

From Words to Meters: Eliciting Metric Spatial Knowledge in Vision-Language Models for Agent Perception

Gianfranco Lombardo^{1,*†}, Andrea Schianchi^{1,†}, Giuseppe Trimigno^{1,†}, Michele Tomaiuolo^{1,†} and Agostino Poggi^{1,†}

¹Department of Engineering and Architecture, University of Parma, Parco Area Delle Scienze 181/A, I-43124 Parma, Italy

Abstract

We investigate whether pre-trained Vision-Language Models (VLMs) possess latent metric spatial knowledge that can be activated through contextual prompting, without any fine-tuning or adaptation. We evaluate five models from the Qwen2.5-VL and LLaVA families in a zero-shot setting on a benchmark of 7,223 annotated object instances derived from NYU Depth V2, framing metric depth estimation as a visual question answering task administered under five query conditions of increasing contextual richness: a bare visual query, a semantic object prior, relational scene structure, and two forms of explicit metric anchor. Each condition is evaluated under both an independent and a cumulative protocol, and results are reported across six metrics including a novel operational definition of referential hallucination, a formally valid response naming the wrong object, tracked independently of metric estimation error. Our results show that explicit metric anchors are the dominant driver of accuracy, reducing RMSE below 1m for Qwen2.5-32B and improving Hit@0.5m by a factor of nearly four for Qwen2.5-7B, without any spatial training. The near and far anchors are not interchangeable: the near anchor more effectively reduces large errors while the far anchor improves tolerance accuracy, suggesting they activate different aspects of metric reasoning. Relational context produces qualitatively different effects across families: for smaller Qwen2.5 models it induces hallucination spikes up to 37.1% while partly confounding apparent accuracy gains, whereas for LLaVA models it reduces hallucination rates substantially without consistently improving metric accuracy. This dissociation is invisible to evaluations that do not track referential errors separately. The two model families exhibit structurally different profiles: LLaVA models carry high baseline hallucination rates (27–38%) that are only partially reduced by contextual grounding, while the Qwen2.5 family shows a strong scaling effect with Qwen2.5-32B occupying a distinct performance tier already at the bare visual query. These findings establish that pre-trained VLMs do encode metric spatial knowledge activatable through prompting, but that reliable activation requires both sufficient model capacity and appropriate linguistic grounding, with direct implications for the design of VLM-based perceptual modules in autonomous agent systems.

Keywords

Vision-Language Models, Metric depth estimation, Contextual prompting, NYU Depth V2, Embodied AI

1. Introduction

The ability to perceive and reason about the spatial structure of an environment is a fundamental prerequisite for any agent that must act within it. Whether navigating a corridor, avoiding an obstacle, or localising an object of interest, an autonomous agent requires reliable estimates of where things are, not merely what they are. Metric depth estimation has been studied extensively as a computer vision task in its own right, with dedicated supervised models such as MiDaS [1], DPT [2], and ZoeDepth [3] achieving strong performance on benchmarks such as NYU Depth V2 [4] and KITTI [5]. These systems are trained specifically for the task and deliver accurate results, but they require dedicated training pipelines and produce depth maps without any semantic grounding or linguistic interface.

The rise of Vision Language Models (VLMs) raises a different question. These models, trained on large multimodal corpora, have demonstrated strong performance on a wide range of visual understanding

WOA 2026: The 27th Workshop from Objects to Agents, June 15–17, 2026, Salerno, IT

*Corresponding author.

†These authors contributed equally.

✉ gianfranco.lombardo@unipr.it (G. Lombardo); andrea.schianchi1@studenti.unipr.it (A. Schianchi);
giuseppe.trimigno@unipr.it (G. Trimigno); michele.tomaiuolo@unipr.it (M. Tomaiuolo); agostino.poggi@unipr.it (A. Poggi)
ID 0000-0003-1808-4487 (G. Lombardo); 0009-0008-0600-2572 (A. Schianchi); 0009-0005-1907-2885 (G. Trimigno);
0000-0002-6030-9435 (M. Tomaiuolo); 0000-0003-3528-0260 (A. Poggi)



Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



Example of a cumulative prompt

Estimate the distance in meters of the *2nd chair from right* in the image from the viewpoint.

Considering its typical real-world dimensions.

Additional context: there are these objects in the image: 2nd cabinet from right, paper towel dispenser, 2nd paper from right, 2nd chair from right, microwave. Between these objects, 2nd cabinet from right is the closest and microwave is the farthest.

Additional information: the 2nd cabinet from right is exactly 2.64 meters away.

Additional information: the microwave is exactly 3.08 meters away.

Output format: {"object": <object_name>, "distance": <float>}

Figure 1: Example indoor scene from NYU Depth V2 (left) paired with the corresponding cumulative prompt for the Q1+Q2+Q3+Q4+Q5 condition (right). The queried object is the *2nd chair from right*. The prompt progressively adds a semantic prior, relational scene context, and two metric anchors to the base visual query.

tasks including image captioning, visual question answering, commonsense reasoning, and embodied planning [6, 7]. However, as Chen et al. [8] observe, most state-of-the-art VLMs still struggle with 3D spatial reasoning, and in particular with quantitative estimation of distances and sizes. This limitation is not necessarily architectural: Chen et al. argue it reflects an absence of metric spatial knowledge in standard training data rather than a fundamental incapacity of the model class. Whether pre-trained VLMs already possess some degree of latent metric spatial understanding that can be activated through appropriate prompting, without any additional training, remains an open and practically important question.

This question matters particularly for the design of autonomous agents that must navigate physical environments. Large Language Models (LLMs) have increasingly become the reasoning engine of choice in multi-agent systems, enabling agents to plan, communicate, act and solve structured tasks through natural language interfaces [9, 10, 11]. An agent equipped with a VLM as its perceptual module could in principle integrate depth estimates with semantic scene understanding and linguistic reasoning within a unified system, operating from a standard camera without specialised sensing hardware. This prospect has motivated recent work on spatially aware VLMs for robotics [8] and on benchmarks for spatial reasoning in embodied contexts [12]. However, existing evaluations of VLM spatial capabilities typically address qualitative or ordinal spatial relationships, or require fine-tuning on spatial data. The capacity of unmodified, pre-trained VLMs to produce reliable metric depth estimates from single indoor images in a zero-shot setting has not been systematically characterised.

A further gap in the existing literature concerns the role of contextual information in eliciting spatial reasoning from VLMs. Standard benchmarks for VLMs, such as MMBench [6] and MMMU [7], assess performance on logical reasoning, domain knowledge, and perception of object attributes, but do not probe metric spatial understanding as a distinct capability. A complementary research direction investigates what knowledge is already latent in pre-trained models, showing through layer-wise analysis that useful predictive knowledge is often already present in intermediate representations and that the final output depends critically on how the model is prompted [13, 14, 15]. This suggests that metric spatial understanding may similarly be encoded in pre-trained VLMs, with the query structure determining whether it surfaces in the response. The question of what types of contextual information, ranging from semantic priors about object dimensions to explicit metric anchors provided by the

environment, are necessary and sufficient to activate metric depth reasoning in pre-trained VLMs has not been addressed.

We address these gaps with a systematic empirical study whose main contributions are the following. We evaluate five pre-trained VLMs from two families, Qwen2.5-VL [16] and LLaVA [17], in a zero-shot inference-only setting on a benchmark of 7,223 annotated object instances derived from NYU Depth V2. We design five query conditions of increasing contextual richness, from a bare visual query to one augmented with semantic object priors, relational scene structure, and explicit metric anchors, and evaluate each condition under both an independent and a cumulative protocol. Our results show that explicit metric anchors are the dominant driver of accuracy, improving Hit@0.5m by a factor of nearly three for mid-size models and reducing RMSE below 1m for the largest model evaluated, without any form of fine-tuning. The near and far anchors are not interchangeable, activating different aspects of metric reasoning that are captured by complementary metrics. We further introduce a precise operational definition of referential hallucination in this setting, a formally valid response that names the wrong object, and show that this failure mode is dissociable from metric estimation error and varies independently across query conditions and model families. In particular, we find that relational context produces qualitatively opposite hallucination effects across families: it inflates hallucination rates in smaller Qwen2.5 models while reducing them in LLaVA models, in both cases generating a confound that is invisible to evaluations that do not track referential errors separately. Finally, we document a qualitative divergence between the Qwen2.5 and LLaVA families that is not reducible to parameter count, with implications for the selection of VLMs as perceptual modules in agent systems.

