TopViewRS: Vision-Language Models as Top-View Spatial Reasoners

Anonymous ACL submission

Abstract

Top-view perspective denotes a typical way in which humans read and reason over different types of maps, and it is vital for localization and navigation of humans as well as of 'non-human' agents, such as the ones backed by large Vision-Language Models (VLMs). Nonetheless, spatial reasoning capabilities of modern VLMs in this setup remain unattested and underexplored. In this work, we study their capability to understand and reason over spatial relations from the top view. The focus on top view also enables controlled evaluations at different granularity of spatial reasoning; we clearly disentangle different abilities (e.g., recognizing particular objects versus understanding their relative positions). We introduce the TopViewRS (**Top**-**View** **R**easoning in **S**pace) dataset, consisting of 11,384 multiple-choice questions with either realistic or semantic top-view map as visual input. We then use it to study and evaluate VLMs across 4 perception and reasoning tasks with different levels of complexity. Evaluation of 10 representative open- and closed-source VLMs reveals the gap of *more than 50%* compared to average human performance, and it is even *lower* than the random baseline in some cases. Although additional experiments show that Chain-of-Thought reasoning can boost model capabilities by 5.82% on average, the overall performance of VLMs remains limited. Our findings underscore the critical need for enhanced model capability in top-view spatial reasoning and set a foundation for further research towards human-level proficiency of VLMs in real-world multimodal tasks.

1 Introduction

Large Language Models (LLMs) such as Llama 2 and 3 (Touvron et al., 2023), Mistral (Jiang et al., 2023), and GPT models (OpenAI, 2022) have delivered impressive performance across a range of text-based tasks and applications such as question answering, language generation, and arithmetic reasoning (Qin et al., 2023a; Zhao et al., 2023). Building on these text-only LLMs, the so-called Vision Language Models (VLMs), equipped with the capability to process and reason over multi-modal vision-language information, have enabled multi-modal processing (Yin et al., 2023; Wu et al., 2023). They ground language reasoning ability of LLMs into the information of different modalities (Chandu et al., 2021). Prominent examples of VLMs such as LLaVA (Liu et al., 2023b), GPT-4V (OpenAI, 2023), and Gemini (Google, 2024), have demonstrated strong performance across applications such as visual question answering (Li et al., 2023d), image captioning (Diesendruck et al., 2024), and object grounding (Zheng et al., 2024).

Spatial reasoning, one of the fundamental desirable properties of and requirements for VLMs, has also gained increased attention recently (Rajabi and Kosecka, 2023; Liu et al., 2023a; Chen et al., 2024). It requires grounding the model's reasoning ability with natural language into its visual perception of the surrounding environment (Freksa, 1991). In particular, it involves two critical steps: (i) *interpreting* the environment visually, and (ii) *reasoning* over spatial relations. As a fundamental ability for the model to recognize, understand, and navigate through the physical world, it plays a crucial role in various downstream tasks such as vision-language generation (Li et al., 2024a) and embodied AI (Cho et al., 2024). However, previous research has focused on exploring spatial reasoning abilities of VLMs only from a conventional first-person perspective view (Liu et al., 2023a). In this work, we aim to study and evaluate spatial understanding and reasoning capability of VLMs from the *top-view perspective*, also referred to as the bird's-eye view (Li et al., 2024b).

When compared to the conventional perspective view, top view offers better *natural alignment*: it is the typical perspective used for reading maps or presenting floor plans. Moreover, it is inherently

