Sparkle: Mastering Basic Spatial Capabilities in Vision Language Models Elicits Generalization to Spatial Reasoning

Yihong Tang[1,♡], **Ao Qu**[2,✉,♡], **Zhaokai Wang**[3,♡], **Dingyi Zhuang**[2,♡], **Zhaofeng Wu**[2]
Wei Ma[4], **Shenhao Wang**[5], **Yunhan Zheng**[2], **Zhan Zhao**[6], **Jinhua Zhao**[2]

[1]McGill University [2]Massachusetts Institute of Technology [3]Shanghai Jiao Tong University
[4]The Hong Kong Polytechnic University [5]University of Florida [6]The University of Hong Kong

yihong.tang@mail.mcgill.ca, qua@mit.edu, wangzhaokai@sjtu.edu.cn, dingyi@mit.edu

Abstract

Vision-language models (VLMs) excel in many downstream tasks but struggle with spatial reasoning, which is crucial for navigation and interaction with physical environments. Specifically, many spatial reasoning tasks rely on fundamental two-dimensional (2D) capabilities, yet our evaluation shows that state-of-the-art VLMs often produce implausible or incorrect solutions for composite spatial problems, including simple pathfinding tasks that humans solve effortlessly at a glance. To address this, we explore an effective approach to enhance 2D spatial reasoning in VLMs by training them solely on basic spatial capabilities. We first disentangle 2D spatial reasoning into three core components: direction comprehension, distance estimation, and localization. Our central hypothesis is that mastering these basic capabilities will significantly boost performance on more complex spatial tasks requiring advanced reasoning and combinatorial problem-solving, as well as generalize to real-world visual-spatial scenarios. To test this hypothesis, we introduce *Sparkle*, a framework that generates synthetic data to provide targeted supervision for VLMs across these three basic spatial capabilities, producing an instruction dataset for each capability. Our experiments demonstrate that VLMs fine-tuned with Sparkle achieve substantial improvements, not only on basic tasks but also in generalizing to composite and out-of-distribution real-world spatial reasoning tasks. These findings highlight that enhancing basic spatial capabilities through synthetic generalization effectively improves complex spatial reasoning, offering insights into systematic strategies for boosting VLMs' spatial understanding.

1 Introduction

Vision language models (VLMs) [OpenAI, 2023; Liu *et al.*, 2023b; Chen *et al.*, 2024c; Hong *et al.*, 2024; Wang *et al.*, 2023] have demonstrated near-human performance in tasks like image captioning [Chen *et al.*, 2015], visual question answering (VQA) [Goyal *et al.*, 2017; Singh *et al.*, 2019] and abundant downstream tasks by combining visual and text inputs to reason about the physical world. However, these models exhibit significant limitations in understanding spatial relationships. For instance, as shown in Figure 1, state-of-the-art (SoTA) VLMs GPT-4o and InternVL2-Pro [OpenAI, 2023; Chen *et al.*, 2024c] generate implausible responses to a shortest path problem that a human could solve at a glance, a simple 2D spatial reasoning task.

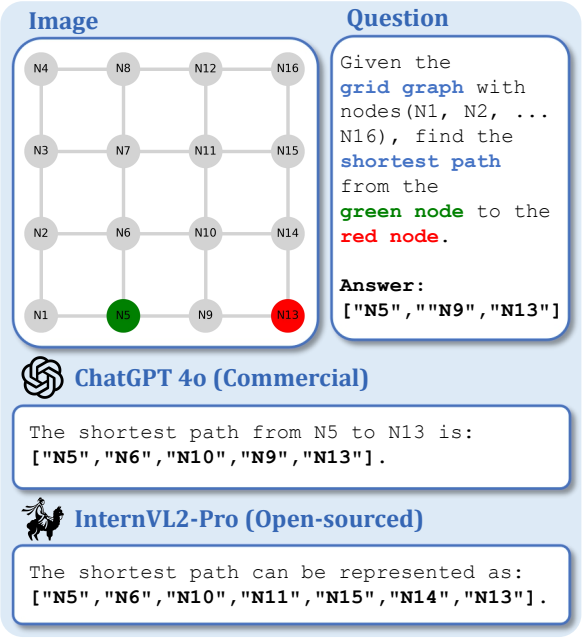

Figure 1: SOTA VLMs fail to solve the pathfinding problem, a simple 2D spatial reasoning task.

Nevertheless, 2D spatial reasoning is essential for VLMs to understand and interact with the physical environments, shaping their ability to solve mazes [Ivanitskiy *et al.*, 2023; Wang *et al.*, 2024], plan routes [Feng *et al.*, 2024; Chen *et al.*, 2024b], and solve geometric problems like humans [Fernandes and de Oliveira, 2009]. These tasks emphasize 2D spatial reasoning, requiring VLMs to process and navigate flat visual planes, interpret spatial relationships, and make deci-

♡ Equal contribution. ✉ Corresponding author.

sions based on geometric understanding. Such capabilities are fundamental in translating visual input into actionable insights. While more and more VLMs are developed with larger training datasets and extensive benchmarks [Ge *et al.*, 2024; Zhang *et al.*, 2024], the focus on enhancing spatial reasoning has received comparatively less attention, despite its importance to the core capabilities of VLMs.

In this paper, we study VLMs' spatial reasoning capabilities in a 2D space by investigating three key questions: (1) How well do existing models perform on 2D spatial reasoning? (2) What fundamental tasks affect spatial reasoning capabilities in 2D? (3) Can mastering basic tasks help improve composite and real-world spatial reasoning?

We begin by providing a systematic breakdown of 2D spatial reasoning, grounded in the principles of coordinate systems that represent 2D space. From this analysis, we identified three basic capabilities fundamental for spatial reasoning in 2D space: direction comprehension, distance estimation, and localization. A systematic evaluation of the performance of existing open-source and closed-source VLMs on these three basic capabilities reveals that even the most advanced VLMs sometimes struggle with these fundamental tasks. For instance, in a simple 2D direction classification task, where a model is asked to determine the relative direction (top left, top right, bottom left, bottom right) of one object relative to another on a straightforward diagram with only two objects, the state-of-the-art VLM GPT-4o can achieve only 76.5% accuracy. In contrast, a human should answer these questions correctly with little effort.

Most real-world spatial reasoning tasks, such as pathfinding [Lester, 2005; Cui and Shi, 2011], inherently require the composition of the basic capabilities identified above. A composite task is often subject to specific constraints that necessitate tailored solutions, unlike improving basic spatial reasoning capabilities, which can exhibit generalizability. In order to effectively improve the model's overall spatial reasoning capabilities in 2D space, we raise a conjecture: whether a VLM that masters the three basic capabilities can generalize and perform better on more complex composite spatial tasks. In other words, can a VLM exhibit compositional generalizability [van Zee, 2020] in spatial reasoning tasks?

To test this, we propose Sparkle, which stands for **SPA**tial **R**easoing through **K**ey capabi**L**ities **E**nhancement. This framework fine-tunes VLMs on these three basic spatial capabilities by programmatically generating synthetic data and providing supervision to form an instruction dataset for each capability. Additionally, Sparkle creates simplified visual representations to reduce recognition errors, allowing us to focus specifically on enhancing and evaluating VLMs' spatial capabilities. Our experimental results show that models trained on Sparkle achieve significant performance gains, not only in the basic tasks themselves (e.g., improving from 35% to 83% for InternVL2-8B on direction comprehension) but also in generalizing to composite and out-of-distribution general spatial reasoning tasks (e.g., improving from 13.5% to 40.0% on the shortest path problem). Additionally, our ablation study confirms the importance of mastering all three basic spatial reasoning capabilities. To summarize, our contributions are:

- We show that existing VLMs struggle with spatial reasoning tasks that humans solve effortlessly.
- We identify three basic spatial reasoning components and propose the Sparkle framework to improve these three fundamental spatial reasoning capabilities.
- Our experiments prove Sparkle's effectiveness in significantly enhancing the basic spatial capabilities of VLMs, with strong generalizability to out-of-distribution composite and real-world spatial reasoning tasks.

