Reasoning Paths with Reference Objects Elicit Quantitative Spatial Reasoning in Large Vision-Language Models

Yuan-Hong Liao[1], Rafid Mahmood[2,3], Sanja Fidler[1,2], David Acuna[2]

[1] University of Toronto, Vector Institute [2] NVIDIA [3] University of Ottawa

Correspondence: andrew@cs.toronto.edu

Abstract

Despite recent advances demonstrating vision-language models' (VLMs) abilities to describe complex relationships in images using natural language, their capability to quantitatively reason about object sizes and distances remains underexplored. In this work, we introduce a manually annotated benchmark, Q-Spatial Bench, with 271 questions across five categories designed for quantitative spatial reasoning and systematically investigate the performance of state-of-the-art VLMs on this task. Our analysis reveals that reasoning about distances between objects is particularly challenging for SoTA VLMs; however, some VLMs significantly outperform others, with an over 40-point gap between the two best performing models. We also make the surprising observation that the success rate of the top-performing VLM increases by 19 points when a reasoning path using a reference object emerges naturally in the response. Inspired by this observation, we develop a zero-shot prompting technique, `SpatialPrompt`, that encourages VLMs to answer quantitative spatial questions using reference objects as visual cues. By instructing VLMs to use reference objects in their reasoning paths via `SpatialPrompt`, Gemini 1.5 Pro, Gemini 1.5 Flash, and GPT-4V improve their success rates by *over* 40, 20, and 30 points, respectively. We emphasize that these significant improvements are obtained without needing more data, model architectural modifications, or fine-tuning.[1]

1 Introduction

Spatial reasoning is essential for humans to interact with the world, such as determining if there is enough room on a desk for a backpack; if there is enough space to navigate through a room without hitting any obstacles; or if an object is placed sufficiently high enough to be inaccessible to a child. Spatial reasoning skills are also essential for visual AI agents in interactive applications such as visual question answering, augmented reality, robotics, or visual assistants (Rocamonde et al., 2024; Du et al., 2023).

Despite this need, the ability for an off-the-shelf vision language model (VLM) to perform fine-grained spatial reasoning remains under-explored. Current benchmarks primarily assess whether these models understand qualitative concepts like "left" versus "right" or "near" versus "far" from monocular images; formulating these tasks as question-answering (Johnson et al., 2016; Krishna et al., 2016; Liu et al., 2023a; Kamath et al., 2023). However, recent studies have revealed that state-of-the-art VLMs struggle with quantitative spatial tasks. While measuring sizes or distances from monocular images is ill-posed (Saxena et al., 2005), humans are surprisingly adept at making such estimations by relying on contextual clues (Sedgwick, 1986), suggesting the potential improvements in VLMs.

To benchmark quantitative spatial reasoning, Spatial VQA (Chen et al., 2024) and SpatialRGPT-bench (Cheng et al., 2024) are proposed. Yet, both benchmarks either heavily relies on metric depth estimation or on the coarse label representations, *e.g.*, 3D cuboids, potentially leading to inaccurate quantitative labels. On the other hand, to improve quantitative spatial reasoning in VLMs, Spatial VLM and SpatialRGPT demonstrate that *fine-tuning* VLMs with additional spatial VQA data and architectural modifications can enhance their performances significantly. However, this requires direct access to the model, as well as a significant amount of data and computational resources.

In this paper, we introduce a new question answering corpora, Q-Spatial Bench, specifically designed to evaluate quantitative spatial reasoning in VLMs with high-precision. We identify the strengths and weakness across models by systematically analyzing the performance of SoTA VLMs. In

[1]Project website: `https://andrewliao11.github.io/spatial_prompt`

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17028–17047

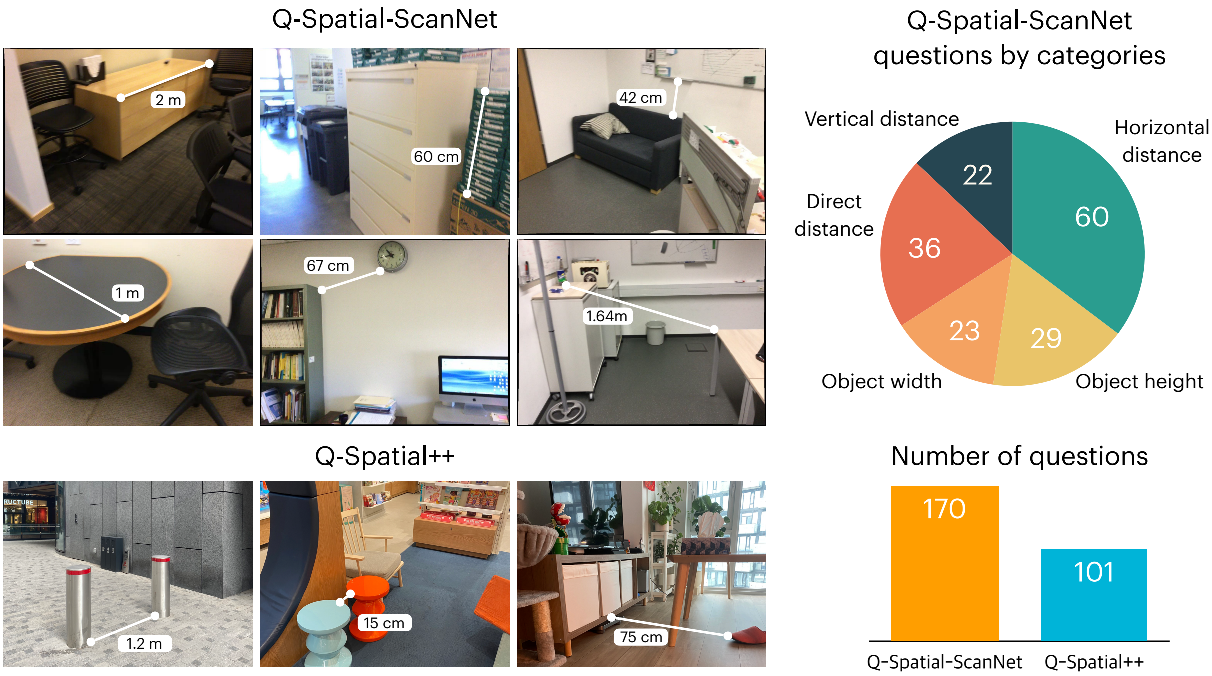

Figure 1: We introduce a human expert-annotated benchmark dedicated to quantitative spatial reasoning: **Q-Spatial Bench**. The benchmark consists of two splits: Q-Spatial-ScanNet and Q-Spatial++. The left panel shows the examples from the two splits. Q-Spatial-ScanNet is repurposed from a subset of images and RGB-D scans in ScanNet (Dai et al., 2017) and the questions are categorized into five categories (top-right). To provide a more robust evaluation in quantitative spatial reasoning, we captured an additional set of images and provide accurately-annotated quantitative spatial questions for Q-Spatial++.

particular, our analysis shows that VLMs become more effective when a reasoning path involving reference objects emerges in the responses. Inspired by this, we propose a prompt-based technique, `SpatialPrompt`, to guide VLMs to effectively solve quantitative spatial tasks. In contrast to prior works (Chen et al., 2024; Cheng et al., 2024), `SpatialPrompt` bypasses the need for direct access to the model, additional task-specific data, architectural modifications, or fine-tuning.

