WHAT VLMS NEED TO SEE: INPUT REPRESENTATIONS FOR SPATIAL REASONING

Matei Gardea, Ziteng (Ender) Ji, Max Yen
University of California, Berkeley
{matei.gardea, zitengji, maxyen}@berkeley.edu

ABSTRACT

Vision-language models (VLMs) reach strong accuracy on semantic perception but fall well below human performance on spatial reasoning tasks (Liu et al., 2023; Kamath et al., 2023; Fu et al., 2024; Li et al., 2025; Jia et al., 2025). A standard explanation attributes this gap to the input representation: the visual encoder fails to preserve 3D structure, and recent work proposes to inject additional spatial signal into the prompt as depth maps (Cai et al., 2024; Chen et al., 2025a), structured scene descriptions (Zhang et al., 2025), or symbolic spatial rationales (Liu et al., 2025; Yang et al., 2024). We test this prescription under matched-input conditions. We construct a procedurally generated synthetic spatial-QA benchmark of approximately $2{,}000$ scenes and $19{,}024$ questions, in which each question is paired with four input representations: RGB alone, RGB with a ground-truth depth image, RGB with a deterministic structured scene description, and RGB with both. We evaluate five open-weights VLMs at the 2–4 B parameter scale (Qwen3-VL-4B, Qwen2.5-VL-3B, PaliGemma 2-3B, Gemma 3-4B, Cosmos-Reason2-2B) without any task-specific fine-tuning, producing roughly $342{,}000$ scored predictions. We find that adding ground-truth spatial signal at the input degrades average accuracy for four of the five models. Appending a depth image hurts every model uniformly, and prepending a perfect structured description reduces overall accuracy for the strong models, even though it lifts accuracy sharply on the tasks for which 2D recognition was the bottleneck (e.g., camera-extremum accuracy gains $+22$ points for Qwen3-VL-4B and $+47$ points for Gemma 3). We further document a sharp competence gradient across spatial axes: lateral left/right is near ceiling for the stronger models, while egocentric depth ordering is near chance for the weaker ones. Several models also display severe yes-bias and counting mode collapse on binary types, which inflates their raw accuracy. We argue that the bottleneck is not the absence of spatial information at the input, but the language side's inability to ground onto the visual evidence already present in the image.

Code and data: https://github.com/mateig/vlm-irp .

1 INTRODUCTION

Recent vision-language models (VLMs) recognise objects, describe scenes, and answer semantic questions with high accuracy, but they continue to perform poorly on spatial reasoning (Liu et al., 2023; Kamath et al., 2023; Fu et al., 2024; Li et al., 2025; Jia et al., 2025). A common explanation in the literature attributes this gap to the input representation (Cai et al., 2024; Chen et al., 2025a; Liu et al., 2025; Qi et al., 2025): the visual encoder fails to preserve 3D structure, the language model never learns to consume what does survive, and accuracy on tasks that hinge on relative position, depth, or perspective remains near chance. The prescription that follows from this diagnosis is to enrich the input. Recent work has proposed augmenting the prompt with a depth image (Cai et al., 2024; Chen et al., 2025a), with a structured scene description (Zhang et al., 2025; Liu et al., 2025), or with explicit textual scaffolds (Yang et al., 2024; Wang et al., 2025), and most of these works report gains and conclude that the representation matters. The question we ask in this paper is simpler: when the model, scenes, and questions are all held fixed, does adding spatial signal at the input actually improve accuracy?

We answer this question with a controlled empirical study. We procedurally generate approximately $2{,}000$ synthetic 3D scenes with full ground-truth geometry, and we emit a suite of templated spatial questions per scene. Because the scenes are synthetic, we can produce four matched input representations at no additional annotation cost: the RGB image alone, the image paired with a ground-truth depth render, the image paired with a deterministic structured text description, and the image paired with both auxiliary signals. The same question, scored by exact match, is presented under all four conditions, so any difference in accuracy is attributable to the input representation. We evaluate five open-weights VLMs spanning two model families and a 2–4 B parameter range against the full sweep, producing roughly 360k predictions, of which 342k pass our scoring filter.

Our findings run counter to the prevailing prescription. Adding ground-truth spatial signal at the input lowers average accuracy for four of the five models we tested. A depth image is harmful in every case: each model loses roughly 1 to 4 accuracy points when its RGB prompt is augmented with a perfect depth render. A structured description is also harmful in expectation for every model except Gemma 3, which gains $+13$ points; we show in §5.4 that this gain is an artifact of a near-degenerate yes-bias in Gemma 3's RGB-only baseline. Underneath the average, however, the description is not a uniform regression but a substitution. It improves tasks whose bottleneck is 2D perception, with camera-extremum and directional-extremum gaining up to $+47$ points for Gemma 3 and $+22$ for Qwen3-VL, while it degrades tasks that the visual stream already solves, such as binary spatial relations and counting. Combined with a sharp axis-of-competence asymmetry (lateral left/right near ceiling for the strong models, egocentric depth ordering near chance for the weak ones) and strong yes-bias on binary types, these results suggest that the representation problem is not a shortage of spatial signal, but the language side's inability to consume the signal already present in the image.

