# SpatialSV: Internalizing Interpretable 3D Spatial Awareness in MLLMs via Task-Oriented Visual Supervision

**Jiayu Tang**, **Yuchen Zhou**, **Chao Gou***
School of Intelligent Systems Engineering, Sun Yat-sen University
tangjy59@mail2.sysu.edu.cn, zhouych37@mail2.sysu.edu.cn, gouchao@mail.sysu.edu.cn

## Abstract

Unlocking the spatial intelligence of multimodal large language models (MLLMs) is crucial for understanding and interacting with the 3D world. Prevailing approaches typically inject spatial priors via external tools, which impose significant inference overhead, or rely on latent feature distillation, which remains uninterpretable and lacks fine-grained geometric constraints. To address these issues, we propose SpatialSV, a framework designed to internalize robust 3D spatial awareness within MLLMs while simultaneously offering inherent interpretability. Deviating from passive feature imitation, SpatialSV employs task-oriented visual supervision, compelling the model to actively lift its 2D visual features into explicit 3D representations, including depth maps, camera poses, and point clouds. Crucially, this 2D-to-3D lifting process provides a transparent window into the model's representations: the resulting 3D reconstructions serve as an intuitive proxy for visualizing and diagnosing the quality of the model's intrinsic spatial knowledge. Extensive experiments across multiple models and benchmarks demonstrate the effectiveness of SpatialSV in enhancing and interpreting MLLMs' spatial intelligence. Furthermore, the framework exhibits strong generalization in semi-supervised settings, validating its potential to leverage unlabeled visual data for scalable, interpretable spatial representation learning.

## 1 Introduction

Spatial intelligence refers to the ability to understand, interpret and reason about the spatial relationships and properties of objects and scenes. This ability is fundamental for real-world tasks ranging from autonomous driving [Zhou *et al.*, 2025; Tang *et al.*, 2026] to robotic manipulation [Song *et al.*, 2025], where understanding and interacting with the 3D environments is essential. Despite significant progress in spatial reasoning and scene understanding, even the most advanced multimodal large language models (MLLMs) face challenges in maintaining cross-view consistency and reasoning about occluded objects. These limitations highlight a fundamental weakness of MLLMs—the reliance on 2D visual-textual data and autoregressive training paradigms—which fails to yield robust and consistent internal 3D spatial representations. This absence of intrinsic 3D spatial awareness is a major bottleneck toward genuine spatial intelligence.

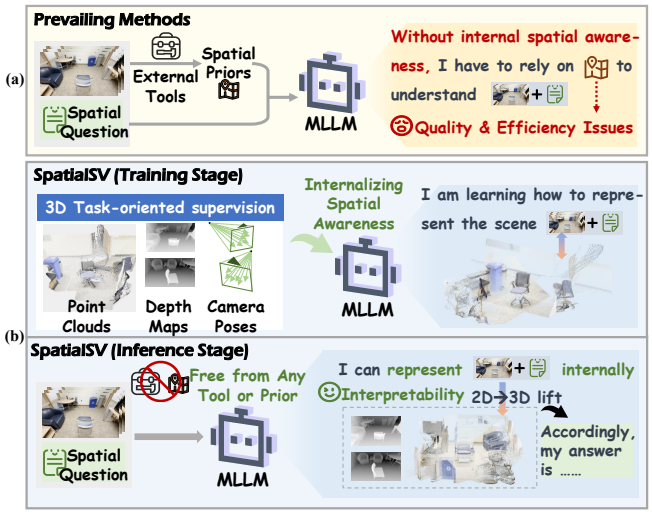


Figure 1: (a) Prevailing methods incorporate external spatial priors as input to MLLMs, suffering from prior quality and inference efficiency issues. (b) SpatialSV internalizes 3D spatial awareness in MLLMs via task-oriented visual supervision, which enables spatial understanding with interpretable intrinsic spatial representations.

To address this limitation, one major line of research focuses on providing MLLMs with external spatial priors to compensate for their insufficient intrinsic representations, as illustrated in Figure 1 (a). Among these, prompt-based methods rely on external models or tools to construct prompts enriched with spatial priors, [Yang *et al.*, 2025b; Li *et al.*, 2025d; Qi *et al.*, 2025; Zhu *et al.*, 2025b; Liu *et al.*, 2025]. However, such methods introduce substantial inference overhead and are sensitive to errors from external models. A more direct alternative is to explicitly encode 3D data such as point clouds and depth maps, and project them to the language-aligned space of large language models [Chen *et al.*, 2024;

*Corresponding authors

Li *et al.*, 2025b; Wang *et al.*, 2025b; Zhu *et al.*, 2025a]. However, these methods typically depend on highly customized encoding and alignment modules and will introduce both training and inference complexity. Moreover, the strong 2D pretraining bias of MLLMs poses an additional obstacle to the explicit integration of 3D data.

Instead of emphasizing the construction and utilization of external spatial representations, we argue that MLLMs should internalize spatial awareness to form an internal spatial mental model [Johnson-Laird, 1980; Yin *et al.*, 2025]. Imaging in a practical embodied AI system where acquiring 3D data online is often costly and challenging with strict real-time constraints, MLLMs must rely on their own spatial awareness to construct internal representations of the environment, while maintaining a lightweight architecture to enable fast inference. To this end, recent works attempt to inject spatial awareness into MLLMs by distilling features from 3D vision foundation models (VFMs) [Huang *et al.*, 2025; Chen *et al.*, 2025b]. However, *the representation learned via distillation is inherently hard to interpret*, limiting a deeper understanding of the internal spatial mental modeling mechanism [Johnson-Laird, 1983] within MLLMs.

Inspired by the probing techniques [El Banani *et al.*, 2024; Chen *et al.*, 2025a], we conduct a pilot study to investigate the quality of spatial representations across different MLLMs under two paradigms: pure autoregressive text supervision and feature distillation. The probing results, as illustrated in Figure 2, reveal the following observations: (1) The quality of an MLLM's intrinsic spatial representation is positively correlated with its level of spatial intelligence. (2) Introducing 3D-aware supervision improves the quality of spatial representations. These findings validate the effectiveness of 3D-aware supervision in enhancing the spatial intelligence of MLLMs. However, the performance gap between the depth probing results and the ground-truth further exposes the limitations of the distillation paradigm. This is because *distilling features from 3D VFMs constitutes a coarse-grained alignment process that lacks explicit and structured spatial constraints*. In addition, aligning feature dimensions during distillation inevitably incurs information loss in the target representations, further weakening the robustness of the supervision signal.

