# Why Far Looks Up: Probing Spatial Representation in Vision-Language Models

Cheolhong Min[1], Jaeyun Jung[1], Daeun Lee[1], Hyeonseong Jeon[1], Yu Su[2], Jonathan Tremblay[3], Chan Hee Song[3,†,‡], and Jaesik Park[1,†]

[1] Seoul National University, Seoul, Republic of Korea
[2] The Ohio State University, Columbus, OH, USA
[3] NVIDIA, Santa Clara, CA, USA
lusong@nvidia.com, {cheolhong.min, jaesik.park}@snu.ac.kr

**Abstract.** Vision-language models (VLMs) achieve strong performance on spatial reasoning benchmarks, yet it remains unclear whether this reflects structured 3D understanding or reliance on statistical shortcuts in natural images. We introduce a representation-level analysis framework that constructs minimal contrastive pairs to measure how spatial axes are organized and disentangled within VLM embeddings. Our analysis across multiple model families reveals a consistent *vertical-distance entanglement*: models conflate vertical image position with distance, mirroring the perspective bias of natural photographs. This bias produces a significant accuracy gap between perspective-consistent and counter-heuristic examples, and intensifies under data scaling even as overall benchmark accuracy improves. We further show that models with similar benchmark scores can exhibit different internal representations, and that these differences predict accuracy and robustness across diverse spatial reasoning benchmarks. To isolate this bias from evaluation-set skew, we introduce `SpatialTunnel`, a synthetic benchmark designed to expose spatial shortcut biases by removing common correlations present in natural images. Experiments suggest that the entanglement is model-intrinsic, and that models with well-separated spatial axes exhibit greater robustness, indicating that well-structured spatial representations lead to more reliable spatial reasoning across diverse benchmarks. Code and benchmark are available on the project website.

**Keywords:** Vision-Language Models · Spatial Understanding · Representation Analysis

## 1 Introduction

Spatial reasoning is a core capability for Vision-Language Models (VLMs), particularly as these systems are increasingly deployed in robotics [19, 22, 26, 35], embodied agents [1, 30, 33], and multimodal assistants [2, 16, 27] that observe

† Co-corresponding author
‡ Project Lead

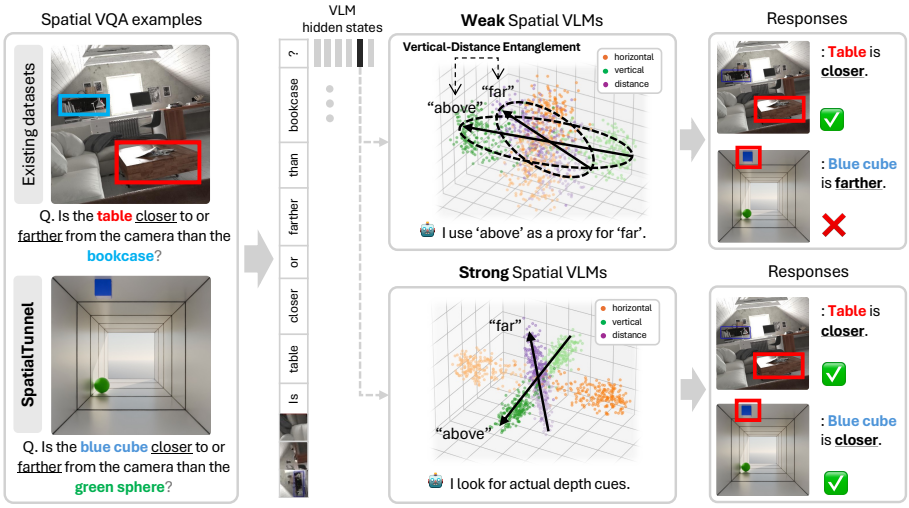



**Fig. 1:** Many VLMs answer spatial questions via a perspective-driven shortcut, *e.g.*, objects located higher in the image are further away in 3D. By confusing 2D vertical position with 3D distance, models fail systematically on counter examples. Our `SpatialTunnel` benchmark and contrastive probing expose this vertical-distance entanglement. In contrast, strong spatial VLMs show disentangled axes and consistent correctness across both real and synthetic settings.

and interact with physical environments. Although modern VLMs are primarily trained on 2D image–text pairs [3, 4, 12, 24], they achieve strong performance on spatial reasoning benchmarks [14, 15, 36], and recent work continues to improve these results through scaling and spatial training data [9, 10, 32, 34, 42]. These advances suggest that current models possess meaningful spatial understanding. However, it remains unclear whether strong benchmark accuracy reflects robust spatial reasoning or the exploitation of statistical regularities in natural images.

Many spatial relations can be partially inferred from correlations that arise naturally in photographic data rather than from explicit reasoning about 3D spatial structure. For example, perspective in everyday photographs introduces a consistent relationship between vertical image position and depth: objects appearing higher in the image are often farther from the camera, as in Figure 1. Such correlations allow models to rely on shortcuts that substitute vertical cues for depth reasoning, achieving high benchmark accuracy while internally conflating distinct spatial dimensions.

This limitation highlights a broader challenge in evaluating spatial understanding in VLMs. Behavioral benchmarks measure whether a model produces correct answers, but they provide limited insight into *how* those answers are obtained. Two models may achieve similar performance while relying on different internal mechanisms: one encoding spatial relations in a structured, separable manner, and another depending on correlated cues present in natural

imagery which become brittle under distribution shift. Distinguishing these possibilities requires examining how spatial information is represented inside the model, rather than relying on output-level performance.

Recent work has revealed persistent spatial reasoning failures through controlled benchmarks [20, 40, 41] and has begun probing internal model behavior such as attention dynamics [8]. However, these efforts primarily assess individual task performance or local mechanisms, leaving the global geometric organization of spatial relations in representation space largely unexplored.

We address this gap from two complementary angles. First, we analyze how spatial relations along three core 3D axes – horizontal (left / right), vertical (above / below), and depth (close / far) – are organized within VLM internal embeddings, using controlled contrastive examples that vary only the spatial relation between objects while holding confounds such as object identity fixed. Second, we introduce `SpatialTunnel`, a synthetic benchmark designed to remove perspective-driven biases in spatial evaluation. Its tunnel geometry decouples vertical image position from depth, enabling balanced assessment beyond the correlations present in natural image benchmarks.