Contributions.

- A controlled synthetic spatial-QA benchmark of approximately $2{,}000$ scenes, six scene regimes, and seven scored question types, with matched RGB, depth, and description conditions for every example.
- A 5-model $\times$ 4-modality $\times$ $\sim$19k-question evaluation ($\sim$342k scored rows) that isolates the effect of input representation from model capacity.
- Evidence that the prevailing prescription of adding more spatial signal to the input fails on average for 2–4 B-parameter open-weights VLMs, together with a per-question-type breakdown that identifies when this prescription fails (binary perception tasks, where the visual stream is already grounded) and when it succeeds (extremum tasks, where 2D recognition was the bottleneck).
- A failure-mode analysis covering axis competence, yes-bias, and counting mode collapse, which provides evidence that the bottleneck for these models is grounding, not signal.

2 Related Work

Spatial reasoning as a weakness of VLMs. A range of benchmarks, including VSR (Liu et al., 2023), What's Up (Kamath et al., 2023), BLINK (Fu et al., 2024), 3DSRBench (Ma et al., 2025), ViewSpatial-Bench (Li et al., 2025), and the recent OmniSpatial suite (Jia et al., 2025), report that current VLMs perform far below humans and often near chance on perspective-sensitive spatial questions. Mechanistic studies of these failures identify two recurring patterns. First, attention is frequently mis-routed onto irrelevant regions during spatial reasoning, and an inference-time intervention that re-shapes attention can recover much of the lost accuracy (Chen et al., 2025b). Second, accuracy degrades when a question requires re-anchoring from the camera viewpoint to an allocentric or person-centred frame (Li et al., 2025; Lee et al., 2025). Qi et al. (2025) trace one mechanism behind these effects to the embedding norms of vision tokens, which overwhelm textual position information and cause the visual stream to be consumed as a bag of semantic tokens rather than a structured spatial layout. The shared conclusion is that current VLMs default to image-plane heuristics rather than to a genuine 3D representation of the scene.

Adding spatial signal to the input. A line of recent work attempts to compensate for this representational shortfall by enriching the input. SpatialBot (Cai et al., 2024) and SD-VLM (Chen et al.,

2025a) pair RGB with a depth image, either as an additional visual input or as a depth-encoded positional signal inside the visual tokens. SSR (Liu et al., 2025) translates depth into textual rationales that are distilled into latent embeddings. SpatialMind (Zhang et al., 2025) compares structured prompts that decompose the scene into 3D maps, 2D grids, and object-centric captions, and Yang et al. (2024) and MINDCUBE (Wang et al., 2025) study text-side scaffolds, including cognitive-map prompting. Along an orthogonal axis, SpatialVLM (Chen et al., 2024) and SpatialRGPT (Cheng et al., 2024) train on large amounts of synthetic spatial QA derived from 3D-lifted images. The common claim across these papers is that augmenting the input with explicit spatial information improves spatial reasoning. The closest counterpoint is Wang et al. (2024), which finds that text-only encodings of synthetic spatial scenes can match or exceed visual encodings for some tasks, and that VLMs sometimes underperform their LLM backbones on the same spatial questions. Our study extends this controlled-comparison perspective by holding the model, scenes, and questions fixed and asking whether the input augmentations that this literature has converged on actually move accuracy in the expected direction.

Synthetic scenes for controlled evaluation. Procedural rendering provides a cheap source of ground-truth geometry, segmentation, and scene metadata for arbitrary camera–object configurations. We render our scenes directly through Blender's bpy Python API and use the resulting ground truth to derive matched RGB, depth, and description conditions for every question, which avoids confounding the input representation we are testing with the noise of human annotation.