The paper is organised as follows. Section 2 revise the state of the art. Section 3 describes the dataset, query conditions, and evaluation framework. Section 4 presents results and analysis. Section 5 summarises findings and their implications for VLM-based agent perception.

2. Related works

A growing line of research positions VLMs as the perceptual and reasoning backbone of autonomous agents, moving beyond traditional pipelines of specialised perception and decision-making modules [18, 19] toward unified models that jointly process visual and linguistic inputs. In [20], the authors demonstrated that a VLM fine-tuned on robotic trajectory data can transfer web-scale visual knowledge to low-level robot control, establishing that general-purpose vision-language representations are compatible with action generation. In [8], the authors explicitly targeted spatial perception for robotics, showing that a VLM augmented with metric spatial training data can answer quantitative distance questions and support chain-of-thought spatial reasoning in embodied settings. These works highlight that the perceptual quality of a VLM, and in particular its ability to estimate metric distances, is a prerequisite for reliable agent behaviour in physical environments. Our work addresses a complementary question: rather than improving this capacity through training, we characterise to what extent it already exists in pre-trained models and under what linguistic conditions it can be reliably activated.

Language-guided depth estimation. In [21], the authors showed that CLIP patch embeddings respond to semantic distance tokens and can be projected to depth bins to produce coarse zero-shot depth maps without any training, suggesting that vision-language representations may encode some spatial information transferable to depth tasks. The approach operates at the pixel-patch level and produces dense maps rather than object-level estimates, and does not investigate whether varying the linguistic content of the query affects estimation quality. In [22], the authors demonstrated that metric depth estimation can be substantially improved by prompting a general-purpose depth foundation model with sparse metric cues, achieving high-resolution accurate results without full fine-tuning; crucially, this work shows that the way a depth model is queried directly impacts estimation quality, a finding that motivates our investigation of analogous prompting effects in VLMs. In [23], the authors proposed supervised fine-tuning of a generative VLM for pixel-level metric depth, reaching performance comparable to dedicated monocular depth models while retaining a linguistic interface. This training-

based approach establishes an upper bound on what VLMs can achieve on the task, but does not address what pre-trained models produce in the absence of spatial training data.

Evaluation of spatial perception in VLMs. In [24], the authors evaluated depth and height perception across multiple VLMs and found that models tend to rely on object-size heuristics rather than geometric inference, with both open and closed-source models exhibiting significant failures on apparently simple spatial tasks. In [12], the authors introduced a benchmark spanning metric depth, proximity, and size reasoning, confirming that unmodified VLMs fall substantially short of fine-tuned systems on metric tasks. In [25], the authors evaluated 3D spatial understanding in multimodal LLMs across a broad range of tasks including depth, distance, and size estimation, finding that models struggle with metric precision when no structured spatial context is provided in the query. In [26], the authors proposed a VLM augmented with explicit spatial reasoning for precise object localisation and metric distance estimation in robotic settings, showing that unmodified VLMs lack the structured representations needed for reliable metric inference. In [27], the authors surveyed the landscape of spatial intelligence in VLMs, identifying metric depth estimation from natural language queries as a largely unexplored capability. None of these works vary the contextual information provided in the query, leaving open the question of which linguistic conditions are necessary to elicit accurate metric responses from pre-trained models.

3. Methodology

3.1. Task Formulation

We frame metric depth estimation as a visual question answering task administered to pre-trained Vision-Language Models (VLMs) in a zero-shot, inference-only setting. No fine-tuning or adaptation is performed: we assess exclusively what spatial knowledge is already encoded in the model weights and how effectively it can be activated through natural language context.

Formally, let $\mathcal{I}$ denote an RGB image of an indoor scene and let $o \in \mathcal{O}$ be a named object instance visible in $\mathcal{I}$, where $\mathcal{O}$ is the set of annotated objects present in the scene. The task requires a VLM $\mathcal{M}$ to produce a scalar distance estimate $\hat{d} \in \mathbb{R}_{>0}$ (in meters) from the camera viewpoint to o , given the image and a natural language query q :

$$\hat{d} = \mathcal{M}(\mathcal{I}, q) \quad (1)$$

The ground-truth depth d^* for each object instance is derived from the per-pixel depth map provided by NYU Depth V2 [4], averaged over the pixels belonging to that instance (avg_depth). Output is constrained to a structured JSON format `{"object": <name>, "distance": <float>}`, which allows unambiguous parsing and enables the detection of referential errors (hallucinations), defined as responses in which the returned object name does not match the queried target o .

3.2. Dataset

We build our evaluation benchmark on NYU Depth V2 [4], a standard indoor RGB-D dataset acquired with a Microsoft Kinect sensor at 640×480 resolution across 464 indoor scenes. The original dataset provides per-pixel semantic labels, instance segmentation maps, and metric depth values, making it suitable for object-level depth evaluation.

Processing pipeline. The raw dataset was processed through a four-stage pipeline to produce a structured, evaluation-ready benchmark. In the first stage, pixel-level extraction, semantically uninformative categories were removed from consideration: ceiling (label 4), floor (label 11), unknown regions (label 20), and walls (label 21) were excluded, as these structural elements do not participate in object-level spatial reasoning. Images were also corrected for the -90° rotation present in the original acquisition format. Each remaining pixel was recorded as a tuple $\langle x, y, \text{depth}, \text{label}, \text{instance} \rangle$.

In the second stage, pixels were grouped by semantic category and instance identifier, and per-object statistics were computed: mean depth (avg_depth), median depth, depth variance, and pixel count. These statistics characterise each object’s spatial properties and visual extent. The third stage disambiguated multiple instances of the same semantic category (e.g., two chairs in the same scene) by assigning progressive indices, yielding descriptions of the form “*2nd chair from right*”, which were used as the object identifiers in the queries.

The fourth and final stage applied a hybrid selection strategy to produce a balanced, diverse set of objects per image. For scenes with five or fewer objects, all instances were retained. For larger scenes, at most one instance per semantic category was kept (the largest by pixel count), then a pixel-size filter excluded objects smaller than 3,000 pixels or larger than 20,000 pixels to remove ambiguous or spatially dominant instances. Up to five objects per image were sampled from the resulting pool, with the constraint that the nearest and farthest objects in each scene were always included, as their distances serve as metric anchors in the experimental conditions described below.

Dataset composition. The final benchmark comprises **7,223 annotated object instances** across a representative subset of NYU Depth V2 images. Depth values range from 0.7m to 9.9m, with a mean of $\mu = 2.83\text{m}$, a median of 2.54m, and a standard deviation of $\sigma = 1.41\text{m}$. The distribution is approximately Gaussian and is representative of typical indoor interaction distances, with the bulk of objects lying in the 1–4m range. Each record encodes: image identifier, semantic label, natural-language object description, instance index, pixel count, average depth, median depth, and depth variance.

3.3. Query Design and Spatial Reasoning Conditions

A central methodological contribution of this work is the design of a set of query conditions that systematically vary the type and amount of contextual information provided to the model. Each condition is intended to probe a distinct aspect of the model’s spatial understanding, from pure visual perception to metric reasoning supported by explicit numerical grounding. The conditions are defined as follows.

(Q1) Visual Reasoning (VR). The base condition. The model receives only the image and a direct request to estimate the distance of the named object in metres, with no additional context. This probes the model’s intrinsic metric spatial knowledge as acquired during pre-training, in the absence of any linguistic or numerical support.

(Q2) Semantic Prior (SP). The query is augmented with the phrase “*Considering its typical real-world dimensions.*” This activates the model’s pre-trained knowledge about the physical scale of object categories, providing a semantic anchor that connects visual appearance to metric scale without supplying any scene-specific information.

(Q3) Relational Context (RC). The model is additionally provided with the full list of objects present in the scene and an explicit statement of which object is nearest and which is farthest from the camera. This condition probes whether the model can exploit relative spatial structure, a form of scene-level understanding, to constrain its estimate of the target object’s distance.

(Q4) Near Metric Anchor (NMA). Instead of relational context, the model is given the exact metric distance of the nearest object in the scene. The model must use this known value as a lower bound and reason about how far the target object lies relative to it. This probes metric interpolation from a near reference point.

(Q5) Far Metric Anchor (FMA). Analogously, the model is given the exact distance of the farthest object. The model must reason about the target’s position relative to an upper bound. This condition is structurally symmetric to NMA but cognitively distinct: it requires constraining a near estimate from above rather than extrapolating upward from a lower bound.