2 Related Work

2.1 Vision Language Models and Applications

Early works on VLMs, such as CLIP [Radford *et al.*, 2021] and ALIGN [Jia *et al.*, 2021], leveraged contrastive learning to align visual and textual embeddings in a shared latent space, demonstrating strong capabilities in linking visual content with corresponding natural language descriptions. With the rapid advancement of Large Language Models (LLMs), modern VLMs increasingly combine pretrained vision models [Dosovitskiy *et al.*, 2021; Chen *et al.*, 2023b] with powerful LLMs [Chiang *et al.*, 2023; Bai *et al.*, 2023a; Jiang *et al.*, 2023; Cai *et al.*, 2024] to facilitate a more cohesive understanding of both modalities [Liu *et al.*, 2023b; Bai *et al.*, 2023a; Chen *et al.*, 2024c]. This approach enables richer visual reasoning, open-ended image captioning, and more interactive multimodal dialogue systems.

VLMs have been applied in various pre-training tasks, such as image-text matching, masked image modeling, and multimodal reasoning [Li *et al.*, 2022; Li *et al.*, 2023b; Wang *et al.*, 2022b]. In downstream tasks, they excel in applications like visual question answering [Antol *et al.*, 2015; Wang *et al.*, 2022a], image captioning [Li *et al.*, 2020; Sidorov *et al.*, 2020; Wang *et al.*, 2021], image generation based on textual prompts [Ramesh *et al.*, 2022; Baldridge *et al.*, 2024], and aiding human-machine interactions in complex real-world settings, showcasing their versatility and potential across a broad range of vision language applications.

2.2 Spatial Reasoning in LLMs and VLMs

Spatial reasoning in LLMs involves understanding and manipulating spatial relationships described in text. Early work focused on extracting spatial information from natural language [Hois and Kutz, 2011; Kordjamshidi *et al.*, 2011]. Recent efforts emphasize improving multi-hop spatial reasoning [Li *et al.*, 2024b], especially in complex scenarios like 2D visual scenes [Shi *et al.*, 2022]. Methods include pretraining on synthetic datasets to better capture spatial patterns [Mirzaee *et al.*, 2021], and using in-context learning to generalize spatial reasoning across tasks, such as transforming spatial data into logical forms or visualizing reasoning trace [Yang *et al.*, 2023b; Wu *et al.*, 2024; Tang *et al.*, 2024].

Building on these foundations, VLMs extend spatial reasoning by integrating visual inputs and often implicitly encode spatial knowledge through large-scale pretraining on visual-text datasets [Radford *et al.*, 2021; Li *et al.*, 2023c]. Early studies on VLMs primarily focus on understanding spatial relationships between objects in front-view images [Liu *et al.*, 2023a], laying the groundwork for 2D spatial reasoning. More

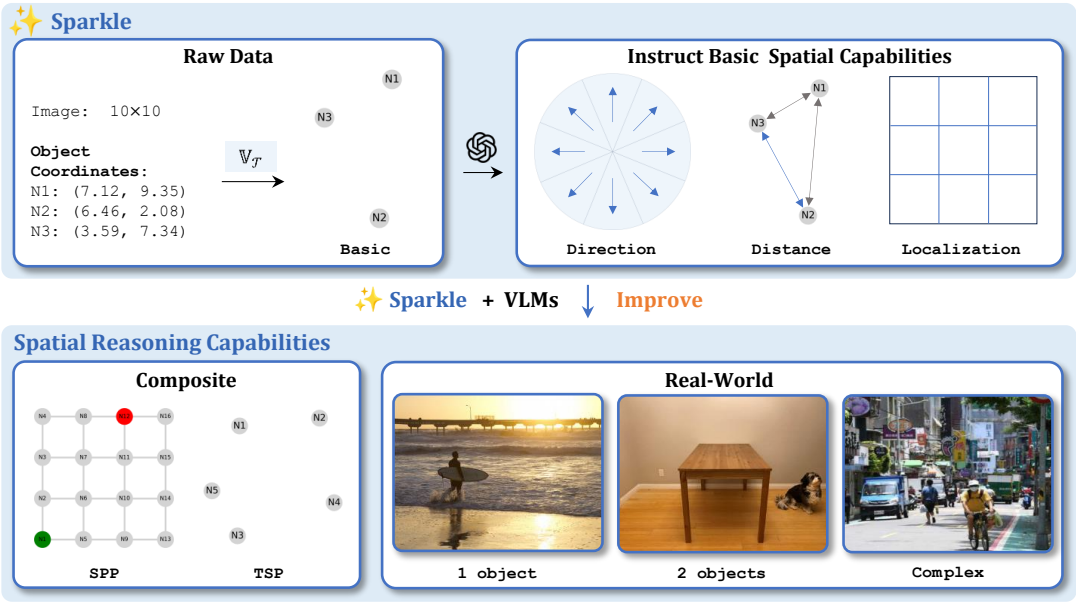

Figure 2: The proposed Sparkle framework.

recently, research on VLMs has expanded to 3D reasoning tasks, which introduce additional challenges such as depth estimation [Chen *et al.*, 2024a] and path planning [Chen *et al.*, 2024b; Deng *et al.*, 2020], as seen in applications like robotic grasping [Xu *et al.*, 2023] and navigation [Shah *et al.*, 2023; Chiang *et al.*, 2024] in the embodied AI field [Li *et al.*, 2024c]. Despite these advances, 2D spatial reasoning remains more fundamental and flexible, as it can be applied to various tasks, including VQA [Ge *et al.*, 2024; Kamath *et al.*, 2023; Li *et al.*, 2024a] and user interface grounding [Rozanova *et al.*, 2021]. Due to its broad applicability and foundational role, this work focuses on exploring 2D spatial reasoning capabilities within VLMs.

3 Methodology

In order to systematically evaluate and enhance the spatial reasoning capabilities of VLMs in 2D environments, we introduce the Sparkle framework, as illustrated in Figure 2. This section is structured as follows:

- Disentangling basic elements: How we identified the basic spatial capabilities of 2D spatial reasoning, and why these elements are foundational.
- Sparkle framework: We present the Sparkle framework to enhance VLMs' performances in 2D spatial reasoning by systematically improving the identified basic spatial capabilities.
- Tasks: We employ three spatial reasoning tasks specifically designed to evaluate both basic and composite spatial reasoning capabilities of VLMs.

3.1 Disentangling Spatial Reasoning

2D spaces are usually represented by coordinate systems, which provide a structured way to describe objects' positions and spatial relationships within a plane [Byrne and Johnson-Laird, 1989]. These systems rely on core principles to articulate an object's position: direction defines orientation, distance represents magnitude, and localization integrates both to precisely describe a location [Just and Carpenter, 1985]. Building on these principles and characteristics of 2D spaces, we identify three foundational components of 2D spatial reasoning: (1) *Direction Comprehension*: The ability to understand the orientation of an object relative to a reference object; (2) *Distance Estimation*: The ability to measure the magnitude of spatial displacement between objects; (3) *Localization*: The ability to determine the precise position of an object in space.

Cognitively, Freksa [Freksa, 1991] identifies orientation, proximity, and the spatial arrangement of objects as universally useful conceptual properties for spatial reasoning. Frank [Frank, 1992] also adopted a similar decomposition to study human reasoning about space and spatial properties. The conceptual neighborhoods theory [Freksa, 1991] demonstrates that simpler conceptual distinctions naturally generalize to broader reasoning contexts (our hypothesis). These evidences support the disentangled spatial capabilities form the foundation of 2D spatial reasoning, offering essential elements required to fully describe, understand, and reason about an object's position and relationships with other objects within a 2D space. This decomposition enables a systematic and comprehensive evaluation of spatial reasoning by disentangling these basic spatial capabilities, enhancing specificity in assessing spatial reasoning capabilities in VLMs.

3.2 Sparkle

To comprehensively investigate our hypothesis, we introduce Sparkle, a simple yet effective framework for constructing an instruction dataset focused on enhancing a model's spatial reasoning abilities. This framework only improves VLMs' ba-

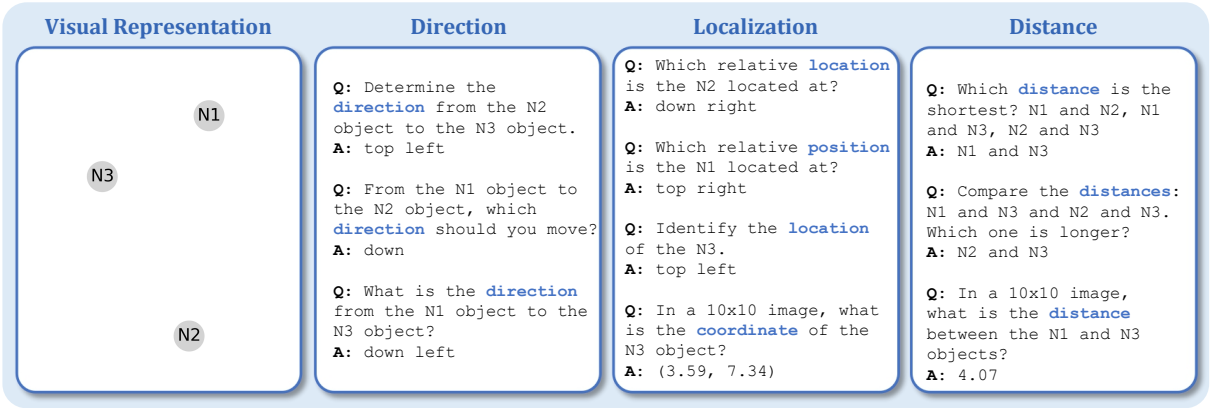

Figure 3: An instruction data sample from Sparkle.

sic spatial capabilities, and this design enables us to evaluate whether models that perform well on basic spatial reasoning tasks can also excel in more complex and composite problems.