These insights raise a central question: *can we identify a more intuitive and fine-grained form of 3D-aware supervision?* In response to this question, we propose SpatialSV, a framework that internalizes interpretable and robust 3D spatial awareness in MLLMs via task-oriented visual supervision, as illustrated in Figure 1 (b). Specifically, we perform 2D-to-3D lifting of MLLMs' visual features and align them with explicit spatial representations such as depth maps and point clouds, so as to unlock their intrinsic spatial intelligence. Inspired by 3D VFMs which leverage overlapping yet complementary downstream 3D tasks to learn geometrically consistent representations, we incorporate a set of complementary 3D tasks, including depth estimation, point-cloud reconstruction, and ray-map prediction. Technically, we extract multi-layer hidden visual features from the MLLM and employ projection layers together with decoupled DPT modules to perform multi-task prediction. By combining the coarse-grained guidance of feature distillation with

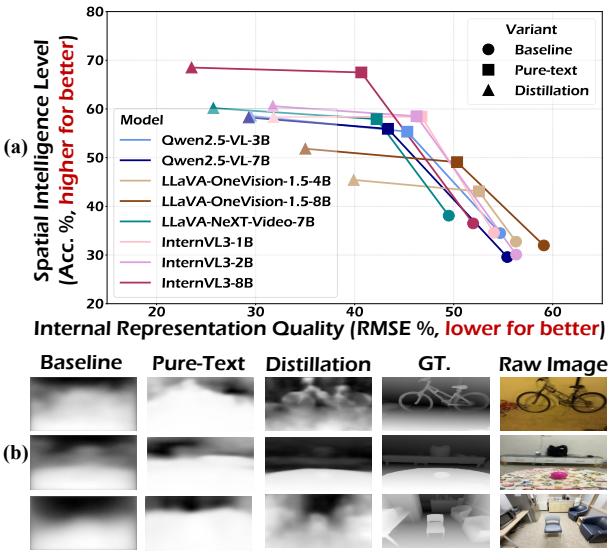


Figure 2: Depth probing results. (a) Quantitative results: the correlation between the quality of internal representations and the level of spatial intelligence. We compare 8 MLLMs under three variants: an untuned baseline, a text supervision variant, and a distillation variant. (b) Qualitative results: visualized depth maps of three variants for Qwen2.5-VL-3B, along with the ground-truths and raw images.

fine-grained, task-oriented constraints, SpatialSV facilitates robust intrinsic spatial representation learning of MLLMs. Extensive experiments across multiple MLLMs and benchmarks demonstrate the superior performance of SpatialSV in enhancing spatial intelligence.

Additionally, the 3D lifting results derived from MLLMs serve as an intuitive and interpretable manifestation of models' internal spatial representations. Through both quantitative and qualitative analyses, we investigate a distinct correlation between 3D lifting results and intra-model spatial intelligence as well as sample difficulties, highlighting the inherent interpretability of SpatialSV. Furthermore, we apply SpatialSV to a semi-supervised setting where $50\%$ of the samples have no textual annotations and achieve performance comparable to full text supervision. This is particularly important for scenarios where 3D text annotations are scarce. In summary, our contributions include:

- We propose SpatialSV, a novel framework that internalizes robust 3D spatial awareness in MLLMs through task-oriented visual supervision, with the 3D lifting results serving as an intuitive and interpretable proxy for intra-model spatial representations.
- We introduce a probe-based representation analysis framework for MLLMs, revealing the limitations of existing supervision paradigms and offering insights into improving supervision signals.
- Extensive experiments demonstrate SpatialSV's strong cross-model and cross-dataset generalization in enhancing the spatial intelligence of MLLMs, and highlight its inherent interpretabiliy, as well as the potential for exploiting unlabeled visual data.

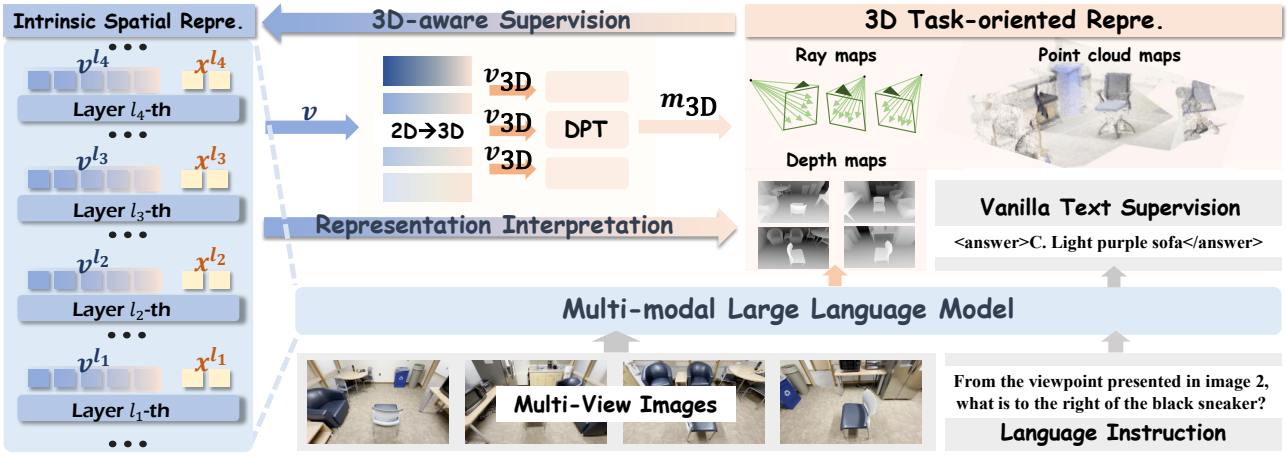


Figure 3: The schematic illustration of SpatialSV, a framework that internalizes interpretable and robust 3D spatial awareness in MLLMs via task-oriented visual supervision. We perform 2D-to-3D lifting of the MLLM's multi-layer hidden visual features and align them with explict 3D representations, including depth maps, ray maps, and point clouds. In this way, SpatialSV facilitates robust spatial representation learning and offers representation interpretability.