3 A Controlled Synthetic Spatial-QA Benchmark

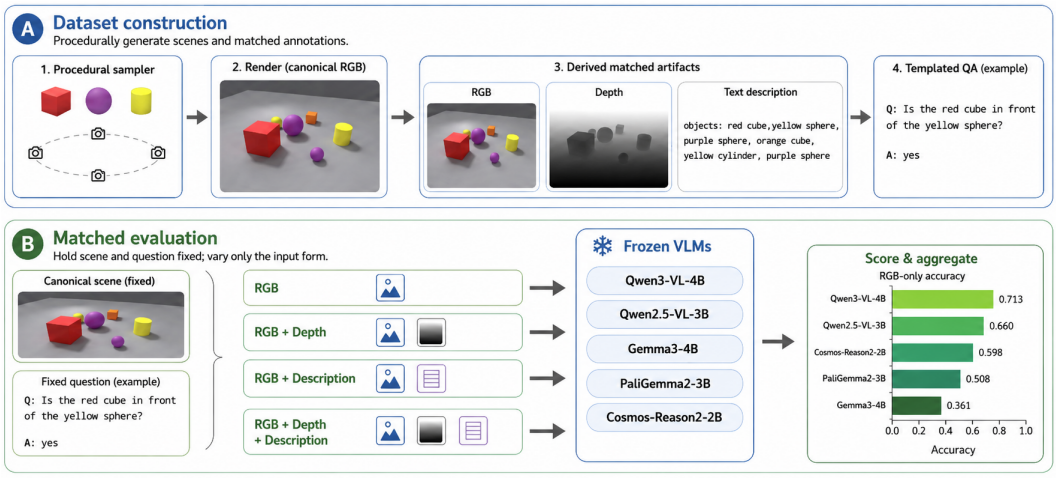


Figure 1: Pipeline. (1) A procedural sampler instantiates a 3-to-6 object scene under one of six regimes. (2) Blender's Cycles renderer, driven through bpy, produces RGB, depth, and instance segmentation; full camera-frame geometry is dumped to JSON. (3) A deterministic serialiser produces a structured text description; templated QA generation produces up to ten balanced spatial questions per scene. (4) For every QA pair we evaluate four matched input conditions against five open-weights VLMs without further fine-tuning.

3.1 Scene Generation

The full pipeline is summarised in Figure 1. We render approximately 2,000 scenes at 1024×1024 resolution with Blender's Cycles renderer, driven directly through the bpy Python API. Generation runs on a single RTX 4090 in roughly three hours. Each scene contains 3 to 6 simple primitives (cube, sphere, cylinder) drawn from a 10-colour, 3-size palette. The scenes are sampled round-robin from six regimes: `sparse` (3 objects), `dense` (5–6), `depth_ambiguous` (4–6 objects, with at least one pair that is laterally close but depth-distant), `size_mixed` (one small and one large object), `same_color_cluster` (two objects forced to share a colour), and `vertical` (a tall cylinder among shorter objects). Cameras are sampled on a hemispherical orbit at 30° elevation and 7–9 m

radius with a 60° field of view, fixing the egocentric viewpoint while varying the scene layout. For each scene we save the RGB render, a Z-pass depth EXR normalised to a $\mathrm{uint8}$ PNG, an instance segmentation, the camera intrinsics and extrinsics, per-object world and camera-frame poses, and the full pairwise spatial-relation table (`left_of`, `right_of`, `in_front_of`, `behind`, `closer_than`, `farther_than`) computed in the camera's right/forward axes with 0.15 m lateral and 0.20 m depth thresholds.

3.2 Question–answer generation

For each scene we generate up to ten spatial QA pairs drawn from eight templates, balanced for yes/no parity and de-duplicated. Two of the eight types (`same_side` and `id_extremum`) score reliably at zero on the text-match scorer for reasons unrelated to model competence (§4) and are dropped, leaving seven scored question types: binary relation (yes/no), existence (yes/no), counting (integer), distance comparison (yes/no), directional extremum (object label), camera extremum (closest or farthest, object label), and between (object label or `none`). Templates restrict referents to objects whose $(\textit{colour}, \textit{shape})$ pair is unique within the scene, and extremum questions require a top-2 gap of at least 0.3 m so that the gold label is unambiguous.

3.3 Four matched input representations

The same scene and question are presented to the model under four input conditions:

- **RGB**: the rendered image alone. This is the default VLM input.
- **RGB + Depth**: the RGB image together with a second image, the ground-truth depth map normalised to grayscale, both passed as image tokens.
- **RGB + Description**: the RGB image together with a deterministic textual scene description, prepended to the question. The description lists per-object world position, camera-frame (x, y, z) coordinates, and one binary relation per ordered pair.
- **RGB + Depth + Description**: both auxiliary signals, image and text.

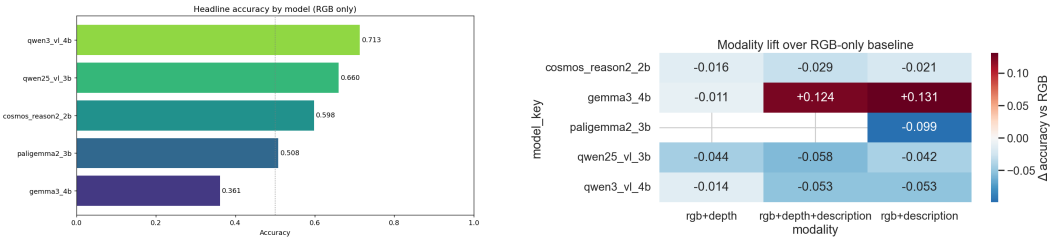
The description is generated from the same ground truth used to score the question, so the text contains the answer to every binary, existence, and extremum question in the scene. This choice is intentional: it sets an upper bound on the benefit that a perfect spatial-text encoder could ever deliver to the model, and the gap between that upper bound and the observed accuracy measures how well the language side consumes the signal.