We demonstrate that by solely using `SpatialPrompt`, Gemini 1.5 Pro, Gemini 1.5 Flash, and GPT-4V improve their success rates by over 47, 22, and 30, respectively, in Q-Spatial Bench *without* the need for more data, model architectural modifications, or fine-tuning. To put these numbers in perspective, we note that the pioneering work, SpatialVLM (Chen et al., 2024), was fine-tuned on a massive dataset with 2 billion direct spatial reasoning QA pairs to achieve a relative improvement of less than 4 points in the same metric. While their evaluation benchmark is not publicly available for a direct comparison, we could still use this number as a proxy for reference. We emphasize our improvement is achieved purely with simple instructions using natural language. There is no external training data, model architectural modifications, or fine-tuning. Most importantly, we believe our approach of analyzing emergent reasoning paths in high-performing models opens new horizons for transferring capabilities to less performant ones in a cost-effective manner.

2 Related Works

2.1 Spatial Reasoning Benchmarks

Many vision-language benchmarks evaluate spatial relationships (Agrawal et al., 2015; Johnson et al., 2016; Krishna et al., 2016; Suhr et al., 2018; Yang et al., 2019; Belz et al., 2018; Goyal et al., 2020; Majumdar et al., 2024). However, spatial reasoning is entangled with other challenges in these datasets. With the promise of deploying vision-language models in the physical world (Brohan et al., 2022, 2023), there are more benchmarks designed to evaluate spatial reasoning alone. Most of them are tasked to evaluate qualitative spatial reasoning where models are tasked to recognize spatial relationships between objects in 3D space, such as “left” versus “right” or “behind” versus “in front

	# Quantitative spatial questions	Human annotated	Publicly available
Spatial VQA (Chen et al., 2024)	215		
SpatialRGPT-Bench (Cheng et al., 2024)	749		✓
Q-Spatial Bench (ours)	271	✓	✓

Table 1: **Comparison of quantitative spatial reasoning benchmark.** Q-Spatial Bench is a human expert-annotated benchmark, specifically designed for quantitative spatial questions.

of". VSR (Liu et al., 2023a) extend this direction by increasing linguistic complexities. What'sUp (Kamath et al., 2023) is designed to remove spatial biases in the real-world.

Quantitative spatial reasoning, where models are tasked to estimate quantitative spatial information such as sizes and distances, is relatively underexplored. Spatial VQA (Chen et al., 2024) contains both qualitative and quantitative spatial questions. However, it is publicly unavailable. SpatialRGPT-Bench (Cheng et al., 2024) is an open-source benchmark and follows the similar data-generation pipeline in Spatial VQA. Both benchmark are first automatically-generated from templates and followed by human verification. Our work takes a step further by providing a human-annotated benchmark dedicated to quantitative spatial reasoning, providing high-precision measurements. Moreover, we capture our own images to minimize the risk of test data leakage (Elangovan et al., 2021).

2.2 Spatial Reasoning in VLMs

Many prior works improve spatial reasoning by leveraging depth information (Fichtl et al., 2014; Rosman and Ramamoorthy, 2011) or sequential RGB images (Kase et al., 2020; Yuan et al., 2021; Migimatsu and Bohg, 2021; Majumdar et al., 2024). Our work focuses on spatial reasoning from a single monocular image. SpatialVLM (Chen et al., 2024) and SpatialRGPT (Cheng et al., 2024) are most aligned to our work. SpatialVLM propose an automatic data engines to generate training data for VLM fine-tuning. SpatialRGPT introduces a plugin module for relative depth input to further boost spatial reasoning in VLMs. Unlike prior works, we propose a prompting technique that consistently improves quantitative spatial reasoning in VLMs, bypassing the need for direct access to the model, additional task-specific data, architectural modifications, or fine-tuning.

3 Quantitative Spatial Reasoning

3.1 Preliminaries

We explore quantitative spatial reasoning where a VLM is tasked to recognize quantitative spatial information of physical objects such as distances or sizes from a 2D image. In particular, we consider *direct* quantitative spatial reasoning, where a VLM predicts the quantitative spatial information *without* accessing any external tools or large models.

Following prior work (Chen et al., 2024), we can formulate quantitative spatial reasoning into a question answering task. For an image I and a question querying spatial quantities in the image Q, the model needs to output the estimated quantities in text format Specifically, given an image and a question querying spatial quantities in the image, the model needs to directly output the estimated quantities in text format A.

3.2 Benchmark: Q-Spatial Bench

Existing benchmarks in quantitative spatial reasoning are either not publicly available (Chen et al., 2024) or automatically-generated (Cheng et al., 2024). We introduce a human expert-annotated benchmark dedicated to quantitative spatial reasoning: Q-Spatial Bench, highlighting its high-quality and high-precision question-answer pairs. Additionally, the benchmark provides a split with freshly captured images which are unseen to the commercial VLMs. Table 1 compares Q-Spatial Bench with existing benchmarks. Though being manually annotated, Q-Spatial Bench is as large as Spatial VQA and will be made publicly available.

Q-Spatial Bench consists of two splits: Q-Spatial-ScanNet and Q-Spatial++. Each split contains a series of image and annotated question tuples. We consider five different categories of spatial questions estimating (i) the width of an object; (ii) the height of an object; (iii) the horizontal distance between two objects on the same plane; (iv) the vertical distance between two objects vertically separated from each other; and (v) the direct distance between two objects in the scene. This categorization allows us to dissect the strengths and weaknesses of VLMs. Fig. 1 shows examples of these questions and overall statistics of Q-Spatial Bench.

Q-Spatial-ScanNet is repurposed from the high-quality RGB-D scans of real-world environments in ScanNet (Dai et al., 2017). Q-Spatial-ScanNet consists of 99 images and 170 human expert-annotated questions. For each question, the annotators are instructed to navigate in the point clouds provided in the dataset to measure the metric distances.

The annotation pipeline consists of three stages: image selection, distance annotation, and QA generation. First, we selected a subset of images from ScanNet as our candidate images. We exclude images that are blurry or with extremely simple layout, *e.g.*, a corner of a room without any objects. Next, with the provided reconstructed point clouds in ScanNet, we perform voxel downsampling to the point clouds and visualize them with Plotly (Inc., 2015). This allows human experts to navigate through the point clouds and annotate the end points of the distances. The experts then describe the annotated distances in natural language. As compared to the prior work (Cheng et al., 2024) that compute the distances from 3D cuboids, allowing human experts to explicitly specify the end points of a distance leads to higher-precision distance measurements With these annotations, we construct a triplet of an image, a distance quantity, and a description of the distance. Finally, we leverage Llama 3-70b (AI@Meta, 2024) to convert the triplets to the question-answering pairs and manually verify the sanity of each pair.

Q-Spatial++ is designed to provide a robust evaluation of VLMs for quantitative spatial reasoning. Because ScanNet is a pre-existing publicly available dataset, it is uncertain whether existing commercial VLMs were trained on these images. To mitigate this uncertainty, Q-Spatial++ uses 101 diverse images captured with an iPhone in various environments, including indoor, outdoor, daytime, and nighttime settings. Each image was captured specifically for spatial question answering, and has several objects within. While capturing these photos, we physically measured the distances between objects in the real-world. For this split, we focus only on horizontal distances between objects, which we observed more challenging for VLMs (see Section 4). Overall, there are 101 annotated questions in this split.