Experimental design. Let $\mathcal{C} = \{Q1, Q2, Q3, Q4, Q5\}$ denote the set of contextual conditions defined above, ordered by increasing specificity of the information provided. Each query administered to a model is a composition $Q1 \oplus S$, where $S \subseteq \{Q2, Q3, Q4, Q5\}$ is a subset of supplementary conditions and $\oplus$ denotes their concatenation into a single natural language prompt. $Q1$ is always present as the base query, since without a target object and a distance request the task is undefined.

Two experimental protocols are applied over this space. The *independent* protocol instantiates each supplementary condition in isolation:

$$\mathcal{E}_{\text{ind}} = \{Q1\} \cup \{Q1 \oplus Qk \mid k \in \{2, 3, 4, 5\}\} \quad (2)$$

yielding five conditions ($Q1, Q1+Q2, Q1+Q3, Q1+Q4, Q1+Q5$). This isolates the marginal contribution of each information type, controlling for interactions among conditions.

The *cumulative* protocol accumulates context in order of increasing specificity, from semantic to relational to metric:

$$\mathcal{E}_{\text{cum}} = \left\{ \bigoplus_{j=1}^k Qj \mid k \in \{1, \dots, 5\} \right\} \quad (3)$$

This protocol measures whether contextual information types interact positively or negatively when combined, and whether the ordering from coarse semantic priors to fine metric anchors constitutes an effective grounding strategy.

Together, $\mathcal{E}_{\text{ind}}$ and $\mathcal{E}_{\text{cum}}$ yield nine distinct experimental conditions, each evaluated across all five models. All queries enforce a structured JSON output format and prohibit free-text explanation, enabling unambiguous extraction of the numeric estimate $\hat{d}$ and the object name for hallucination detection. An example of a fully cumulative prompt, corresponding to the $Q1+Q2+Q3+Q4+Q5$ condition, is shown in Figure 1.

3.4. Hallucination Detection

We define a *hallucination* as a response that is formally valid, i.e., the output is correctly structured JSON containing a distance value, but in which the returned object name does not correspond to the queried target. This definition is operationally precise and distinguishes a referential grounding failure from a metric estimation error: a hallucinated response may still contain a metrically accurate distance estimate for some object in the scene, even if that object is not the one requested. We exploit this distinction throughout our analysis by reporting metrics both inclusive and exclusive of hallucinated responses, and by examining how the delta between the two conditions varies across query configurations.

3.5. Evaluation Metrics

Let d_i^* denote the ground-truth depth of the i -th object instance and $\hat{d}_i$ the model’s estimate, over a set of N non-hallucinated responses. We evaluate performance using four complementary metrics, each capturing a distinct aspect of estimation quality:

- **Root Mean Squared Error (RMSE).** The primary metric for overall estimation accuracy, consistent with the standard evaluation protocol of NYU Depth V2:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{d}_i - d_i^*)^2} \quad (4)$$

RMSE is expressed in metres and penalises large deviations quadratically, making it sensitive to outlier estimates. This property is particularly relevant in our setting, where hallucinated or poorly grounded responses may produce estimates that are far from the true depth. The gap between RMSE and Mean Absolute Error provides a complementary signal on the prevalence of such large errors.

- **Mean Absolute Error (MAE).** The primary metric for overall estimation accuracy:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{d}_i - d_i^*| \quad (5)$$

MAE is expressed in metres and is interpreted relative to the dataset distribution ($\mu = 2.83\text{m}$, $\sigma = 1.41\text{m}$): an MAE of 1m corresponds to $\approx 0.71\sigma$, making errors directly comparable to the natural depth variability of the benchmark.

- **Hit@ τ (tolerance accuracy).** The proportion of estimates whose absolute error does not exceed a threshold τ (expressed in metres):

$$\text{Hit@}\tau = \frac{1}{N} \sum_{i=1}^N \mathbf{1} [|\hat{d}_i - d_i^*| \leq \tau] \quad (6)$$

where $\mathbf{1}[\cdot]$ is the indicator function, equal to 1 when the condition holds and 0 otherwise. Intuitively, Hit@ τ answers the question: *what fraction of estimates land within τ metres of the true depth?* We report two instantiations: Hit@0.5m, which requires the estimate to be within half a metre of ground truth, and Hit@1m, which allows up to one metre of error. Given the dataset mean of $\mu = 2.83\text{m}$, a tolerance of 0.5m corresponds to approximately 18% of the average depth, while 1m corresponds to approximately 35%, making Hit@1m a relatively permissive threshold and Hit@0.5m a stringent one. These two thresholds together characterise the sharpness of the model’s accuracy distribution around the ground truth.

- **Range 70–130% (relative accuracy).** The proportion of estimates whose ratio to the ground-truth depth falls within the interval $[0.70, 1.30]$:

$$\text{R}_{70-130} = \frac{1}{N} \sum_{i=1}^N \mathbf{1} \left[0.70 \leq \frac{\hat{d}_i}{d_i^*} \leq 1.30 \right] \quad (7)$$

This metric captures whether the model’s estimate is proportionally accurate relative to the true distance, regardless of absolute scale. Concretely, an object at 2m is correctly estimated if $\hat{d}_i \in [1.4\text{ m}, 2.6\text{ m}]$; an object at 5m requires $\hat{d}_i \in [3.5\text{ m}, 6.5\text{ m}]$. Unlike Hit@ τ , which applies the same absolute tolerance regardless of depth, Range 70–130% imposes a tolerance that scales with the true distance, penalising proportionally large errors uniformly across the depth range. This makes the two metrics complementary: Hit@ τ is more informative at short distances (where absolute precision matters), while Range 70–130% is more meaningful at longer distances where a fixed absolute threshold would be overly strict.

- **Hallucination rate.** The proportion of model responses that are formally valid but reference the wrong object:

$$\text{Hall} = \frac{|\{i : \text{object}_i \neq o_i\}|}{N_{\text{total}}} \quad (8)$$

where o_i is the queried target and N_{total} includes all valid JSON responses. All other metrics are reported both over the full response set and over the non-hallucinated subset, to disentangle metric estimation accuracy from referential grounding failures.

4. Results

We present results for five pre-trained VLMs evaluated across ten experimental conditions under both the independent ($\mathcal{E}_{\text{ind}}$) and cumulative ($\mathcal{E}_{\text{cum}}$) protocols. The independent protocol isolates the marginal

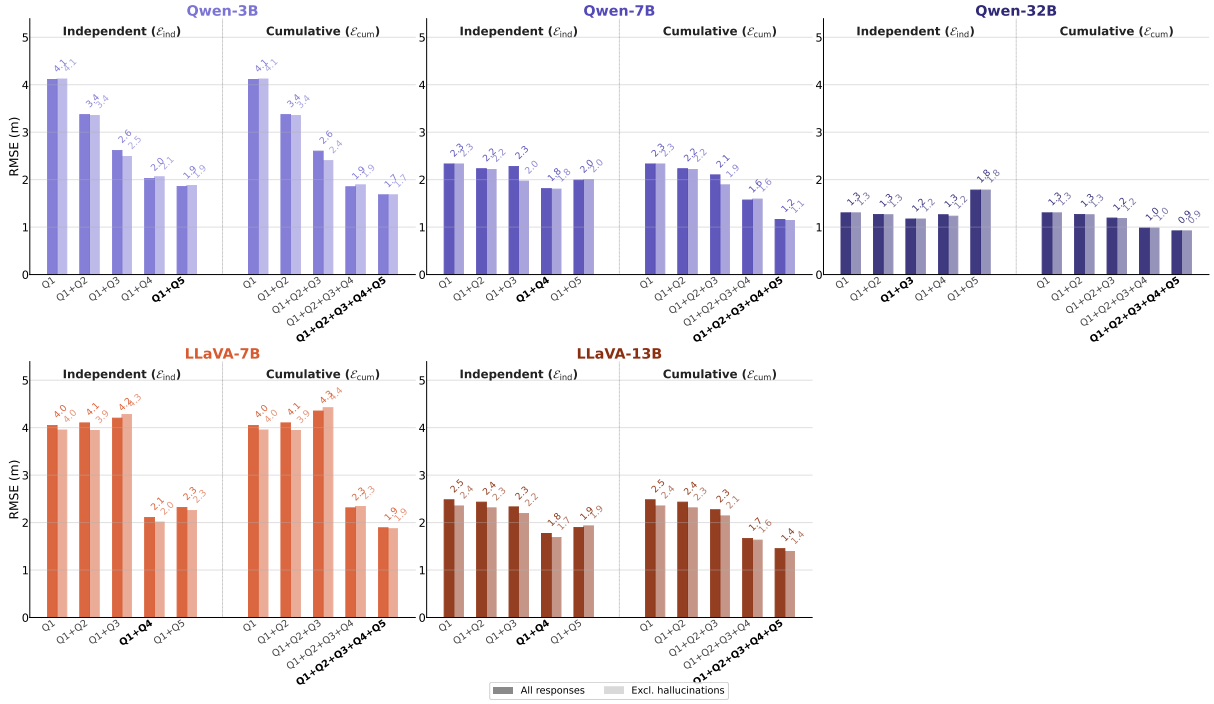


Figure 2: Root Mean Squared Error (RMSE, in metres) across all five models, under the independent ($\mathcal{E}_{\text{ind}}$, left panel per model) and cumulative ($\mathcal{E}_{\text{cum}}$, right panel per model) protocols. Solid bars report results over all valid responses; hatched bars report results excluding hallucinated responses. Lower is better. The best-performing condition per model and protocol is highlighted in bold on the x-axis.