4 Experimental Setup

Models. We evaluate five open-weights VLMs in the 2–4 B parameter range, without any task-specific fine-tuning: Qwen2.5-VL-3B-Instruct (Bai et al., 2025), Qwen3-VL-4B-Instruct (Qwen Team, 2025), Gemma 3-4B-IT (Gemma Team, 2025), PaliGemma 2-3B-mix-224 (Steiner et al., 2024), and Cosmos-Reason2-2B (NVIDIA, 2025). PaliGemma 2 mix accepts only a single image per prompt, so we evaluate it on the `RGB` and `RGB + Description` conditions.

Inference. For every (*model*, *scene*, *question*, *modality*) tuple, we issue a single prompt with the system message *"Answer with one word, a number, or a short phrase."* Decoding is greedy with `max_new_tokens=64` and `do_sample=False`. Non-PaliGemma models receive the standard chat template; the user content interleaves $\{\textsc{rgb}, (\textsc{depth}), \textsc{text}\}$ as available. Inference runs on a single NVIDIA RTX PRO 6000 (Blackwell) GPU, takes roughly 16 hours, and amounts to about 360k forward passes (5 models $\times$ 2,000 scenes $\times$ 10 questions $\times$ either 2 or 4 modalities). Responses are appended to a JSONL with a resume index keyed on the tuple, so individual failures can be retried without re-running successful rows.

Scoring. Predictions are scored with a type-dispatched lenient text-match. Yes/no types look for the first yes/no token in the response. Counting extracts the first integer or number-word (`zero` through `ten`, with `none` mapped to 0). Object-label types (extrema and between) tokenise the response for colour and shape mentions, applying shape aliases (`ball`/`orb` $\rightarrow$ sphere; `box`/`block` $\rightarrow$



(a) Overall accuracy on the matched RGB subset.

(b) Δ accuracy versus the RGB-only baseline.

Figure 2: Overall results. (a) Qwen3-VL-4B leads the RGB-only ranking at 71.3 % and Gemma 3-4B trails at 36.1 %. (b) Adding depth is uniformly harmful; adding description hurts every model on average except Gemma 3, whose RGB-only baseline is driven by a degenerate yes-bias.

cube; `can/tube` $\rightarrow$ cylinder), and require either a strict colour-and-shape match against the gold label, or that the scene's objects filtered by the response's mentions reduce to the unique gold object. The `between` type accepts any `no/none/nothing` token when the gold answer is `none`. We drop `same_side` and `id_extremum`: the former ends with "(left or right)?" and is read as multiple-choice by every model, returning a side rather than yes/no, and the latter has a non-unique parenthetical disambiguator that no colour-and-shape scorer can resolve. Both types score near zero across all models and modalities, and we treat them as evaluation noise relative to the comparisons of interest.

Matched-modality comparison. A naive per-model average across modalities is biased against any model evaluated on fewer modalities, and PaliGemma 2 sees only two of the four conditions. We therefore report aggregate accuracies on the `RGB` subset, which contains 19,024 questions per model and is identical across all five models. All representation-effect plots are reported as Δ accuracy with respect to the same `RGB` baseline.

5 Results

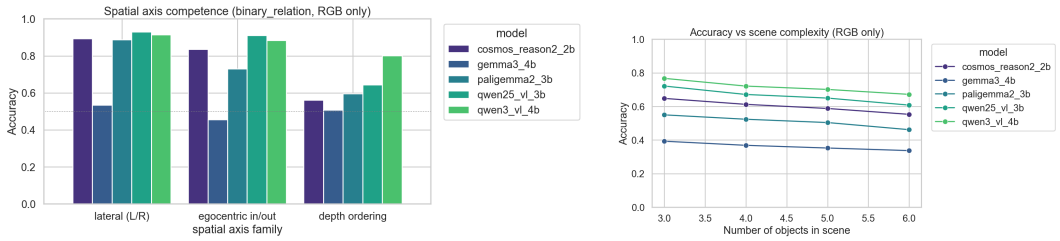
5.1 RGB-only is the best input on average for four of five models

Figure 2 reports the overall accuracies and the modality lift. On RGB-only inputs, Qwen3-VL-4B leads at 71.3 %, followed by Qwen2.5-VL-3B (66.0 %), Cosmos-Reason2-2B (59.8 %), PaliGemma 2-3B (50.8 %), and Gemma 3-4B (36.1 %). The modality-lift heatmap in Figure 2b is the more informative panel: the depth-only column is negative for every model, ranging from -1.1 points (Gemma 3) to -4.3 points (Qwen2.5), and the description-only and depth+description columns are negative for four of the five models. The single positive row is Gemma 3, which gains $+13.1$ points from description and $+12.4$ from description+depth. As we show in §5.4, this gain is a direct consequence of Gemma 3's RGB-only baseline being driven by an always-yes response on the binary types.