3.3 Evaluation metrics

In Q-Spatial Bench, a VLM must provide the answer in text format to facilitate regex extraction with a fixed script. We adopt a consistent output format for every VLM. We measure the performance in success rate by thresholding the maximum ratio between an estimation and a ground truth value. Given a distance estimation $\hat{d}$ and a ground truth value d^*, the maximum ratios between the estimation and ground truth can be expressed as $\delta = \max(\frac{\hat{d}}{d^*}, \frac{d^*}{\hat{d}})$. Following existing works (Chen et al., 2024; Cheng et al., 2024), we set the thresholds to 1.25 and 2 and denote the evaluation metrics as $\delta_{\leq 1.25}$ and $\delta_{\leq 2}$, respectively. In other words, an accurate answer should be within $[0.8\times, 1.25\times]$ or $[0.5\times, 2\times]$ the ground truth value, respectively. We use $\delta_{\leq 2}$ as the main evaluation metrics in this paper and report $\delta_{\leq 1.25}$ in the Appendix.

	Gemini 1.5 Pro*	Gemini 1.5 Flash	GPT-4V	GPT-4o
Q-Spatial-ScanNet	0.59	20	28.24	**69.41**
Q-Spatial++	0.99	26.73	18.81	**61.06**
Overall	0.79	23.36	23.52	**65.23**

Table 2: **GPT-4o outperforms other commercial VLMs in quantitative spatial reasoning.** We evaluates the success rate $\delta_{\leq 2}$ on each split of Q-Spatial Bench. *Gemini 1.5 Pro consistently refuses to provide the measurements.

	Gemini 1.5 Pro*	Gemini 1.5 Flash	GPT-4V	GPT-4o
Object width	0.87	43.48	44.93	**81.16**
Object height	0	31.82	30.3	**93.94**
Horizontal distance	0.33	13.33	10	**49.44**
Vertical distance	0.69	20.69	33.33	**71.26**
Direct distance	1.11	8.33	42.59	**78.7**

Table 3: **Success rate $\delta_{\leq 2}$ breakdown by question categories.** Among five question categories, measuring the distance between two objects is more challenging for SoTA VLMs, particularly when measuring horizontal distances. *Gemini 1.5 Pro consistently refuses to provide the measurements.

4 Analysis

We first test four commercial VLMs including Gemini 1.5 Pro, Gemini 1.5 Flash (Gemini, 2024), GPT-4V (OpenAI, 2024), and GPT-4o, to identify which question categories are more challenging or easier for VLMs in Sec. 4.1. We then statistically analyze the potential factors leading to the strong GPT-4o performances on all tasks in Sec. 4.2.

4.1 Strengths and Weaknesses

Table 2 evaluates the success rate $\delta_{\leq 2}$ in Q-Spatial Bench. GPT-4o outperforms all other models on both splits with success rates of 69.41 and 62.44, respectively. The performance gap between GPT-4o and the second-best model, GPT-4V, is significant by a margin of over 41 points.

To further understand the strengths and weakness within each VLM, we break down Q-Spatial-ScanNet by question categories and demonstrate

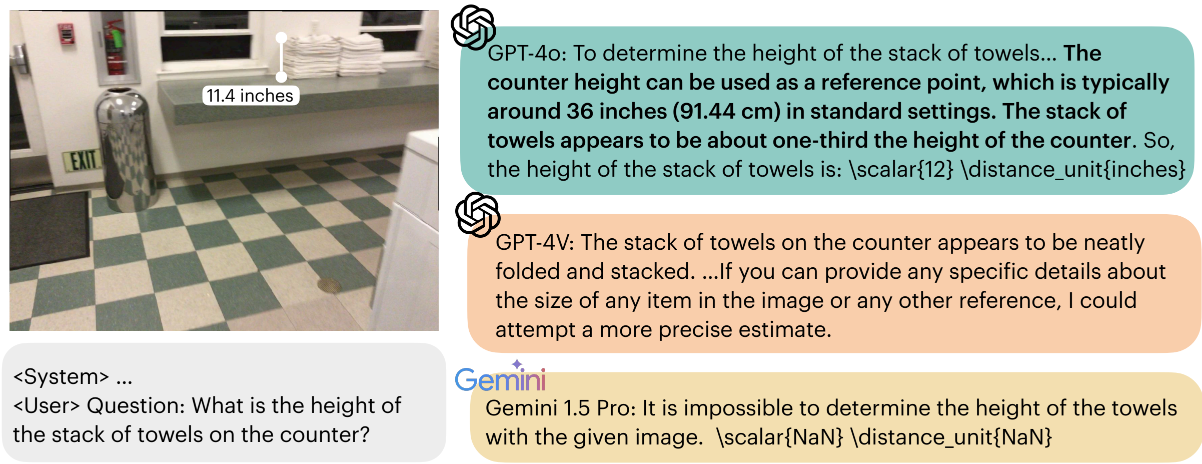

Figure 2: **Qualitative analysis of different VLMs responses.** As compared to GPT-4V and Gemini 1.5 Pro, we observe that for many correct instances, a reasoning path using a reference object naturally emerges in GPT-4o responses. A *reference object* is another object in the scene for whose spatial dimensions can be easily inferred via commonsense reasoning. Exploiting this observation, we design a zero-shot prompting technique that encourages both underperforming VLMs *and* GPT-4o itself to reason this way.

the success rate $\delta_{\leq 2}$ in Fig. 3. In particular, "Horizontal distance" questions are challenging for most VLMs, and usually achieve the lowest success rates. GPT-4o leads in all question types, showing the highest success rates, especially in "Object width" and "Object height" with 81.16 and 93.94 points, respectively. The overall challenge of estimating vertical and horizontal distances is evident, as these categories have the lowest success rates across models, particularly for Gemini 1.5 Pro, which we observe consistently refuses to provide measurements. We consider the responses with unspecified quantities as incorrect answers. Notably, the gap between GPT-4o and the second-best model on "Horizontal distance" question is over 35 points.

4.2 Understanding the dominance of GPT-4o

A qualitative analysis of the output text generated by GPT-4o reveals that for many correct instances, the output text suggests a reasoning path using one or more reference objects to guide the spatial estimation of the target object in each scene. A reference object is another object in the scene for whose spatial dimensions can be easily inferred via commonsense reasoning. As shown in Fig. 2, when measuring the height of the stack of towels, the counter near it can act as a reference object. While the height of the stack of towels can have variable height, the counter appears to have a standard height (36 inches suggested by GPT-4o).

From a reasoning perspective, the height of the counter introduces additional facts about the scene, and analyzing the relative position of the stack of towels with respect to the counter can lead to a more accurate estimate than by naïvely estimating the size of the stack of towels in isolation.

	Correct ($\delta_{\leq 2}$)	Incorrect ($\delta_{>2}$)
w/ references	45	9
w/o references	114	63

Table 4: **Contingency table** of whether GPT-4o's responses use any reference objects as guidance and the success rate of the responses. GPT-4o used a reference object in 23.4% of instances. However, for the 54 questions where a reference object was used, the success rate $\delta_{\leq 2}$ is 83% (45/54), compared to 64% (114/177) when no reference object was used.

To quantify the above observation, we first count the number of instances where the GPT-4o output involves reasoning via a reference object to answer the question. To do this, we feed the sampled GPT-4o responses to a separate GPT-4o instance which we prompt to determine if a model response uses any reference object as guidance (see Appendix C for details). Table 4 gives the contingency table of the number of instances where at least one reference object was used versus whether the maximum ratio δ is less 2. Note that GPT-4o used a reference object for 23.4% of instances. However, of the 54 questions where a reference object was used, the success rate $\delta_{\leq 2}$ is 83%, versus the success rate $\delta_{\leq 2}$ of only 64% over instances when a reference

	Coefficient β	SE
Use reference object (X_r)	1.0179*	0.406
From Q-Spatial++ (X_d)	0.0276	0.318
Ground truth distance (cm) (X_g)	0.0017	0.002

Table 5: **Logistic regression to analyze the effectiveness of GPT-4o.** * denotes a p-value < 0.05. Using a reference object in reasoning increases the likelihood of generating a response with relative error δ less than 2, statistically significantly.

object was not used. This motivates the following hypothesis:

Reasoning via a reference object directly improves overall accuracy on quantitative spatial questions.

