. Dist. | Rel. Dir. | Route Plan | Appr. Order |
| *Proprietary Models (API)* | | | | | | | | | |
| GPT-5 [46] | 55.0 | 53.3 | 34.4 | **73.3** | 47.5 | **63.7** | 48.6 | 50.2 | **68.9** |
| Gemini-2.5 Pro [21] | 53.5 | 46.0 | 37.3 | 68.7 | 54.3 | 61.9 | 43.9 | 47.4 | 68.7 |
| Gemini-3 Pro [22] | **56.0** | 49.0 | **42.8** | 71.5 | 41.8 | 56.6 | **57.5** | **61.9** | 60.0 |
| Grok-4 [67] | 47.9 | 37.1 | 32.9 | 60.8 | 45.4 | 53.1 | 39.6 | 47.4 | 66.8 |
| *Open-source VLMs* | | | | | | | | | |
| InternVL3-78B [80] | **48.5** | **71.2** | **53.7** | 44.4 | **39.5** | **55.9** | 39.5 | 28.9 | **54.5** |
| LLaVA-NeXT-Video-7B [75] | 35.6 | 48.5 | 14.0 | 47.8 | 24.2 | 43.5 | **42.4** | 34.0 | 30.6 |
| LLaVA-NeXT-Video-72B [75] | 40.9 | 48.9 | 22.8 | 57.4 | 35.3 | 42.4 | 36.7 | **35.0** | 48.6 |
| QWen2.5VL-7B [4] | 33.0 | 40.9 | 14.8 | 43.4 | 10.7 | 38.6 | 38.5 | 33.0 | 29.8 |
| LLaVA-OneVision-7B [31] | 32.4 | 47.7 | 20.2 | 47.4 | 12.3 | 42.5 | 35.2 | 29.4 | 24.4 |
| LLaVA-OneVision-72B [31] | 40.2 | 43.5 | 23.9 | **57.6** | 37.5 | 42.5 | 39.9 | 32.5 | 44.6 |
| *Specialized Spatial Reasoning Models* | | | | | | | | | |
| Spacer [47] | 45.5 | 57.8 | 28.2 | 59.9 | 47.1 | 40.1 | 45.4 | 33.5 | 52.1 |
| ViLaSR [66] | 45.4 | 63.5 | 34.4 | 60.6 | 30.9 | 48.9 | 45.2 | 30.4 | 49.2 |
| Spatial-MLLM [65] | 48.4 | 65.3 | 34.8 | 63.1 | 45.1 | 41.3 | 46.2 | 33.5 | 46.3 |
| VLM-3R [15] | 60.9 | 70.2 | 49.4 | 69.2 | 67.1 | 65.4 | 80.5 | 45.4 | 40.1 |
| **SpaceMind (Ours)** | **70.2** | **73.9** | **61.5** | **77.6** | **74.8** | **67.7** | **88.6** | **46.9** | **70.7** |

parison, following prior work such as Spatial-MLLM [65]. This shows that the performance gains stem from architectural design rather than backbone capacity.

**Training setup.** The model is fine-tuned for two epochs on SpaceMind-700K, which aggregates VLM-3R-data [15], ViCA-322K [17], and the training split of SQA3D [40]. During fine-tuning, we freeze both the visual and spatial encoders, fully update the proposed CGMF module, and apply Low-Rank Adaptation (LoRA) [24] with rank 256 and scaling factor 512 to the InternVL3-8B language model backbone. We employ a cosine learning-rate scheduler with an initial learning rate of $2 \times 10^{-5}$ and a warm-up ratio of 0.03, and train with a global batch size of 64. The full training process required roughly 25 hours on a cluster of 64 NVIDIA H100 (80GB) GPUs.

**Data preprocessing.** For both training and inference, SpaceMind uniformly samples 34 frames from each scene, discards the first and last frames, and uses the remaining 32 frames as input. Each image is resized to $448 \times 448$ for InternViT and zero-padded to $518 \times 518$ for VGGT, following its original configuration.

## 4.2. Evaluations

We evaluate SpaceMind on three benchmarks, including VSI-Bench [68], SQA3D [40] (in-domain), and SPBench [33] (out-of-domain). We will release code and model checkpoints to support future research.