Why depth as an image hurts. The depth render adds visual tokens that compete with the RGB image for the context budget but carry no colour or shape information. For object-label questions, the model still grounds identifiers in RGB while paying an attention cost on uninformative depth tokens, yielding the uniform 1 to 4 point regression we observe.

Why a perfect description hurts on average but helps in places. The structured description literally contains the gold answer for most binary and extremum questions, so a model that read it carefully would score near ceiling. None do. We attribute the regression to two factors. First, the description is long (25–60 tokens of pose plus pairwise relations) and re-orients the model toward language-side reasoning, the weaker pathway for tasks the visual stream already solves. Second, the description contains many distractor relations (every ordered pair), and locating the relevant one is a needle-in-a-haystack task the visual input typically wins.

Figure 3: Per-question-type accuracy by modality, per model. The structured description swaps perceptual competence for symbolic competence: `camera_extremum` and `directional_extremum` rise sharply with description, while `binary_relation`, `counting`, and `distance_comparison` fall.



(a) Spatial-axis competence on `binary_relation`.

(b) Accuracy versus scene complexity.

Figure 4: (a) Lateral left/right is near ceiling for every model except Gemma 3, while depth ordering is barely above chance for the weaker models. (b) Accuracy decays monotonically with object count for every model.

5.2 The description is a substitution, not an addition

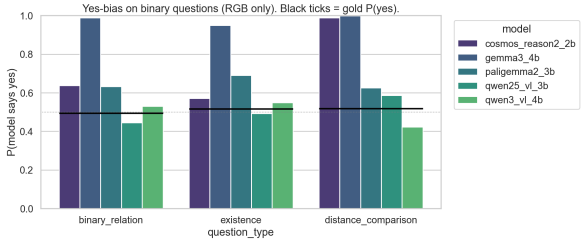
Figure 3 resolves the average lift into per-question-type cells and tells a different story than Figure 2b. Adding the description *trades* performance across question types in a consistent direction:

- **Extremum tasks gain substantially.** The largest single-cell jump in our study is Gemma 3 on `camera_extremum`, $0.187 \rightarrow 0.659$ ($+47.2$ points), when the description is added. Qwen3-VL-4B gains $+21.6$ points on the same task ($0.653 \rightarrow 0.869$). Directional extrema gain $+35.9$ points on Gemma 3 and $+13.0$ on Qwen3-VL.
- **Binary perception tasks lose.** On Qwen3-VL, `binary_relation` falls -8.2 points ($0.866 \rightarrow 0.784$); `counting` falls -16.9 ($0.771 \rightarrow 0.602$); and `distance_comparison` falls -18.4 ($0.754 \rightarrow 0.570$). The same direction holds for Qwen2.5 and PaliGemma.
- **The exchange rate depends on the model's RGB starting point.** Where the RGB-only accuracy is already strong (Qwen3 binary at $87\,\%$), the description hurts; where it is weak (Gemma 3 camera-extremum at $19\,\%$), the description is a near-decisive aid.

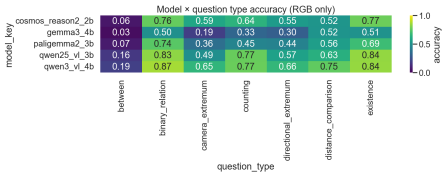
We interpret this as a substitution effect: the description re-routes inference through the language model's text-comprehension capability, which helps when the visual stream was failing (extremum tasks require a global 2D reasoning step that small VLMs perform poorly) and hurts when the visual stream was already grounded (direct binary lookups). The gain is bounded by the model's ability to consume long structured text rather than by the information content of that text: the description contains the answer to every binary question, yet binary accuracy still falls.

5.3 A sharp axis-of-competence gradient explains where models break

Aggregating accuracy by question-type alone hides a more telling pattern. Figure 4a regroups `binary_relation` by spatial axis: **lateral** (`left_of`, `right_of`), **egocentric in/out** (`in_front_of`,



(a) $P(\text{model says yes})$. Black ticks mark the gold yes rate.



(b) Model × question-type accuracy (RGB-only).

Figure 5: (a) Gemma 3 says yes 99 % of the time on `binary_relation` and 100 % on `distance_comparison`; Cosmos collapses on distance comparisons, and PaliGemma is biased toward yes on both binary types. (b) The same pattern is visible in the qtype heatmap as inflated accuracy on yes-heavy types.