Instruction Data Generation

The design of our instruction dataset focuses on three basic spatial capabilities: direction, distance, and localization, based on insights provided in Section 3.1. The proposed finetuning pipeline does not require manual labeling, as all data can be programmatically generated.

We use $\mathbb{G}$ to denote a data generator that can generate a set of objects, $P = \{N_i\}_{i=1}^{n}$, representing a training sample of basic spatial capabilities. Each object $N_i = (x_i, y_i) \in \mathbb{R}^2$ consists of randomly sampled coordinates within a bounded region. For each basic capability $\mathcal{T} \in \{\text{dir., dist., loc.}\}$, we construct a dataset $D_{\mathcal{T}}$ containing input-output pairs $(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}})$, where $\mathcal{X}^{\mathcal{T}}$ represents the inputs and $\mathcal{Y}^{\mathcal{T}}$ represents the corresponding ground truth outputs. Each input $\mathcal{X}^{\mathcal{T}}$ consists of: (1) A visual input $\mathcal{X}_V^{\mathcal{T}}$: A labeled diagram representing the spatial configuration of a sample of objects through a visual representation function $\mathbb{V}_{\mathcal{T}}(P)$, (2) A language prompt $\mathcal{X}_L^{\mathcal{T}}$: A question querying some aspects of the spatial properties for P.

For example, to craft a training sample for direction comprehension, two objects, N_1 and N_2, are selected from P, and a question such as "What is the direction of N_2 relative to N_1?" is posed. The corresponding correct answer $Y^{\mathcal{T}}$ can be easily computed since we can access the exact coordinates of these objects, e.g., we can obtain the answer to the above question by calculating the vector from N_1 to N_2 based on their coordinates and map it to the corresponding directional label. Details about these generation processes can be found in Appendix §A.

The resulting training dataset consists of these generated questions and answers, paired with the corresponding visual representations, as shown in Figure 3. Specifically, the training pairs are represented as $\{(\mathcal{X}_L^{\text{train}}, \mathcal{X}_V^{\text{train}}, \mathcal{Y}^{\text{train}})\}$, where $\mathcal{X}_L^{\text{train}}$ represents the language-based queries, $\mathcal{X}_V^{\text{train}}$ represents the visual representations, and $\mathcal{Y}^{\text{train}}$ represents the corresponding answers. We provide a complete training sample from Sparkle in Appendix §D.1.

Instruction Finetuning for Basic Tasks

To enhance the spatial reasoning capabilities of VLMs, we use the Sparkle training set, denoted as $\mathcal{X}^{\text{train}} = \{(\mathcal{X}_L^{\text{train}}, \mathcal{X}_V^{\text{train}})\}$. The objective is to minimize the negative log-likelihood of the predicted answers. Specifically, the loss function $\mathcal{L}$ is defined as:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(\mathcal{X}^{\text{train}}, \mathcal{Y}^{\text{train}})} \left[\log p(\mathcal{Y}^{\text{train}} \mid \mathcal{X}_V^{\text{train}}, \mathcal{X}_L^{\text{train}}; \theta)\right]$$

where θ represents the parameters of the VLM. The training aims to improve the model's proficiency in basic spatial reasoning tasks, which subsequently allows for evaluation of its performance on more complex spatial challenges.

3.3 Tasks

The goal of the employed tasks is to evaluate the 2D spatial reasoning capabilities of VLMs and provide a foundation for studying how acquiring basic spatial capabilities can enhance performance on complex tasks. To achieve this, we follow key design criteria: (1) focus on spatial reasoning, and (2) progression from basic to composite tasks.

Basic Tasks

As shown in Figure 4 (left), the basic tasks in Sparkle are designed to assess the model's understanding of three basic spatial capabilities: (1) direction comprehension, (2) distance estimation, (3) localization.

In each basic task, the VLM is provided with an image containing several labeled data objects and a multiple-choice question about the spatial properties of these objects, with the goal of having the model answer these questions correctly. We first generate labeled diagrams that serve as visual inputs, then generate the questions (in multiple-choice format) and corresponding answer pairs, similar to the process in Section 3.2, to obtain the basic task test set.

Composite Tasks

Building on the basic spatial relationships, the composite tasks introduce greater complexity. The objective is to assess whether the model can apply basic spatial skills to solve problems requiring a combination of these skills or whether it has merely learned each skill in isolation without being able to

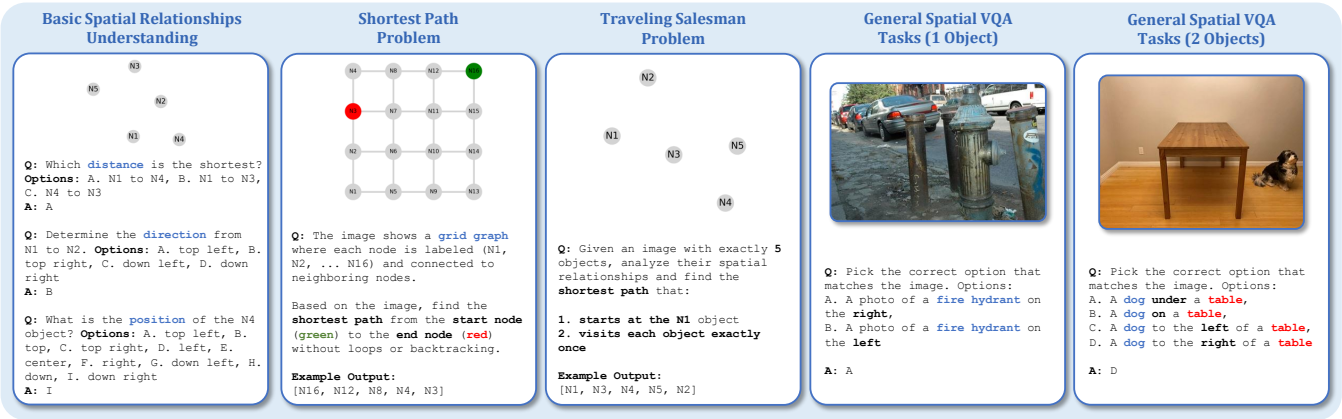

Figure 4: Evaluation samples used in our experiments.

generalize effectively. We choose the *Shortest Path Problem (SPP)* and the *Traveling Salesman Problem (TSP)* as composite tasks for evaluation.

Shortest Path Problem (SPP) SPP evaluates the ability to compute the most efficient route between two objects on a 2D grid, requiring a combination of distance estimation and spatial planning.

Consider a grid G of size $n \times n$, with two special objects: the start object $N_{\text{start}} = (x_s, y_s)$ and the end object $N_{\text{end}} = (x_e, y_e)$. We employ a language model LM generates the prompt $\mathcal{X}_L^{\text{spp}}$ using a predefined prompt template $\mathbb{P}_{\text{spp}}$, expressed as: $\mathcal{X}_L^{\text{spp}} = \text{LM}(\mathbb{P}_{\text{spp}}(G, N_{\text{start}}, N_{\text{end}}))$. The visual input is produced similar to basic tasks: $\mathcal{X}_V^{\text{spp}} = \mathbb{V}_{\text{spp}}(G, N_{\text{start}}, N_{\text{end}})$.

The combined input for the VLM is $\mathcal{X}^{\text{spp}} = (\mathcal{X}_V^{\text{spp}}, \mathcal{X}_L^{\text{spp}})$, and the model is expected to predict the shortest path $\hat{\mathcal{Y}}^{\text{spp}}$, which is evaluated against the true shortest path, $\mathcal{Y}^{\text{spp}}$, computed using standard algorithms.