Across multiple VLM families, our experiments reveal that horizontal relations form stable, opposing directions in representation space, whereas vertical and depth relations are frequently entangled, suggesting reliance on perspective-driven cues. Moreover, models with more structured spatial representations perform better across diverse spatial reasoning benchmarks, including EmbSpatial-Bench [14], CV-Bench [36], and BLINK [15]. Evaluations on `SpatialTunnel` further expose biases hidden under standard benchmark settings, and models with more structured representations exhibit greater robustness when these correlations are removed. Together, these results suggest that benchmark accuracy alone may overestimate the spatial reasoning capabilities of current VLMs. Our contributions are threefold:

- **Representation-level analysis of spatial reasoning.** We introduce a framework for analyzing how spatial relations are organized within VLM embeddings, diagnosing whether models encode structured spatial reasoning or rely on shortcut cues.
- **Spatial representations predict robustness.** We show that models with similar benchmark performance can exhibit markedly different internal spatial representations, and that models with more structured spatial representations exhibit greater robustness and generalization.
- **A bias-controlled synthetic benchmark for spatial reasoning.** We construct a synthetic dataset that decouples vertical image position from depth, revealing shortcut biases hidden under standard benchmark settings.

## 2 Related Work

***Spatial Understanding Datasets and Benchmarks.*** Recent benchmarks have revealed persistent weaknesses in VLM spatial reasoning despite strong semantic performance. Controlled evaluations such as What's Up [20] and COM-

FORT [41] show that models frequently fail on basic positional distinctions and frame-of-reference consistency. To probe deeper spatial competence, subsequent work has expanded along several axes: egocentric and cross-video reasoning [14, 36], 6DoF diagnostic tasks [37], and multi-step spatial referring [42]. In parallel, simulation-based datasets [28, 32, 35] provide large-scale supervision for physical dynamics, yet spatial performance often plateaus with data scaling [40]. While these efforts effectively measure *whether* models succeed or fail, they do not examine *what cues* models rely on internally—in particular, none isolate the entanglement between vertical image position and perceived depth that arises from perspective projection. Our work targets this gap by constructing controlled synthetic environments and contrastive splits that systematically expose this bias.

***Probing Internal Representations of Vision-Language Models.*** Recent work has moved beyond behavioral evaluation to examine the internal states of VLMs. Linear probing studies show that vision encoders inherently represent monocular depth cues [11] and bind geometric coordinates to object activations in early layers [21], while unified extraction frameworks [29] facilitate systematic comparison across model families. On the mechanistic side, ADAPTVIS [8] analyzes attention dynamics during spatial reasoning, and Spatial Forcing [23] explicitly aligns intermediate layers with 3D structure. However, these approaches primarily detect the *presence* of individual spatial primitives or adjust local attention behavior; they do not examine how different spatial dimensions are *jointly organized*—in particular, whether depth and vertical cues occupy separable or entangled directions in representation space. We address this gap through controlled contrastive analysis of internal embeddings, directly measuring the geometric relationship between spatial axes to reveal entanglement that isolated probing cannot detect.

## 3 Perspective Projection Bias in Spatial Understanding

Vision-language models are increasingly expected to reason about 3D spatial relationships from a single RGB image, *e.g.*, answering questions such as "Is the chair closer to the camera than the table?" However, monocular images provide only a 2D projection of the 3D scene, requiring models to infer spatial structure from indirect visual cues. A central question is whether current VLMs genuinely learn such 3D reasoning, or instead rely on the visual cues that happen to correlate with depth in image space.

In this section, we analyze how VLMs perform spatial reasoning across multiple model families and benchmarks. Our analysis reveals a systematic bias arising from perspective projection: models frequently use an object's vertical position in the image as a proxy for its distance from the camera. We term this phenomenon **vertical-distance entanglement**, where image-plane vertical position becomes conflated with depth. Across multiple models and benchmarks, we show that this bias consistently emerges and leads to systematic errors in spatial reasoning.

### 3.1 What is Vertical-Distance Entanglement?

***Perspective projection and vertical position.*** From the observer's viewpoint, objects farther away on a common ground surface tend to appear higher in the image. This phenomenon gives rise to the classical elevation cue: for objects lying on the ground plane, those nearer to the horizon line are perceived as being farther from the observer [11] (see Appendix A for details).

***Entanglement as a shortcut.*** We hypothesize that VLMs exploit this correlation as a shortcut: when asked about the relative depth, they partially rely on the vertical positions of the objects rather than robustly reasoning about 3D structure. We refer to this phenomenon as **vertical-distance entanglement**, indicating the tendency of a model to treat *above* $\approx$ *far* and *below* $\approx$ *close* when answering depth-related questions.

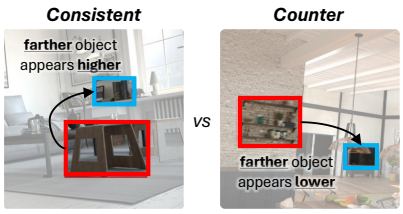



**Fig. 2:** ***Consistent*** **vs.** ***counter*** **examples.** ***Consistent***: Farther one appears higher in image; ***Counter***: Farther one appears lower.

***Consistent and counter-heuristic examples.*** To systematically analyze this entanglement, we categorize depth-related samples into two groups, *consistent* and *counter* (Figure 2). The classification is based on whether the ground-truth spatial relationship aligns with the vertical-position heuristic. We implement this by comparing the vertical center coordinates of the two queried objects in pixel space: if the farther object has a smaller $y$-coordinate (*i.e.*, higher in the image), the example is consistent; otherwise, it is counter. If a model exhibits no entanglement, its accuracy should be comparable on both groups. Conversely, a systematic accuracy gap between the two groups would constitute evidence that the model relies on the vertical-position shortcut.

### 3.2 Experimental Setup

***Models.*** We evaluate three VLM families spanning different architectures: Molmo-7B-O-0924 [12], NVILA-Lite-2B [25], and Qwen2.5-VL-3B-Instruct [4]. To analyze how spatial fine-tuning affects entanglement, we train variants of each model at multiple data scales (80k, 400k, 800k, and 2M samples); base models refer to the original pretrained weights without additional fine-tuning. We also include RoboRefer-2B-SFT [42], which shares the NVILA-Lite-2B base but is trained on more than 20M samples including RGB and RGB-D images, and Qwen3-VL-235B-A22B-Instruct [3] as a large-scale reference.