`behind`), and **depth ordering** (`closer_than`, `farther_than`). Lateral left/right is at or near ceiling for every model except Gemma 3, with accuracies in the range 0.89 to 0.93 for the others. Egocentric in/out accuracies are comparable for the strong models. Depth ordering, however, collapses: Qwen3-VL holds 0.79, but Qwen2.5-VL sits at 0.64, PaliGemma at 0.59, Cosmos at 0.55, and Gemma 3 at 0.50 (random). Depth ordering is the only axis that demands genuine 3D inference rather than image-plane comparison, and the drop matches the literature's claim that VLMs default to reasoning on the 2D projection (Yang et al., 2024; Wang et al., 2024; Qi et al., 2025). It also explains why a depth image alone fails to help: the bottleneck is fusing depth into a relational judgement, not seeing depth.

Figure 4b shows the second structural pattern: accuracy decays monotonically with object count for every model. Qwen3-VL drops -11.1 points from 3-object scenes (0.745) to 6-object scenes (0.634); Gemma 3 drops -9.4 points ($0.472 \rightarrow 0.378$). The decay rates are roughly parallel across models, suggesting a shared scaling problem rather than a model-specific artifact: every additional object adds candidates, distractors, and disambiguation steps to the same visual encoder.

5.4 Yes-bias and counting mode collapse inflate raw accuracy

Three of the five models exhibit a degenerate response distribution on binary question types, visible in Figure 5a. Gemma 3 says yes $98.9\,\%$ of the time on `binary_relation` and $99.9\,\%$ of the time on `distance_comparison`; Cosmos says yes $98.8\,\%$ of the time on `distance_comparison`; and PaliGemma sits at $63\,\%$ on both types. Because the gold yes-rate is approximately $50\,\%$, an always-yes responder achieves roughly $50\,\%$ accuracy on these types from the prior alone, so much of Gemma 3's RGB-only score reflects a constant baseline rather than spatial reasoning. This effect also explains why Gemma 3 alone gains so much from the description: the description re-anchors its responses, replacing the always-yes prior with text-matched extraction, which is harder to collapse onto a single token. By contrast, Qwen3-VL-4B and Qwen2.5-VL-3B are well-calibrated, with yes rates within ± 4 points of gold.

Counting mode collapse. Counting confusion (Appendix Figure 8) reveals the same phenomenon on a non-binary axis. Gemma 3's predictions concentrate almost entirely on $\{1, 2\}$ regardless of the gold count: when the answer is 0 it predicts 1 on $57\,\%$ of cases and never predicts 0; when the answer is 4 it predicts 2 on $56\,\%$ of cases. The Qwen models are sharply diagonal (Qwen3-VL achieves at least 0.62 on the diagonal for every gold count from 0 to 4), and PaliGemma is strongly biased toward 1.

5.5 Scene-regime effects are small and consistent

Across scene regimes (Appendix Figure 7), the accuracy ranking is preserved, every model peaks on `sparse` scenes, and the `depth_ambiguous` regime sits within ± 2 points of `dense` for every model, suggesting that depth confusions are pervasive rather than layout-triggered.

5.6 THE HARDEST TASK IS THE NON-RECOGNITION TASK: `between`

Every model scores below $20\,\%$ on `between` under every modality condition (Qwen3-VL tops at $18.9\,\%$ on RGB, Gemma 3 bottoms at $3.4\,\%$). Approximately $90\,\%$ of `between` questions have gold `none`, yet the models almost never reply "none". `between` is the only template that requires the model to deny the existence of a satisfying object, and the failure is a recognition-bias mismatch rather than a perception failure.

6 DISCUSSION

The bottleneck is grounding, not signal. If the input-representation thesis were correct, the depth-augmented and description-augmented conditions should both lift accuracy. Neither does on average. The single case in which an auxiliary signal helps a single model (the description for Gemma 3) does so by suppressing a degenerate language prior rather than by delivering new spatial information. The strong models lose accuracy when spatial signal is added because their language side cannot integrate it without disturbing the visual pathway that was already solving the task.

When does a structured spatial input pay off? Structured text helps when the task requires a global 2D or 3D summary the visual encoder cannot produce within its token budget (extremum tasks), and hurts on localised perceptual lookups the visual stream already grounds (binary relations, existence, counting). A useful spatial preprocessor should be task-conditioned rather than dumped wholesale into the prompt.

Implications for VLA perception design. VLA systems commonly attach a spatial preprocessor that emits depth or 3D scene structure into the action policy's prompt (Qu et al., 2025). Our results suggest that for 2–4 B-parameter backbones this is at best neutral and at worst a regression, unless (i) the spatial output is a summary rather than a full pose dump, and (ii) the policy is calibrated against the yes-bias and mode-collapse failure modes that small VLMs exhibit even when the input is correct.