To test our hypothesis, we fit a logistic regression model on the GPT-4o responses to whether the estimate is accurate. For any response, let $X_r \in \{0, 1\}$ indicate whether the response used any reference object. To isolate the effect of using any reference object, we control for two task-related factors. First, it may be that one of the two dataset splits is substantially easier than the other, thereby leading to more accurate estimates; here, let $X_d \in \{0, 1\}$ indicate whether the question belongs to the Q-Spatial++ split or Q-Spatial-ScanNet split. Second, tasks where the ground truth distance measured is large inherently have a larger absolute range to achieve $\delta_{\leq 2}$. To ensure that GPT-4o is not just better at answering questions involving larger distances, let X_g be the ground truth distance (in cm) for each question. Thus, our logistic regression model fits:

$$p(\delta_{\leq 2}) \sim \beta_0 + \beta_r X_r + \beta_d X_d + \beta_g X_g$$

where $\beta_0, \beta_r, \beta_g, \beta_d$ are regression coefficients.

Table 5 summarizes the results. We first find that when GPT-4o uses a reference object in the responses, the odds of an accurate estimate increase by a factor of $e^{1.0179} \approx 2.7$ (p-value < 0.05). Furthermore, there is no statistically significant difference in terms of GPT-4o performance on Q-Spatial++ or Q-Spatial-ScanNet, *i.e.*, the model performs roughly equally well on both splits. Finally, the accuracy of the distance estimate does not statistically significantly depend on the actual ground truth distance. That is, even if the question involves measuring a large distance, this does not change the likelihood of generating an accurate response, after factoring for whether a reference object is used. *We conclude by verifying our hypothesis that reference objects can improve the overall reasoning of the model.*

5 `SpatialPrompt`: Eliciting Reasoning Paths with Reference Objects

Inspired by the analysis in Sec. 4.2, we propose a prompting technique, `SpatialPrompt`, for quantitative spatial reasoning that explicitly encourages VLMs to identify reference objects within the images. Orthogonal to prior work (Chen et al., 2024; Cheng et al., 2024), our prompting techniques involve no additional data generation, fine-tuning or architectural modifications.

5.1 Prompt Design

`SpatialPrompt` is specifically designed to trigger a reasoning path that uses reference objects in the image as visual cues. We propose two variants of text prompts: `SpatialPrompt-Single` and `SpatialPrompt-Steps` as shown in Fig. 3. The former is inspired by zero-shot CoT (Kojima et al., 2022). We specifically design `SpatialPrompt-Single` to keep it as compact and easy to remember as possible. In contrast, `SpatialPrompt-Steps` provides a more detailed breakdown of the steps to solve the problem. We found that some VLMs perform better when given explicit step-by-step instructions. For instance, Gemini 1.5 Pro follows the detailed instructions more effectively, whereas GPT-4V's performance is less dependent on the complexity of the prompt as long as it is encouraged to find a reference object.

5.2 Experiments

In this section, we evaluate the effectiveness of the proposed text prompting techniques in success rate $\delta_{\leq 2}$ in Sec. 5.2.1. Since our initial motivation is to encourage VLMs to use reference objects within the images more frequently, we report the frequency of whether the VLMs use reference objects in the responses in Sec. 5.2.2. Lastly, we discuss the attempt to apply our prompting techniques to open-source VLMs in Sec. 5.2.3.

To gauge the effectiveness of `SpatialPrompt`, we adopt two baselines: **1)** Standard prompt and **2)** Zero-shot CoT (Kojima et al., 2022). The standard prompt only involves the spatial question. Zero-shot CoT is a task agnostic prompting technique and appends "Let's think step by step" afters the user query. The full prompt templates are shown in the Appendix C.

	Q-Spatial-ScanNet				Q-Spatial++			
	Gemini 1.5 Pro	Gemini 1.5 Flash	GPT-4V	GPT-4o	Gemini 1.5 Pro	Gemini 1.5 Flash	GPT-4V	GPT-4o
Standard prompt	0.59	20	28.24	69.41	0.99	26.73	18.81	61.06
Zero-shot CoT	2.94	18.53	**56.47**	70.2	4.16	26.11	46.53	59.74
`SpatialPrompt`	**53.65**	**52.71**	54.9	**71.96**	**43.17**	**38.42**	**53.47**	**62.71**

Table 6: **Success rate $\delta_{\leq 2}$ of different VLMs and prompting techniques.** The proposed prompt `SpatialPrompt` consistently leads to higher success rates across different VLMs. We bold font the best numbers across different prompting techniques and highlight their performances as compared to the performances of the standard prompt.

`SpatialPrompt-Single`

```
User: Question: [SPATIAL QUESTION]
Let's think step by step and start by finding good reference objects
↪  or object parts in the image.
Assistant: [OUTPUT]
```

`SpatialPrompt-Steps`

```
User: Question: [SPATIAL QUESTION]
Use the following 4 steps sequentially to answer the question:
Step 1 **Analyze the question**

Step 2 **Identify up to 10 reference scales in the image, ranging from
↪  large to small sizes, and list them in the specified format**
[details omitted]

Step 3 **Propose a robust step-by-step plan to answer the question by
↪  using the reference scales in Step 2**
[details omitted]

Step 4 **Focus on the image and follow the plan in Step 3 to answer
↪  the question**
Assistant: [OUTPUT]
```

Figure 3: We propose `SpatialPrompt`, a specialized text prompt designed to improve quantitative spatial reasoning capabilities in VLMs by encourage them to reason by identifying reference objects within the images.

5.2.1 Main results

We evaluate the success rate $\delta_{\leq 2}$ in Table 6. For `SpatialPrompt`, we apply `SpatialPrompt-Single` on GPT-4V and GPT-4o and `SpatialPrompt-Steps` on Gemini 1.5 Pro and Gemini 1.5 Flash. As shown before in Sec. 4, GPT-4o outperforms all the other VLMs out of the box. Zero-shot CoT significantly improves GPT-4V over standard prompt, but fail to boost the performances in Gemini 1.5 Pro and Gemini 1.5 Flash with only marginal or even negative improvements. On the contrary, `SpatialPrompt` improves the success rate across almost all VLMs, resulting in 47, 22, 30, and 2 points in Gemini 1.5 Pro, Gemini 1.5 Flash, GPT-4V, and GPT-4o, respectively. This drastically closes the gap between the GPT-4o and other VLMs. In Sec. 4, we observe a 41 point gap between GPT-4o and the second-best VLM in Q-Spatial Bench when receive the standard prompt. With `SpatialPrompt`, the gap between GPT-4o and the second-best VLM is reduced to 13 points.

5.2.2 Does `SpatialPrompt` encourage the use of references?

The motivation behind `SpatialPrompt` is to encourage VLMs to use reference objects within the scene to answer the quantitative spatial questions. We, therefore, compare the frequency of whether a response involves reasoning via reference objects to answer the question. Same as the analysis in Sec. 4.2, we adopt a separate GPT-4o instance to determine if the model responses involves reference objects. Fig. 4 visualizes the relations between the frequency of using any reference objects and the success rates $\delta_{\leq 2}$. The Spearman correlation coefficients of the two variables are 0.69 and 0.61 in Q-Spatial-ScanNet and Q-Spatial++, respectively. This empirically corroborates the hypothesis in Sec. 4.2.