contribution of each contextual condition added to the base query Q1, while the cumulative protocol measures their joint effect when combined in order of increasing specificity. Results are reported across six metrics, RMSE, MAE, Hit@0.5m, Hit@1m, R70-130, and hallucination rate, and for each condition under two reporting modes: over all valid responses and over the non-hallucinated subset exclusively. This dual reporting allows metric estimation accuracy to be cleanly disentangled from referential grounding failures, which we show to be a distinct and independently varying failure mode.

The results are structured around four emergent phenomena. First, explicit metric anchors emerge as the dominant driver of accuracy gains across all models and metrics. Second, relational context produces a systematic trade-off between accuracy and hallucination that depends critically on model capacity. Third, the two model families, Qwen2.5 and LLaVA, exhibit qualitatively different response profiles that are not reducible to parameter count alone. Fourth, the cumulative protocol reveals interaction effects between context types that are invisible to the independent protocol, and that determine whether accumulating context is beneficial or harmful.

4.1. Semantic priors provide a weak but consistent grounding signal

The simplest form of contextual augmentation, the semantic prior Q1+Q2, augments the base query with a reminder of the object’s typical real-world dimensions. This condition produces small but consistent improvements over Q1 across most models: for Qwen2.5-3B, RMSE drops from 4.11m to 3.37m and MAE from 2.66m to 2.30m; for Qwen2.5-7B the gains are more modest (RMSE 2.33m to 2.23m); for LLaVA-13B, Hit@0.5m improves from 8.2% to 9.5% and R70-130 from 11.2% to 12.8% (Figures 2, 3, 4, 6). Qwen2.5-32B is the exception, showing negligible variation (RMSE 1.30m to 1.27m), consistent with a model that already operates near its capacity ceiling at Q1.

These gains confirm that activating pre-trained knowledge about object scale through language provides a weak but non-trivial grounding signal. However, the improvements are uniformly small relative to those produced by metric anchors, and the hallucination rate is not meaningfully affected in either direction (Figure 7), indicating that semantic priors influence estimation quality without altering

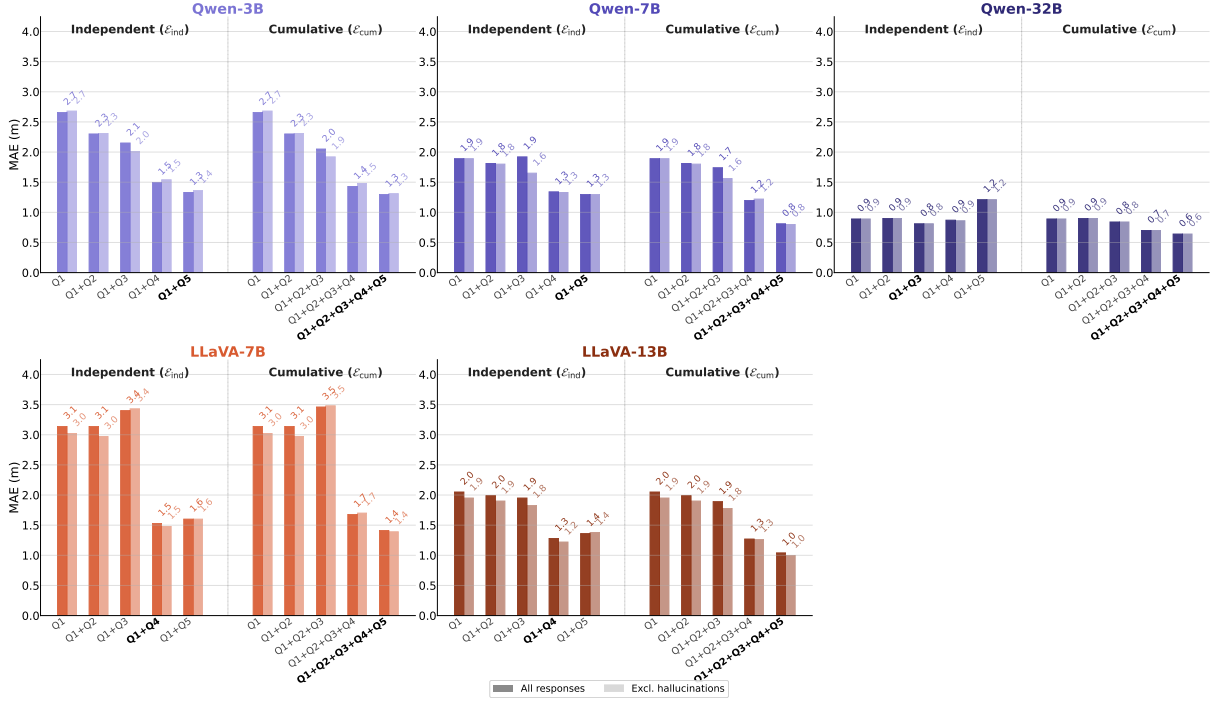


Figure 3: Mean Absolute Error (MAE, in metres) across all five models, under the independent ($\mathcal{E}_{ind}$) and cumulative ($\mathcal{E}_{cum}$) protocols. Solid bars report results over all valid responses; hatched bars report results excluding hallucinated responses. Lower is better. MAE is complementary to RMSE (Figure 2): the gap between the two metrics signals the presence of large outlier estimates.

referential grounding. The semantic prior therefore occupies a distinct role in the contextual grounding hierarchy: useful as a soft constraint on plausible depth ranges, but insufficient as a substitute for scene-specific metric information.

4.2. Metric anchors are the primary driver of accuracy

The most consistent finding across all models and metrics is that conditions providing an explicit metric anchor, Q1+Q4 (near anchor) and Q1+Q5 (far anchor) in the independent protocol, and their cumulative counterparts, produce substantially larger gains than any other contextual manipulation. This holds regardless of model family or size.

For Qwen2.5-7B, the independent evaluation of Q1+Q4 reduces RMSE from 2.33m (Q1) to 1.81m and MAE from 1.89m to 1.34m, while Q1+Q5 achieves the best Hit@0.5m (31.1%) and R70-130 (43.2%) among all independent conditions (Figures 2, 3, 4, 6). The effect is even more pronounced in the cumulative protocol, where the full Q1+Q2+Q3+Q4+Q5 condition drives RMSE down to 1.16m and Hit@0.5m up to 45.6%, representing a fourfold improvement over Q1 on the latter metric.

For Qwen2.5-32B, which already achieves a strong baseline (RMSE 1.30m, Hit@0.5m 38.7% at Q1), metric anchors provide further gains in the independent protocol: Q1+Q4 reaches Hit@0.5m 44.3%, while the cumulative protocol yields the global best performance with RMSE 0.92m and Hit@0.5m 54.6% at the full Q1+Q2+Q3+Q4+Q5 condition, confirming that metric grounding is beneficial even for the strongest model in our evaluation.