7 LIMITATIONS

Our scenes use a deliberately simple visual vocabulary (10 colours, 3 shapes, 3 sizes) on a uniform ground plane, which isolates spatial reasoning from semantic perception but means absolute accuracy numbers should not be compared with natural-image benchmarks. Our model range (2–4 B) was constrained by single-GPU inference, and our negative result for the description condition may invert at larger scales where long-context reasoning is more reliable. The system prompt is a single short sentence; a modality-specific prompt could plausibly reduce yes-bias without changing what the model sees, but we leave it out of scope to keep the representation comparison clean. Finally, the structured description is noiseless; a preprocessor-derived description in a real system would be lossier and would likely yield a smaller, possibly opposite, effect.

8 CONCLUSION

Holding the model, scenes, and questions fixed, we found that adding ground-truth spatial signal at the input, whether as a depth image or as a perfect structured description, degrades the average accuracy of 2–4 B open-weights VLMs on a controlled synthetic spatial-QA benchmark. The exception is task-conditioned: a structured description helps extremum questions, where the bottleneck is 2D global reasoning rather than perception, and hurts binary perception tasks, where the visual stream was already grounded. Underneath this modality comparison, the models display a sharp axis-of-competence asymmetry, with near-perfect lateral left/right accuracy and near-chance egocentric depth ordering, and weaker models fall back on degenerate language priors that inflate raw accuracy on yes/no types. These findings push back on the prevailing prescription that VLMs need richer spatial inputs: at this scale of model, the bottleneck is the language side's ability to ground onto signal that is already present in the image, and adding more signal to the prompt does not solve it.

CONTRIBUTION STATEMENT AND GENAI STATEMENT

The three authors contributed equally to dataset generation, inference, evaluation, and writing. We used Claude and Gemini for code completion and debugging, in accordance with the GenAI policy of CS 288 (Spring 2026, UC Berkeley; `https://cal-cs288.github.io/sp26/`).

REFERENCES

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.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *International Conference on Machine Learning (ICML)*, 2025b.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. *arXiv preprint arXiv:2504.17207*, 2025.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics (TACL)*, 11:635–651, 2023.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. SSR: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M. de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

NVIDIA. Cosmos-Reason2: Reasoning vision-language models for physical ai. Model card, `https://huggingface.co/nvidia/Cosmos-Reason2-2B`; code: `https://github.com/nvidia-cosmos/cosmos-reason2`, 2025. Post-trained from Qwen3-VL-2B-Instruct; successor to the Cosmos-Reason1 line (arXiv:2503.15558).

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. PaliGemma 2: A family of versatile VLMs for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025.

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. *arXiv preprint arXiv:2412.14171*, 2024.

Haoyu Zhang, Meng Liu, Zaijing Li, Haokun Wen, Weili Guan, Yaowei Wang, and Liqiang Nie. Spatial understanding from videos: Structured prompts meet simulation data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A Appendix

A.1 Supplementary plots

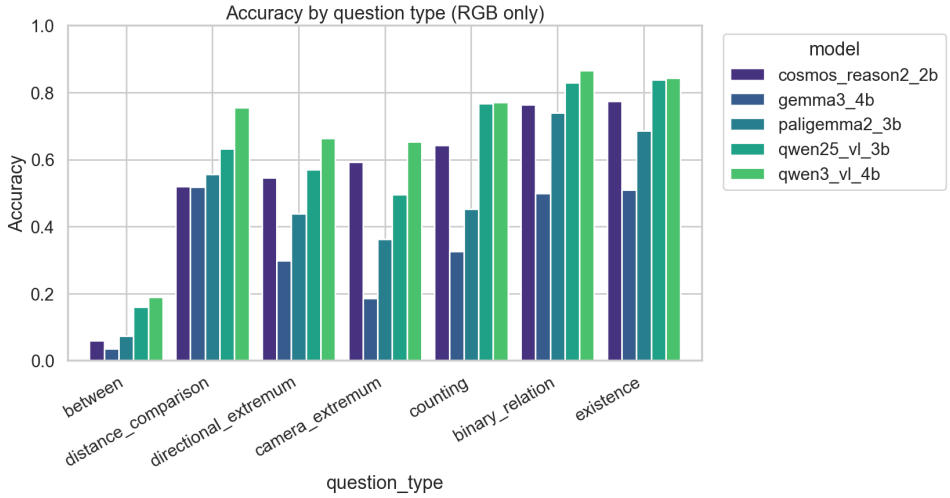


Figure 6: Accuracy by question type (RGB only). The `between` type is the hardest for every model, while `binary_relation` and `existence` are the easiest.