5.2.3 Open-sourced VLMs

In this work we mainly focus on SoTA VLMs (*i.e.*, commercial VLMs) as larger models are known for their superior reasoning capabilities and their ability to follow instruction prompts more closely. Here, however, we also investigate the off-the-shelf performance and capabilities of open-source VLMs, *i.e.*, LLaVA-v1.6-34b (Liu et al., 2024).

Table 7 shows the performances of LLaVA in Q-Spatial Bench using standard prompt. Surprisingly, LLaVA achieves 60.59 in Q-Spatial-ScanNet, outperforming most commercial VLMs including Gemini 1.5 Pro, Gemini 1.5 Flash, and GPT-4V. However, LLaVA falls more than 20 points when evaluated on Q-Spatial++. Interestingly, when qualitatively analyzing the outputs of the models, we find that they essentially "predict" numbers rather than demonstrating any intuitive reasoning behavior, as seen in more powerful VLMs. We initially speculated that this might indicate the model was trained on some collection of datasets including ScanNet data. However, we found no reference suggesting that LLaVA directly uses ScanNet in its

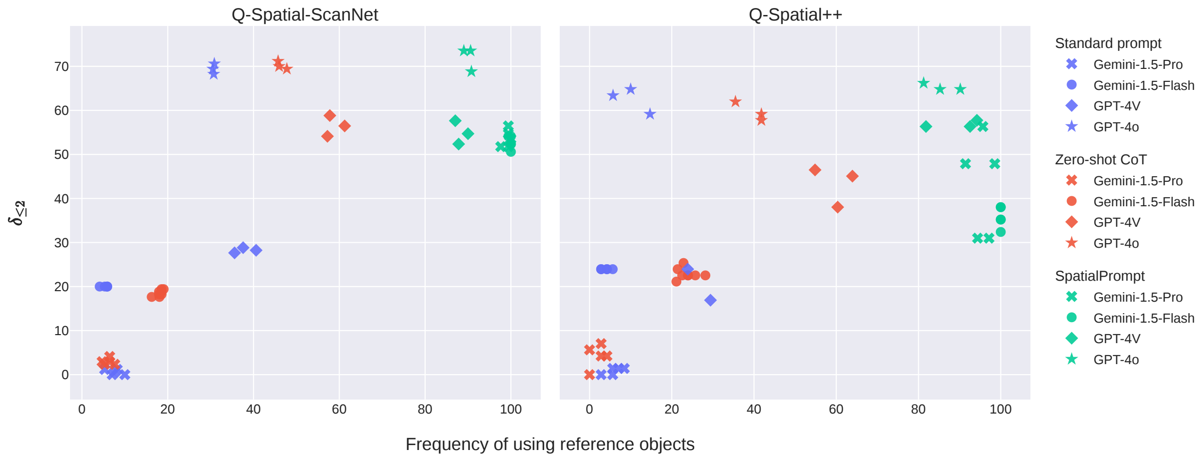

Figure 4: **Success rates versus the frequencies of using reference objects.** Green corresponds to `SpatialPrompt`, Red corresponds to Zero-shot CoT, and Blue corresponds to standard prompt. Across all VLMs (represented by distinct symbols) and text prompts (indicated by distinct colors), the two variables show a strong correlation, with the values of 0.69 and 0.91 in Q-Spatial-ScanNet and Q-Spatial++, respectively. This result is aligned with our hypothesis in Sec. 4.2: *Reasoning via a reference object directly improves overall accuracy on quantitative spatial questions*.

	Q-Spatial-ScanNet	Q-Spatial++
Standard prompt	**60.59**	36.62
Zero-shot CoT	40	32.39
`SpatialPrompt`	51.76	**45.07**

Table 7: **Success rate $\delta_{\leq 2}$ of LLaVA in Q-Spatial-ScanNet and Q-Spatial++.**

training dataset.

Additionally, we found LLaVA does not perform well when receiving zero-shot CoT prompts. When receiving zero-shot CoT prompts, the performances drop by over 20 and 4 points in Q-Spatial-ScanNet and Q-Spatial++. For `SpatialPrompt`, we adopt `SpatialPrompt-Steps`, leading to a decrease around 9 points in Q-Spatial-ScanNet. In sharp contrast, `SpatialPrompt-Steps` improves 9 points in Q-Spatial++. We hypothesize several reasons for this: (i) LLaVA's capabilities are not as developed as those in very large commercial models, and inducing reasoning structures in-context via prompts do not yet lead to the same level of improvements; (ii) LLaVA may fail to follow the instruction prompts effectively; (iii) LLaVA may not be proficient at visual comparison, so even if it increases the use of reference objects, it does not yield better performance. This observation is aligned with LLMs in language tasks (Wei et al., 2022; Kojima et al., 2022), suggesting that the reasoning capabilities usually emerge when the model size exceeds 100B.

6 Limitations

In this section, we discuss the limitations of our work from two aspects: the proposed benchmark and the proposed prompting technique.

Q-Spatial Bench is limited by its size due to the time-consuming questions-answering annotation, especially the time spent to measure high-precision distances. In cases where the end points of the distance is ambiguous, the annotators are required to clicking around the nearby point cloud to ensure the measurement quality. For example, when annotating the shortest distance between two objects, to find the end points of the shortest distance can be time-consuming. As compared to prior works (Chen et al., 2024; Cheng et al., 2024), our benchmark is around four times smaller.

`SpatialPrompt` is essentially trading tokens to performances since the appending text prompt encourages VLMs to generate the reasoning paths in a detailed manner. This makes it less likely to be applied in the application requires real-time inference. Additionally, the motivation of `SpatialPrompt` is that we want to encourage VLMs to use the context information to improve quantitative spatial reasoning. However, when the image consists of a clean background, it is possible that there is little visual cues in the image; therefore, adding `SpatialPrompt` hurts or confuses VLMs. Finally, as shown in Sec. 5.2.3, `SpatialPrompt` results in inconsistent improvements for open-source VLMs. It is unclear that if the proposed `SpatialPrompt` only works when the model scale is large enough.

To summarize, Q-Spatial Bench is mainly limited by its sizes due to the costly labeling process. On the other hand, `SpatialPrompt` can be limited by inference speed, image context and maybe model scales.

7 Conclusion

In conclusion, our work introduces Q-Spatial Bench, a novel benchmark tailored for evaluating VLMs on quantitative spatial reasoning and that incorporates 5 different question categories. Through systematic analysis, we have identified the strengths and weaknesses of current SoTA VLMs in this task. Furthermore, we have developed a prompt-based technique that leverages naturally emerging reasoning paths to enhance VLM performance on quantitative spatial tasks. Our findings demonstrate that instructing VLMs to use reference objects in their reasoning pathways significantly boosts their performance without the need for additional data or fine-tuning.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31.

AI@Meta. 2024. Llama 3 model card.

Anja Belz, Adrian Muscat, Pierre Anguill, Mouhamadou Sow, Gaétan Vincent, and Yassine Zinessabah. 2018. SpatialVOC2K: A multilingual dataset of images with annotations and features for spatial relations between objects. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 140–145, Tilburg University, The Netherlands. Association for Computational Linguistics.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2022. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *Preprint*, arXiv:2401.12168.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision language model. *Preprint*, arXiv:2406.01584.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443.

Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. 2023. Vision-language models as success detectors. *Preprint*, arXiv:2303.07280.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Severin Fichtl, Andrew McManus, Wail Mustafa, Dirk Kraft, Norbert Krüger, and Frank Guerin. 2014. Learning spatial relationships from 3d vision using histograms. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 501–508.

Team Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. 2020. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. *ArXiv*, abs/2012.01634.

Plotly Technologies Inc. 2015. Collaborative data science.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*.

Kei Kase, Chris Paxton, Hammad Mazhar, Tetsuya Ogata, and Dieter Fox. 2020. Transferable task execution from pixels through deep planning domain learning. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10459–10465.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Toki Migimatsu and Jeannette Bohg. 2021. Grounding predicates through actions. *2022 International Conference on Robotics and Automation (ICRA)*, pages 3498–3504.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. 2024. Vision-language models are zero-shot reward models for reinforcement learning. *Preprint*, arXiv:2310.12921.

Benjamin Rosman and Subramanian Ramamoorthy. 2011. Learning spatial relationships between objects. *The International Journal of Robotics Research*, 30:1328 – 1342.

Ashutosh Saxena, Sung H. Chung, and A. Ng. 2005. Learning depth from single monocular images. In *Neural Information Processing Systems*.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Henry A Sedgwick. 1986. Space perception. *Sensory processes and perception*.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *ArXiv*, abs/1811.00491.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. 2019. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2051–2060.

Wentao Yuan, Chris Paxton, Karthik Desingh, and Dieter Fox. 2021. Sornet: Spatial object-centric representations for sequential manipulation. *ArXiv*, abs/2109.03891.

A Table of Contents

We organize the Appendix in the following structures:

1. Sec. B provides the full experiment results of Gemini 1.5 Pro, Gemini 1.5 Flash, GPT-4V, and GPT-4o in Table 8, Table 9, Table 10.
2. Sec. C shows all the text prompts used in this work.
3. Sec. D provide the qualitative results.
4. Sec. E shows the implementation details including generation parameters and computational resources used in this work.
5. Sec. F shows the license related to this work

B Full experimental results

This section provides the additional experimental results:

Full version of Table 6. Table 8 shows the full results of Gemini 1.5 Pro, Gemini 1.5 Flash, GPT-4V, and GPT-4o along with their standard deviations. We find that Gemini 1.5 Pro only works when prompted with detail procedures and Gemini 1.5 Flash works pretty well when prompted with abstract and detailed procedures. On the other hand, prompting GPT-4V and GPT-4o with detail procedures hurts the performances in Q-Spatial Bench. We hypothesize that since GPT-4V and GPT-4o are more competent models for quantitative spatial reasoning, constraining the reasoning path with detail procedures limits the performances.

Full version of Table 7. Table 9 includes different variants of LLaVA (Liu et al., 2024, 2023b). As show in the main paper, LLaVA-v1.6-34b demonstrate strong performances on Q-Spatial-ScanNet, but falls more than 20 points in Q-Spatial++. From Table 9, we find that different LLaVA variants has very different preferred prompts in Q-Spatial Bench. These inconsistent preferences might result from the findings in the prior works (Wei et al., 2022; Kojima et al., 2022) where models under 100B parameters exhibit unreliable reasoning paths.

Evaluation in $\delta_{\leq 1.25}$. Table 10 shows the performances in $\delta_{\leq 1.25}$. By comparing Table 8 that evaluates in $\delta_{\leq 2}$ and Table 10, we find that the standard deviations generally increases when evaluated in $\delta_{\leq 1.25}$ due to a smaller threshold.

Full table of the frequency of using reference objects for each VLM.

In Table 11, we show the frequency of different VLMs and different text prompts. While zero-shot CoT increases the chances of involving reference objects in the responses, `SpatialPrompt` increases the most by explicitly asking the model to use the reference objects in the image.

Human evaluation. To gauge the difficulty of the introduced benchmark, Q-Spatial Bench, we ask three humans to perform quantitative spatial reasoning. Each human answer a sampled set of 40 quantitative spatial reasoning questions randomly draw from both splits. Table 12 show the success rate $\delta_{\leq 2}$ of 3 humans. Though GPT-4o demonstrates strong performances as compared to other VLMs, it still talls behind humans by more than 30 points. This suggests that the curated benchmark is easy to humans but difficult to VLMs.

C Full text prompt

This section provides the full text prompts used in our work,

1. Fig. 5: The text prompt used to determine whether the VLM responses involves using any reference objects. It is used in Table 4, Table 5, and Table 11.
2. Fig. 6: The text prompt is used in Table 2 and Table 3. It is referred to as "Standard prompt" in Sec. 5.
3. Fig. 7: The text prompt is slightly modified from the prompt in Fig. 6 since LLaVA (Liu et al., 2024) constantly fails to follow the right output format. Following prior work (Kojima et al., 2022), we adopt a two-stage generation strategy to encourage the correct formats in the responses.
4. Fig. 8: The text prompt is referred to as `SpatialPrompt-Steps` in Sec. 5.
5. Fig. 9: The text prompt is used to convert a triplet of an image, a distance, and a description of the distance to a question-answering pair.

D Qualitative results

GPT-4V responses with different text prompts. In Fig. 11, we compare GPT-4V responses with different text prompts. When using standard prompt,

	Q-Spatial-ScanNet				Q-Spatial++			
	Gemini 1.5 Pro	Gemini 1.5 Flash	GPT-4V	GPT-4o	Gemini 1.5 Pro	Gemini 1.5 Flash	GPT-4V	GPT-4o
Standard prompt	0.59 $\pm$ 0.52	20 $\pm$ 0.	28.24 $\pm$ 0.48	69.41 $\pm$ 0.96	0.99 $\pm$ 0.62	23.73 $\pm$ 0.	18.81 $\pm$ 1.98	61.06 $\pm$ 1.86
Zero-shot CoT	2.94 $\pm$ 0.64	18.53 $\pm$ 0.65	**56.47** $\pm$ **1.92**	70.2 $\pm$ 0.7	4.16 $\pm$ 1.91	26.11 $\pm$ 0.68	46.53 $\pm$ 2.8	59.74 $\pm$ 1.23
`SpatialPrompt-Single`	6.47 $\pm$ 1.17	49.76 $\pm$ 1.8	54.9 $\pm$2.16	**71.96** $\pm$ **2.21**	5.35 $\pm$ 1.05	**55.77** $\pm$ **2.45**	**53.47** $\pm$ **3.23**	**62.71** $\pm$ **1.68**
`SpatialPrompt-Steps`	**53.65** $\pm$ **1.79**	**52.71** $\pm$ **1.32**	53.14 $\pm$ 0.55	63.33 $\pm$ 1.54	**43.17** $\pm$ **7.21**	38.42 $\pm$ 2.37	35.68 $\pm$ 1.75	58.69 $\pm$ 1.75

Table 8: **Full table of the success rate $\delta_{\leq 2}$ of Gemini 1.5 Pro, Gemini 1.5 Flash, GPT-4V, and GPT-4o.** All numbers are averaged over 5 different runs, except for GPT-4V and GPT-4o, which are run on three seeds. Each number is followed by their standard deviations.