**Evaluation on VSI-Bench.** VSI-Bench [68] contains over 5,000 questions curated from ARKitScenes [5], ScanNet [10], and ScanNet++ [70]. It comprises two answer formats—Numerical Answer (NA) and Multiple-Choice Answer (MCA)—covering eight subtasks: object counting, relative direction, absolute distance, route planning, object size, room size, relative distance, and appearance order. Following the official protocol, we evaluate NA with relative accuracy and MCA with mean accuracy. As shown in Table 1, **SpaceMind** attains the best score on all eight subtasks and raises the overall average to **70.2**, an improvement of **9.3** over the strongest prior baseline (VLM-3R-7B, 60.9). Gains appear in both NA and MCA groups: measurement-oriented tasks (absolute/relative distance, object/room size) improve consistently, while view-integration tasks show especially large gains on *relative direction* and *appearance order*. Data-centric RGB-only baselines such as SpaceR [47] and VILASR [66], which keep the visual backbone largely fixed and focus on curriculum design, achieve averages of only 45.5 and 45.4 respectively—barely matching a generic VLM and still performing poorly on geometry-heavy subtasks like absolute/relative distance and appearance order—indicating that training and data tricks alone are insufficient to elicit robust 3D reasoning from 2D features. Geometry-augmented models such as Spatial-MLLM [65] and VLM-3R [15] narrow the gap by introducing explicit spatial tokens, but their one-stage fusion of camera and scene representations still leaves substantial headroom: they remain notably weaker than SpaceMind on view-integration metrics such as *relative direction* and especially *appearance order*. In particular, *appearance or-*

Table 2. **Evaluation on SQA3D [40] Test Split.** SpaceMind achieves the best performance across most question types and establishes a new state of the art on both EM@1 and EM@R1, despite using video-only inputs while many existing methods rely on richer modalities.

| Method | TestSet | | | | | | EM@1 | EM@R1 | **Video-Input Only** |
|---|---|---|---|---|---|---|---|---|---|
| | What | Is | How | Can | Which | Others | | | |
| PQ3D [82] | 37.1 | 61.3 | 44.5 | 60.9 | 47.0 | 45.1 | 47.1 | 49.3 | No |
| 3D-VisTA [81] | 34.8 | 63.3 | 45.4 | 69.8 | 47.2 | 48.1 | 48.5 | 50.9 | No |
| LEO [41] | 39.0 | 63.9 | 44.9 | 66.2 | 47.7 | 51.1 | 50.0 | 52.4 | No |
| SIG3D [43] | 35.6 | 67.2 | 48.5 | 71.4 | 49.1 | 45.8 | 52.6 | 54.4 | No |
| Scene-LLM [61] | 40.9 | 69.1 | 45.0 | 70.8 | 47.2 | 52.3 | 54.2 | 56.2 | No |
| ChatScene [25] | 45.4 | 67.0 | 52.0 | 69.5 | 49.9 | 55.0 | 54.6 | 57.5 | No |
| Video-3D LLM [77] | 51.1 | 72.4 | 55.5 | 69.8 | 51.3 | **56.0** | 58.6 | 60.8 | No |
| **SpaceMind (Ours)** | **54.1** | **74.8** | **61.7** | **71.0** | **51.9** | 53.6 | **61.1** | **63.8** | Yes |

Table 3. **Evaluation on SPBench [33].** All models are evaluated without using SPBench training data. SpaceMind achieves the best performance on both SPBench-SI and SPBench-MV, outperforming general-purpose VLMs and prior spatial models by a clear margin.