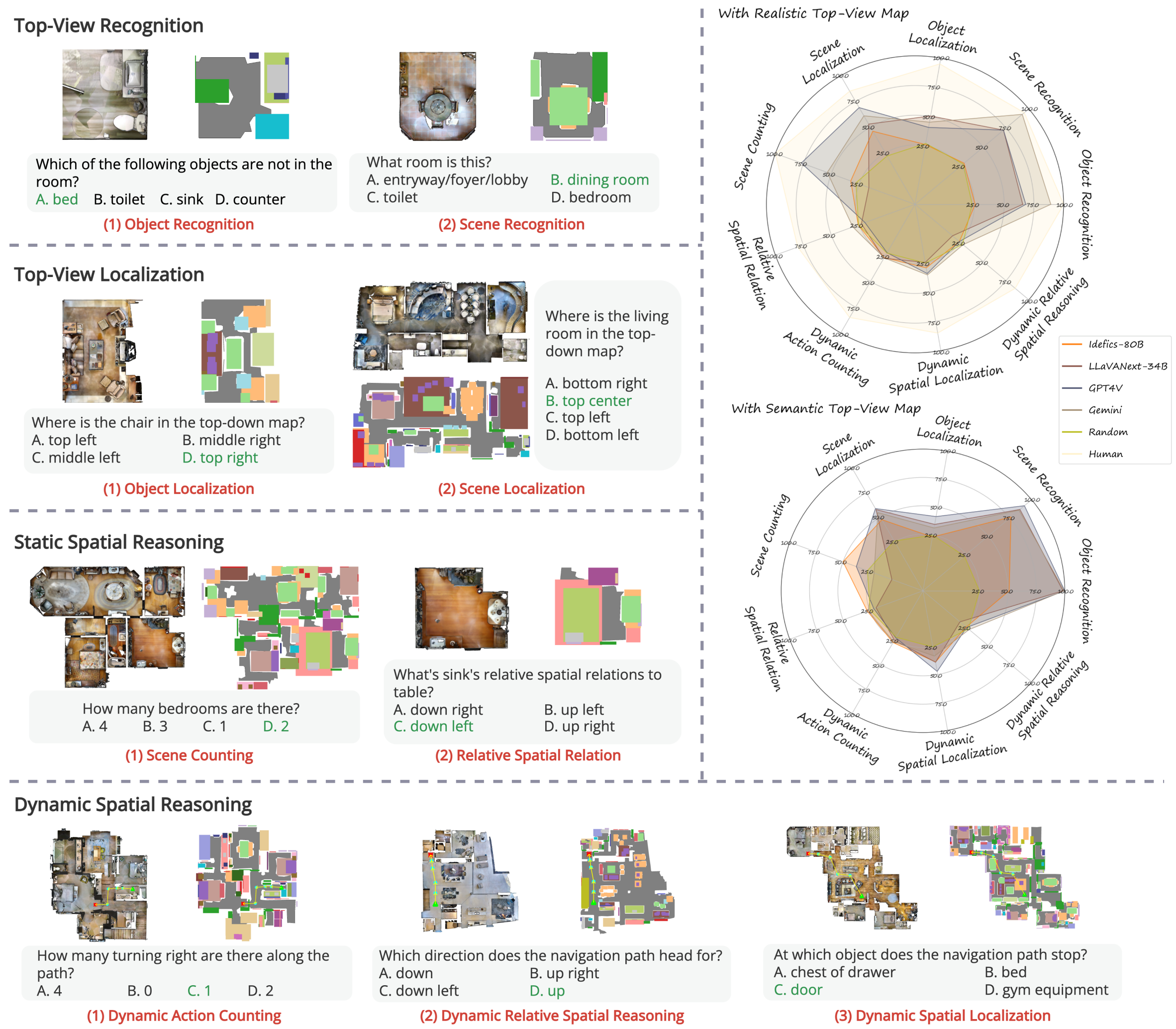


Figure 1: Illustration of the four evaluation tasks with an incremental level of complexity on the two types of top-view maps (photo-realistic versus semantic maps), covering top-view perception and spatial reasoning abilities, with 9 sub-tasks in total (red font), focusing on different, well-defined VLM abilities. The radar graphs (top right) compare the representative models' performance on all sub-tasks, indicating *a large gap with human performance*.

more *complex*: top-view maps encapsulate a wealth of information about different scenes, locations, objects and their relationships in the environment based on a single image. In addition to the *photo-realistic* top-view maps, *semantic* top-view maps (Nanwani et al., 2023; Li et al., 2024a) use different colors to represent different types of objects; we run experiments with both map types, see Figure 1.

One advantage of top-view maps is that they define a controlled and interpretable experimental framework. Indoor scenes, which are the focus of this work, typically feature a relatively stable set of objects and layouts, making them ideal for controlled studies. This allows us to disentangle and investigate various aspects of spatial reasoning and VLMs' capabilities in a controlled manner.[1]

In this work, we thus investigate the basic top-view spatial understanding and reasoning abilities of current state-of-the-art VLMs across four tasks of gradually increasing complexity, and their finer-grained sub-tasks. The tasks are as follows. *1) Top-View Recognition* assesses whether the model can recognize concrete objects and scenes in top-view maps. *2) Top-View Localization* evaluates the ability to localize objects or regions on a map based on textual descriptions. *(3) Static Spatial Reasoning* investigates whether the model can reason about spatial relationships among localized objects and regions within the map. *(4) Dynamic Spatial Reasoning* evaluates reasoning about spatial relations along the points of a dynamic navigation path. Figure 1 illustrates all the tasks with concrete examples. As one key finding of this study, con-

[1]For instance, we can apply different interventions (e.g., drawing a navigation trajectory in a realistic map, or changing the color-object mapping in a semantic top-view map).

ducted evaluations reveal that current VLMs lack sufficient capability to effectively tackle top-view spatial reasoning challenges, indicating substantial room for improvement in future research.