***Training data.*** Recent work has attributed VLMs' limited spatial understanding to a lack of spatial reasoning data during training, motivating several spatial-focused datasets [6, 13, 28, 32, 39, 42]. To study the effect of data scaling within

**Table 1: Distribution of consistent, counter, and ambiguous examples.** Existing spatial benchmarks are skewed toward consistent examples, mirroring the natural statistics of perspective projection in real-world images.

| Type | EmbSpatial-Bench | CV-Bench-3D | Definition |
|---|---|---|---|
| Consistent | 976 (80.9%) | 363 (60.5%) | GT aligns with heuristic |
| Counter | 129 (10.7%) | 65 (10.8%) | GT contradicts heuristic |
| Ambiguous | 101 (8.4%) | 172 (28.7%) | $\Delta y < 5\%$ of image height |

and across model families, we uniformly mix five existing spatial understanding datasets (*i.e.*, SAT [28], RoboSpatial [32], SPAR-7M [39], RefSpatial [42], PRISM [13]) and subsample at four target scales (80k, 400k, 800k, and 2M) for supervised fine-tuning (see Appendix B.2 and B.3 for details).

### 3.3 Evidence from Existing Benchmarks

We first examine whether vertical-distance entanglement is observable on established spatial reasoning benchmarks that use real-world images: EmbSpatial-Bench [14] and the 3D-spatial split of CV-Bench [36].

***Data distribution is skewed toward consistent examples.*** We classify all depth-related questions in both benchmarks into consistent, counter, and ambiguous categories following the criteria defined in Section 3.1. As shown in Table 1, consistent examples account for 80.9% of EmbSpatial-Bench and 60.5% of CV-Bench-3D, while counter examples constitute only about 10% in each. This heavy skew reflects the natural statistics of real-world photographs: in most everyday scenes, farther objects do appear higher in the image.

***Models systematically fail on counter examples.*** We evaluate a range of VLMs spanning different architectures and scales on the two benchmarks, reporting accuracy separately for consistent and counter subsets (Table 2). Across *all* models and *all* training scales, accuracy on consistent examples significantly exceeds that on counter examples. For instance, Qwen2.5-VL fine-tuned on 2M samples achieves 60.9% on the consistent split of EmbSpatial-Bench but only 24% on counter examples, yielding a 36.9 percentage-point gap. This pattern holds regardless of model family (Molmo, NVILA, or Qwen2.5-VL), model size, or the amount of spatial fine-tuning data, suggesting that vertical-distance entanglement is a widespread phenomenon rather than an artifact of any single architecture, training recipe, or data scale.

## 4 Behavioral Analysis with a Synthetic Dataset

The accuracy gap in Section 3 indicates that VLMs systematically fail on counter examples in real-world datasets. However, real photographs conflate multiple

**Table 2: Accuracy on consistent vs. counter examples across models and benchmarks.** All models exhibit a substantial accuracy gap, with consistent examples outperforming counter examples. Results are reported on depth-related questions from EmbSpatial-Bench and CV-Bench-3D. Indented rows denote fine-tuned variants at the given spatial data scale.

| | EmbSpatial-Bench | | CV-Bench-3D | |
|---|---|---|---|---|
| **Model** | Consistent | Counter | Consistent | Counter |
| Molmo-7B-O-0924 [12] | 63.5 | 34.9 (-28.6) | 93.1 | 75.4 (-17.7) |
| + 80k | 60.6 | 29.5 (-31.1) | 80.2 | 56.9 (-23.3) |
| + 400k | 62.7 | 27.1 (-35.6) | 89.5 | 56.9 (-32.6) |
| + 800k | 65.2 | 34.1 (-31.1) | 88.7 | 70.8 (-17.9) |
| + 2M | 65.3 | 39.5 (-25.8) | 90.6 | 72.3 (-18.3) |
| NVILA-Lite-2B [25] | 49.0 | 27.1 (-21.9) | 74.4 | 40.0 (-34.4) |
| + 80k | 57.7 | 15.5 (-42.2) | 71.6 | 50.8 (-20.8) |
| + 400k | 61.1 | 34.1 (-27.0) | 81.3 | 58.5 (-22.8) |
| + 800k | 63.2 | 38.8 (-24.4) | 84.6 | 67.7 (-16.9) |
| + 2M | 60.7 | 41.1 (-19.6) | 97.2 | 93.8 (-3.4) |
| RoboRefer-2B-SFT [42] | 87.0 | 59.7 (-27.3) | 98.9 | 95.4 (-3.5) |
| Qwen2.5-VL-3B-Instruct [4] | 54.7 | 32.6 (-22.1) | 75.5 | 55.4 (-20.1) |
| + 80k | 50.6 | 30.2 (-20.4) | 69.7 | 60.0 (-9.7) |
| + 400k | 52.6 | 27.1 (-25.5) | 65.8 | 58.5 (-7.3) |
| + 800k | 55.8 | 26.4 (-29.4) | 61.2 | 58.5 (-2.7) |
| + 2M | 60.9 | 24.0 (-36.9) | 62.0 | 53.8 (-8.2) |
| Qwen3-VL-235B-A22B-Instruct [3] | 73.3 | 41.7 (-31.6) | 98.1 | 90.8 (-7.3) |

depth cues (*e.g.*, vertical position, apparent size, and occlusion), making it difficult to isolate the contribution of any single cue. To enable controlled interventions, we introduce `SpatialTunnel`, a synthetic dataset that decouples an object's vertical image-plane position from its 3D depth by design, allowing the two factors to be manipulated independently.

### 4.1 `SpatialTunnel` Benchmark

To evaluate spatial relations in a controlled manner, we require an environment with two key properties: (i) objects can be positioned arbitrarily, enabling queries over any spatial relation (*e.g.*, left/right, above/below, near/far); and (ii) an object's vertical position can be adjusted independently of its depth, allowing us to construct image groups that differ only in vertical placement while preserving depth ordering.