The two anchor conditions are not symmetric in their effects, and their relative performance depends on the metric considered. For Qwen2.5-7B, Q1+Q4 achieves lower RMSE than Q1+Q5 (1.81m versus 1.99m), while Q1+Q5 yields slightly lower MAE (1.30m versus 1.34m), higher Hit@0.5m (31.1% versus 26.7%), and higher R70-130 (43.2% versus 34.6%). This suggests that the two anchors activate different aspects of metric reasoning: the near anchor constrains the scale of the estimate more tightly, reducing large errors, while the far anchor improves the proportion of estimates falling within a strict absolute tolerance. For Qwen2.5-32B, Q1+Q4 dominates Q1+Q5 on all metrics in the independent protocol (RMSE

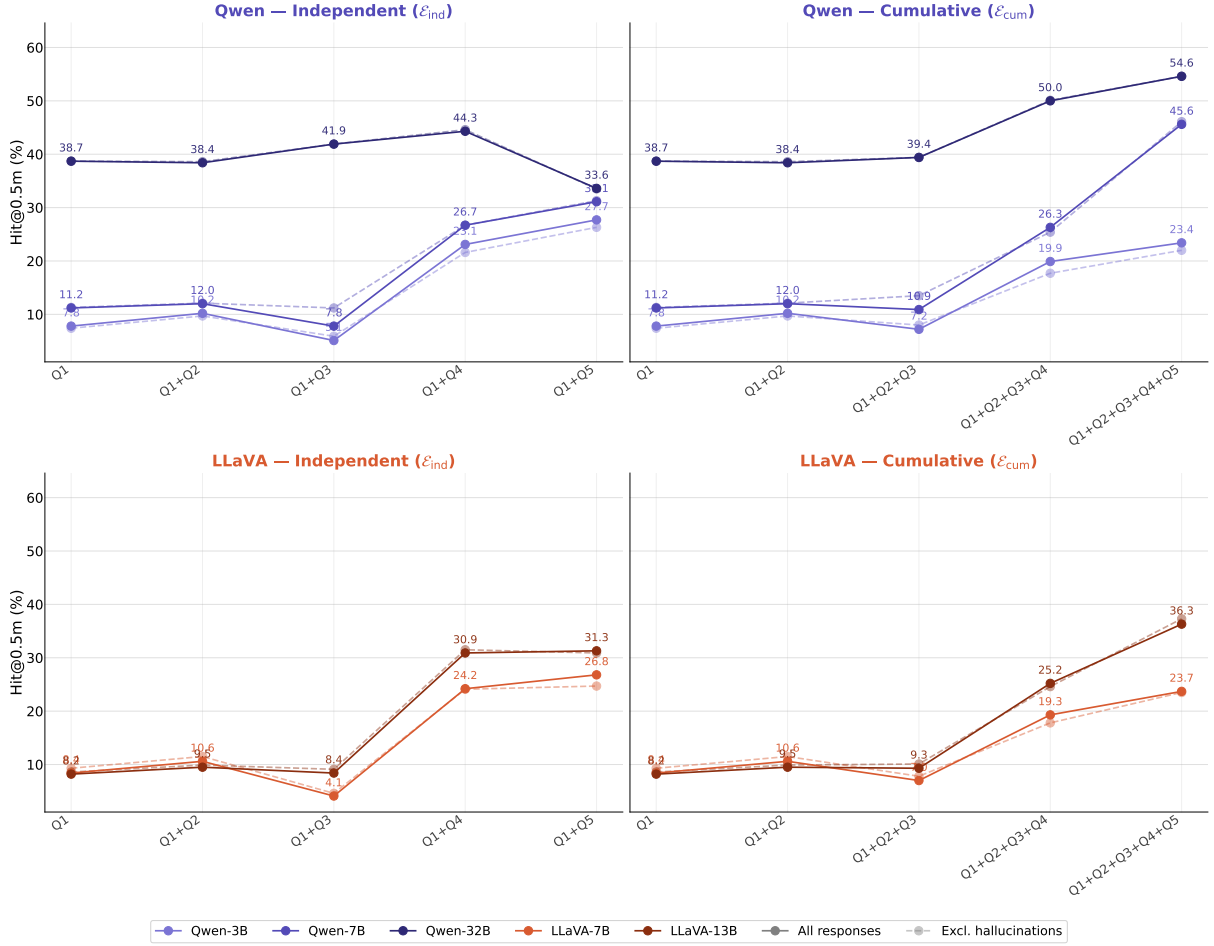


Figure 4: Tolerance accuracy Hit@0.5m (%) as a function of query condition, shown separately for the Qwen2.5 family (top row) and the LLaVA family (bottom row), under the independent ($\mathcal{E}_{ind}$, left column) and cumulative ($\mathcal{E}_{cum}$, right column) protocols. Solid lines report results over all valid responses; dashed lines report results excluding hallucinated responses. Higher is better. All four panels share the same y axis. Hit@0.5m requires the estimate to fall within 0.5m of the ground-truth depth, corresponding to approximately 18% of the dataset mean depth of 2.83m.

1.26m versus 1.78m, Hit@0.5m 44.3% versus 33.6%), suggesting that for a model already operating at high accuracy the near anchor is a more effective constraint. For smaller or weaker models the differences between the two anchor conditions are less consistent.

The gap between all-responses and excl.-hallucinations metrics is notably small under anchor conditions (Figures 2, 3), indicating that metric anchors simultaneously reduce estimation error, and in some cases also reduce referential confusion, although this effect is not uniform across models.

4.3. Relational context improves accuracy but induces hallucination

The independent evaluation of Q1+Q3 presents a more nuanced picture that differs substantially across model families. On the accuracy side, Q1+Q3 delivers modest but consistent RMSE reductions relative to Q1 for the Qwen2.5 family: Qwen2.5-3B improves from 4.11m to 2.62m (all responses), and to 2.49m excluding hallucinations; Qwen2.5-7B from 2.33m to 2.28m (Figure 2). For the LLaVA family the picture is split: LLaVA-13B shows a small improvement (RMSE 2.48m to 2.33m in Q1+Q3, further improving to 2.27m in the cumulative setting), while LLaVA-7B is the only model for which relational context is unambiguously harmful on error metrics, with RMSE increasing from 4.05m to 4.20m (independent Q1+Q3 condition), with a larger degradation under the cumulative setting (4.35m).

The effect of Q1+Q3 on hallucination rate is equally heterogeneous across families (Figure 7). For the