Traveling Salesman Problem (TSP) As shown in Figure 4 (middle), the TSP represents a more challenging spatial reasoning task, involving combinatorial optimization. The model must find the shortest possible route that visits each object exactly once and returns to the starting object.

Given n objects $P^{\text{tsp}} = \{N_i\}_{i=1}^n$ sampled from $\mathbb{G}$, the ground truth solution $\mathcal{Y}^{\text{tsp}}$ is computed using a TSP solver $\mathbb{M}_{\text{tsp}}(P^{\text{tsp}})$. Similarly, the input to VLMs consists of a visual representation $\mathcal{X}_V^{\text{tsp}} = \mathbb{V}_{\text{tsp}}(P^{\text{tsp}})$ and a corresponding language prompt $\mathcal{X}_L^{\text{tsp}}$. The complete input query is $\mathcal{X}^{\text{tsp}} = (\mathcal{X}_V^{\text{tsp}}, \mathcal{X}_L^{\text{tsp}})$. Similarly, the model's predicted order of visiting all objects, $\hat{\mathcal{Y}}^{\text{tsp}}$, is then evaluated against the ground truth solution $\mathcal{Y}^{\text{tsp}}$.

Discussion Given that the SPP can be solved in polynomial time, we expect that if the model can effectively combine its knowledge of basic spatial concepts, it will show significant improvements in solving this task efficiently. On the other hand, the TSP is an NP-hard problem, requiring combinatorial optimization to obtain the exact solution. We include the TSP to push the limits of the model's spatial reasoning capabilities, aiming to investigate how well the model can manage more complex problem-solving tasks beyond the basic integration of spatial skills.

General Visual-Spatial Tasks
Sparkle uses simplified visual representations to focus on improving the spatial reasoning abilities of vision-language models (VLMs). The goal is for these enhanced spatial capabilities to generalize across different visual distributions. To evaluate this, we incorporate general visual-spatial tasks with real-world images from standard VQA datasets.

4 Experiments

In this section, we provide our findings and results to demonstrate the effectiveness of the Sparkle framework. Specifically, the experiments are designed to answer the following research questions: **RQ1**: Can mastering basic 2D spatial components enhance overall spatial reasoning capability in VLMs? **RQ2**: What insights from the results of evaluations (Section 4.2), enhancements (Section 4.2), and spatial components (Section 4.3) can guide improvements in model design, training strategies, and data collection for spatial reasoning in VLMs?

4.1 Settings

Models We tested open-source and commercial models to evaluate and enhance VLMs' spatial reasoning capabilities. For commercial VLMs, we used GPT-4o from OpenAI [Yang *et al.*, 2023a] and Google-Gemini [GeminiTeam *et al.*, 2023]. We included LLaVA1.6 [Liu *et al.*, 2024], Qwen-VL [Bai *et al.*, 2023b], ChatGLM-4V [GLM *et al.*, 2024], MiniCPM-V2.5 [Yao *et al.*, 2024] and InternVL2 [Chen *et al.*, 2024c] for open-source models. For all adopted tasks, we report accuracy as the evaluation metric. We use the MS-Swift library [Zhao *et al.*, 2024] and apply the LoRA [Hu *et al.*, 2022] fine-tuning strategy, with low-rank dimension of 32. We set a constant learning rate of 1e-4 and a batch size of 1. All training and evaluation tasks are performed on GPU clusters with 8 × NVIDIA A100 machines. Further details can be found in Appendix §A.

Table 1: VLM performance on spatial reasoning tasks before and after Sparkle enhancement. Δ indicates the relative improvement.

Model	Basic Tasks			Composite Tasks				General Tasks				
	Loc.	Dist.	Dir.	SPP		TSP		What's Up	COCO-Spatial		GQA-Spatial	
				4Grid	5Grid	4Obj	5Obj		1Obj	2Obj	1Obj	2Obj
GPT-4o	68.2	43.2	77.2	75.2	76.2	23.4	21.5	95.9	88.2	49.7	89.4	63.6
Gemini	61.4	41.2	56.2	66.4	64.2	14.3	16.4	69.4	50.8	34.1	42.9	21.7
LLaVA1.6-7B	25.2	37.3	30.8	1.7	0.9	12.0	4.0	44.9	82.3	68.5	82.7	80.4
+ Sparkle	40.7	57.2	75.9	6.2	2.3	15.8	6.8	51.4	86.8	84.2	92.3	84.1
Δ	+61.5%	+53.4%	+146.4%	+264.7%	+155.6%	+31.7%	+70.0%	+14.5%	+5.5%	+22.9%	+11.6%	+4.6%
Qwen-VL-7B	25.0	37.6	24.4	2.2	1.2	11.7	3.7	42.7	89.8	74.3	98.5	94.0
+ Sparkle	59.6	61.3	64.8	5.4	4.6	18.4	12.0	49.6	96.8	87.1	99.0	96.4
Δ	+138.4%	+63.0%	+165.6%	+145.5%	+283.3%	+57.3%	+224.3%	+16.2%	+7.8%	+17.2%	+0.5%	+2.6%
ChatGLM-4V-8B	49.7	45.7	41.6	15.8	8.7	9.8	4.4	96.4	75.9	66.5	78.3	75.5
+ Sparkle	72.6	70.3	67.9	36.3	17.1	20.4	8.6	98.4	85.4	82.9	90.5	81.8
Δ	+46.1%	+53.8%	+63.2%	+129.7%	+96.6%	+108.2%	+95.5%	+2.1%	+12.5%	+24.7%	+15.6%	+8.3%
MiniCPM-Llama3-V2.5-8B	42.5	26.2	44.2	16.4	11.4	14.7	4.2	76.2	70.1	73.1	80.3	53.3
+ Sparkle	66.0	82.0	79.6	31.9	14.0	17.2	13.9	80.2	88.0	88.7	91.8	79.4
Δ	+55.3%	+213.0%	+80.1%	+94.5%	+22.8%	+17.0%	+231.0%	+5.2%	+25.5%	+21.3%	+14.3%	+49.0%
InternVL2-8B	61.3	44.2	34.6	15.4	13.9	17.1	9.6	92.7	92.5	71.3	97.5	85.3
+ Sparkle	74.4	83.8	83.2	38.8	39.0	21.6	14.4	94.9	94.2	78.7	99.0	90.4
Δ	+21.4%	+89.6%	+140.5%	+151.9%	+180.6%	+26.3%	+50.0%	+2.3%	+1.8%	+10.4%	+1.5%	+6.0%

Data We built the Sparkle training dataset by generating 10K images, each with 17 instruction–answer pairs that describe the spatial relationships between objects, resulting in a total of 170K samples. Among these pairs, 3 focus on directions between objects, 7 on distances (including 4 for comparing distances and 3 for estimating numerical distances), and 6 on localization (with 3 for identifying object locations and 3 for estimating exact positions). The final pair describes the overall spatial relationships in the image in natural language. This setup ensures that the VLM maintains its ability to follow instructions effectively. Numerical values are learned using a standard autoregressive cross-entropy loss, as is standard for grounding tasks in VLM training [Chen *et al.*, 2023a; Liu *et al.*, 2023c]. A complete sample can be found in Figure 10 in Appendix §D.1. Our evaluation includes tasks of: (1) shortest path problem (SPP), (2) traveling salesman problem (TSP), and (3) basic spatial relationship understanding. For each of them, we generated 2000 samples, which together make up the evaluation set. For SPP and TSP, we use LLaMA 3.1 [Dubey *et al.*, 2024] to process the VLMs' responses into list formats to enable metric computation. For the basic tasks, we structured them in a multiple-choice question format. In addition, for SPP and TSP, we designed experiments that vary by grid size and the number of objects involved. Detailed data statistics and sample data are provided in Appendix §B. To further assess the generalizability of the improved spatial reasoning capabilities, we evaluated VLMs on existing general spatial reasoning-related benchmarks to examine their out-of-distribution performance. The general benchmarks we used include What's Up [Kamath *et al.*, 2023], COCO-spatial [Lin *et al.*, 2014], and GQA-spatial [Hudson and Manning, 2019], which feature real-world images and spatial reasoning questions.

4.2 Main Results

Evaluation of Existing VLMs

From Table 1, we observe that even the state-of-the-art commercial VLMs cannot obtain satisfactory results on composite tasks like SPP and TSP. Open-source models achieve even worse performance (≤25% accuracy) on these tasks.