To satisfy these requirements, we build a tunnel-shaped synthetic scene in Blender [5] (Figure 3). Each scene consists of a single-point-perspective corridor whose walls, ceiling, and floor are symmetric about the camera's optical axis, where objects are placed anywhere on the interior tunnel surfaces. Because ob-

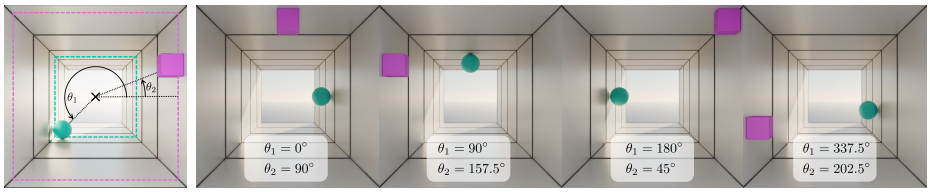



**Fig. 3: `SpatialTunnel` visualization.** `SpatialTunnel` holds the two objects at fixed depths while sweeping their angular positions around the tunnel cross-section, so that 2D image-plane layout varies independently of depth ordering.

jects near the top and bottom of the image can be equidistant from the camera, the common heuristic "higher in the image ⇒ farther" no longer holds. We parameterize each object by its depth $z$ and an angular position $\theta$ on the tunnel cross-section. Holding $z$ fixed while varying $\theta$ moves the object up/down and left/right in the image without changing its depth ordering, enabling matched counterfactual pairs that flip vertical arrangement while preserving depth.

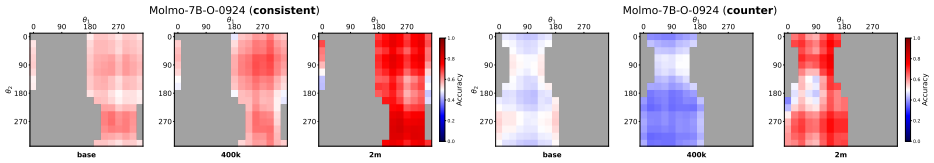
We construct a synthetic benchmark suite, `SpatialTunnel`, that enables controlled spatial interventions in a single-point-perspective corridor. Specifically, we place two objects at predetermined depths and sweep each object along the tunnel cross-section, discretizing the interior into 16 angular positions (see Figure 3). This yields a $16 \times 16$ Cartesian grid over $(\theta_1, \theta_2)$, enabling heatmap-style diagnostics of model behavior across configurations (see Figure 4). To increase visual diversity and improve robustness, we randomize object appearance (color, size, and shape) and scene lighting across renders. Additional synthetic variants for other spatial cues (*e.g.*, object size) and auxiliary analyses are provided in the Appendix C.4.

### 4.2 Experimental Setup on `SpatialTunnel`

Given a rendered RGB image containing two objects, the model is asked a binary depth-comparison question. In our setup, an object is always placed farther from the camera than the other, and the VLM is asked to answer the questions like "*Is {obj$_1$} closer to / farther from the camera than {obj$_2$}?*" Following prior work [18, 38, 41], we define a local probability by extracting the logits for `Yes` and `No` at the first generated token. We then compute the predicted probability as

$$p = \sigma\big(\ell_{\texttt{Yes}} - \ell_{\texttt{No}}\big).$$

The correctness score for a single query is defined as $v = p$ if the ground-truth answer is `Yes`, and $v = 1 - p$ if it is `No`. We report the following metrics for all VLMs described in Section 3.2. Following the definition in Section 3.1, samples are partitioned into *consistent* and *counter* subsets. We report four metrics: (1) Mean accuracy ($v$), the mean correctness score across all images and questions; (2) Consistent accuracy ($v_{\text{cons}}$), the mean correctness score on *consistent* examples; (3) Counter accuracy ($v_{\text{ctr}}$), the mean correctness score on *counter*




**(a)** Results on *consistent* samples **(b)** Results on *counter* samples.

**Fig. 4: Mean accuracy heatmaps on `SpatialTunnel` for Molmo-7B.** Each cell indexes a joint angular configuration $(\theta_1, \theta_2)$ of the two objects (red = higher accuracy; blue = lower). Gray indicates configurations outside the subset. From base → 400k → 2M training samples, accuracy on (a) perspective-consistent cells improves steadily. In contrast, (b) counter cells remain substantially harder, with the largest drop at 400k and a partial recovery at 2M.

**Table 3: Consistent vs. Counter accuracy on `SpatialTunnel`.** $v$: mean correctness score; $v_{\text{cons}}$ and $v_{\text{ctr}}$: scores on consistent and counter subsets; $\Delta = v_{\text{cons}} - v_{\text{ctr}}$.

| **Model** | $\mathbf{v}$ | $\mathbf{v}_{\text{cons}}$ | $\mathbf{v}_{\text{ctr}}$ | $\boldsymbol{\Delta}$ |
|---|---|---|---|---|
| Molmo-7B-O-0924 | 0.528 | 0.565 | 0.487 | +0.078 |
| + 80k | 0.496 | 0.507 | 0.486 | +0.021 |
| + 400k | 0.501 | 0.593 | 0.409 | +0.184 |
| + 800k | 0.531 | 0.628 | 0.430 | +0.198 |
| + 2M | 0.666 | 0.703 | 0.630 | +0.073 |
| NVILA-Lite-2B | 0.488 | 0.504 | 0.471 | +0.033 |
| + 80k | 0.499 | 0.562 | 0.438 | +0.124 |
| + 400k | 0.669 | 0.804 | 0.538 | +0.267 |
| + 800k | 0.646 | 0.728 | 0.571 | +0.157 |
| + 2M | 0.812 | 0.875 | 0.749 | +0.127 |
| RoboRefer-2B-SFT | 0.793 | 0.816 | 0.770 | +0.046 |
| Qwen2.5-VL-3B-Instruct | 0.570 | 0.776 | 0.360 | +0.416 |
| + 80k | 0.512 | 0.585 | 0.440 | +0.145 |
| + 400k | 0.503 | 0.588 | 0.418 | +0.171 |
| + 800k | 0.499 | 0.600 | 0.398 | +0.202 |
| + 2M | 0.500 | 0.648 | 0.353 | +0.295 |
| Qwen3-VL-235B-A22B-Instruct | 0.908 | 0.948 | 0.880 | +0.068 |

examples; and (4) Accuracy gap ($\Delta = v_{\text{cons}} - v_{\text{ctr}}$), the accuracy difference between the two subsets, quantifying the vertical-distance entanglement. A model with no directional bias would yield $\Delta \approx 0$.