| Methods | SPBench-SI | | | SPBench-MV | | | Overall |
|---|---|---|---|---|---|---|---|
| | NQ | MCQ | Avg. | NQ | MCQ | Avg. | |
| *Proprietary Models* | | | | | | | |
| GPT-4o [26] | 24.5 | 60.3 | 42.4 | 40.7 | 59.4 | 50.1 | 46.2 |
| Gemini-2.0-Flash [12] | 49.0 | 60.4 | 54.7 | 51.9 | 50.7 | 51.3 | 53.0 |
| *Open-Source Models* | | | | | | | |
| InternVL-2.5-4B [8] | 31.8 | 53.3 | 42.6 | **37.7** | 51.4 | **44.6** | 43.6 |
| InternVL-2.5-8B [8] | 28.3 | 56.3 | 42.3 | 37.3 | 47.5 | 42.4 | 42.3 |
| Kimi-VL-A3B [14] | 25.7 | 44.9 | 35.3 | 23.3 | **57.6** | 40.5 | 37.9 |
| LLaVA-OneVision-7B [31] | 25.4 | 41.0 | 33.2 | 20.6 | 49.6 | 35.1 | 34.2 |
| Qwen2.5-VL-7B [4] | **36.3** | **60.5** | **48.4** | 28.9 | 49.8 | 39.3 | **43.9** |
| Qwen2.5-VL-3B [4] | 24.3 | 56.2 | 40.3 | 25.6 | 53.2 | 39.4 | 39.8 |
| *Spatial Reasoning Models* | | | | | | | |
| Video-R1 [16] | 27.7 | 62.0 | 44.9 | 32.5 | 53.0 | 42.8 | 43.8 |
| SpaceR-7B [47] | 35.7 | 61.5 | 48.6 | 63.2 | 53.7 | 58.5 | 53.5 |
| VILASR-7B [66] | 36.6 | **63.7** | 50.2 | 56.2 | 59.6 | 57.9 | 54.0 |
| Spatial-MLLM-4B [65] | 38.1 | 49.3 | 43.7 | 63.7 | 58.9 | 61.3 | 52.5 |
| **SpaceMind (Ours)** | **66.3** | 53.2 | **59.7** | **76.2** | **70.5** | **73.8** | **67.3** |

*der* shows a lead of nearly **30.6** points, which is the largest increase among all subtasks. This suggests that explicitly conditioning spatial tokens on the camera signal helps consolidate cross-view evidence and stabilize ordering judgments across viewpoints, while *route planning* also exhibits a smaller but consistent increase.

**Evaluation on SQA3D.** SQA3D [40] is a situated 3D question-answering benchmark built from reconstructed indoor scenes in ScanNet [10]. It provides textual situations and questions that require localizing an egocentric viewpoint and reasoning about nearby objects, relations, and navigation targets. The test split covers diverse question templates (*What*, *Is*, *How*, *Can*, *Which*, and *Others*). We follow the standard protocol and report exact-match accuracy (EM@1) and its refined variant (EM@R1). As shown in Table 2, SpaceMind achieves the best performance across most question types and sets a new state of the art on both metrics, reaching 61.1 on EM@1 and 63.8 on EM@R1. Unlike prior methods that rely on depth, point clouds, meshes, or other auxiliary modalities, SpaceMind uses only video inputs. This demonstrates that our camera-guided fusion can recover strong 3D spatial cues directly from RGB videos, enabling robust situated reasoning under realistic video-only settings where dense 3D supervision is unavailable.

**Evaluation on SPBench.** We assess out-of-domain generalization on SPBench [33], which is derived from the ScanNet [10] validation split via the SpatialLadder-26k pipeline and contains two subsets: SPBench-SI (single-view) and SPBench-MV (multi-view). Each subset includes numerical questions (NQ) and multiple-choice questions (MCQ) spanning absolute/relative distance, relative direction, object size, and (for SPBench-MV) cross-view object counting. Following the official protocol, we report mean relative

Table 4. **Ablation Study on VSI-Bench.** We systematically analyze SpaceMind's design choices: (1) the contribution of spatial encoding via VGGT, (2) different strategies for integrating camera tokens as a guiding modality, and (3) the components of our proposed Camera-Guided Modality Fusion (CGMF). Results demonstrate that treating the camera as an independent modality, rather than a passive conditioning signal, is essential for robust spatial reasoning.