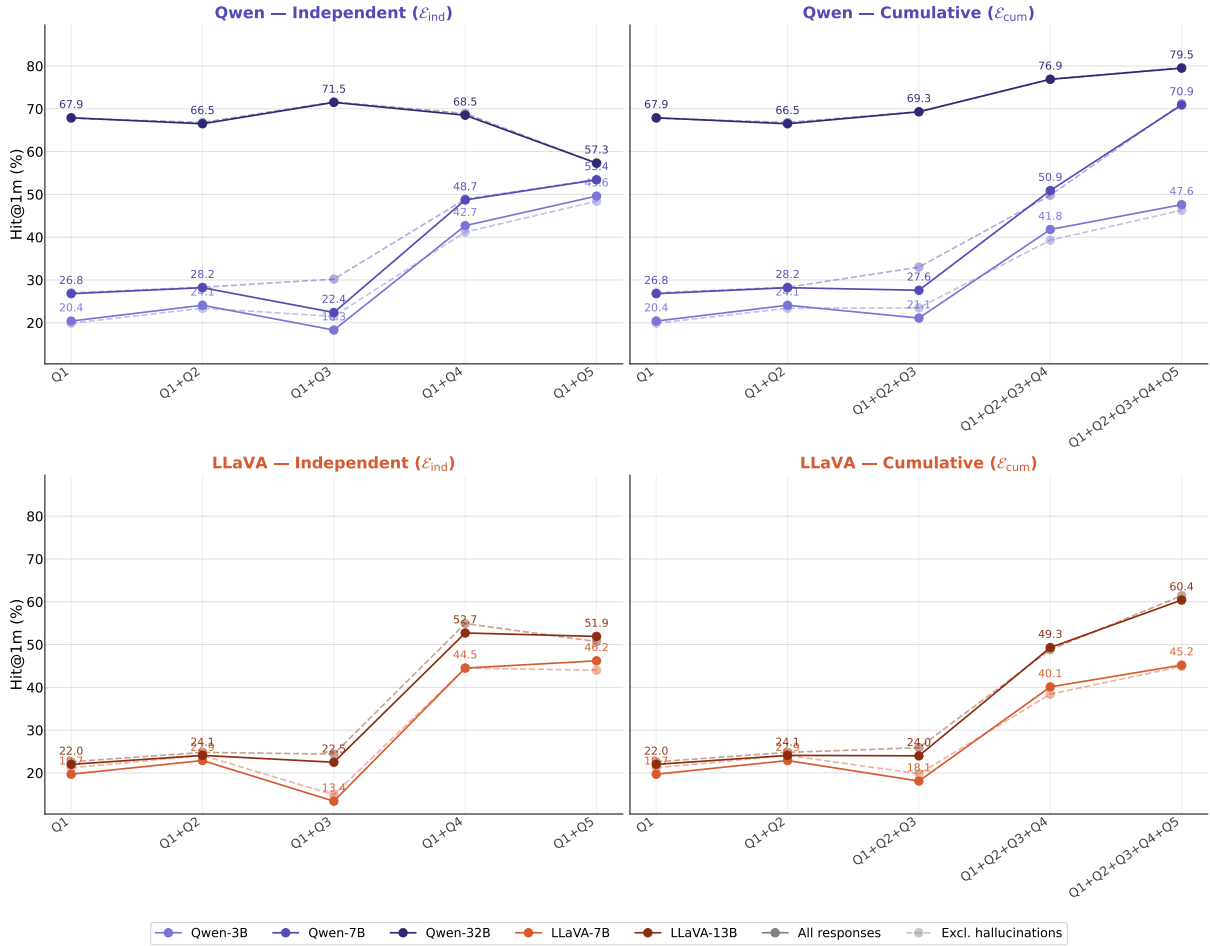


Figure 5: Tolerance accuracy Hit@1m (%) as a function of query condition, shown separately for the Qwen2.5 family (top row) and the LLaVA family (bottom row), under the independent ($\mathcal{E}_{ind}$, left column) and cumulative ($\mathcal{E}_{cum}$, right column) protocols. Solid lines report results over all valid responses; dashed lines report results excluding hallucinated responses. Higher is better. All four panels share the same y axis. Hit@1m allows up to 1m of absolute error, corresponding to approximately 35% of the dataset mean depth, and constitutes a more permissive threshold than Hit@0.5m.

smaller Qwen2.5 models, Q1+Q3 induces a sharp increase: Qwen2.5-7B rises from 3.4% to 37.1% and Qwen2.5-3B from 4.2% to 25.9% in the independent protocol. This pattern reveals a fundamental trade-off for these models: enriching the prompt with a list of all scene objects and their ordinal roles appears to confuse the referential grounding mechanism, causing the model to respond about a different object than the one queried. The accuracy improvements are largely visible only after excluding hallucinated responses, suggesting that part of the apparent gain is driven by estimates referring to the wrong object.

For the LLaVA models the hallucination response to Q1+Q3 is reversed: LLaVA-7B drops from 27.5% to 17.6% and LLaVA-13B from 37.9% to 15.2%. This is a structurally different behaviour, the object list appears to help these models identify the queried target more reliably, possibly because their high baseline hallucination rate reflects a systematic difficulty with object identification from the image alone, which the explicit enumeration of scene objects partially resolves. However, this referential benefit does not translate into metric accuracy gains for LLaVA-7B, whose RMSE worsens under Q1+Q3 despite the lower hallucination rate.

Crucially, the dissociation between hallucination and metric accuracy is only detectable because we track the two failure modes independently. When results are filtered to exclude hallucinated responses, the RMSE improvement under Q1+Q3 is attenuated for Qwen2.5-3B and Qwen2.5-7B, and the gap between the two reporting modes widens relative to anchor conditions. This confirms that the accuracy signal in Q1+Q3 is contaminated by referential errors for the Qwen2.5 small models, and underscores the

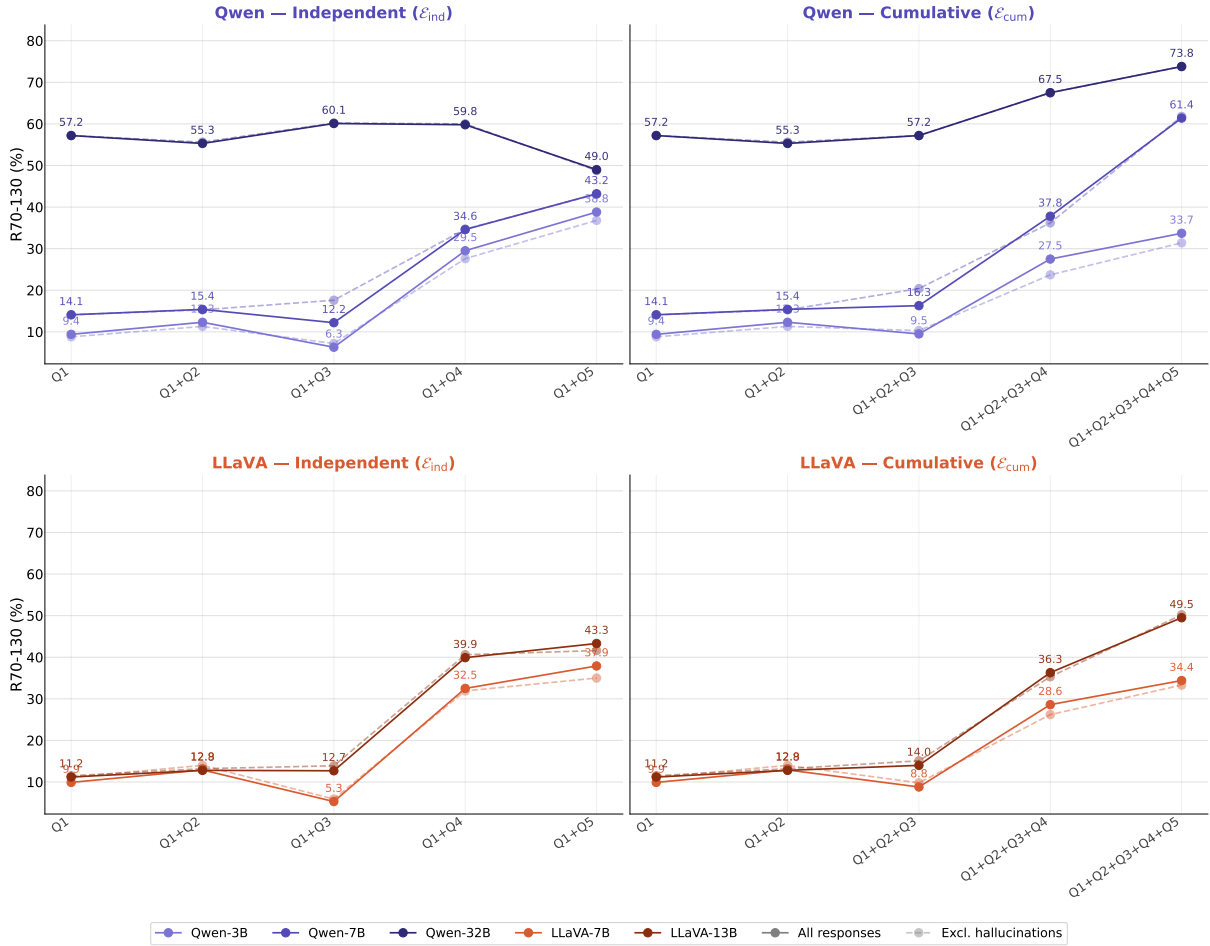


Figure 6: Relative accuracy R70-130 (%) by query condition for Qwen2.5 (top) and LLaVA (bottom), under independent ($\mathcal{E}_{ind}$, left) and cumulative ($\mathcal{E}_{cum}$, right) protocols. Solid lines: all responses; dashed: excluding hallucinations. Higher is better. R70-130 measures estimates within [0.70, 1.30] of ground truth.

importance of treating hallucination as a distinct failure mode rather than subsuming it into aggregate error metrics.

Qwen2.5-32B stands apart from all other models: it achieves the lowest hallucination rate under Q1+Q3 (0.2%, well below its Q1 baseline of 2.3%), and shows a clean accuracy improvement with RMSE dropping from 1.30m to 1.19m. This suggests that sufficient model capacity allows relational context to be exploited without sacrificing referential precision.

The contrasting behaviour of Qwen2.5-3B and Qwen2.5-7B under Q1+Q3 suggests a more fundamental asymmetry in the capabilities acquired during pre-training. These models appear to possess a degree of visual spatial understanding sufficient to produce metrically plausible estimates when the query is simple and unambiguous, as evidenced by their reasonable baseline performance at Q1. However, their ability to parse and correctly ground a linguistically richer prompt, one that introduces a list of competing object referents and assigns them relational roles — appears to lag substantially behind their visual competence. The hallucination spike induced by Q1+Q3 is, under this interpretation, not a failure of spatial perception but a failure of referential language understanding: the model loses track of which object it has been asked about when the prompt introduces multiple named entities in a relational structure. Qwen2.5-32B’s immunity to this effect suggests that the linguistic grounding capacity necessary to handle complex structured prompts emerges primarily as a function of scale, and is not yet present at the 3B and 7B parameter range in this model family.