### 4.3 Results on `SpatialTunnel`: Vertical-Distance Entanglement

Consistent with Section 3.3, we observe that vertical-distance entanglement is universal. Across all base and fine-tuned models, accuracy is consistently higher

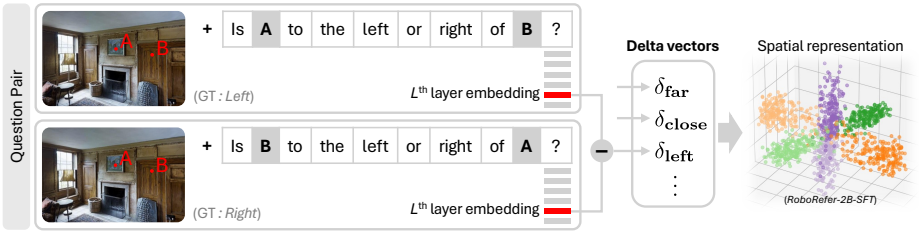



**Fig. 5: Contrastive probing for representation-level spatial analysis.** Given a spatial-relation VQA, we construct a minimal question pair by swapping the object order, which flips the ground-truth relation. We extract the final-token hidden state at an intermediate layer for each question and compute a *delta vector* as their difference, isolating the relational displacement in embedding space. Aggregated across samples, these vectors summarize the model's internal spatial representations and enable diagnosing systematic confounds among spatial cues.

on the consistent subset than on the counter subset, yielding a positive accuracy gap $\Delta$. Table 3 and Figure 4 summarize model behavior on `SpatialTunnel`.

For example, base Qwen2.5-VL-3B-Instruct achieves $v_{\text{cons}} = 0.776$ but only $v_{\text{ctr}} = 0.360$, indicating strong reliance on the vertical-position shortcut. While base NVILA-Lite-2B produces a narrower gap, its sub-0.5 overall accuracy suggests near-random performance rather than meaningful depth understanding. Figure 4 visualizes positional bias at the cell level for Molmo variants. If predictions were insensitive to 2D placement, accuracy would be approximately uniform across the grid. Instead, most models show pronounced contrast between consistent and counter regions. The results suggest that large-scale spatial training reduces this reliance. RoboRefer [42], trained on more than 20M QA pairs, achieves the smallest gap ($\Delta = +0.046$) among models performing above chance. Qwen3-VL-235B attains the highest mean accuracy ($v = 0.908$) with a similarly small gap ($\Delta = +0.068$), indicating that very large-scale pretraining can substantially alleviate this bias even without targeted spatial fine-tuning.

## 5 Representation Analysis via Contrastive Probing

Sections 3–4 established vertical-distance entanglement as a model-intrinsic phenomenon through behavioral evaluation. We now turn to internal representations to examine how spatial axes are encoded and what distinguishes models that exhibit robust spatial reasoning.

### 5.1 Beyond Benchmark Accuracy

Behavioral accuracy alone can be a misleading indicator of spatial understanding. Table 4 reports performance across five spatial reasoning tasks spanning different formats, dimensionalities, and difficulty levels. Beyond EmbSpatial-Bench and the 3D-spatial split of CV-Bench (CV-3D) used in Table 2, we additionally

**Table 4: Performance comparison across spatial understanding benchmarks.** Fine-tuned variants of Molmo, NVILA, and Qwen2.5-VL exhibit inconsistent performance fluctuations depending on the benchmark, while RoboRefer and Qwen3-VL-235B, which show strong representation structure under our probing framework (Section 5.4), achieve consistently high performance across all evaluated tasks. **Bold** denotes best. BLINK confidence intervals are reported in Appendix B.5.

| | **EmbSpatial** | **CV-2D** | **CV-3D** | | **BLINK** | |
|---|---|---|---|---|---|---|
| **Model** | Overall | Relation | Depth | Distance | Rel. Depth | Spat. Rel. |
| Molmo-7B-O-0924 [12] | 60.7 | 76.3 | 84.5 | 68.5 | 78.2 | 70.6 |
| + 80k | 52.9 | 62.3 | 71.0 | 67.5 | 72.6 | 60.8 |
| + 400k | 64.9 | 84.3 | 80.0 | 70.8 | 72.6 | 68.5 |
| + 800k | 69.1 | 90.0 | 82.0 | 70.8 | 75.0 | 61.5 |
| + 2M | 74.3 | 93.7 | 87.3 | 81.3 | 71.0 | 69.2 |
| NVILA-Lite-2B [25] | 54.0 | 58.6 | 69.2 | 52.3 | 64.5 | 67.1 |
| + 80k | 65.1 | 78.9 | 66.2 | 60.8 | 53.2 | 74.1 |
| + 400k | 62.1 | 83.2 | 74.3 | 67.0 | 71.8 | 63.6 |
| + 800k | 69.7 | 85.5 | 78.2 | 71.3 | 57.3 | 65.0 |
| + 2M | 69.4 | 91.4 | 93.8 | 87.2 | 70.2 | 62.9 |
| RoboRefer-2B-SFT [42] | **92.0** | **96.5** | **95.7** | 90.5 | **84.7** | 79.7 |
| Qwen2.5-VL-3B-Instruct [4] | 62.3 | 67.4 | 70.3 | 60.2 | 68.6 | 83.9 |
| + 80k | 57.3 | 59.7 | 64.7 | 61.5 | 58.1 | 79.7 |
| + 400k | 58.6 | 58.2 | 62.0 | 54.5 | 58.9 | 78.3 |
| + 800k | 60.9 | 59.4 | 58.7 | 51.2 | 58.1 | 79.0 |
| + 2M | 65.7 | 68.8 | 58.5 | 52.2 | 53.2 | 78.3 |
| Qwen3-VL-235B-A22B-Instruct [3] | 82.0 | **96.5** | 93.3 | **91.0** | **84.7** | **90.2** |

include the 2D-spatial split of CV-Bench (CV-2D) and the spatial relationship and relative depth splits of BLINK [15]. Detailed task descriptions are provided in the Appendix B.4.

Fine-tuned variants of Molmo, NVILA, and Qwen2.5-VL show inconsistent patterns across benchmarks. For example, NVILA (2M) achieves 93.8% on CV-3D Depth but only 62.9% on BLINK Spatial Relation, while Qwen2.5-VL (2M) scores 78.3% on BLINK Spatial Relation but drops to 52.2% on CV-3D Distance. No single accuracy figure reliably indicates how well these models have internalized 3D spatial concepts. In contrast, RoboRefer and Qwen3-VL-235B-A22B-Instruct achieve consistently high performance across all benchmarks. As we show in Section 5.4, these models also exhibit the most structured internal representations under our probing framework, including high axis coherence, well-separated PCA clusters, and low entanglement. This suggests that representation quality underlies robust spatial reasoning across benchmarks, motivating us to look beyond accuracy and analyze model-internal representations directly.