Contributions. **1)** We define the top-view spatial reasoning challenge for VLMs via 4 carefully designed tasks of increasing complexity, also encompassing 9 distinct fine-grained sub-tasks with a structured design of the questions focusing on different model abilities. **2)** We collect the TopViewRS dataset, comprising 11,384 multiple-choice questions with either photo-realistic or semantic top-view maps of real-world scenarios through a pipeline of automatic collection followed by human alignment. **3)** We use TopViewRS to evaluate and study 10 VLMs from different model families and sizes, highlighting the substantial performance gap compared to humans.[2]

2 Related Work

Top-View Map Understanding. There are only limited studies in NLP focused on the use of top-view maps, though considerable research has been conducted within the broader AI community on the so-called *bird's-eye view*, which is an instance of top view. This body of work has explored applications in autonomous driving (Unger et al., 2023; Li et al., 2024c), with contributions on fusing different types of views (Qin et al., 2023b) and working with arbitrary camera setups (Peng et al., 2023). In other application scenarios, Yan et al. (2021) introduce a bird's-eye view person re-identification task.

Efforts to bridge top-view images with natural language in applications beyond the above are less diverse. The WAY dataset, proposed by Hahn et al. (2020), contains 6,154 dialogs aimed at localizing an observer's position on a top-view map through conversations between an observer and a locator. This dataset has inspired follow-up research focusing on merging vision with dialog information (Zhang et al., 2024a) and leveraging pretraining strategies to enhance performance (Hahn and Rehg, 2022). In general, prior research does not assess VLMs' basic spatial reasoning abilities with top-view images and lacks fine-grained and controllable analyses of these fundamental abilities.

Spatial Reasoning on Multi-Modal Vision-Text. There has been a body of work on text-only spatial reasoning with the advancement of LLMs (Yamada et al., 2024), within the context of relative spatial relation recognition (Mirzaee et al., 2021; Shi et al., 2022), natural language navigation (Yamada et al., 2024), and planning (Momennejad et al., 2023) (see Appendix A for a more complete overview).

Cross-modal spatial reasoning puts forward higher requirements for the models in terms of language grounding (Rozanova et al., 2021; Rajabi and Kosecka, 2023). Liu et al. (2023a) investigate spatial reasoning with 2D natural realistic front-view images and Chen et al. (2024) extend the analysis to 3D point clouds. The environmental contexts become more diverse compared to synthetic symbols in text-only spatial reasoning, ranging from indoor environments (Koch et al., 2024) to outdoor street views (Chen et al., 2019). Regarding typical tasks, visual QA (VQA) is the mainstream task for benchmarking spatial reasoning abilities (Dong et al., 2021; Banerjee et al., 2021; Liu et al., 2023a; Li et al., 2023a,b; Kamath et al., 2023), while other tasks include vision-language navigation (Chen et al., 2019; Li et al., 2024a) and user interface grounding (Rozanova et al., 2021).[3]

We stress that none of the prior research efforts allows for *disentangled evaluation* of models' spatial reasoning abilities. Prior work typically conflates object recognition with spatial reasoning. We thus design a dataset and conduct a study that not only offers insight into fundamental abilities but also allows for easier interpretation of results (§4).

3 Task Definition

Following prior work (Li et al., 2023a), we frame all tasks as multiple-choice QA tasks. Given a top-view (realistic or semantic) map of a room M, the model must choose the correct option o_i from the four options provided $O = \{o_0, o_1, o_2, o_3\}$ that answers the question.[4] This format simplifies the evaluation and interpretation of the results.

Top-View Maps. We provide two different types of top-view maps to the models: realistic maps M_{Real} and semantic maps M_{Sem}. Realistic maps are constructed by placing a simulated orthographic camera above the scene to capture a photo-realistic top-view image. Semantic maps represent objects

[2] We publicly release (the part of) the dataset and code online at URL-ANONYMOUS.

[3] Research on multi-modal spatial reasoning also intersects with efforts from the computer vision community on scene understanding (Teney et al., 2017), simultaneous localization and mapping (Cadena et al., 2016), and combining LLMs with representations of the 3D physical world (Hong et al., 2023).

[4] For simplicity, for each question, there is always a *single correct answer*.

in the scene with colored bounding boxes. Each object is assigned a specific color and labeled at the same relative coordinates on the map to preserve the object's semantic information and spatial allocation. In comparison to realistic maps, semantic maps simplify the initial step of spatial reasoning (i.e., environment interpretation) by labeling the object types with corresponding colors and excluding irrelevant additional details such as shape and texture found in realistic top-view maps. Given the customizable and flexible nature of color-object mapping, the semantic map can also serve as an ideal testbed for evaluating models' out-of-distribution (OOD) performance, thereby encouraging further exploration beyond the scope of this work. Example maps are in Figure 1.