Specifically, LLaVA performs poorly particularly on SPP compared to TSP, which may be attributed to the grid data structure in SPP being more complex for VLMs to perceive, understand, and generate valid paths grounded on the grid compared to ordering just a few objects in TSP, indicating that these VLMs struggle with visual representations involving intricate spatial structures. Performance on the TSP task worsens as the number of objects increases across most models, highlighting the growing difficulty of spatial reasoning with more objects. However, in SPP, we discover that increasing the grid size has little impact on performance, indicating that a larger grid does not increase the difficulty of reasoning. This result aligns with our initial design principles, where SPP was intended to combine basic spatial understanding with straightforward spatial planning. For general tasks involving real-world images, VLMs still struggle to identify correct spatial relationships and perform spatial reasoning, leaving a significant gap compared to human capability. To delve into how VLMs behave poorly on spatial reasoning tasks, we further examine their performance on basic spatial relationship understanding, i.e. direction, location and localization comprehension. As shown in Table 1, even the state-of-the-art VLM GPT-4o struggles with basic spatial relationship understanding, achieving only 68.2%, 43.2%, and 77.2% accuracy on the direction, distance, and localization tasks, respectively. This investigation explains why VLMs underperform on composite and general tasks, as their weak basic spatial capabilities

directly hinder their ability to handle more complex spatial challenges.

Effectiveness of Sparkle

To demonstrate the effectiveness of Sparkle, we present results from fine-tuning all selected open-source VLMs using this method. The results reveal significant improvements in both basic and composite tasks, with generalized improvements to general tasks, indicating that 2D spatial reasoning capabilities can be significantly improved when a model masters the basic components of spatial reasoning. When combining these enhanced spatial abilities with the inherent generalizability of VLMs, the performance gains can be effectively extended to complex spatial reasoning tasks in real-world image domains. Specifically, Sparkle only contains instructions for basic spatial relationship understanding. However, after fine-tuning with this data, VLMs improved in basic spatial reasoning (around 90%) and showed significant gains (around 120%) in composite tasks and general tasks (around 12%). This justifies that improving these basic spatial reasoning capabilities could effectively enhance VLMs' overall spatial reasoning, enabling them to tackle more complex tasks and comprehend more sophisticated visual representations. This outcome also justifies the rationality of adopting a simplified visual representation, with the hope of helping VLMs acquire inherent spatial reasoning capabilities that can transfer to more complex visual representations (e.g., real-world images).

It is worth noting that the TSP involves more complex spatial reasoning than the SPP. However, VLMs find the SPP more challenging because their outputs must be precisely aligned with the grid. In contrast, solving the TSP only requires determining the optimal order of objects. When comparing the improvements of VLMs on SPP and TSP, we observe that the gains (around 90%) on TSP are much smaller than those on the SPP task (150%). One possible explanation is that the TSP involves more complex optimization challenges, which may not be as easily addressed by simply improving basic spatial reasoning skills, as discussed in Section 3.3. This underscores the need for further research into the optimization capabilities of language models, a topic we hope our findings will inspire.

Generalizability

In the previous subsection, we have shown that spatial reasoning improvements can generalize from simple tasks to more complex ones. In this section, we evaluate this generalization further by testing spatial reasoning performance in an out-of-distribution visual setting to assess whether these enhanced capabilities extend to broader VLM spatial tasks.

Specifically, we explore whether the enhanced spatial reasoning capabilities transfer to other general VLM spatial tasks. As shown in Table 1 , there are consistent gains across general VLM benchmarks related to spatial reasoning. For instance, the COCO-spatial and GQA-spatial benchmarks illustrate that current VLMs often struggle to accurately capture spatial relationships between two objects. With our Sparkle framework, this capability is greatly improved. This generalized improvement demonstrates that Sparkle enhances the inherent spatial reasoning capabilities of VLMs, supporting the effectiveness of using simplified visual representations. These findings indicate that the Sparkle framework offers a simple yet powerful method for enhancing spatial reasoning capabilities in VLMs. Future VLM research could benefit from incorporating Sparkle's approach by decomposing spatial tasks into foundational skills and systematically improving them in pretraining and fine-tuning stages, thereby enhancing overall model performance on complex and general spatial reasoning tasks.

4.3 Ablation Studies

Impact of Training Components

Table 2: Random perturbation results for InternVL2-8B.

Perturbation	What's Up	COCO-1	COCO-2	GQA-1	GQA-2
Direction	85.4	90.9	62.4	96	76.8
Distance	90.5	91.4	64.6	96.4	78.8
Localization	87.6	89.6	65.8	94.6	80.5
N/A	94.9	94.2	78.7	99.0	90.4

To evaluate the impact of different training components, we first conduct random perturbation (i.e., perturb training labels randomly) to InternVL2-8B on each spatial capabilities to justify the derivation of disentangled spatial capabilities. As shown in Table 2, the VLM's performance degrades drastically after perturbation, confirming the critical role of each identified basic spatial component.

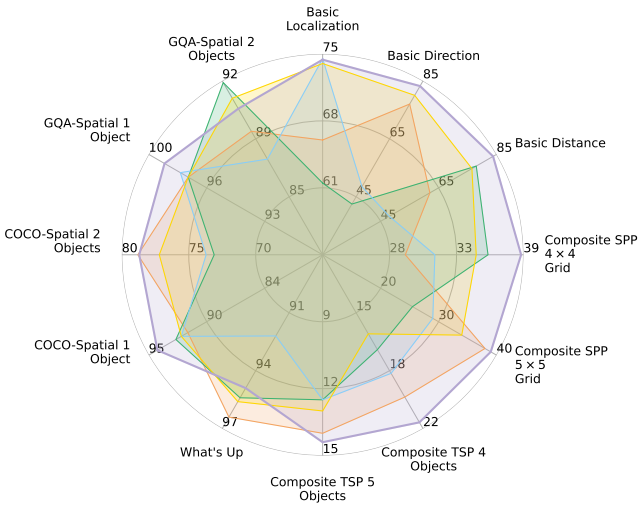

Figure 5: Ablations for Sparkle variants: Sparkle ; Sparkle without numerical information ; Sparkle (Localization) ; Sparkle (Distance) ; Sparkle (Direction) .

We also trained InternVL2-8B on individual spatial reasoning tasks with our Sparkle framework, resulting in *Sparkle(Direction, Distance, Localization)*. We also tested a version called *Sparkle w/o Num* that excludes numerical information (i.e., distance and location estimation) in Sparkle. All the four variants are trained with the same number of total samples as the full Sparkle model. The results shown in Figure 5 reveal two key insights: First, *Sparkle w/o Num* consistently underperforms compared to the full Sparkle model, particularly in tasks that require strong distance reasoning, such

as TSP. This suggests that incorporating numerical information during training significantly enhances the model's capability in tasks involving distance reasoning and other related composite challenges. Second, training on specific spatial reasoning subsets can sometimes yield optimal performance for certain tasks. For example, *Sparkle (Direction)* achieves 96.4% accuracy on the What's Up benchmark, indicating that task-specific training can be highly effective. This highlights the importance of tailoring the training process to the unique characteristics of individual tasks. When a task emphasizes a particular spatial reasoning capability, focusing the training data on that aspect can improve performance on the targeted task. Overall, the full Sparkle framework consistently delivers the best results across the majority of benchmarks, demonstrating the effectiveness of a more comprehensive approach to training.

Impact of Training Sample Size

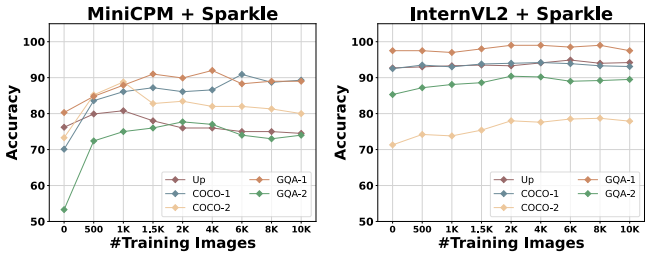

Figure 6: Results of Sparkle on InternVL2 and MiniCPM with varying training sample sizes.

We varied the training sample size in Sparkle and evaluated its impact on general spatial reasoning tasks. The results are shown in Figure 6. We observe a general improvement in VLM performance as the training sample size increases despite some fluctuations in the curve. However, a noteworthy finding is the existence of task-specific sweet spots, beyond which performance gains taper off or degrade.

Performances on Common VLM Benchmarks

Table 3: Performance on common VLM benchmarks.