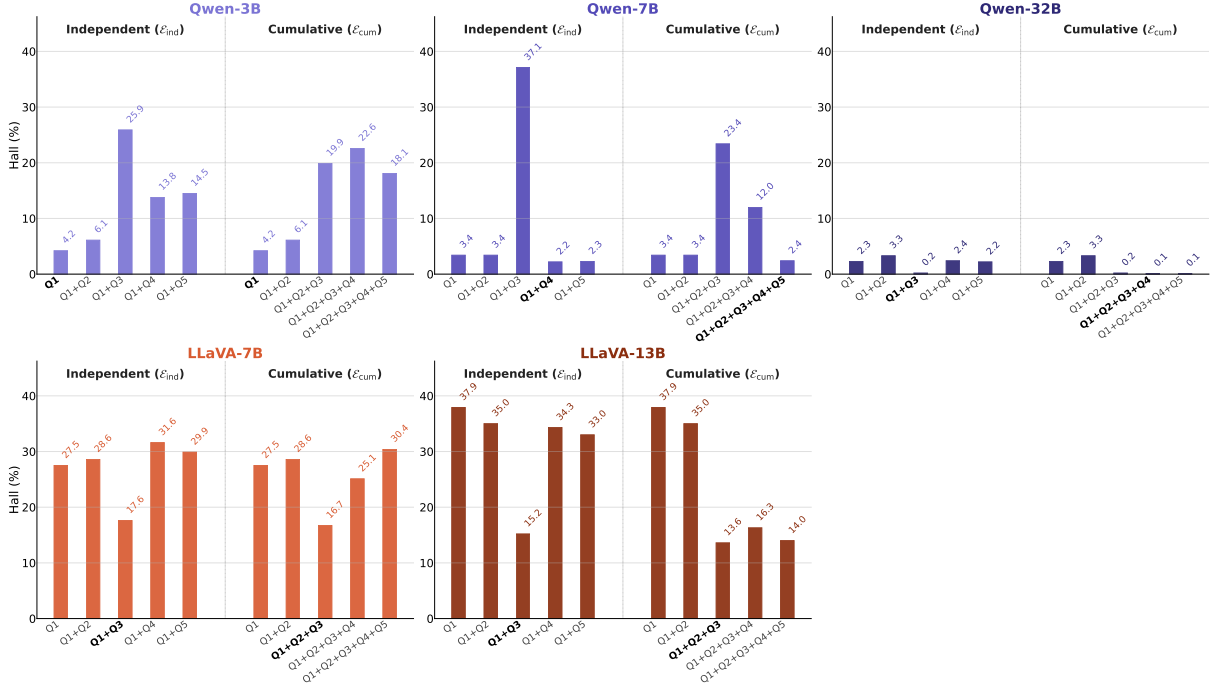


Figure 7: Hallucination rate (%) across all five models, under the independent ($\mathcal{E}_{ind}$) and cumulative ($\mathcal{E}_{cum}$) protocols. A hallucination is defined as a formally valid JSON response in which the returned object name does not match the queried target. Lower is better. The hallucination rate is reported over all valid responses and is used throughout to disentangle referential grounding failures from metric estimation errors.

4.4. A systematic divergence between model families

Across all conditions and metrics, Qwen2.5 and LLaVA models exhibit qualitatively different response profiles that go beyond differences attributable to parameter count alone.

The LLaVA models enter the evaluation with a structural hallucination problem. LLaVA-7B produces hallucinated responses in 27.5% of Q1 queries; LLaVA-13B in 37.9%, rates that are six to eleven times higher than those of comparably sized Qwen2.5 models at baseline (Figure 7). This floor-level difference persists across nearly all conditions and both protocols for LLaVA-7B, while for LLaVA-13B it drops substantially under Q1+Q3 (15.2%) and reaches its minimum at the full cumulative condition (14.0%). The accuracy metrics for LLaVA models reported over all responses are therefore substantially confounded by referential errors, and the excl.-hallucinations series should be taken as the primary reference for their metric estimation capability.

On the accuracy metrics, LLaVA-13B demonstrates meaningful spatial reasoning capability once hallucinations are excluded. Under Q1+Q5 (independent), it achieves RMSE 1.90m, Hit@0.5m 31.3%, and R70-130 43.3%. Under the full cumulative condition (Q1+Q2+Q3+Q4+Q5), these improve to RMSE 1.39m, Hit@0.5m 37.3%, and R70-130 50.2% excluding hallucinations. These values compare favourably with Qwen2.5-7B under the same condition on all-responses metrics (RMSE 1.99m, Hit@0.5m 31.1%, R70-130 43.2%), suggesting that the latent metric estimation capacity of LLaVA-13B is competitive once referential grounding failures are removed. LLaVA-7B, by contrast, shows persistently high error with RMSE remaining above 1.9m even under the best cumulative condition, and is the only model for which relational context worsens both RMSE and MAE simultaneously (Figure 2, 3).

Within the Qwen2.5 family, the scaling effect is strong and monotone. Qwen2.5-32B consistently occupies a separate performance tier: at Q1 it already achieves RMSE 1.30m and Hit@0.5m 38.7%, values that smaller Qwen2.5 models reach only under the most informative anchor conditions (Figure 4). Notably, Qwen2.5-32B is also the only model that benefits cleanly from relational context without incurring referential confusion, achieving its best independent RMSE under Q1+Q3 (1.17m) and its best Hit@1m there as well (71.5%). This suggests that a large fraction of the metric spatial knowledge required

for this task is acquired during pre-training as a function of model capacity, and that prompt-level grounding is most valuable as a complement to, rather than a substitute for, scale.

4.5. Cumulative context and interaction effects

The cumulative protocol ($\mathcal{E}_{\text{cum}}$) reveals how contextual information types interact when combined, beyond the marginal effects isolated by the independent protocol.

For Qwen2.5-7B, the cumulative protocol produces the best results of any condition in our evaluation: the full Q1+Q2+Q3+Q4+Q5 condition achieves RMSE 1.16m, Hit@0.5m 45.6%, and R70-130 61.4%, all substantially above the best independent condition (Figures 2, 4, 6). The hallucination rate at the full cumulative condition is 2.4%, comparable to the Q1 baseline of 3.4%, indicating that the hallucination spike induced by Q1+Q3 alone (37.1%) is fully mitigated when relational context is embedded within a richer prompt that also supplies metric anchors. This suggests a positive interaction: metric anchors serve not only as estimation supports but also as referential anchors that counteract the grounding confusion introduced by the object list.

For Qwen2.5-32B, the cumulative protocol yields the global best performance across all models and conditions, with RMSE reaching 0.92m and Hit@0.5m 54.6% at the full cumulative condition. Hit@1m follows the same trend, reaching 79.5% for Qwen2.5-32B and 60.4% for LLaVA-13B under the full cumulative condition, confirming that the accuracy gains extend across both tolerance thresholds (Figure 5) The improvement is not strictly monotone across all metrics, Hit@0.5m shows a slight dip at Q1+Q2 (38.4%) relative to Q1 (38.7%) before recovering strongly, but the overall trend is consistently positive, with the largest gains occurring when metric anchors are introduced at Q1+Q2+Q3+Q4. For Qwen2.5-3B, gains are present but more modest across the accumulation sequence, reflecting the limited capacity of the model to exploit dense contextual information.

The picture is less favourable for LLaVA. For LLaVA-7B, cumulative context does not yield consistent gains: RMSE at Q1+Q2+Q3 (4.35m) exceeds Q1 (4.05m), and hallucinations increase at the full condition (30.4% vs 27.5%), indicating destabilisation. LLaVA-13B performs better, reaching RMSE 1.39m, Hit@0.5m 36.3%, and R70-130 49.5%, with hallucinations reduced to 14.0%, its lowest level.

Overall, metric anchors drive performance, while semantic and relational context help only when paired with metric references. The progression Q2 $\rightarrow$ Q3 $\rightarrow$ Q4/Q5 is effective for capable models, but induces referential instability in weaker ones when relational context precedes metric anchoring.

5. Conclusion

This paper has presented a systematic empirical study of metric depth estimation in pre-trained Vision-Language Models under zero-shot, inference-only conditions. By evaluating five models from two families across ten experimental conditions designed to vary the type and amount of contextual information provided in the query, we have characterised four phenomena that jointly determine when and how metric spatial knowledge can be elicited from pre-trained VLMs through prompting alone.