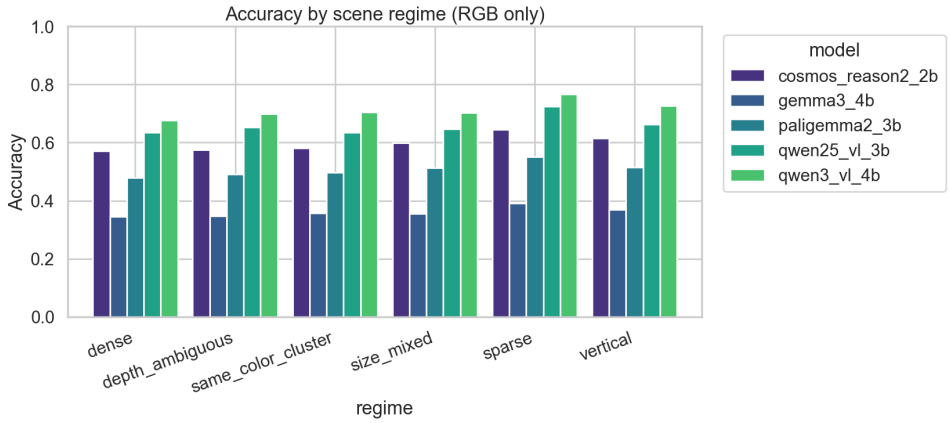


Figure 7: Accuracy by scene regime (RGB only). Sparse scenes are easiest for every model, and the `depth_ambiguous` regime is comparable to `dense`, indicating that depth confusion is pervasive rather than layout-specific.

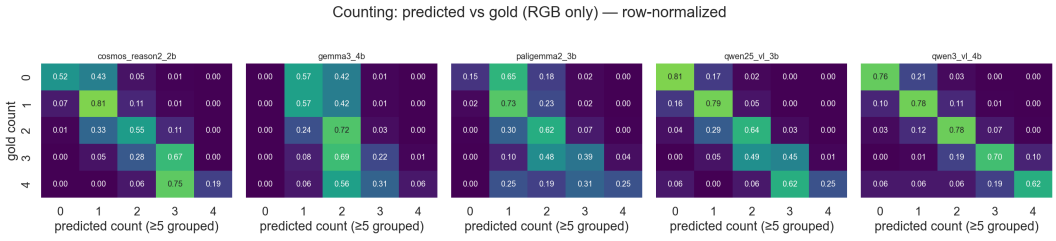


Figure 8: Counting confusion (RGB-only), row-normalised. Gemma 3 collapses onto $\{1, 2\}$ regardless of gold; PaliGemma is biased toward 1; the Qwen models are sharply diagonal.

A.2 PER-MODALITY OVERALL ACCURACY (FULL TABLE)

Table 1: Full per-modality accuracy. Cross-model comparisons require the same modality column across all models, and PaliGemma 2 supports only `rgb` and `rgb+description`. Bold indicates the best modality for each model.

Model	RGB	RGB+Depth	RGB+Desc	RGB+Depth+Desc
qwen3_vl_4b	**0.713**	0.699	0.661	0.660
qwen25_vl_3b	**0.660**	0.617	0.619	0.603
cosmos_reason2_2b	**0.598**	0.583	0.577	0.569
paligemma2_3b	**0.508**	–	0.409	–
gemma3_4b	0.361	0.350	0.492	**0.485**

A.3 SELECTED PER-QUESTION-TYPE MODALITY DELTAS

Table 2: Δ accuracy versus the RGB-only baseline, for question types where representation matters most. The structured description exhibits a clear substitution pattern: it lifts extremum tasks at the cost of binary perception tasks.

	Camera extremum		Binary relation	
Model	RGB	+Desc	RGB	+Desc
qwen3_vl_4b	0.653	0.869 (+0.22)	0.866	0.784 (−0.08)
qwen25_vl_3b	0.495	0.503 (+0.01)	0.828	0.749 (−0.08)
cosmos_reason2_2b	0.593	0.662 (+0.07)	0.763	0.681 (−0.08)
paligemma2_3b	0.363	0.308 (−0.06)	0.739	0.540 (−0.20)
gemma3_4b	0.187	0.659 (+0.47)	0.499	0.558 (+0.06)

A.4 INFERENCE LATENCY

Per-question wall-clock latency on the inference GPU (RTX PRO 6000 Blackwell), including all four modality conditions. PaliGemma is fastest because it accepts only one image per prompt; Qwen2.5-VL-3B is slowest due to a longer visual-token sequence. These numbers contextualise but do not affect accuracy comparisons.

Table 3: Mean / median per-question latency in seconds.

Model	Mean (s)	Median (s)
paligemma2_3b	0.023	0.022
gemma3_4b	0.091	0.095
cosmos_reason2_2b	0.116	0.130
qwen3_vl_4b	0.171	0.192
qwen25_vl_3b	0.262	0.276