Model	SEED-I		MME		BLINK		MMBench	
	All	SR	All	Pos	All	SR	All	SR
InternVL2	75.4	62.1	1641	143	50.3	80.4	82.4	46.7
+Sparkle	75.5	64.5	1644	151	52.3	81.1	83.2	53.3

While VLMs show significant improvements in spatial reasoning with Sparkle enhancement, we also evaluate them on common benchmarks. As shown in Table 3, Sparkle-trained models show substantial improvement in spatially related subdimensions, while maintaining or improving overall performance, demonstrating that Sparkle does not negatively affect the overall abilities of VLMs. This suggests that incorporating Sparkle into the pretraining process could further enhance these general capabilities.

4.4 Discussion

The analysis and results confirm that mastering basic 2D spatial reasoning capabilities through Sparkle can significantly enhance VLMs' overall spatial reasoning in composite tasks (e.g., spatial planning) and general spatial tasks. This directly addresses RQ1 and supports the assumption presented in the methods section.

Turning to RQ2, the evaluation results revealed the limitations of existing VLMs, particularly in their capability to perceive complex spatial structures, as evidenced in tasks like SPP. This highlights the need for improved model and training designs to support more detailed spatial reasoning. Moreover, introducing synthetic data focusing on basic spatial relationships has proven to enhance overall VLM spatial reasoning performance, offering a clear path for future spatial data collection. Lastly, our ablation study suggests that training specific spatial reasoning capabilities in isolation yields the best results for tasks that demand focused spatial abilities. Therefore, in terms of training strategy, our findings suggest adopting a pre-train and fine-tune approach (i.e., using diverse spatial data in pretraining and fine-tuning specific spatial capabilities tailored to particular tasks) to improve VLMs' performances on corresponding tasks.

5 Conclusion

We present the Sparkle framework to address the limited spatial reasoning ability of Vision Language Models (VLMs). It is designed to enhance spatial reasoning by focusing on three basic capabilities: direction, distance and localization. Experiments show that fine-tuning on these basic capabilities leads to substantial improvements in the basic tasks and composite tasks, showcasing its compositional generalizability. It also leads to generalization on broader tasks, strengthening VLMs' ability to interact with the physical world.

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[Bai *et al.*, 2023a] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv: 2309.16609*, 2023.

[Bai *et al.*, 2023b] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[Baldridge *et al.*, 2024] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

[Byrne and Johnson-Laird, 1989] Ruth MJ Byrne and Philip N Johnson-Laird. Spatial reasoning. *Journal of memory and language*, 28(5):564–575, 1989.

[Cai *et al.*, 2024] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, and et al. Internlm2 technical report. *arXiv: 2403.17297*, 2024.

[Chen *et al.*, 2015] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[Chen *et al.*, 2023a] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[Chen *et al.*, 2023b] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv: 2312.14238*, 2023.

[Chen *et al.*, 2024a] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[Chen *et al.*, 2024b] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9796–9810, 2024.

[Chen *et al.*, 2024c] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv:2404.16821*, 2024.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[Chiang *et al.*, 2024] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, et al. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. *arXiv preprint arXiv:2407.07775*, 2024.

[Cui and Shi, 2011] Xiao Cui and Hao Shi. A*-based pathfinding in modern computer games. *International Journal of Computer Science and Network Security*, 11(1):125–130, 2011.

[Deng *et al.*, 2020] Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. Evolving graphical planner: Contextual global planning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 33:20660–20672, 2020.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[Duan *et al.*, 2024] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong,

Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Feng *et al.*, 2024] Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. Citygpt: Empowering urban spatial cognition of large language models. *arXiv preprint arXiv:2406.13948*, 2024.

[Fernandes and de Oliveira, 2009] Leandro Augusto Frata Fernandes and Manuel Menezes de Oliveira. Geometric algebra: a powerful tool for solving geometric problems in visual computing. In *2009 Tutorials of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, pages 17–30. IEEE, 2009.

[Frank, 1992] Andrew U Frank. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing*, 3(4):343–371, 1992.

[Freksa, 1991] Christian Freksa. Qualitative spatial reasoning. In *Cognitive and linguistic aspects of geographic space*, pages 361–372. Springer, 1991.

[Fu *et al.*, 2023] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv: 2306.13394*, 2023.

[Fu *et al.*, 2024] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[Ge *et al.*, 2024] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making LLaMA SEE and draw with SEED tokenizer. In *The Twelfth International Conference on Learning Representations*, 2024.

[GeminiTeam *et al.*, 2023] GeminiTeam, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv: 2312.11805*, 2023.

[GLM *et al.*, 2024] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334, 2017.

[Hois and Kutz, 2011] Joana Hois and Oliver Kutz. Towards linguistically-grounded spatial logics. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2011.

[Hong *et al.*, 2024] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[Hudson and Manning, 2019] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019.

[Ivanitskiy *et al.*, 2023] Michael Igorevich Ivanitskiy, Rusheb Shah, Alex F Spies, Tilman Räuker, Dan Valentine, Can Rager, Lucia Quirke, Chris Mathwin, Guillaume Corlouer, Cecilia Diniz Behn, et al. A configurable library for generating and manipulating maze datasets. *arXiv preprint arXiv:2309.10498*, 2023.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, volume 139, pages 4904–4916, 2021.

[Jiang *et al.*, 2023] Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, et al. Think and retrieval: A hypothesis knowledge graph enhanced medical large language models. *arXiv preprint arXiv:2312.15883*, 2023.

[Just and Carpenter, 1985] Marcel A Just and Patricia A Carpenter. Cognitive coordinate systems: accounts of mental rotation and individual differences in spatial ability. *Psychological review*, 92(2):137, 1985.

[Kamath *et al.*, 2023] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023.

[Kordjamshidi *et al.*, 2011] Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):1–36, 2011.

[Lester, 2005] Patrick Lester. A* pathfinding for beginners. *online]. GameDev WebSite. http://www. gamedev. net/reference/articles/article2003. asp (Acesso em 08/02/2009)*, 2005.

[Li *et al.*, 2020] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference,*

Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, pages 121–137. Springer, 2020.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICLR*, volume 162, pages 12888–12900, 2022.

[Li *et al.*, 2023a] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv: 2307.16125*, 2023.

[Li *et al.*, 2023b] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202, pages 19730–19742, 2023.

[Li *et al.*, 2023c] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

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

[Li *et al.*, 2024b] Fangjun Li, David C Hogg, and Anthony G Cohn. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18500–18507, 2024.

[Li *et al.*, 2024c] Hao Li, Xue Yang, Zhaokai Wang, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. Auto mc-reward: Automated dense reward design with large language models for minecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16426–16435, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[Liu *et al.*, 2023a] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[Liu *et al.*, 2023c] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023.

[Liu *et al.*, 2023d] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv: 2307.06281*, 2023.

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

[Mirzaee *et al.*, 2021] Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. Spartqa:: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*, 2021.

[OpenAI, 2023] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763, 2021.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[Rozanova *et al.*, 2021] Julia Rozanova, Deborah Ferreira, Krishna Dubba, Weiwei Cheng, Dell Zhang, and Andre Freitas. Grounding natural language instructions: Can large language models capture spatial information? *arXiv preprint arXiv:2109.08634*, 2021.

[Shah *et al.*, 2023] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR, 2023.

[Shi *et al.*, 2022] Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11321–11329, 2022.

[Sidorov *et al.*, 2020] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020.

[Singh *et al.*, 2019] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019.

[Tang *et al.*, 2024] Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Kebing Hou, Xiaotong Guo, Jinhua Zhao, et al. Itinera: Integrating spatial optimization with large language models for open-domain urban itinerary planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1413–1432, 2024.

[van Zee, 2020] Marc van Zee. Measuring compositional generalization, 2020.

[Wang *et al.*, 2021] Zhaokai Wang, Renda Bao, Qi Wu, and Si Liu. Confidence-aware non-repetitive multimodal transformers for textcaps. In *Proceedings of the AAAI Con-*

ference on Artificial Intelligence, volume 35, pages 2835–2843, 2021.

[Wang *et al.*, 2022a] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022.

[Wang *et al.*, 2022b] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv: 2208.10442*, 2022.

[Wang *et al.*, 2023] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[Wang *et al.*, 2024] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *arXiv preprint arXiv:2406.14852*, 2024.

[Wu *et al.*, 2024] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arXiv:2404.03622*, 2024.

[Xu *et al.*, 2023] Jinxuan Xu, Shiyu Jin, Yutian Lei, Yuqian Zhang, and Liangjun Zhang. Reasoning tuning grasp: Adapting multi-modal large language models for robotic grasping. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[Yang *et al.*, 2023a] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv: 2309.17421*, 9, 2023.