First, explicit metric anchors, the provision of the known distance of the nearest or farthest object in the scene, are the dominant driver of accuracy gains, producing improvements that are consistent across all models, all metrics, and both experimental protocols. The near anchor (Q1+Q4) and the far anchor (Q1+Q5) are not interchangeable: for most models Q1+Q4 achieves lower RMSE while Q1+Q5 yields higher tolerance accuracy, suggesting that the two anchors activate different aspects of metric reasoning. In the cumulative protocol, metric anchors also suppress the referential confusion induced by relational context, acting simultaneously as estimation supports and grounding stabilisers: for Qwen2.5-7B, the hallucination spike of 37.1% induced by relational context alone collapses to 2.4% at the full cumulative condition.

Second, relational context, providing the list of scene objects and their ordinal depth roles without metric values, produces qualitatively different effects across model families. For Qwen2.5-3B and Qwen2.5-7B it induces a sharp hallucination spike (up to 37.1%) while producing apparent accuracy improvements that are largely artefactual, driven by metrically plausible estimates for the wrong object.

For the LLaVA family the effect is reversed: hallucination rates drop substantially under relational context (LLaVA-13B from 37.9% to 15.2%), yet this referential benefit does not translate into metric accuracy gains for LLaVA-7B, whose RMSE worsens. Qwen2.5-32B is the only model immune to both failure modes, exploiting relational context to improve accuracy without referential confusion. This dissociation is only detectable through independent hallucination tracking, and underscores the importance of treating referential grounding failures as a distinct failure mode rather than subsuming them into aggregate error metrics.

Third, the two model families exhibit qualitatively different response profiles that are not reducible to parameter count alone. LLaVA models carry a structural hallucination problem at baseline (27–38%) that persists across most conditions, while the Qwen2.5 family exhibits a strong and monotone scaling effect. Qwen2.5-32B occupies a distinct performance tier already at Q1 (RMSE 1.30m, Hit@0.5m 38.7%), values that smaller Qwen2.5 models approach only under the most informative anchor conditions. Once hallucinations are excluded, LLaVA-13B demonstrates competitive metric estimation capability under anchor conditions, suggesting that its spatial knowledge is present but systematically misdirected by referential failures.

Fourth, the cumulative protocol reveals non-trivial interaction effects between context types. For capable models, the full accumulation of semantic, relational, and metric context produces results that substantially exceed the best independent condition: Qwen2.5-32B reaches RMSE 0.92m and Hit@0.5m 54.6% at the full cumulative condition, its global best. For LLaVA-7B, however, the full cumulative condition produces substantially lower RMSE (4.05m to 1.89m) but a higher hallucination rate (27.5% to 30.4%), indicating a trade-off between metric accuracy and referential grounding.

Implications for VLM-based agent perception. These findings have direct consequences for the design of autonomous agents that use VLMs as perceptual modules. The results establish that pre-trained VLMs do possess latent metric spatial knowledge that can be activated through prompting, but that the reliability of this activation depends critically on model capacity and prompt structure. For agents operating in environments where at least one metric reference distance is available, a common condition in robotics, where fiducial markers, known object sizes, or range sensors can supply a single anchor — our results suggest that even moderately sized VLMs can produce useful depth estimates without any spatial fine-tuning. Conversely, the hallucination patterns we document warn against naive deployment of richer contextual prompts with smaller models: introducing scene object lists without metric grounding can silently degrade referential accuracy while appearing to improve metric accuracy in aggregate evaluations that do not track hallucination separately. The asymmetry between the near and far anchor further suggests that the choice of which metric reference to supply to the agent is not arbitrary and should be informed by the target depth range of the objects of interest.

Limitations and future work. Several limitations of the present study point to directions for future investigation. Our evaluation is restricted to a zero-shot, inference-only setting with two model families; it remains to be determined whether the phenomena we observe generalise to other architectures, including closed-source models and models specifically trained for spatial tasks. The benchmark is derived exclusively from indoor scenes; outdoor environments, with their larger depth ranges and different object statistics, may elicit qualitatively different response profiles. The query conditions we design are hand-crafted and evaluated at fixed levels of contextual richness; future work could explore adaptive prompting strategies that select the appropriate context level as a function of the model and scene. Finally, the interaction between hallucination and metric accuracy we document raises the question of whether targeted fine-tuning on referential grounding, distinct from spatial accuracy training, could decouple the two failure modes and unlock the latent spatial capability of models such as LLaVA-13B more reliably.

Acknowledgments

This research benefits from the HPC (High Performance Computing) facility of the University of Parma, Italy.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude-AI in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, *IEEE transactions on pattern analysis and machine intelligence* 44 (2020) 1623–1637.
- [2] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12179–12188.
- [3] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, M. Müller, Zoedepth: Zero-shot transfer by combining relative and metric depth, *arXiv preprint arXiv:2302.12288* (2023).
- [4] P. K. Nathan Silberman, Derek Hoiem, R. Fergus, Indoor segmentation and support inference from rgb-d images, in: *ECCV*, 2012.
- [5] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, *The international journal of robotics research* 32 (2013) 1231–1237.
- [6] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al., Mmbench: Is your multi-modal model an all-around player?, in: *European conference on computer vision*, Springer, 2024, pp. 216–233.
- [7] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al., Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 9556–9567.
- [8] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, F. Xia, Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14455–14465.
- [9] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al., The rise and potential of large language model based agents: A survey, *Science China Information Sciences* 68 (2025) 121101.
- [10] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: A survey of progress and challenges., in: *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, IJCAI; Cornell arxiv, 2024.
- [11] G. Trimigno, et al., LLMs in staging: An orchestrated LLM workflow for structured augmentation with fact scoring, *Future Internet* 17 (2025) 535. doi:10.3390/fi17120535.
- [12] P. Xu, S. Wang, Y. Zhu, J. Li, Y. Zhang, Spatialbench: Benchmarking multimodal large language models for spatial cognition, *arXiv preprint arXiv:2511.21471* (2025).
- [13] C. Neo, L. Ong, P. Torr, M. Geva, D. Krueger, F. Barez, Towards interpreting visual information processing in vision-language models, *arXiv preprint arXiv:2410.07149* (2024).
- [14] G. Trimigno, G. Lombardo, S. Cagnoni, Low-rank lens for scalable LLMs interpretability, in: *Proceedings of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2026. doi:10.14428/esann/2026.ES2026-221.
- [15] N. Jiang, A. Kachinthaya, S. Petryk, Y. Gandelsman, Interpreting and editing vision-language representations to mitigate hallucinations, *arXiv preprint arXiv:2410.02762* (2024).

- [16] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al., Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, *arXiv preprint arXiv:2409.12191* (2024).
- [17] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, *Advances in neural information processing systems* 36 (2023) 34892–34916.
- [18] D. Weyns, A. Omicini, J. Odell, Environment as a first class abstraction in multiagent systems, *Autonomous agents and multi-agent systems* 14 (2007) 5–30.
- [19] G. Angiani, P. Fornaciari, G. Lombardo, A. Poggi, M. Tomaiuolo, Actors based agent modelling and simulation, in: *International Conference on Practical Applications of Agents and Multi-Agent Systems*, Springer, 2018, pp. 443–455.
- [20] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al., Rt-2: Vision-language-action models transfer web knowledge to robotic control, in: *Conference on Robot Learning*, PMLR, 2023, pp. 2165–2183.
- [21] R. Zhang, Z. Zeng, Z. Guo, Y. Li, Can language understand depth?, in: *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6868–6874.
- [22] H. Lin, S. Peng, J. Chen, S. Peng, J. Sun, M. Liu, H. Bao, J. Feng, X. Zhou, B. Kang, Prompting depth anything for 4K resolution accurate metric depth estimation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 17070–17080.
- [23] Z. Cai, C.-F. Yeh, H. Xu, Z. Liu, G. Meyer, X. Lei, C. Zhao, S.-W. Li, V. Chandra, Y. Shi, Depthlm: Metric depth from vision language models, *arXiv preprint arXiv:2509.25413* (2025).
- [24] S. Azad, Y. Jain, R. Garg, V. Vineet, Y. Rawat, Understanding depth and height perception in large visual-language models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 3620–3629.
- [25] E. Daxberger, N. Wenzel, D. Griffiths, H. Gang, J. Lazarow, G. Kohavi, K. Kang, M. Eichner, Y. Yang, A. Dehghan, P. Grasch, MM-Spatial: Exploring 3D spatial understanding in multimodal LLMs, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.
- [26] W. Cai, Y. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, B. Zhao, SpatialBot: Precise spatial understanding with vision language models, in: *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 9490–9498.
- [27] D. Liu, T. Liang, Z. Hu, J. Peng, Y. Lu, Y. Xu, Y. Fu, Y. Yin, Spatial intelligence in vision-language models: A comprehensive survey (2026).