### 5.2 Contrastive Probing

Given an image, we construct a pair of questions that differ only in the ordering of the queried objects, such as swapping *"Is A to the left or right of B?"* with *"Is B to the left or right of A?"*. Consequently, the ground-truth answer for the swapped query becomes the spatial inverse of the original; for instance, a *left* relationship is inverted to *right*. For probing, we extract hidden states at a fixed intermediate layer $L^*$ per model family. Let $h_q \in \mathbb{R}^d$ denote the final-token hidden state at layer $L^*$ for question $q$. For a question pair $(q_1, q_2)$, we define **delta vector** $\delta = h_{q_2} - h_{q_1}$ as the representation-space displacement induced by the swap. By repeating this across many images, we obtain a set of delta vectors per spatial category (*e.g.*, *above*, *below*, *far*, *close*, *left*, *right*). This procedure aims to isolate the latent encoding of spatial directions by neutralizing common visual components. We define two metrics over these delta vectors:

***Axis coherence.*** For each spatial axis (horizontal, vertical, distance), we pool the delta vectors from both opposing categories (*e.g.*, *far* and *close* for the distance axis). To align directions, we negate the deltas from the opposing category so that all vectors point toward the canonical direction:

$$\tilde{\delta}^{(i)} = \begin{cases} \delta^{(i)} & \text{if category is canonical (\textit{e.g.}, \textit{far})} \\ -\delta^{(i)} & \text{if category is opposite (\textit{e.g.}, \textit{close})} \end{cases} \tag{1}$$

Axis coherence is the mean pairwise cosine similarity over the sign-corrected set:

$$\mathrm{Coh}_{\mathrm{axis}} = \frac{2}{N(N-1)} \sum_{i<j} \cos(\tilde{\delta}^{(i)},\ \tilde{\delta}^{(j)}). \tag{2}$$

High coherence indicates that the model encodes the axis as a stable, consistent direction in representation space.

***VD-Entanglement Index.*** To quantify the degree of vertical-distance entanglement at the representation level, we compute the mean delta vector $\mu_c$ for each category $c \in \{\textit{above}, \textit{below}, \textit{far}, \textit{close}\}$ and define the **VD-Entanglement Index** (VD-EI):

$$\begin{aligned} \text{VD-EI} = \tfrac{1}{4}\big[ &\cos(\mu_{\text{above}}, \mu_{\text{far}}) + \cos(\mu_{\text{below}}, \mu_{\text{close}}) \\ &- \cos(\mu_{\text{above}}, \mu_{\text{close}}) - \cos(\mu_{\text{below}}, \mu_{\text{far}})\big]. \end{aligned} \tag{3}$$

The first two terms measure the similarity between perspective-aligned pairs (above↔far, below↔close); the last two measure perspective-opposing pairs. A positive value indicates that vertical and distance representations are directionally coupled in the manner predicted by perspective projection; zero indicates independence. We extract hidden states from EmbSpatial-Bench [14] images at a fixed intermediate layer per model family, following previous approaches [7,17,31] (see Appendix D.3 and D.4 for details of layer selection).

### 5.3 Distance Coherence and Counter Accuracy

***Distance coherence is the weakest axis.*** Across all models and training scales in Table 5, $\text{Coh}_\text{D}$ is the lowest among the three axes. Fine-tuning substantially increases vertical coherence (*e.g.*, Molmo: 0.23 → 0.57; Qwen2.5-VL: 0.29 → 0.59), but $\text{Coh}_\text{D}$ grows by a comparatively smaller margin.

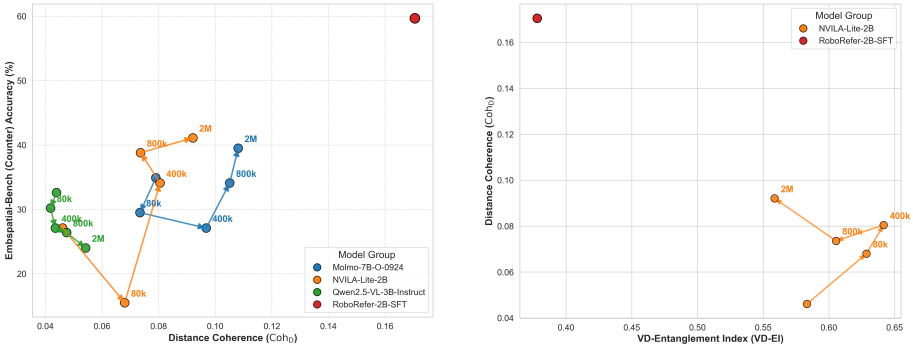
***Distance coherence growth accompanies counter accuracy improvement.*** Figure 6a plots $\text{Coh}_\text{D}$ against counter accuracy on EmbSpatial-Bench, the same dataset from which the coherence values are derived. At early scaling steps (*e.g.*, 80k), counter accuracy sometimes drops before $\text{Coh}_\text{D}$ has meaningfully increased. However, once spatial fine-tuning data reaches sufficient scale, a consistent pattern emerges across NVILA (from 80k onward) and Molmo (from 400k onward): as $\text{Coh}_\text{D}$ increases, counter accuracy rises in tandem, tracing an upward trajectory in Figure 6a. In contrast, Qwen2.5-VL's $\text{Coh}_\text{D}$ remains nearly flat throughout scaling (0.043 → 0.052), and its counter accuracy declines, widening the consistent-counter gap. This suggests that as models form a more coherent distance representation through spatial data scaling, they become more robust to the vertical-position shortcut. Conversely, when distance coherence stagnates, continued scaling does not resolve the entanglement.

***Cross-domain validity of distance coherence.*** To examine whether $\text{Coh}_\text{D}$ reflects a reusable representation rather than a benchmark-specific artifact, we measure $\text{Coh}_\text{D}$ on `SpatialTunnel` and compare it against *counter* accuracy on two other benchmarks. $\text{Coh}_\text{D}$ computed on `SpatialTunnel` correlates with *counter* accuracy on both EmbSpatial-Bench and CV-Bench-3D ($\rho = 0.759$ and 0.804, respectively; both $p < 10^{-3}$). This cross-domain alignment supports the view that $\text{Coh}_\text{D}$ captures predictive signal beyond in-domain computation artifacts (see Appendix D.5 for additional details).