## 2 Method

### 2.1 Preliminaries

A typical MLLM consists of a visual encoder $\mathcal{E}_v$ and a decoder-only large language model (LLM) $\mathcal{F}$. In the context of multi-view spatial understanding, the multimodal input comprises $N$ multi-view images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ along with a language instruction. The visual encoder transforms the images into visual features $\boldsymbol{v}^0 = \mathcal{E}_v(\mathcal{I}) \in \mathbb{R}^{N_v \times d}$ aligned with the LLM's input space, where $N_v$ is the number of visual tokens, and $d$ is the feature dimension. The LLM then performs cross-modal feature interaction between the visual features and instruction tokens, producing $L$ layers of multimodal representations $\{\boldsymbol{v}^i, \boldsymbol{x}^i\}_{i=1}^{L}$, where $\boldsymbol{v}^i$ and $\boldsymbol{x}^i$ denote the $i$-th layer of visual and textual features, respectively.

During the fine-tuning stage, the LLM applies causal modeling to the final-layer multimodal features and autoregressively predicts the next text token. The model is optimized by minimizing the standard cross-entropy loss:

$$\mathcal{L}_{\text{text}} = -\sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_t^L \mid \boldsymbol{x}_{<t}^L, \boldsymbol{v}^L), \tag{1}$$

where $p_\theta$ is the probability distribution conditioned on the previous multimodal context, and $T$ is the length of the text token sequence. Notably, only the textual features are explicitly supervised under this training paradigm.

Following prior work [Huang *et al.*, 2025], we inject spatial awareness into the MLLM by explicitly supervising its visual features. Specifically, we introduce a 3D vision foundation model $\mathcal{F}_{3D}$ to extract target features that encodes geometric priors, and apply 2D average pooling to match the size of MLLM's visual features: $\boldsymbol{v}_{\text{3D}} = f_{\text{pool}}(\mathcal{F}_{\text{3D}}(\mathcal{I})) \in \mathbb{R}^{N_v \times d_{\text{3D}}}$, where $d_{\text{3D}}$ is the dimension of target features. In addition, we project MLLM's last-layer visual features into the target feature space to ensure dimension compatibility: $\boldsymbol{v}_{\text{proj}} = \phi(\boldsymbol{v}^L) \in \mathbb{R}^{N_v \times d_{\text{3D}}}$. Finally, the model is optimized by maximizing the cosine similarity between $\boldsymbol{v}_{\text{proj}}$ and $\boldsymbol{v}_{\text{3D}}$ in a distillation paradigm:

$$\mathcal{L}_{\text{distill}} = -S(\boldsymbol{v}_{\text{proj}}, \boldsymbol{v}_{\text{3D}}), \tag{2}$$

where $S(\cdot, \cdot)$ denotes cosine similarity. However, this distillation supervision suffers from inherent limitations in the interpretability of representations. In contrast, an intuitive analysis of MLLMs' internal representation quality and its correlation with spatial intelligence is crucial for understanding the spatial mental modeling mechanism within MLLMs and uncovering their capability boundaries.

### 2.2 Probe-based Spatial Representation Analysis

To acquire deeper insights into the intrinsic spatial intelligence of MLLMs, we investigate the correlation between representation quality and intelligence level under different supervision paradigms. Inspired by prior work [El Banani *et al.*, 2024; Chen *et al.*, 2025a] which investigate the 3D-awareness within visual models, we introduce a probe-based representation analysis framework for MLLMs.

Specifically, we evaluate multiple MLLMs, including Qwen2.5-VL [Bai *et al.*, 2025], LLaVA-NeXT-Video [Zhang *et al.*, 2024], InternVL3 [Zhu *et al.*, 2025c], and LLaVA-OneVision-1.5 [Li *et al.*, 2024a], on the multi-view spatial understanding benchmark MindCube-Tiny [Yin *et al.*, 2025]. For each MLLM, we consider three variants: (i) an untuned version, (ii) a version fine-tuned solely with autoregressive text supervision, and (iii) a version that additionally incorporates visual distillation supervision beyond text supervision.

To quantify spatial intelligence, we compute each model's question-answering accuracy on MindCube-Tiny. To access the quality of intrinsic representations, we attach a DPT [Ranftl *et al.*, 2021] module on top of MLLM's visual features for depth estimation. During this process, all MLLM parameters are frozen and only the DPT module is trainable. The depth map annotations for multi-view images in MindCube are obtained using DepthAnything-v3 [Lin *et al.*, 2025], and the

DPT head is trained on the training split. After training, we evaluate the depth estimation quality on MindCube-Tiny using the RMSE metric, which serves as a proxy for the quality of MLLM's internal spatial representations.

Figure 2 presents both quantitative and qualitative results, from which we draw two key observations: (1) for a given MLLM, introducing visual distillation supervision substantially improves both intrinsic representation quality and spatial intelligence; (2) as the quality of intrinsic representations improves, the model's spatial intelligence consistently increases. These findings strongly validate the effectiveness of 3D-aware visual supervision in enhancing the spatial intelligence of MLLMs. However, from the depth visualizations in Figure 2, we can observe a noticeable performance gap between the probing depth maps obtained under the distillation paradigm and the ground-truths. Specifically, the probing results appear overly blurred and suffer from significant loss of geometric details. This can be attributed to the inherent limitations of feature distillation supervision: (1) distillation at the feature level is inherently coarse-grained and lacks explicit, structured geometric constraints; (2) feature dimension alignment (e.g., via 2D average pooling) inevitably leads to information loss in the target representations. Collectively, these insights motivate a central question: **can we identify a fine-grained and intuitive form of 3D-aware visual supervision that can not only enhance but also interpret the intrinsic spatial intelligence of MLLMs?**

### 2.3 SpatialSV

To overcome the inherent limitations of feature distillation supervision, we propose SpatialSV, a framework that internalizes interpretable and robust 3D spatial awareness in MLLMs via task-oriented visual supervision, as depicted in Figure 3.