	Q-Spatial-ScanNet				Q-Spatial++			
	LLaVA-v1.6-34b	LLaVA-v1.6-13b	LLaVA-v1.5-13b	LLaVA-v1.5-7b	LLaVA-v1.6-34b	LLaVA-v1.6-13b	LLaVA-v1.5-13b	LLaVA-v1.5-7b
Standard prompt	**60.59**	55.88	42.24	40	36.62	38.03	35.21	29.58
Zero-shot CoT	40	**58.24**	41.76	44.12	32.39	**47.89**	23.94	**35.21**
`SpatialPrompt-Single`	47.65	44.12	35.29	47.06	39.44	45.07	35.21	**35.21**
`SpatialPrompt-Steps`	51.76	35.29	**44.71**	**49.41**	**45.07**	39.15	**46.68**	33.8

Table 9: **Full table of the success rate $\delta_{\leq 2}$ of LLaVA at different versions and model sizes.** All numbers are averaged over 5 different runs and followed by their standard deviations.

GPT-4V tends to perform a very rough estimate directly. On the other hand, when prompted with zero-shot CoT and `SpatialPrompt`, GPT-4V leverages surrounding objects for quantitative spatial reasoning.

Gemini 1.5 Pro responses with different text prompts. In Fig. 12, we compare Gemini 1.5 Pro responses with different text prompts. Among all the evaluated prompts, SpatialPropmt is the only one that successfully elicit quantitative spatial reasoning in Gemini-1.5-Pro.

Failure cases of GPT-4o. Fig. 13 shows the common failure case of GPT-4o. We find that GPT-4o have a tendency to use floor tiles as references, leading to inaccurate estimation at times.

E Implementation details

In this work, we employ deterministic sampling and conduct each experiment using five different random seeds, except for GPT-4V and GPT-4o, which are run on three seeds. The specific endpoint for GPT-4o is `gpt-4o-2024-05-13`, for GPT-4V is `gpt-4-turbo-2024-04-09`, for Gemini 1.5 Pro is `gemini-1.5-pro-001`, and for Gemini 1.5 Flash is `gemini-1.5-flash-001`. All GPT-4o, GPT-4V, and Gemini experiments are run with the provided API.

It is worth noted that most VLM APIs do not support fully reproducible results. That is, even when the temperature is set of 0, there is still randomness in the results. We empirically find that setting top_p to a very small value 1^{-14} and temperature to 1 produces better than setting top_p to 1 and temperature to 0. For all experiment, we set top_p to 1^{-14} and temperature to 1.

For LLaVA experiments, we set temperature to 0 and top_p to 1.. For LLaVA, we run on a machine with 2 A100 GPUs.

For the logistic regression analysis, we use the default parameters in statsmodels (Seabold and Perktold, 2010).

F License

In Q-Spatial-ScanNet, we use ScanNet images under the license at `https://kaldir.vc.in.tum.de/scannet/ScanNet_TOS.pdf`

	Q-Spatial-ScanNet				Q-Spatial++			
	Gemini 1.5 Pro	Gemini 1.5 Flash	GPT-4V	GPT-4o	Gemini 1.5 Pro	Gemini 1.5 Flash	GPT-4V	GPT-4o
Standard prompt	0.35 ± 0.28	10.59 ± 0.	13.92 ± 0.99	32.55 ± 1.2	0. ± 0.	16.83 ± 0.	8.91 ± 0.99	22.77 ± 3.7
Zero-shot CoT	1.41 ± 0.47	7.06 ± 0.58	**23.53** ± **1.73**	33.53 ± 0.83	0.4 ± 0.48	11.51 ± 0.84	20.79 ± 2.13	20.46 ± 3.26
`SpatialPrompt-Single`	3.41 ± 0.86	18.47 ± 0.95	22.94 ± 0.48	**35.69** ± **1.99**	1.41 ± 1.25	**22.25** ± **2.25**	**22.44** ± **5.13**	21.78 ± 4.2
`SpatialPrompt-Steps`	**38** ± **0.59**	**21.29** ± **1.76**	19.8 ± 2.46	26.67 ± 2.26	**12.08** ± **3.33**	12.67 ± 1.58	13.15 ± 3.31	**27.7** ± **2.65**

Table 10: **Full table of the success rate $\delta_{\leq 1.25}$ of Gemini 1.5 Pro, Gemini 1.5 Flash, GPT-4V, and GPT-4o.** All numbers are averaged over 5 different runs, except for GPT-4V and GPT-4o, which are run on three seeds. Each number is followed by their standard deviations.

	Q-Spatial-ScanNet				Q-Spatial++			
	Gemini 1.5 Pro	Gemini 1.5 Flash	GPT-4V	GPT-4o	Gemini 1.5 Pro	Gemini 1.5 Flash	GPT-4V	GPT-4o
Standard prompt	7.64	5.42	37.9	30.7	5.91	3.94	26.64	10.14
Zero-shot CoT	6.11	18.12	58.75	46.48	1.97	27.71	59.7	39.68
`SpatialPrompt`	**99.17**	**99.8**	**88.28**	**90.14**	**95.42**	**100**	**89.51**	**85.55**

Table 11: **Frequency of whether the responses involve using reference objects of different VLMs and prompting techniques.** The proposed prompt `SpatialPrompt` consistently lead to higher chances to have reference objects involved in the responses.

	Sampled Q-Spatial-Bench
Human A	90
Human B	82.5
Human C	97.5
Human Avg.	90

Table 12: **Human performances in Q-Spatial Bench.** Though GPT-4o demonstrates strong performances as compared to other VLMs, it still falls behind humans by more than 30 points. This suggests that the curated benchmark is easy to humans but difficulty to VLMs.

```
System: You are an AI assistant evaluating responses to questions about
↪  measuring distances in 3D space using 2D images. Check if the response
↪  uses reference objects or scales to answer the question. Answer "YES" if
↪  it does, "NO" if it does not, and "UNSURE" if uncertain. Begin your answer
↪  with "YES", "NO", or "UNSURE".
User: Question: [SPATIAL QUESTION]
Response: [VLM RESPONSE]
Assistant: [OUTPUT]
```

Figure 5: **Text prompt for determining whether the VLM responses involves using any reference objects.** The text prompt is used in Table 4, Table 5, and Table 11.

```
System: You will be provided with a question and a 2D image. The question
↪  involves measuring the precise distance in 3D space through a 2D image.
↪  You will answer the question by providing a numeric answer consisting of a
↪  scalar and a distance unit in the format of """\scalar{scalar}
↪  \distance_unit{distance unit}""" at the end of your response.
User: Question: [SPATIAL QUESTION]
Assistant: [OUTPUT]
```

Figure 6: **Standard prompt for quantitative spatial reasoning.** The text prompt is used in Table 2 and Table 3. The prompt is referred to as "Standard prompt" in Section 5.

```
System: You will be provided with a question and a 2D image. The question
↪  involves measuring the precise distance in 3D space through a 2D image.
↪  You will answer the question by providing a numeric answer consisting of a
↪  scalar and a distance unit in the format of """\scalar{scalar}
↪  \distance_unit{distance unit}""" at the end of your response.
User: Question: [SPATIAL QUESTION]
Answer by providing a numeric answer consisting of a scalar and a distance
↪  unit in the format of """\scalar{scalar} \distance_unit{distance unit}"""
↪  at the end of your response.
Assistant: [FIRST STAGE OUTPUT] In conclusion, the final answer in the
↪  specified format is: """\scalar{[SECOND STAGE OUTPUT]
```