[Yang *et al.*, 2023b] Zhun Yang, Adam Ishay, and Joohyung Lee. Coupling large language models with logic programming for robust and general reasoning from text. *arXiv preprint arXiv:2307.07696*, 2023.

[Yao *et al.*, 2024] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[Zhang *et al.*, 2024] Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *arXiv preprint arXiv:2406.07057*, 2024.

[Zhao *et al.*, 2024] Yuze Zhao, Jintao Huang, Jinghan Hu, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift: A scalable lightweight infrastructure for fine-tuning. *arXiv preprint arXiv:2408.05517*, 2024.

Appendix

A Implementation Details

In addition to the experimental settings outlined in Section 4.1, we provide the following categorized implementation details for this work.

For model specifications, the GPT-4o model used in our experiments and demonstrations is based on the gpt-4o-2024-05-13 version, while the Gemini model is Gemini 1.5 Flash. For TSP data generation, we used an open-source Python TSP solver[1] to obtain the ground truth visiting order of the given object coordinates.

For VLM evaluations, we focused on four directional categories (top left, top right, bottom left, and bottom right) to make it easier for VLMs to distinguish between directions. To discretize object locations for localization learning in VLM, the 2D space is proportionally divided using 40% and 60% thresholds along both the x and y axes, creating nine distinct regions (center, top, bottom, left, right, top-left, top-right, bottom-left, bottom-right). Detailed data statistics and distribution visualizations are provided in Section B.

To extract and format the VLMs' responses, we used the LLaMA 3.1 language model [Dubey *et al.*, 2024], which converts the results into the required format for metric calculations. The specific prompts used for each task are detailed in Section C. The evaluation for basic spatial relationship understanding is intuitive, as it follows a multiple-choice question format. For the SPP evaluation, we check two criteria: (1) whether the solved path is valid on the grid, and (2) whether the length of the solved path is indeed the shortest between the given start and end objects. For the TSP evaluation, a path is considered "correct" only if it exactly matches the solution from the TSP solver mentioned above. To reduce the difficulty for VLMs in solving TSP, we explicitly specify the starting object in our implementation.

For the benchmark evaluation of Vision-Language Models (VLMs), we utilized the following benchmark datasets: MMBench *dev* [Liu *et al.*, 2023d], SeedBench [Li *et al.*, 2023a], MME [Fu *et al.*, 2023], and BLINK [Fu *et al.*, 2024] Additionally, we employed the VLMEvalKit [Duan *et al.*, 2024], an open-source evaluation toolkit, to ensure standardized and reproducible evaluation of the VLMs.

B Data Statistics

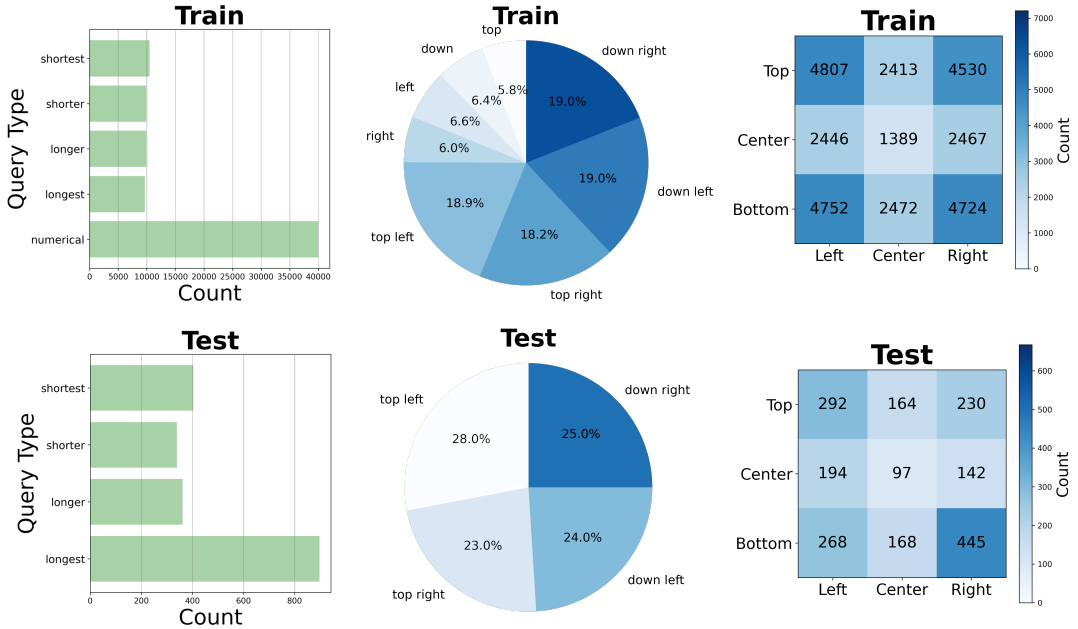

Figure 7: Data statistics of basic spatial relationships (from left to right: distance, direction, and localization statistics).

To complement Section 4.1, this section provides detailed statistics of the data from Sparkle training set and evaluation. We begin by discussing data related to basic spatial relationships (i.e., distance, direction, localization), covering both Sparkle training set and the spatial relationship understanding task in the evaluation set.

[1] https://github.com/fillipe-gsm/python-tsp

Figure 7 illustrates various statistics. In the left column, we see the distribution of questions and instructions related to the *distance* between objects, which includes comparative expressions (e.g., shortest, shorter, longer, longest) and numerical distance estimations considered only in Sparkle training set. The training set shows a fairly even distribution of comparison queries, while in the test set, queries involving the "shortest" and "longest" distances occur more frequently than those involving "shorter" and "longer".

The middle column of Figure 7 presents the data concerning *directional* relationships between objects. We divided the 2D space into direction sectors: four sectors for testing and eight for training. The directional relationships of "bottom-right", "bottom-left", "top-right", and "top-left" each make up about 19% of the training data, while "top", "bottom", "left", and "right" each account for roughly 6%. In the test set, the four main directional relationships are distributed evenly.

Lastly, the right column in Figure 7 shows the *localization* data. Objects are most frequently located in the corners of the space (i.e., top-left, top-right, bottom-left, and bottom-right) in both the training and test sets. The number of objects placed in "top", "bottom", "left", and "right" positions is about half that of those in the corners, while the fewest objects are placed in the center. This is due to the intentional narrowing of the center area as we explained in Section A, which reduces the likelihood of randomly generated objects being placed there. Since there is no clear distinction between regions like "left" and "top-left", this narrowed design encourages VLMs to accurately distinguish specific areas such as the "center", "top", "bottom", "left", and "right".

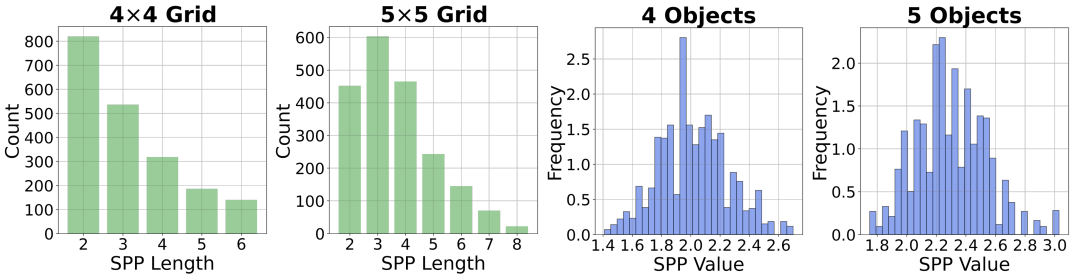

Figure 8: Data statistics of composite spatial reasoning tasks in the evaluation set.

Figure 8 presents data statistics for composite spatial reasoning tasks. The two left subfigures show the distribution of ground truth shortest path lengths in 4×4 and 5×5 grids, while the two right subfigures depict the distribution of total distances for the optimal path in the TSP with 4 and 5 objects.

C Prompts for Extracting Inference Results from VLMs

In this section, we provide the designed prompts for a language model to extract results from VLMs' responses.

C.1 Multi-choice Questions

Prompt for Extracting Results from VLMs' Responses to Multiple-Choice Questions

```
Extract the option capital letter from the result and return it as \\boxed{{X}}, where X is the letter.

Provide no additional content. The result is: ```{result}```.

Make sure your response is in the \\boxed{{X}} format.
```

The above prompt is adopted for all evaluations that in a Multi-choice Questions format.