3D VFMs [Wang *et al.*, 2025a; Lin *et al.*, 2025] facilitate robust and geometrically consistent representation learning via overlapping yet complementary 3D downstream tasks, spanning point cloud reconstruction, camera pose estimation, depth estimation, and point tracking. Motivated by these work, we select three types of explicit spatial representations that are highly compatible with MLLM's visual feature maps, namely depth maps, ray maps and point cloud maps. These representations encode per-pixel spatial semantics, e.g., depth, camera pose and 3D coordinates, and are therefore more intuitive, structured and fine-grained than abstract feature maps. To obtain the ground-truth data, we leverage 3D VFMs to estimate multi-view depth maps and camera parameters, which are further transformed into ray maps and point clouds (see *Supp.A.1* for detailed computations). We align MLLM's visual features with these 3D task outputs to improve the quality of internal spatial representations.

Concretely, we employ a set of two-layer projectors $\Phi_{\text{3D}}$ to lift MLLM's multi-layer visual features into a shared 3D feature space. Among the hidden visual features of an MLLM, higher layers capture richer cross-modal semantics, whereas lower layers preserve more low-level visual information. Accordingly, we select visual features from four different layers to capture complementary spatial and semantic representations, with the set of selected layer indexes as $\{l_i\}_{i=1}^4$, and the projector set $\Phi_{\text{3D}} = \{\phi_i\}_{i=1}^4$. The projectors map the selected layer-wise features into multi-task shared 3D features:

$$\boldsymbol{v}_{\text{3D}}^i = \phi_i(\boldsymbol{v}^{l_i}), i \in [1, 4], \tag{3}$$

where $\boldsymbol{v}_{\text{3D}}^i \in \mathbb{R}^{N_v \times d_{\text{3D}}^i}$ is the $i$-th selected layer of 3D-lifted features, and $d_{\text{3D}}^i$ is the corresponding dimension. Subsequently, we introduce task-decoupled DPT modules $\mathcal{F}_{\text{DPT}}$ [Ranftl *et al.*, 2021] which receive the lifted features as input and independently predict per-pixel depth, camera pose, and 3D coordinates:

$$\{\boldsymbol{m}_{\text{3D}}^t, \boldsymbol{c}^t\} = \mathcal{F}_{\text{DPT}}^t(\{\boldsymbol{v}_{\text{3D}}^i\}_{i=1}^4), \tag{4}$$

where $t \in \{\text{depth}, \text{ray}, \text{pointcloud}\}$ denotes the task type; $\boldsymbol{m}_{\text{3D}}^t$ is the 3D outputs of task $t$; and $\boldsymbol{c}^t \in \mathbb{R}^{N_v}$ denotes the confidence of $\boldsymbol{m}_{\text{3D}}^t$. Specifically, $\boldsymbol{m}_{\text{3D}}^{\text{dep.}} \in \mathbb{R}^{N_v \times 1}$ corresponds to per-pixel depth values; $\boldsymbol{m}_{\text{3D}}^{\text{ray}} \in \mathbb{R}^{N_v \times 6}$ denotes per-pixel ray representations, where the first three channels encode the ray origin and the last three channels encode the ray direction; and $\boldsymbol{m}_{\text{3D}}^{\text{poi.}} \in \mathbb{R}^{N_v \times 3}$ represents per-pixel 3D spatial coordinates. Following [Wang *et al.*, 2025a; Lin *et al.*, 2025], we then compute task-specific losses between the MLLM's 3D-lifted predictions and the corresponding ground-truth annotations:

$$\begin{aligned}\mathcal{L}_{\text{3D}}^t = \mathcal{L}_2(\boldsymbol{y}_{\text{3D}}^t, \boldsymbol{m}_{\text{3D}}^t) + \mathcal{L}_{\text{conf}}(\boldsymbol{y}_{\text{3D}}^t, \boldsymbol{m}_{\text{3D}}^t; \boldsymbol{c}^t)) + \\ \mathbb{1}_{\{t=\text{dep.}\}} \mathcal{L}_{\text{grad}} + \mathbb{1}_{\{t=\text{poi.}\}} \mathcal{L}_{\text{norm}},\end{aligned} \tag{5}$$

where $\boldsymbol{y}_{\text{3D}}^t$ is the ground-truth of task $t$; $\mathbb{I}_{\{\cdot\}}$ denotes the indicator function; $\mathcal{L}_{\text{conf}}$ is the confidence loss; $\mathcal{L}_{\text{grad}}$ is the gradient loss applied to the depth residual, penalizing blurred, misplaced, or spurious depth transitions; and $\mathcal{L}_{\text{norm}}$ is the surface normal loss that enforce the alignment between the local surface geometry of predicted points with the ground-truth. See *Supp.A.2* for the detailed loss computation procedure.

Consequently, the overall training objective integrates an autoregressive text loss, a coarse-grained feature distillation loss, and fine-grained 3D-aware task-oriented losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{3D}}^{\text{dep.}} + \mathcal{L}_{\text{3D}}^{\text{ray}} + \mathcal{L}_{\text{3D}}^{\text{poi.}}. \tag{6}$$

This approach combines coarse-grained guidance with fine-grained geometric constraints, which promotes robust intrinsic spatial representation learning within MLLMs. Moreover, during inference, the 3D-lifted predictions derived from the MLLM serve as an intuitive proxy to assess representation quality and delineate the model's capability boundaries, as discussed in Sec. 3.3.

## 3 Experiments

### 3.1 Experimental Settings

**Datasets**. For training, we use 10k training samples from MindCube [Yin *et al.*, 2025] to fine-tune the model. For evaluation, we focus on two benchmarks, MindCube-Tiny [Yin *et al.*, 2025] and VSI-Bench [Yang *et al.*, 2025a], both of which target spatial understanding from limited ego-centric views. We evaluate models on MindCube-Tiny and the multiple-choice split of VSI-Bench.