Tasks and Sub-Tasks. We define 4 different tasks which cover a total of 9 finer-grained sub-tasks, with concrete examples shown in Figure 1. The tasks are designed to have an increasing level of complexity, where each subsequent task depends on the abilities measured in the preceding one(s).

(1) Top-View Recognition evaluates the fundamental ability to interpret the input map, and covers two sub-tasks: *Object Recognition* and *Scene Recognition*. It does not require the model to identify specific locations of objects and rooms.

(2) Top-View Localization investigates whether the model can localize objects or rooms in the top-view map based on textual descriptions, including *Object Localization* and *Scene Localization* as two sub-tasks. Beyond understanding the top-view map as a whole, it requires the model to ground entities in the map, representing the model's ability to align spatial descriptions with corresponding locations.

(3) Static Spatial Reasoning aims to evaluate the model's spatial reasoning ability with more complex questions. It includes two sub-tasks: reasoning over *Scene Counting* and *Relative Spatial Relations* between different objects and rooms. These questions require the model to perform multi-step reasoning based on the recognition and localization of entities in the top-view map.

(4) Dynamic Spatial Reasoning. Finally, we introduce a novel task that involves dynamic spatial reasoning over top-view maps in the context of agent navigation. It requires the model to understand the sequential relations along the points of the navigation path (sub-task *Dynamic Action Counting*) and answer spatial questions with regard to the dynamic navigation path (sub-task *Dynamic Relative Spatial Reasoning*) and the circumstantial environments (*Dynamic Spatial Localization*).

4 TopViewRS Dataset

In order to study and evaluate the abilities of state-of-the-art VLMs on the 4 tasks spanning 9 sub-tasks from §3, we now introduce a novel evaluation dataset, TopViewRS, which focuses on *top-view maps of indoor scenes* (i.e., houses and rooms), discussed in what follows.

Dataset Features. It introduces several advancements and innovative features that distinguish it from all prior visual spatial reasoning datasets.

1) Multi-Scale Top-View Maps: The selected top-view maps of indoor scenes (see Figure 1) provide a more natural representation of spatial environments that aligns with human cognitive map (Epstein et al., 2017). This makes benchmarking spatial awareness more straightforward and meanwhile mitigates spurious correlations in the positions between objects commonly found in realistic front-view images. Compared to the front view, the multi-scale top-view maps of single rooms and full houses add more divergence in the granularity of the entities (objects or rooms) in spatial reasoning. Meanwhile, we provide both realistic maps and semantic maps for more comprehensive evaluation.

2) Realistic Environmental Scenarios with Rich Object Sets: We provide real-world environments from indoor scenes, with 80 objects per scene on average, ensuring a natural distribution and complexity of object locations. This also sets it apart from existing front-view spatial reasoning datasets, which typically contain only a handful of objects.

3) Structured Question Framework: Unlike previous datasets (Li et al., 2023a; Liu et al., 2023a; Kamath et al., 2023), which conflate spatial reasoning with object recognition, our dataset clearly defines 4 tasks including 9 sub-tasks in total using diverse question templates. This structured approach allows for a fine-grained evaluation and analysis of models' capabilities from various perspectives and levels of granularity.

Dataset Collection. We employ a two-stage data collection strategy that includes *automatic collection from a simulator* and *alignment through human judgment*. First, to approximate real-life scenarios, we use the Matterport3D dataset (Chang et al., 2017), which includes 90 building-scale scenes with instance-level semantic and room-level region