C.2 Shortest Path Problem

Prompt for Extracting Results from VLMs' Responses to Shortest Path Problems

```
Extract the sequence of node labels from the given input and return it as a Python list.

**Return Format:**
- Do not include any additional text or explanations.
- Ensure that the response is a single list containing only the node text labels (N1, N2, ...).
- If no valid action sequence is found, return 'None'.

**Example Output format:**
```
[node1 text label, node2 text label, ...]
```

Now, extract the result from the following input: ```{result}```. Strictly adhere to the return format.
```

C.3 Traveling Salesman Problem

Prompt for Extracting Results from VLMs' Responses to Traveling Salesman Problems

```
Extract the sequence of movements from the given input and return it as a Python list of object names.

**Return Format:**
- Do not include any additional text or explanations.
- Ensure that the response is a single list containing only the object names.

**Expected Output Format:**
```
{output_format}
```

Now, extract the result from the following response: ```{result}```. Strictly adhere to the output format.
```

D Sample Data Demonstration

In this part, we provide detailed data sample from our experiments.

D.1 Data Sample from Sparkle Training Set and Evaluation

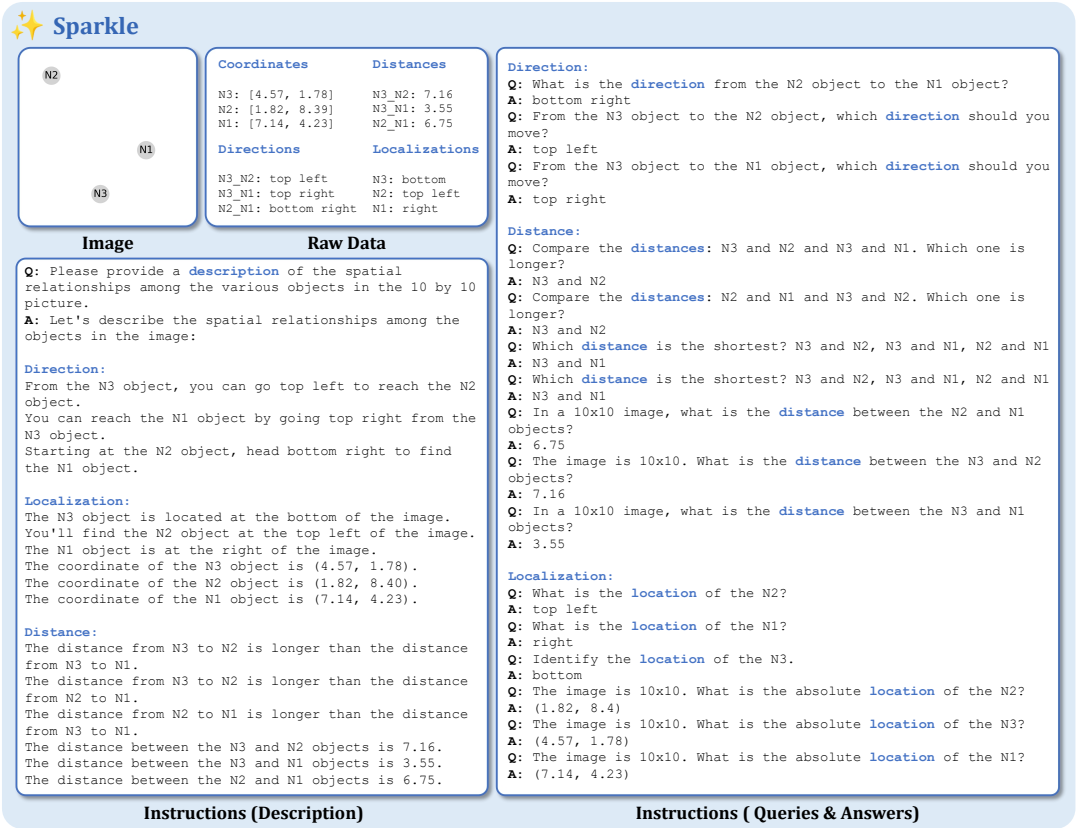

Figure 9: A data sample from the Sparkle training set.

D.2 Data Sample from the Basic Spatial Relationship Understanding task

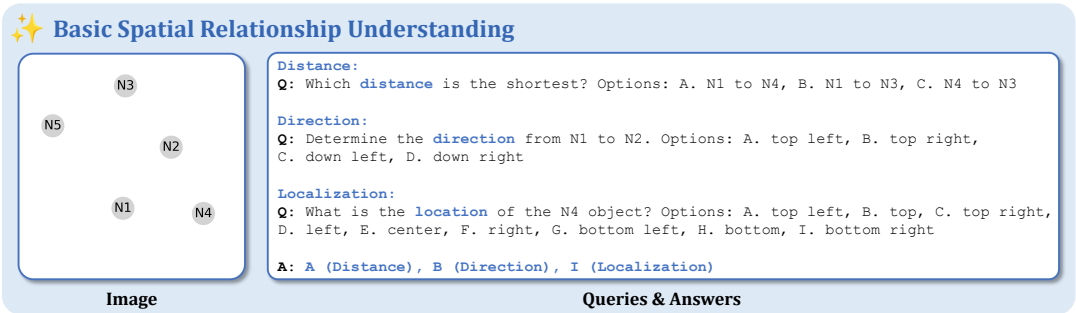

Figure 10: A data sample for Basic Spatial Relationship Understanding

D.3 Data Sample from the Shortest Path Problem

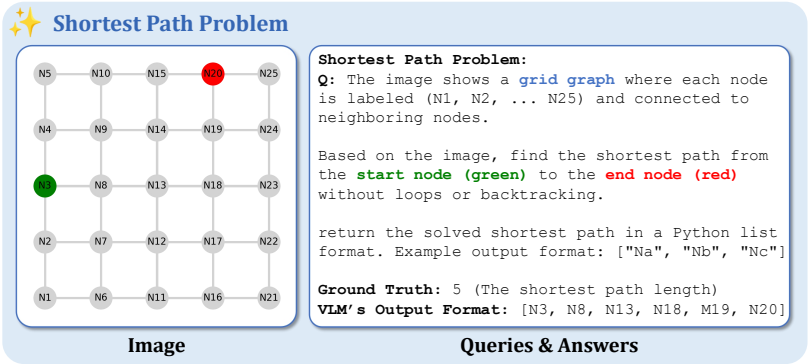

Figure 11: A data sample from the Shortest Path Problem.

D.4 Data Sample from the Traveling Salesman Problem

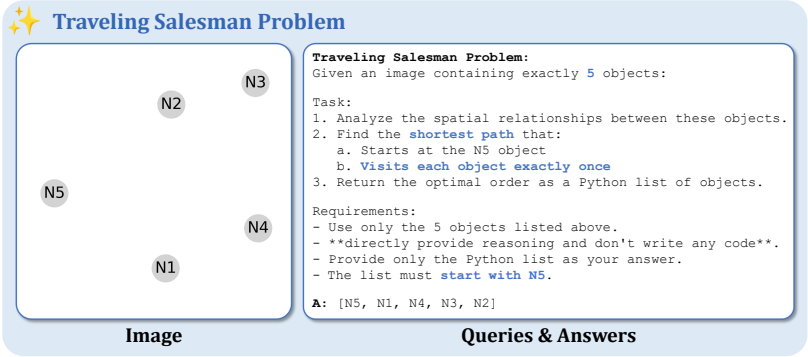

Figure 12: A data sample from the Traveling Salesman Problem.

D.5 Data Sample from General Spatial Tasks

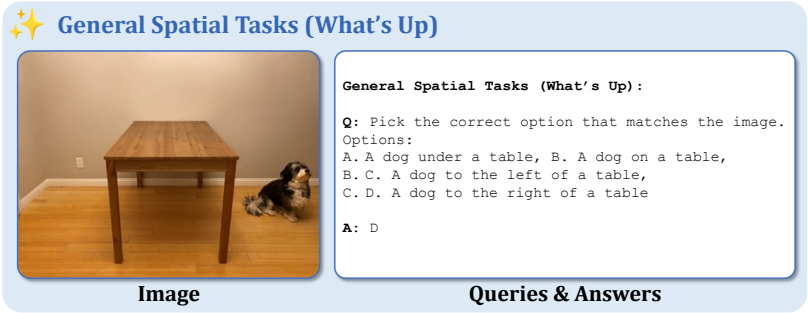

Figure 13: A data sample from the General Spatial Tasks ("What's Up").