**Evaluation Metrics**. We use Accuracy to evaluate the overall performance on multiple-choice questions, computed by exact matching between model predictions and ground-truths.

| Method | MindCube-Tiny | | | | VSI-Bench | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rotation | Among | Around | Overall↑ | Rel.Dir. Hard | Rel.Dir. Medium | Rel.Dir. Easy | Rel. Dist. | App. Order | Route Plan | Overall↑ |
| *Baseline* | | | | | | | | | | | |
| Chance Level (*Random*) | 26.5 | 24.5 | 23.8 | 24.6 | 26.5 | 25.7 | 24.9 | 24.1 | 23.3 | 28.4 | 27.7 |
| Chance Level (*Frequency*) | 34.5 | 34.8 | 33.3 | 34.3 | 25.2 | 33.6 | 50.2 | 25.1 | 25.2 | 29.4 | 29.0 |
| *Qwen2.5-VL Family* | | | | | | | | | | | |
| Qwen2.5-VL-3B | 34.5 | 36.0 | 32.3 | 34.5 | 34.3 | 33.3 | 29.0 | 31.7 | 22.7 | 27.8 | 29.6 |
| Qwen2.5-VL-3B + *Text.* | 33.5 | 49.7 | 74.8 | 55.3 | 35.7 | 35.7 | 26.3 | 31.1 | **23.6** | 29.9 | 30.1 |
| Qwen2.5-VL-3B + *Distillation.* | 34.5 | 51.5 | 81.3 | 58.6 | **35.9** | 34.1 | 30.9 | 32.1 | 22.8 | 32.0 | 30.6 |
| Qwen2.5-VL-3B + *SpatialSV* | **36.0** | **56.2** | **84.5** | **62.3** | 34.6 | **37.8** | **34.6** | **33.5** | 23.3 | **37.1** | **32.2** |
| Qwen2.5-VL-7B | 35.0 | 29.7 | 26.8 | 29.6 | 28.4 | 24.9 | 45.2 | 31.8 | 29.5 | 31.4 | 30.8 |
| Qwen2.5-VL-7B + *Text.* | 35.5 | 51.0 | 73.5 | 55.9 | 27.4 | 31.2 | **51.2** | 31.4 | **30.1** | 34.5 | 32.4 |
| Qwen2.5-VL-7B + *Distillation.* | 35.0 | 53.8 | 76.5 | 58.3 | 29.0 | 32.5 | 50.2 | 34.8 | 29.1 | 37.6 | 33.7 |
| Qwen2.5-VL-7B + *SpatialSV* | **37.5** | **58.5** | **81.8** | **62.8** | 29.5 | **35.2** | 48.4 | **37.2** | **30.1** | **42.8** | **35.4** |
| *LLaVA-OneVision-1.5 Family* | | | | | | | | | | | |
| LLaVA-OneVision-1.5-8B | 33.5 | 31.0 | 32.5 | 32.0 | 28.7 | 34.9 | 48.9 | 36.1 | 28.8 | 28.9 | 33.5 |
| LLaVA-OneVision-1.5-8B + *Text.* | 35.0 | 49.5 | 55.5 | 49.1 | 34.1 | 35.5 | **52.1** | 37.3 | **29.9** | 30.4 | 35.5 |
| LLaVA-OneVision-1.5-8B + *Distillation.* | 36.5 | 52.7 | 58.3 | 51.8 | 36.5 | 35.5 | 50.2 | 38.3 | 28.3 | 32.0 | 35.7 |
| LLaVA-OneVision-1.5-8B + *SpatialSV* | **38.0** | **57.5** | **60.3** | **55.2** | **38.3** | **36.5** | 49.3 | **42.1** | 29.6 | **40.2** | **38.1** |
| *LLaVA-NeXT-Video Family* | | | | | | | | | | | |
| LLaVA-NeXT-Video-7B | 34.5 | 41.7 | 34.5 | 38.1 | 27.1 | 32.5 | 50.7 | 35.9 | 27.5 | 31.4 | 32.9 |
| LLaVA-NeXT-Video-7B + *Text.* | 33.0 | 52.3 | 78.8 | 57.9 | 27.9 | **35.7** | 45.6 | 35.5 | 29.8 | 32.5 | 33.6 |
| LLaVA-NeXT-Video-7B + *Distillation.* | 36.5 | 55.2 | 79.5 | 60.2 | **29.5** | 33.6 | 48.4 | 37.5 | 29.5 | 34.5 | 34.4 |
| LLaVA-NeXT-Video-7B + *SpatialSV* | **39.5** | **61.8** | **82.3** | **64.9** | 29.2 | 33.9 | **52.5** | **37.8** | **30.4** | **44.9** | **35.9** |
| *InternVL3 Family* | | | | | | | | | | | |
| InternVL3-2B | 31.0 | 33.3 | 24.8 | 30.1 | 26.5 | 31.5 | **48.4** | 30.1 | 24.4 | 34.0 | 30.3 |
| InternVL3-2B + *Text.* | 32.0 | 54.3 | 78.0 | 58.5 | 29.0 | 34.1 | 45.6 | 32.0 | **30.1** | 29.9 | 32.4 |
| InternVL3-2B + *Distillation.* | 35.0 | 56.5 | 79.5 | 60.6 | 29.5 | 33.3 | 47.0 | 33.0 | 28.0 | 31.4 | 32.4 |
| InternVL3-2B + *SpatialSV* | **37.5** | **57.2** | **82.8** | **62.4** | **31.4** | **35.2** | 47.9 | **35.1** | 28.8 | **41.2** | **34.6** |
| InternVL3-8B | 34.5 | 39.3 | 33.3 | 36.5 | 23.6 | 34.1 | 46.1 | 34.1 | 32.9 | 32.5 | 33.1 |
| InternVL3-8B + *Text.* | 33.5 | 68.8 | 82.5 | 67.5 | 25.7 | 33.6 | 50.2 | 34.5 | 32.0 | 29.9 | 33.5 |
| InternVL3-8B + *Distillation.* | 35.5 | 70.8 | 81.5 | 68.5 | 28.4 | 36.0 | 47.5 | 35.5 | 31.7 | 34.0 | 34.5 |
| InternVL3-8B + *SpatialSV* | **38.5** | **77.2** | **86.0** | **73.7** | **30.6** | **37.3** | **50.7** | **36.2** | **33.2** | **42.8** | **36.6** |

Table 1: Comparison of our approach (*SpatialSV*) with the baseline models, the *pure-text supervision* variants, and the *feature distillation* varients on MindCube-Tiny and VSI-Bench. The best results within each model type are **bolded**. More detailed results across **8 MLLMs** are provided in *Supp.C.1*.

For the 3D-lifted depth estimation task, we follow [El Banani *et al.*, 2024] to report RMSE as the evaluation metric.