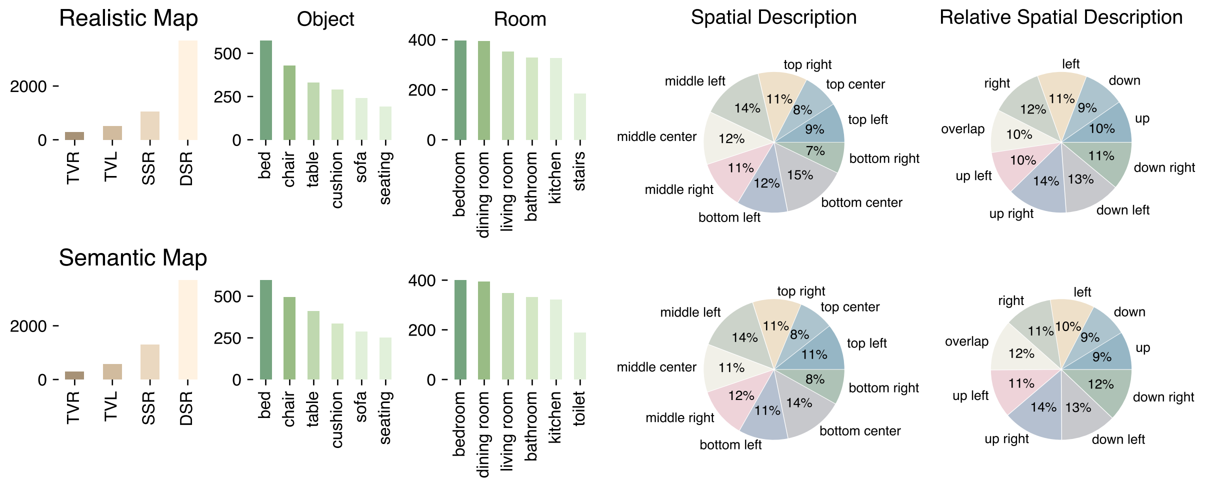


Figure 2: TOPVIEWRS data statistics, showing distribution of task sizes, objects, regions, spatial and relative spatial descriptions in realistic and semantic map settings, where the tasks are described with their initials for visualization.

annotations in 3D meshes. We filter these to exclude multi-floor and low-quality scenes, selecting 7 scenes with an average of 80 objects and 12 rooms each. Realistic top-view maps are extracted using orthographic cameras, and semantic top-view maps are constructed using the Habitat (Manolis Savva* et al., 2019; Szot et al., 2021) simulation environment. We then design a structured question framework with 15 templates to minimize human labor and standardize the data collection process. To ensure quality, a second stage of manual *human judgment* aligns and verifies the data, ensuring questions are natural and correct. Participants are encouraged to discard or modify data points to improve quality, maintaining alignment with human judgments. We refer readers to Appendix B for further details regarding the data collection process.

Dataset Statistics. TOPVIEWRS comprises a total of 11,384 multiple-choice questions after human verification, with 5,539 questions associated with realistic top-view maps, and 5,845 with semantic top-view maps. Human verification keeps 587/784 questions from the automatic collection phase for Top-View Recognition, 1,077/1,384 for Top-View Localization, 2,340/3,080 for Static Spatial Reasoning. The choices are uniformly distributed over choices A (*25.5%*), B (*24.6%*), C (*24.5%*) and D (*25.4%*). Figure 2 shows the distribution of different tasks, objects, regions and spatial descriptions. The size of each task aligns with its corresponding difficulty level, where the easier task comprises fewer examples. We provide further insights and technical details in Appendix B.4.

5 Experiments and Results

Models and Implementation. We test a representative selection of both open-sourced and close-sourced models which achieve state-of-the-art performance on a range of multimodal benchmarks (Liu et al., 2023c; Li et al., 2023a) in a zero-shot inference setup. Regarding open-sourced models, we study and evaluate Idefics (9B & 80B) (Laurençon et al., 2023), LLaVA-Next (7B, 13B & 34B) (Liu et al., 2024), InternLM-XComposer2 (7B) (Dong et al., 2024), Qwen-VL (7B) (Bai et al., 2023). The chosen close-sourced models are GPT-4V (OpenAI, 2023) and Gemini (Google, 2024).[5] All the models are implemented within the VLMEvalKit framework (OpenCompass Contributors, 2023).

Prompts. For realistic maps, we provide the VLMs with the task description along with the multiple-choice question. For semantic maps, in addition to the information above, we also introduce the concept of a semantic map to the model and provide the color-object mapping in the prompt in order to facilitate its understanding of the abstract map. We only provide the color-object mappings of the colors that are presented in the semantic map as a preprocessing strategy in order to exclude irrelevant information. For the specific prompting templates used in this paper, we refer to Appendix C.2.

Evaluation Measures. We measure multiple-choice QA accuracy via *Exact Match (EM)* and *Partial Match (PM)*. EM measures whether the predicted option indices are exactly the same as the label indices. However, there may be cases where

[5]We use *GPT-4-turbo-2024-04-09* of GPT-4V and latest stable *gemini-pro-vision 1.0* of Gemini.

the correct answer to the question can be considered partially correct, e.g., the answer is *top right* while the prediction is *top left*. PM then calculates the proportion of overlapping words between the predicted answer and the gold answer. It is calculated based on the correctness of the text spans (or words) of predicted options, as given by:

$$PM = \frac{|\{\text{labels}\} \cap \{\text{predictions}\}|}{\max\left(|\{\text{labels}\}|, |\{\text{predictions}\}|\right)}$$

5.1 Results and Discussion

We first discuss the models' performance across our four tasks, with results summarized in Table 1, and fine-grained sub-task performance illustrated in Figure 3. We find that the performance of current state-of-the-art VLMs is *unsatisfactory* on the proposed TopViewRS benchmark with model-wise average EM and PM over all tasks below 50%. Gemini is the best-performing model for realistic maps, while GPT-4V excels in semantic maps. For some models, such as Qwen-VL, the results are sometimes much worse than the random baseline. This issue primarily arises from the models' difficulty in following the instructions to choose from the four provided options.