| Methods | Avg. | Numerical Question | | | | Multiple-Choice Question | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Obj. Cnt. | Abs. Dist. | Obj. Size | Room Size | Rel. Dist. | Rel. Dir. | Route Plan | Appr. Order |
| *Part 1: Spatial Encoding Baseline* | | | | | | | | | |
| InternVL3-8B ft | 63.7 | 71.0 | 51.1 | 74.7 | 65.4 | 64.9 | 73.2 | 42.8 | 66.6 |
| InternVL3-8B + VGGT | 64.6 | 70.1 | 54.9 | 72.2 | 67.4 | 59.7 | 82.6 | 44.6 | 65.2 |
| *Part 2: Camera Integration Strategies* | | | | | | | | | |
| InternVL3-8B + VGGT + ConcatCam | 67.4 | 72.4 | 59.1 | 74.8 | 70.6 | 65.8 | 85.4 | 44.5 | 66.6 |
| InternVL3-8B + VGGT + MultiAttn | 67.9 | 73.3 | 59.9 | 75.4 | 71.0 | 66.0 | 85.9 | 44.3 | 67.3 |
| *Part 3: CGMF Components* | | | | | | | | | |
| InternVL3-8B + VGGT + twMLP | 67.8 | 73.3 | 58.3 | 75.4 | 72.7 | 66.4 | 85.2 | 44.0 | 67.2 |
| InternVL3-8B + VGGT + twMLP + geoMLP | 69.4 | 73.5 | 60.7 | 76.7 | 74.4 | **67.9** | 87.9 | 45.6 | 68.4 |
| **SpaceMind (full)** | **70.2** | **73.9** | **61.5** | **77.6** | **74.8** | 67.7 | **88.6** | **46.9** | **70.7** |

accuracy for NQ and accuracy for MCQ; the SI/MV scores are the arithmetic mean of their NQ and MCQ results, and the overall score is the average of SI and MV. Importantly, SPBench is not included in our training data and thus serves as a genuine out-of-domain evaluation. As shown in Table 3, SpaceMind substantially outperforms both general-purpose VLMs and prior specialized spatial models, establishing a new state of the art. Notably, on the *single-image* subset, SpaceMind outperforms all competing methods by a clear margin *despite being trained exclusively on 32-frame clips per QA*, indicating strong cross-regime transfer and that our camera-guided fusion remains effective even when the input collapses to a single view.

### 4.3. Ablation Studies

Ablation studies are conducted on VSI-Bench to quantify the contribution of each SpaceMind component to 3D spatial reasoning. All variants follow the same training protocol described in Sec. 4.1, and the corresponding results are reported in Table 4. For clarity, the analysis is organized into three parts: the spatial encoding baseline, camera integration strategies, and CGMF components.

**Part 1: Spatial Encoding Baseline.** Directly finetuning InternVL3-8B already achieves a solid 63.7% accuracy. Adding VGGT spatial tokens via shallow cross-attention further improves performance to 64.6%, with notably stronger gains on absolute distance and relative direction. This clearly establishes that geometric cues are to 2D visual features and provide measurable benefits.

**Part 2: Camera Integration Strategies.** We next examine how camera information should be integrated with spatial and visual features. *ConcatCam* concatenates camera tokens with spatial tokens, achieving 67.4%. *MultiAttn* applies sequential cross-attention, reaching 67.9%. Both treat camera as a homogeneous feature stream without explicitly modeling the distinct roles of viewpoint and scene, resulting in only modest improvements over the baseline.

**Part 3: CGMF Components.** We introduce three key ingredients to make the camera's role explicit and structured. Adding token-weight MLP yields 67.8%, thereby providing a learned importance prior over spatial tokens. Introducing geometric MLP raises accuracy to 69.4%, with particularly substantial gains on viewpoint-sensitive tasks like relative direction. Finally, the camera-conditioned SwiGLU gate achieves 70.2%, with consistent improvements across nearly all subtasks.

**Summary.** Naive camera integration provides modest gains, while CGMF's explicit treatment of camera as a guiding modality yields substantially larger improvements. This validates our core principle: camera and scene should be treated as distinct modalities, with the camera actively guiding spatial information injection.

## 5. Conclusion

We propose **SpaceMind**, a multimodal large language model for 3D spatial reasoning that revisits fusion from a camera-centric perspective. Rather than collapsing geometry-related signals into a single stream, SpaceMind treats the camera representation as a guiding modality and integrates it via the proposed Camera-Guided Modality Fusion (CGMF) module. CGMF applies a camera-conditioned biasing, query-independent weighting over spatial evidence, and camera-conditioned gating while preserving the interface of standard VLMs and operating purely on RGB inputs. Empirically, SpaceMind achieves a new state of the art on VSI-Bench with a clear margin and delivers comparable or superior results on other spatial reasoning benchmarks, indicating that how viewpoint and geometry are fused is a key factor in visuospatial intelligence.

## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 7

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *Advances in Neural Information Processing Systems*, 2021. 6

[6] Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *arXiv preprint arXiv:2305.00948*, 2023. 1

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. 1

[8] Zhe Chen, Weiyun Wang, Yue Cao, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 7

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 6, 7

[11] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, 2017. 5

[12] Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era. Google Blog, 2024. 7

[13] Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3

[14] Angang Du, Bohong Yin, Bowei Xing, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 7

[15] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 2, 3, 6

[16] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025. 7

[17] Qi Feng and Hidetoshi Shimodaira. Visuospatial cognitive assistant. *arXiv preprint arXiv:2505.12312*, 2025. 6

[18] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 3

[19] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[20] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[21] Google DeepMind. Gemini 2.5 pro: Access google's latest preview ai model, 2025. Accessed: 2026-03-21. 6

[22] Google DeepMind. Gemini 3 pro: the frontier of vision ai, 2025. Accessed: 2026-03-21. 6

[23] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2020. 3

[24] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[25] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao,

Jiangmiao Pang, and Zhou Zhao. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *Advances in Neural Information Processing Systems*, 2024. 2, 3, 7

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 7

[27] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, and Feng Zhao. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025. 2

[28] Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schütze, and Peter Clark. Language models with rationality. In *Conference on Empirical Methods in Natural Language Processing*, 2023. 1

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. 3

[30] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R. In *European Conference on Computer Vision*, 2024. 3

[31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 6, 7

[32] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 2

[33] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025. 6, 7

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 1, 2

[35] Zhiqi Li, Chengrui Dong, Yiming Chen, Zhangchi Huang, and Peidong Liu. Vicasplat: A single run is all you need for 3d gaussian splatting and camera estimation from unposed video frames. *arXiv preprint arXiv:2503.10286*, 2025. 2

[36] 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. *arXiv preprint arXiv:2311.10122*, 2023. 2

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 2

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2024. 1

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

[41] Xiaojian Ma, Jiangyong Huang, Silong Yong, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. Leo: An embodied generalist agent in 3d world. In *International Conference on Machine Learning*, 2024. 2, 3, 7

[42] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2

[43] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 7

[44] 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. In *European Conference on Computer Vision*, 2020. 3

[45] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023. 1

[46] OpenAI. Gpt-5 system card, 2025. Accessed: 2026-03-21. 6

[47] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 3, 6, 7

[48] Alec Radford. Improving language understanding by generative pre-training. OpenAI Blog, 2018. 1

[49] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019. 1

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2

[51] Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[52] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 3

[53] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 5

[54] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 2

[55] Ilias Marios Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025. 2

[56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[58] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3, 4, 5

[59] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. CUT3R. 2

[60] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 4

[61] Tai Wang, Chenming Zhu, Haoxuan Deng, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Scene-llm: Extending language model for 3d visual understanding and reasoning. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. 2, 3, 7

[62] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. $\pi^3$: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 3

[63] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. 1

[64] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. *arXiv preprint arXiv:2404.03384*, 2024. 2

[65] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems*, 2025. 2, 3, 6, 7

[66] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025. 3, 6, 7

[67] xAI. Grok 4 model card. Technical report, 2025. Accessed: 2026-03-21. 6

[68] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 2, 3, 6

[69] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2018. 3

[70] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *IEEE/CVF International Conference on Computer Vision*, 2023. 6

[71] Zhihao Yuan, Shuyi Jiang, Chun-Mei Feng, Yaolun Zhang, Shuguang Cui, Zhen Li, and Na Zhao. Scene-r1: Video-grounded large language models for 3d scene reasoning without 3d annotations. *arXiv preprint arXiv:2506.17545*, 2025. 2, 3

[72] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, Lucas Beyer, Joan Puigcerver, Jessica Yung, Michael Tschannen, Fabian Mentzer, and Neil Houlsby. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 2

[73] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[74] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3

[75] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. LLaVA Blog, https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024. 2, 6

[76] Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. Unveiling linguistic regions in large language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024. 1

[77] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3, 7

[78] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *IEEE/CVF International Conference on Computer Vision*, 2025. 2, 3

[79] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

[80] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3, 5, 6

[81] 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 *IEEE/CVF International Conference on Computer Vision*, 2023. 2, 3, 7

[82] 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*, 2024. 2, 3, 7