### 5.4 What Characterizes Strong Spatial Representations

We compare the NVILA-Lite-2B scaling series and RoboRefer-2B-SFT [42], which share the same base architecture, to identify what representation profile accompanies robust spatial reasoning. Figure 7 visualizes the delta vectors via PCA. In the base model (*i.e.*, NVILA-Lite-2B), distance delta vectors are collapsed near the origin without forming a distinguishable axis. Fine-tuning with 2M samples initiates directional spreading, but vertical and distance clusters remain overlapped. RoboRefer exhibits three cleanly separated clusters, each aligned with a distinct principal component. Figure 6b quantifies this contrast. The NVILA scaling trajectory yields only marginal gains in $\text{Coh}_\text{D}$ while VD-EI remains high (0.54–0.64). RoboRefer occupies a distinct region: the highest $\text{Coh}_\text{D}$ (0.182) and the lowest VD-EI (0.362) in the family, corresponding to 59.7% counter accuracy on EmbSpatial-Bench versus 41.1% for NVILA (2M).

These results suggest that high $\text{Coh}_\text{D}$, together with low VD-EI as a complementary signal, accompanies robust spatial reasoning across benchmarks. Within




**(a)** Counter Accuracy vs. Distance Coherence **(b)** Distance Coherence vs. VD-EI

**Fig. 6: Internal probing analysis of spatial representations.** (a) Positive correlation between behavioral accuracy on counter examples and internal distance coherence ($\text{Coh}_\text{D}$). (b) Comparing distance coherence ($\text{Coh}_\text{D}$) against geometric entanglement (VD-EI) within the NVILA family; RoboRefer occupies a unique region of high coherence and low entanglement. Unlabeled points denote base models, and numeric labels (e.g., 80k) indicate data-mix fine-tuned variants.

**Table 5: Axis coherence and VD-Entanglement Index.** Axis coherence measures directional consistency within axes. Distance coherence (*i.e.*, $\text{Coh}_\text{D}$) is consistently the lowest among the three axes across all models. **Bold** denotes best.

| Model | $\text{Coh}_\text{H}$ | $\text{Coh}_\text{V}$ | $\text{Coh}_\text{D}$ | VD-EI |
|---|---|---|---|---|
| Molmo-7B-O-0924 | 0.143 | 0.228 | 0.075 | 0.279 |
| + 80k | 0.122 | 0.332 | 0.072 | 0.388 |
| + 400k | 0.236 | 0.597 | 0.096 | 0.459 |
| + 800k | 0.247 | 0.559 | 0.107 | 0.514 |
| + 2M | 0.239 | 0.574 | 0.112 | 0.474 |
| NVILA-Lite-2B | 0.323 | 0.289 | 0.052 | 0.539 |
| + 80k | 0.295 | 0.497 | 0.070 | 0.606 |
| + 400k | 0.242 | 0.574 | 0.095 | 0.589 |
| + 800k | 0.278 | 0.498 | 0.089 | 0.591 |
| + 2M | 0.241 | 0.553 | 0.104 | 0.550 |
| RoboRefer-2B-SFT | **0.649** | **0.830** | **0.182** | 0.362 |
| Qwen2.5-VL-3B-Instruct | 0.367 | 0.293 | 0.043 | 0.457 |
| + 80k | 0.386 | 0.315 | 0.040 | 0.456 |
| + 400k | 0.450 | 0.452 | 0.042 | 0.451 |
| + 800k | 0.473 | 0.538 | 0.045 | 0.429 |
| + 2M | 0.485 | 0.586 | 0.052 | 0.472 |

the NVILA scaling series, incremental increases in fine-tuning scale yield only modest gains in $\text{Coh}_\text{D}$. In contrast, RoboRefer and Qwen3 exhibit well-separated axes and strong overall performance, showing such structure can emerge un-

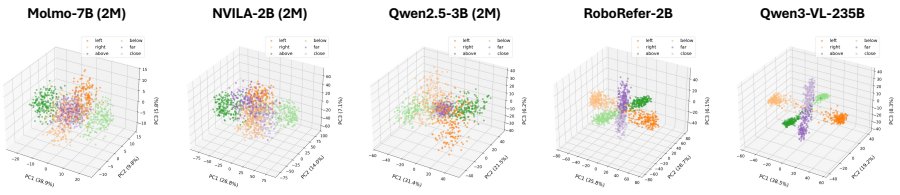



**Fig. 7: PCA of delta vectors across models.** Each point is a delta vector colored by axis (orange: horizontal, green: vertical, purple: distance), with darker/lighter shades distinguishing opposing categories within each axis (*e.g.*, left vs. right). Molmo (2M), NVILA (2M), and Qwen2.5 (2M) show separation along horizontal and vertical axes, but distance delta vectors remain poorly distinguished. RoboRefer and Qwen3 exhibit three clearly separated clusters, each aligned with a distinct principal component.

der substantially richer training regimes. RoboRefer reflects a different training regime (*e.g.*, additional supervision and much larger training), so we treat it as an illustrative reference rather than attributing its gains to a single factor. Overall, $\mathrm{Coh}_{\mathrm{D}}$ can serve as a practical diagnostic for whether training improves spatial representations.

## 6 Conclusion

We introduced a representation-level diagnostic framework that reveals vertical-distance entanglement as a pervasive, model-intrinsic bias across VLM families and scales. Our analysis shows that models with more structured spatial representations – characterized by high distance coherence and low VD-Entanglement Index – not only exhibit stronger counter-heuristic robustness but also achieve higher accuracy across diverse spatial reasoning benchmarks. To isolate this bias from evaluation-set confounds, we introduced the synthetic benchmark `SpatialTunnel`, which removes perspective-driven correlations present in natural images. Together, these results demonstrate that representational structure, rather than benchmark accuracy alone, provides a reliable indicator of robust spatial reasoning in vision-language models.

## Acknowledgments

This work was supported by IITP grants funded by the Korea government (MSIT) (RS-2021-II211343: AI Graduate School Program at SNU (5%), 02-26-01-0285: the Advanced GPU Utilization Support Program (5%), RS-2024-00509257: Global AI Frontier Lab (40%), and RS-2025-25442338: AI Star Fellowship Support Program at SNU (50%)).