Models perform better on recognition and localization tasks compared to reasoning tasks. Top-View Recognition consistently demonstrates the highest performance across all models. Gemini shows human-comparable performance with the EM score over 90%. Top-View Localization exhibits lower performance compared to Top-View Recognition, followed by Static Spatial Reasoning. The performance difference of various tasks with different levels of complexity underscores *the advantage of our benchmark to capture well-defined and disentangled phenomena*, which allows for controlled studies in controlled environments.

Regarding Dynamic Spatial Reasoning, models perform better on this task than on the previous tasks. Fine-grained performance in Figure 3 indicates that the improved performance primarily stems from high accuracy in dynamic action counting and spatial localization, which constitute 18% and 66% of the data respectively for this task. We attribute the high accuracy in these areas to the equivalence between navigation path symbols and visual prompting (Shtedritski et al., 2023). Despite these advancements, the overall EM accuracy remains below 40%, and *models still struggle with reasoning over dynamic relative spatial relations*.

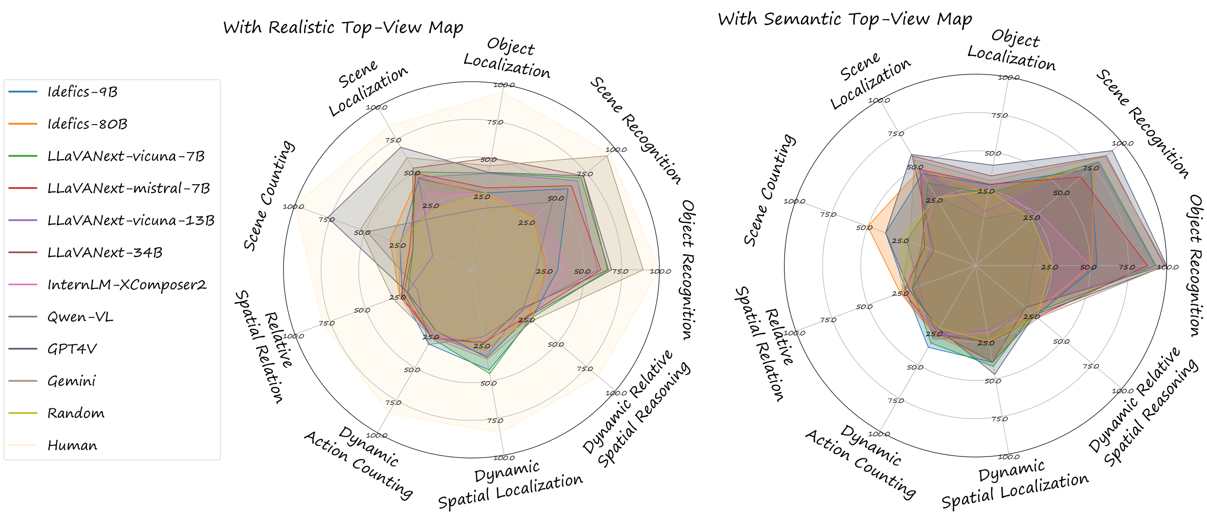
Larger models do not always show better spatial awareness. Surprisingly, our results reveal that larger model sizes do not consistently translate to better performance. In Top-View Recognition, closed-source models outperform open-source models by 31.10% EM with realistic maps and 29.33% EM with semantic maps. However, the performance gap narrows as the task complexity increases. Using realistic maps as the visual input, Gemini stands out by achieving a minimum of 5.53% higher EM accuracy in Static Spatial Reasoning compared to other models, while GPT-4V performs worse than Idefics-9B on both Static and Dynamic Spatial Reasoning tasks. This indicates a lack of significant difference in spatial awareness between closed-source and open-source models for tasks with higher complexity, despite the disparity in their model sizes. This trend holds true within open-sourced models as well. Both Idefics and LLaVANext model families in some cases show comparable or worse performance with larger model variants than with smaller ones. Similar observations have been made by previous studies (Zhong et al., 2021; Shi et al., 2024). We conjecture that this might be caused by inadequate evidence of the scaling law (Kaplan et al., 2020) in the computer vision community (Tian et al., 2024). The results on TopViewRS thus advocate for further investigation and analysis in this area.