**Implementation Details**. We adopt multiple MLLMs as the baselines for SpatialSV, including Qwen2.5-VL [Bai *et al.*, 2025], LLaVA-OneVision-1.5 [Li *et al.*, 2024a], LLaVA-NeXT-Video [Zhang *et al.*, 2024], and InternVL3 [Zhu *et al.*, 2025c]. For all models, we fine-tune the vision–language projector, the large language model, all 2D-to-3D projectors, and the task-decoupled DPT modules, while freezing the visual encoder. All models are optimized using AdamW with a batch size of 16 and a warm-up ratio of 0.03. We use DepthAnything-v3 [Lin *et al.*, 2025] by default to obtain the ground-truth supervision signals. More detailed hyperparameter settings of all models are provided in *Supp.B.2*. All experiments are conducted on 8 NVIDIA H800 GPUs.

## 3.2 Comparison with Baselines

**Comparison with Other Supervision Paradigms**. We evaluate our approach against the baselines, the pure-text supervision variants, and the feature distillation variants on 6 MLLMs spanning 4 model families (more detailed results present in *Supp.C.1*). As shown in Table 1, SpatialSV consistently achieves the best performance, demonstrating its strong **cross-model generalization** capability. Specifically, on MindCube-Tiny, SpatialSV yield overall gains ranging from 3.42% to 12.66% over the pure-text variants, and from 2.97% to 7.81% over the distillation variants. On VSI-Bench, we observe an overall improvement by 4.36% to 6.79% compared with the distillation variants. Notably, SpatialSV achieves an average improvement of 10.0% over the pure-text variants on the *relative distance* task, while on the more challenging *route planning* task, the improvement reaches 32.6%. This indicates that SpatialSV effectively enhances the intrinsic spatial representations for 3D scenes within MLLMs, which is critical for estimating spatial attributes such as distance and direction. Moreover, it is worth noting that no VSI-Bench-related sample is incorporated in training data; nevertheless, SpatialSV still achieves significant improvements, highlighting its remarkable **cross-dataset generalization** ability.

**Comparison on Multiple Benchmarks**. We further evaluate our approach on several additional datasets, including 6 spatial understanding benchmarks, namely Ego3D-Bench [Gholami *et al.*, 2025], Spatial457 [Wang *et al.*, 2025c], ViewSpatial-Bench [Li *et al.*, 2025a], 3DSR-Bench [Ma *et al.*, 2025], SP-Bench [Li *et al.*, 2025c], TopViewRS [Li *et al.*, 2024b], and 2 general benchmarks, CVBench [Zhu *et al.*, 2025d] and MMBench [Liu *et al.*, 2024]. Detailed information about these datasets is provided in *Supp.B.1*. As shown in Table 2, our method consistently outperforms baseline mod-

| Method | Ego3D-Bench | Spatial457 | ViewSpatial-Bench | 3DSR-Bench | SPBench | TopViewRS | CVBench | MMBench | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | 32.4 | 33.9 | 35.8 | 49.6 | 39.3 | 43.3 | 70.7 | 76.3 | 47.7 |
| Qwen2.5-VL-3B+*Text.* | 31.8 | 33.4 | 36.3 | 49.3 | 40.2 | 43.4 | 70.2 | 76.2 | 47.6 |
| Qwen2.5-VL-3B+*SpatialSV* | **34.7** | **34.8** | **39.5** | **52.3** | **44.5** | **43.8** | **71.4** | **76.9** | **49.8** |
| Qwen2.5-VL-7B | 35.8 | 43.7 | 37.2 | 53.9 | 45.3 | 43.4 | 78.0 | 77.4 | 51.8 |
| Qwen2.5-VL-7B+*Text.* | 36.2 | 43.4 | 37.6 | 53.7 | 46.4 | 43.4 | 77.6 | 77.7 | 52.0 |
| Qwen2.5-VL-7B+*SpatialSV* | **39.3** | **44.3** | **40.1** | **55.2** | **47.8** | **43.6** | **78.2** | **79.3** | **53.5** |

Table 2: Comparison of our approach (*SpatialSV*) with the baseline models and the *pure-text supervision* variants on multiple spatial understanding and general benchmarks. The best results are **bolded**.

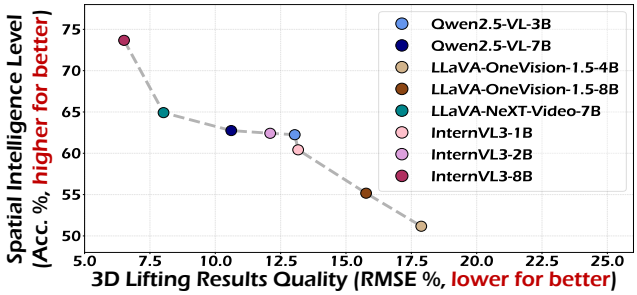


Figure 4: Correlation between the quality of the SpatialSV-based 3D lifting results and intra-model spatial intelligence level.

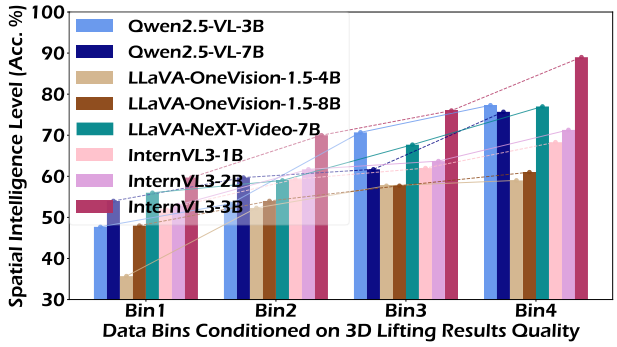


Figure 5: Correlation between the SpatialSV-based 3D lifting results with model-specific sample perferences. Samples are partitioned into 4 bins based on the quality of 3D lifting results.

els and pure-text variants across all benchmarks, indicating that SpatialSV **enhances spatial intelligence while also preserving the general understanding capability**. This further underscores SpatialSV's strong cross-dataset generalization.

## 3.3 Interpretability Analysis

In this section, we investigate the inherent interpretability of SpatialSV-based 3D lifting results. We conduct both quantitative and qualitative analyses to reveal their correlations with intra-model spatial intelligence and model-specific sample preferences.