## References


1. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al.: Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691 (2022)
2. Anthropic: Claude opus 4 & claude sonnet 4 system card. Tech. rep., Anthropic (May 2025), https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Blender Online Community: Blender - a 3D modelling and rendering package. Blender Foundation (2025), https://www.blender.org/
6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
7. Chen, H., Lin, J., Chen, X., Fan, Y., Jin, X., Su, H., Dong, J., Fu, J., Shen, X.: Rethinking visual layer selection in multimodal llms. arXiv e-prints pp. arXiv–2504 (2025)
8. Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J.C., Geva, M., He, J., Wu, J., Li, M.: Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=k7vcuqLK4X
9. Chen, S., Uy, M.A., Song, C.H., Ladhak, F., Murali, A., Qu, Q., Birchfield, S., Blukis, V., Tremblay, J.: SpaceTools: Tool-augmented spatial reasoning via double interactive RL. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026), https://arxiv.org/abs/2512.04069, to appear
10. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
11. Danier, D., Aygün, M., Li, C., Bilen, H., Mac Aodha, O.: Depthcues: Evaluating monocular depth perception in large vision models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20049–20059 (2025)
12. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 91–104 (2025)
13. Deshpande, A., Deng, Y., Ray, A., Salvador, J., Han, W., Duan, J., Zeng, K.H., Zhu, Y., Krishna, R., Hendrix, R.: Graspmolmo: Generalizable task-oriented grasping via large-scale synthetic data generation. arXiv preprint arXiv:2505.13441 (2025)
14. Du, M., Wu, B., Li, Z., Huang, X.J., Wei, Z.: Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 346–355 (2024)

15. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
16. Google: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
17. Gurnee, W., Tegmark, M.: Language models represent space and time. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 2483–2503 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/0a6059857ae5c82ea9726ee9282a7145-Paper-Conference.pdf
18. Hu, J., Levy, R.: Prompting is not a substitute for probability measurements in large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 5040–5060 (2023)
19. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Galliker, M.Y., Ghosh, D., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A.Z., Shi, L.X., Smith, L., Springenberg, J.T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., Zhilinsky, U.: $\pi_{0.5}$: a vision-language-action model with open-world generalization (2025), https://arxiv.org/abs/2504.16054
20. Kamath, A., Hessel, J., Chang, K.W.: What's "up" with vision-language models? investigating their struggle with spatial reasoning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023)
21. Kang, R., Chen, H., Gkioxari, G., Perona, P.: Linear mechanisms for spatiotemporal reasoning in vision language models. arXiv preprint arXiv:2601.12626 (2026)
22. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: Openvla: An open-source vision-language-action model. In: Agrawal, P., Kroemer, O., Burgard, W. (eds.) Proceedings of the 8th Conference on Robot Learning (CoRL). Proceedings of Machine Learning Research, vol. 270, pp. 2679–2713. PMLR (06–09 Nov 2025), https://proceedings.mlr.press/v270/kim25c.html
23. Li, F., Song, W., Zhao, H., Wang, J., Ding, P., Wang, D., Zeng, L., Li, H.: Spatial forcing: Implicit spatial representation alignment for vision-language-action model. arXiv preprint arXiv:2510.12276 (2025)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
25. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4122–4134 (2025)
26. NVIDIA, :, Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L.J., Fang, Y., Fox, D., Hu, F., Huang, S., Jang, J., Jiang, Z., Kautz, J., Kundalia, K., Lao, L., Li, Z., Lin, Z., Lin, K., Liu, G., Llontop, E., Magne, L., Mandlekar, A., Narayan, A., Nasiriany, S., Reed, S., Tan, Y.L., Wang, G., Wang, Z., Wang, J., Wang, Q., Xiang, J., Xie, Y., Xu, Y., Xu, Z., Ye, S., Yu, Z., Zhang, A., Zhang, H., Zhao, Y., Zheng, R., Zhu, Y.: Gr00t n1: An open foundation model for generalist humanoid robots (2025), https://arxiv.org/abs/2503.14734
27. OpenAI: Openai gpt-5 system card (2025), https://arxiv.org/abs/2601.03267

28. Ray, A., Duan, J., II, E.L.B., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., Zeng, K.H., Saenko, K.: SAT: Dynamic spatial aptitude training for multimodal language models. In: Second Conference on Language Modeling (2025), https://openreview.net/forum?id=DW8U8ZWa1U
29. Sheta, H., Huang, E.H., Wu, S., Alenabi, I., Hong, J., Lin, R., Ning, R., Wei, D., Yang, J., Zhou, J., et al.: From behavioral performance to internal competence: Interpreting vision-language models with vlm-lens. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 886–895 (2025)
30. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: Progprompt: Generating situated robot task plans using large language models. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11523–11530 (2023). https://doi.org/10.1109/ICRA48891.2023.10161317
31. Skean, O., Arefin, M.R., Zhao, D., Patel, N.N., Naghiyev, J., LeCun, Y., Shwartz-Ziv, R.: Layer by layer: Uncovering hidden representations in language models. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=WGXb7UdvTX
32. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15768–15780 (2025)
33. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.L., Su, Y.: Llm-planner: Few-shot grounded planning for embodied agents with large language models. In: ICCV (2023)
34. Tan, H., Zhou, E., Li, Z., Xu, Y., Ji, Y., Chen, X., Chi, C., Wang, P., Jia, H., Ao, Y., Cao, M., Chen, S., Li, Z., Liu, M., Wang, Z., Rong, S., Lyu, Y., Zhao, Z., Co, P., Li, Y., Han, Y., Xie, S., Yao, G., Wang, S., Zhang, L., Yang, X., Jiao, Y., Shi, D., Xie, K., Nie, S., Men, C., Lin, Y., Wang, Z., Huang, T., Zhang, S.: Robobrain 2.5: Depth in sight, time in mind. arXiv preprint arXiv:2601.14352 (2026)
35. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
36. Tong, S., II, E.L.B., Wu, P., Woo, S., IYER, A.J., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., Pan, X., Fergus, R., LeCun, Y., Xie, S.: Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=Vi8AepAXGy
37. Wang, X., Ma, W., Zhang, T., de Melo, C.M., Chen, J., Yuille, A.: Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24669–24679 (2025)
38. Wang, Y., Shi, F.: Logical forms complement probability in understanding language model (and human) performance. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 16862–16877 (2025)
39. Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. arXiv preprint arXiv:2503.22976 (2025)

40. Zhang, W., Huang, Y., Xu, Y., Huang, J., Zhi, H., Ren, S., Xu, W., Zhang, J.: Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture (2025), https://arxiv.org/abs/2509.02359
41. Zhang, Z., Hu, F., Lee, J., Shi, F., Kordjamshidi, P., Chai, J., Ma, Z.: Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=84pDoCD4lH
42. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., Zhang, S.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=OGxalNUHbJ