Models perform better in easier tasks with semantic maps. In simple tasks such as Top-View Recognition, models generally perform better with semantic maps than with realistic maps, except for Qwen-VL, showing an improvement of 20.35%. However, this advantage decreases in more complex tasks. For Top-View Localization and Static Spatial Reasoning, models struggle to utilize semantic top-view maps, yielding performances akin to random baselines in both EM and PM accuracy. One possible explanation is that the semantic top-view image and the input prompt with color-object mapping deviate too much from the models' training data distribution. This is further evidenced by the predictions from open-sourced models such as Qwen-VL, which fail to respond to instructions and answer with numbers or RGB values 91.25% of the time for Top-View Localization and 47.65% of the time for Static Spatial Reasoning.

Fine-Grained Insights with Sub-Tasks. Models using realistic maps excel more in the sub-task of Scene Recognition, which involves larger entities,

Model		Idefics		LLaVANext				XComposer2	Qwen-VL	GPT-4V	Gemini
Model Size		9B	80B	vicuna 7B	mistral 7B	vicuna 13B	34B	7B	7B	API	API
Realistic Map											
Top-View Recognition	EM	41.10	26.71	67.47	61.30	61.64	67.81	37.67	27.05	69.52	**90.41**
	PM	41.10	26.88	67.64	61.47	61.99	67.81	37.67	27.26	69.86	**90.58**
Top-View Localization	EM	30.39	30.00	42.16	33.92	41.18	**50.98**	27.84	16.27	46.27	48.24
	PM	46.42	46.08	56.67	48.63	54.31	**61.76**	41.86	26.31	60.39	60.98
Static Spatial Reasoning	EM	24.07	26.07	19.87	24.36	20.25	22.73	25.79	14.71	22.16	**31.61**
	PM	33.68	38.52	34.40	37.34	36.26	35.56	38.73	21.15	35.59	**45.22**
Dynamic Spatial Reasoning	EM	38.10	27.94	**38.81**	24.31	29.08	23.79	24.07	22.11	30.29	32.60
	PM	40.88	30.68	**42.15**	26.69	32.89	27.28	26.86	24.65	33.86	35.80
Semantic Map											
Top-View Recognition	EM	60.68	59.32	88.81	80.00	88.14	94.58	43.05	19.66	**97.29**	94.92
	PM	60.68	59.32	88.81	80.00	88.49	94.58	43.05	20.05	**97.29**	94.92
Top-View Localization	EM	31.21	27.34	25.40	32.10	17.28	38.45	24.87	9.70	**44.44**	35.27
	PM	47.62	45.41	44.27	47.80	23.66	53.79	41.09	13.99	**58.55**	49.91
Static Spatial Reasoning	EM	23.82	**28.07**	18.72	24.28	16.63	18.41	23.05	14.85	21.73	26.22
	PM	34.13	38.17	30.57	37.26	29.94	31.22	35.50	21.99	33.09	**39.12**
Dynamic Spatial Reasoning	EM	36.67	34.55	37.45	26.23	19.92	33.12	21.60	23.55	**39.30**	31.41
	PM	39.92	37.75	40.69	28.89	23.63	36.86	24.32	26.09	**43.20**	34.86

Table 1: Comparison of 10 models on both realistic and semantic top-view maps. Performance is analysed according to four tasks with EM and PM. The best performance for each task is illustrated in **bold**.



(a) Performance with realistic top-view maps

(b) Performance with semantic top-view maps

Figure 3: Visualization of fine-grained comparison with 10 models and humans on 9 sub-tasks using realistic and semantic top-view maps, demonstrating that *most current models perform on par with random baseline in spatial reasoning and has a large gap with human performance*. Exact numbers are reported in Table 15 in the Appendix.

compared to Object Recognition. This gap is also evident in a 12.66% and 19.73% performance difference between object-level and scene-level localization with both map types. Conversely, with semantic maps, the model struggles more with scene-level recognition than with realistic maps, showing an 11.09% lower performance than object-level recognition among closed-source models. Most models perform similarly to a random baseline in reasoning over spatial relations but show higher accuracy in scene counting. This likely occurs because 95% of the correct room counts are within a narrow range (1 or 2), reflecting real-life distributions. Thus, models leverage commonsense knowledge as the shortcut for counting, as seen in the 54.73% performance gap (with GPT-4V) between counting scenes and actions. However, the spatial localization and reasoning abilities of both open-source and closed-source models still remain unsatisfactory, even at the level of sub-tasks.