**Correlation with Intra-Model Spatial Intelligence**. We use the depth estimation performance to quantify the quality of 3D lifting results, and VQA performance to measure MLLMs' intrinsic spatial intelligence, evaluating 8 MLLMs

| feat. | dep. | poi. | ray. | Qwen2.5VL-3B | | LLaVA-NV-7B | |
|---|---|---|---|---|---|---|---|
| | | | | MC. | VSI. | MC. | VSI. |
| – | – | – | – | 55.3 | 30.1 | 57.9 | 33.6 |
| ✓ | – | – | – | 57.4 | 30.2 | 60.3 | 34.0 |
| – | ✓ | – | – | 59.4 | 31.3 | 62.2 | 34.8 |
| – | – | ✓ | – | 58.2 | 31.5 | 62.4 | 33.9 |
| – | – | – | ✓ | 56.3 | 30.0 | 60.2 | 34.2 |
| – | ✓ | ✓ | ✓ | 61.4 | **32.3** | 63.7 | 35.6 |
| ✓ | ✓ | ✓ | ✓ | **62.3** | 32.2 | **64.9** | **35.9** |

Table 3: Ablation of different supervision signals on MindCube-Tiny (MC.) and VSI-Bench (VSI.). *Feat.*, *dep.*, *poi.*, and *ray.* denote features, depth maps, ray maps and point cloud maps, respectively.

on MindCube-Tiny. In Figure 4, we observe a strong correlation between the quality of 3D lifting results and intra-model spatial intelligence, which is aligned with the findings in Sec.2.2. Qualitatively, Figure 6 shows that MLLMs exhibit varying representation capability under the same spatial context. In the first case, the depth maps and point clouds derived from LLaVA-OneVision-8B lack necessary details of the key object *small table*, leading to its deviation from the correct option involving this object. In the second case, the failure of Qwen2.5-VL-7B can be partly attributed to its ignorance of *chairs*, which could help establish cross-view correspondences. In contrast, models that answer correctly can represent these critical objects in their 3D lifting results. These strongly validate the effectiveness of SpatialSV-based 3D lifting results as a proxy of intra-model spatial representations.

**Correlation with Model-Specific Sample Preferences**. For each MLLM, we partition the samples in MindCube-Tiny into four bins according to the quality of estimated depth maps. From Bin1 to Bin4, the samples correspond to progressively high-quality 3D lifting results. In Figure 5, QA accuracy exhibit a clear upward trend across all models, suggesting that **samples with unfaithful 3D lifting results tend to be more challenging for the models**. As shown in Figure 6, Qwen2.5-VL-7B succeeds in the first sample—accurately representing the spatial scene and providing the correct answer—but fails in the second sample in both aspects. This further highlights the correlation between SpatialSV-based 3D lifting results and model-specific sample preferences.

## 3.4 Ablation Study

**Different 3D-Aware Visual Supervision Signals**. As reported in Table 3, we examine different 3D-aware visual su-

| Layer Indexes | Projector Numbers | MindCube-Tiny | | | |
|---|---|---|---|---|---|
| | | Rotation | Among | Around | Overall |
| (36, 36, 36, 36) | 4 | 32.5 | 54.7 | 81.3 | 59.8 |
| (25, 26, 27, 28) | 4 | **36.5** | 55.7 | 84.3 | 62.0 |
| (4, 12, 20, 28) | 1 | 35.0 | 55.2 | 81.8 | 60.7 |
| (4, 12, 20, 28) | 4 | 36.0 | **56.2** | **84.5** | **62.3** |

Table 4: Ablation of different visual layers and projector numbers on MindCube-Tiny for Qwen2.5-VL-3B. The best is **bolded**.

| 50% data | 50% data | MindCube-Tiny | | | |
|---|---|---|---|---|---|
| | | Rotation | Among | Around | Overall |
| $\mathcal{L}_{\text{text}}$ | – | 27.5 | 42.0 | 64.8 | 47.2 |
| $\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{spatial}}$ | – | 29.5 | 46.7 | 70.5 | 51.8 |
| $\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{spatial}}$ | $\mathcal{L}_{\text{spatial}}$ | 29.0 | 49.5 | 73.0 | 53.9 |
| $\mathcal{L}_{\text{text}}$ | $\mathcal{L}_{\text{text}}$ | 33.5 | 49.7 | 74.8 | 55.3 |

Table 5: Semi-supervised learning with SpatialSV.

pervision signals, including 3D VFM features, depth maps, point maps and ray maps. Each supervision signal yields performance gains to varying degrees, with depth maps contributing the most, likely due to their intuitive spatial semantics and relatively low alignment difficulty. When all task-oriented supervision signals are jointly applied, the model significantly outperforms any single-task supervision, highlighting **the effectiveness of overlapping yet complementary multi-task supervision for learning robust spatial representations**. Furthermore, incorporating VFM features brings additional gains, suggesting that coarse-grained feature alignment and fine-grained task-oriented spatial constraints can mutually guide and reinforce each other.

**Different Supervised Visual Feature Layers and Numbers of 2D-to-3D Projectors**. As shown in Table 4, we explore different configurations of visual feature layers and projector numbers on Qwen2.5-VL-3B, which contains 37 visual feature layers (e.g., $L = 37$). Regarding the selection of visual layers, choosing $\{l_i\}_{i=1}^{4} = \{4, 12, 20, 28\}$ yields the best performance, achieving a $4.18\%$ improvement over supervising only the final visual layer. This is because overly deep layers tend to lack low-level visual details crucial for 3D reconstruction, thereby hindering the learning of 3D spatial representations. In contrast, combining shallow visual details with deep cross-modal semantics which are complementary proves feasible. For projector numbers, employing layer-wise decoupled projectors leads to a $2.64\%$ performance gain compared to using a single shared projector. This can be attributed to the semantic heterogeneity across different visual layers, which necessitates distinct 2D-to-3D mapping functions.