Figure 7: **Standard prompt for LLaVA.** The text prompt is slightly modified from the prompt in Fig. 6 since LLaVA (Liu et al., 2024) constantly fails to follow the right output format. Following prior work (Kojima et al., 2022), we adopt a two-stage generation strategy to encourage the correct formats in the responses.

```
System: You will be provided with a question and a 2D image. The question
↪ involves measuring the precise distance in 3D space through a 2D image.
↪ You will answer the question by providing a numeric answer consisting of a
↪ scalar and a distance unit in the format of """\scalar{scalar}
↪ \distance_unit{distance unit}""" at the end of your response.
User: Question: [SPATIAL QUESTION]

---

Use the following 4 steps sequentially to answer the question:

Step 1 **Analyze the question**

Step 2 **Identify up to 10 reference scales in the image, ranging from large
↪ to small sizes, and list them in the specified format**
- A reference scale must be typical in size.
- A reference scale can be the dimensions of an object or an object part.
- A reference scale must NOT be floor tiles or floor planks.
- Formulate the reference scales using the format: """The [choose from
↪ front-to-back, side-to-side, left-to-right, diameter, height (top to
↪ bottom edge), or mounting height (bottom edge to floor)] of [object or
↪ object part] is approximately [dimension estimate]."""

Step 3 **Propose a robust step-by-step plan to answer the question by using
↪ the reference scales in Step 2**
- A robust step-by-step plan performs the estimation in a coarse-to-fine
↪ manner.
  - First, use a reliable and large-sized reference scale as the primary
  ↪ reference for estimation.
  - Then, gradually use a reliable and smaller-sized reference scale for
  ↪ adjustment.
  - Repeat until the estimation is precise enough.
- When performing visual comparison, be aware of perspective distortion.
- Do NOT rely on pixel measurements from the images.

Step 4 **Focus on the image and follow the plan in Step 3 to answer the
↪ question**
Assistant: [OUTPUT]
```

Figure 8: **Text prompt in `SpatialPrompt-Steps`**. The text prompt is used in Sec. 5

```
User: You are an intelligent agent that can help me. You will be given a distance or a surface area spcified in an image and based on this information, you need to generate a visual question answer pair.

For example,
User input: ("the minimum distance between the red mug and the blue fork", "30 cm")
Output: ("What is the minimum distance between the red mug and the blue fork in the image? Answer in centimeters.", "30")

User input: ("the available surface area on the floor", "400 sqft")
Output: ("What is the available surface area on the floor? Answer in sqft", "400")

User input: ("the available surface area on the kitchen counter", "2 meters in length and 40 cm in depth")
Output: ("What is the available surface area on the kitchen counter? Answer by square meters", "0.8 square meter")

Now it's your turn,
User input: ("[DISTANCE DESCRIPTION]", "[DISTANCE VALUE]")
Output:
Assistant: [OUTPUT]
```

Figure 9: **Text prompt used to convert an annotated triplet (image, distance, description) to a question-answering pair.** The text prompt is used to convert a triplet of an image, a distance, and a description of the distance to a question-answering pair. It is used in Sec. 3.2.

(A) (B) (C)

Figure 10: **Sampled images from Q-Spatial++**

```
[Image A in Figure 10]
Question: What is the minimum distance between the two orange chairs next to
↪  the window?
Ground truth: 47 centimeters

# Response with Standard prompt
Assistant: The minimum distance between the two orange chairs next to the
↪  window appears to be very small, possibly around a few centimeters, as
↪  they are placed quite close to each other. However, without a reference
↪  for scale or the ability to measure directly, it's difficult to provide an
↪  exact measurement. Based on visual estimation, the distance might be
↪  around (\scalar{5} \distance_unit{cm}).

# Response with Zero-shot CoT
Assistant: To estimate the minimum distance between the two orange chairs next
↪  to the window, we can use visual cues from the image. The chairs are
↪  positioned close to each other, with a small table between them. The table
↪  appears to be narrow, suggesting that the chairs are not far apart.

Given the typical width of such small tables, which is usually around 50 to 60
↪  centimeters, and considering the slight gap between the edge of the table
↪  and the chairs, we can estimate the minimum distance between the chairs.
↪  Assuming each gap is about 5 to 10 centimeters, the total distance between
↪  the two chairs, including the table and gaps, would likely be around 60 to
↪  80 centimeters.

Therefore, the minimum distance between the two orange chairs is approximately
↪  (\scalar{70} \distance_unit{cm}).

# Response with SpatialPrompt
Assistant: To estimate the minimum distance between the two orange chairs next
↪  to the window, we can use the black table between them as a reference
↪  object. The table appears to be a standard small side table, which
↪  typically measures around 50 cm in diameter.

Observing the image, the table fits snugly between the two chairs, suggesting
↪  that the distance between the chairs is slightly more than the diameter of
↪  the table, to accommodate the table without pressing against the chairs.
↪  Therefore, we can estimate the minimum distance between the chairs to be
↪  slightly more than 50 cm.

Considering the visual information and typical furniture dimensions, the
↪  minimum distance between the two orange chairs is approximately
↪  (\scalar{55} \distance_unit{cm}).
```

Figure 11: **GPT-4V responses with different text prompts.**

```
[Image B in Figure 10]
Question: What is the minimum distance between the coffee grinder and the blue
↪  cloth in the image?
Ground truth: 15 centimeters

# Reponse with Standard prompt
It is impossible to determine the distance between the coffee grinder and the
↪  blue cloth from a 2D image.  A 2D image does not provide depth
↪  information.  \scalar{} \distance_unit{}

# Reponse with Zero-shot CoT
It is impossible to measure the distance in 3D space from a 2D image.  We do
↪  not know how far away the camera is from the coffee grinder and the blue
↪  cloth.  We also do not know the angle at which the picture was taken.  All
↪  of these factors would affect the perceived distance between the two
↪  objects.  Without more information, we cannot answer this question.

# Response with SpatialPrompt
Step 1 **Analyze the question**

We need to find the shortest distance between the coffee grinder and the blue
↪  cloth. This is the perpendicular distance from the front edge of the
↪  grinder's base to the edge of the cloth.

Step 2 **Identify up to 10 reference scales in the image, ranging from large
↪  to small sizes, and list them in the specified format**

1. """The height (top to bottom edge) of the coffee grinder is approximately
↪  10 inches."""
2. """The diameter of the kettle's base is approximately 8 inches."""
[text ignored]

Step 3 **Propose a robust step-by-step plan to answer the question by using
↪  the reference scales in Step 2**
[text ignored]

Step 4 **Focus on the image and follow the plan in Step 3 to answer the
↪  question**

1. The distance between the coffee grinder and the cloth appears to be
↪  slightly less than half the height of the coffee grinder. This suggests an
↪  initial estimate of around 4-5 inches.
2. Comparing the distance to the kettle's base diameter, it seems closer to 4
↪  inches than 5 inches.

Therefore, the minimum distance between the coffee grinder and the blue cloth
↪  is approximately  \scalar{4} \distance_unit{inches}.
```

Figure 12: **Gemini 1.5 Pro responses with different text prompts.**

```
[Image C in Figure 10]
Question:What is the minimum distance between the garbage bin and the stool in
↪  the image?
Ground truth: 7 centimeters

# Reponse with Standard prompt
To determine the minimum distance between the garbage bin and the stool in the
↪  image, I will estimate based on the visual perspective and relative sizes
↪  of the objects.

The garbage bin and the stool appear to be relatively close to each other.
↪  Given the perspective and the floor tiles as a reference, the minimum
↪  distance between the garbage bin and the stool is approximately:

\scalar{30} \distance_unit{cm}
```

Figure 13: **Common failure cases of GPT-4o**