**Semi-Supervised Learning with SpatialSV**. Given the scarcity of high-quality 3D vision-language data, we explore directly exploiting unlabeled visual data. Since task-oriented visual supervision is independent of text autoregressive supervision, SpatialSV is naturally suited to semi-supervised learning. Specifically, we split the training data into two non-overlapping subsets and consider 4 configurations: (i) applying pure-text supervision ($\mathcal{L}_{\text{text}}$) to $50\%$ data; (ii) applying $\mathcal{L}_{\text{text}}$ and SpatialSV ($\mathcal{L}_{\text{spatial}}$) to $50\%$ data; (iii) applying both

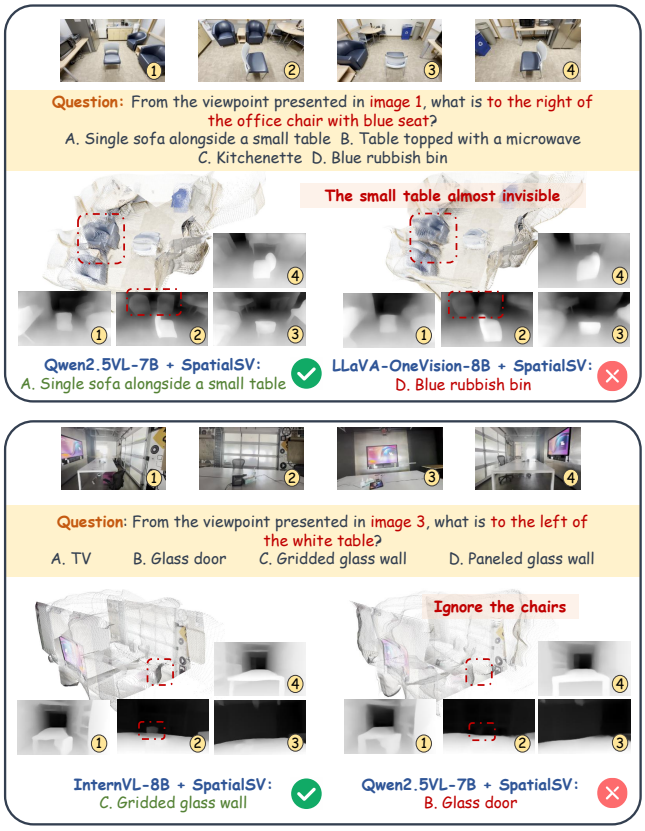


Figure 6: Qualitative results on MindCube-Tiny.

losses to $50\%$ data and only $\mathcal{L}_{\text{spatial}}$ to the remaining unlabeled $50\%$; (iv) applying $\mathcal{L}_{\text{text}}$ to $100\%$ data. From Table 5, We observe that the semi-supervised setting in (iii) achieves performance close to full text supervision (53.9 vs. 55.3), while yielding a remarkable $14.19\%$ improvement over setting (i). These results **highly validate the potential of SpatialSV for effectively exploiting unlabeled raw visual data**. Notably, SpatialSV can directly leverage off-the-shelf 3D VFMs to generate 3D task annotations as semi-supervised signals, which is much less costly and easier to obtain than constructing spatial question-answering datasets.

## 4 Conclusion

We propose SpatialSV, a novel framework that internalizes robust and interpretable 3D spatial awareness into MLLMs through task-oriented visual supervision. SpatialSV performs 2D-to-3D lifting of MLLMs' multi-layer visual features to align with explicit 3D representations, with the 3D-lifting results serving as intuitive and interpretable proxy of models' internal representations. Extensive experiments across multiple models and benchmarks demonstrate SpatialSV's effectiveness in enhancing and interpreting spatial intelligence. Semi-supervised learning results further validate SpatialSV's potential to leverage unlabeled visual data for scalable spatial representation learning.

## Acknowledgments

This work is supported in part of the National Natural Science Foundation of China (Grant 62373387) and Shenzhen Fundamental Research Program (Grant No. JCYJ20240813151301003).

## References

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

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

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

[Chen *et al.*, 2025b] Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025.

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

[Gholami *et al.*, 2025] Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025.

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

[Johnson-Laird, 1980] Philip N Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980.

[Johnson-Laird, 1983] Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

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

[Li *et al.*, 2024b] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. *arXiv preprint arXiv:2406.02537*, 2024.

[Li *et al.*, 2025a] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

[Li *et al.*, 2025b] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025.

[Li *et al.*, 2025c] 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.

[Li *et al.*, 2025d] Pengteng Li, Pinhao Song, Wuyang Li, Weiyu Guo, Huizai Yao, Yijie Xu, Dugang Liu, and Hui Xiong. See&trek: Training-free spatial prompting for multimodal large language model. *arXiv preprint arXiv:2509.16087*, 2025.

[Lin *et al.*, 2025] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

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

[Liu *et al.*, 2025] Yifan Liu, Fangneng Zhan, Kaichen Zhou, Yilun Du, Paul Pu Liang, and Hanspeter Pfister. Abstract 3d perception for spatial intelligence in vision-language models. *arXiv preprint arXiv:2511.10946*, 2025.

[Ma *et al.*, 2025] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6924–6934, 2025.

[Qi *et al.*, 2025] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025.

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

[Song *et al.*, 2025] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780, 2025.

[Tang *et al.*, 2026] Jiayu Tang, Yuchen Zhou, Chen Xiong, and Chao Gou. Letp: Coupling attention localization and cognitive reasoning for ego-centric multi-task driving scene perception. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19797–19801. IEEE, 2026.

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

[Wang *et al.*, 2025b] Xiaoyan Wang, Zeju Li, Yifan Xu, Jiaxing Qi, Zhifei Yang, Ruifei Ma, Xiangde Liu, and Chao Zhang. Spatial 3d-llm: Exploring spatial awareness in 3d vision-language models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[Wang *et al.*, 2025c] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24669–24679, 2025.

[Yang *et al.*, 2025a] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[Yang *et al.*, 2025b] Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025.

[Yin *et al.*, 2025] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

[Zhang *et al.*, 2024] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[Zhou *et al.*, 2025] Yuchen Zhou, Jiayu Tang, Xiaoyan Xiao, Yueyao Lin, Linkai Liu, Zipeng Guo, Hao Fei, Xiaobo Xia, and Chao Gou. Where, what, why: Towards explainable driver attention prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2675–2685, 2025.

[Zhu *et al.*, 2025a] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4295–4305, 2025.

[Zhu *et al.*, 2025b] Fangrui Zhu, Hanhui Wang, Yiming Xie, Jing Gu, Tianye Ding, Jianwei Yang, and Huaizu Jiang. Struct2d: A perception-guided framework for spatial reasoning in large multimodal models. *arXiv preprint arXiv:2506.04220*, 2025.

[Zhu *et al.*, 2025c] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[Zhu *et al.*, 2025d] Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. Cvbench: Benchmarking cross-video synergies for complex multimodal reasoning. *arXiv preprint arXiv:2508.19542*, 2025.