5.2 Further Discussion

Gap to Human Performance. We now study how humans perform on this dataset and the gap between current models and human performance. To this end, we recruited 4 human participants who were not involved in dataset creation for human evaluation. A total of 60 data points with realistic

Task	Ability	Size	Human	GPT-4V
TVR	Object Recognition	5	95	**100**
	Scene Recognition	5	**100**	80
TVL	Object Localization	5	**95**	20
	Scene Localization	10	**85**	60
SSR	Scene Counting	5	**100**	80
	Relative Spatial Relation	10	**80**	0
DSR	Dynamic Action Counting	5	**85**	0
	Dynamic Spatial Localization	10	**85**	40
	Dynamic Relative Spatial Reasoning	5	**85**	0
Average Score			**90.0**	42.2

Table 2: Performance (EM) between human and GPT-4V on all the sub-tasks, demonstrating a huge *gap* between GPT-4V and human.

Model	GPT-4V			Gemini		
	w/o. CoT	w. CoT	Δ	w/o.CoT	w. CoT	Δ
RGB Overall	22.16	26.74	**+4.58**	31.61	40.02	**+8.41**
Scene Counting	*76.74*	*25.58*	*-51.16*	*53.49*	*48.84*	*-4.65*
Relative Spatial Relations	*19.82*	*26.79*	*+6.97*	*30.68*	*39.64*	*+8.96*
Semantic Overall	21.73	28.07	**+6.34**	26.22	30.16	**+3.94**
Scene Counting	*37.50*	*47.92*	*+10.42*	*20.83*	*29.17*	*+8.34*
Relative Spatial Relations	*21.12*	*27.31*	*+6.19*	*26.43*	*30.20*	*+3.77*

Table 3: Comparison of model performance (EM) w/ and w/o Chain of Thought (CoT) on Static Spatial Reasoning, showing that *CoT helps elicit spatial reasoning*.

top-view maps are randomly selected from the sub-tasks, covering all fine-grained question types.[6] We use Fleiss Kappa as the measure of inter-annotator agreement. The kappa score is 0.747, indicating substantial agreement shared by the human participants according to Landis and Koch (1977). The average performance of the human participants is shown in Table 2. The experimental results show that there is still a large gap with human performance by over 50% across all the sub-tasks that involve spatial awareness. We also observe that with GPT-4V, human performs 47.8% higher than the model on average. The gap between human and model performance is larger on complex reasoning tasks compared to the recognition tasks, indicating plenty of room for improvement.

Chain-of-Thought Helps Elicit Spatial Reasoning. Due to the compositionality of Static Spatial Reasoning based on Top-View Recognition and Localization in task design, the model is supposed to answer the question based on the locations of the entities in the top-view map. Inspired by this requirement, we explored whether Chain-of-Thought (CoT) reasoning (Wei et al., 2022) could facilitate spatial reasoning by initially prompting the model to localize entities before producing the final answer to the question. To implement this, we modified the instruction to include: *"You should first localize the entity and then answer the question based on the locations"*, thereby encouraging the model to process information and think step by step. Considering that CoT has shown effectiveness in larger models (Wei et al., 2022; Li et al., 2023c), we conducted experiments with GPT-4V and Gemini to evaluate this hypothesis. As shown in Table 3, incorporating CoT into the reasoning process notably enhances performance. Specifically, the models' accuracy improved by 4.58% when using realistic maps and 6.34% with semantic maps for GPT-4V. This improvement underscores the potential of step-by-step reasoning in enhancing the efficacy of spatial reasoning tasks, but there is still a substantial performance gap to the human ceiling.

[6] We did not run human evaluation on semantic maps because they are inherently easier to reason over; they skip the process of recognizing the objects before reasoning, which makes the task simpler but with more sufficient and accurate information for reasoning.

6 Conclusion

In this study, we designed four tasks to examine the capabilities of VLMs as top-view spatial reasoners, progressing from basic top-view map comprehension to dynamic spatial reasoning along navigation paths. To enable investigation into top-view spatial reasoning abilities of VLMs, we collected a novel dataset, TOPVIEWRS, which includes 11,384 multiple-choice questions, featuring photo-realistic and semantic top-view maps as the visual input. Our extensive experiments involved evaluating 10 VLMs across various model families and sizes on TOPVIEWRS. The results highlight a critical observation: particularly in complex reasoning tasks, VLMs frequently perform only as well as a random baseline, with even more pronounced deficits when handling tasks with semantic maps. Moreover, there is a noticeable performance gap compared to human annotators, underscoring the significant potential for further improvements in this area. In response to these findings, we discovered that employing chain-of-thought reasoning enhances model performance in spatial reasoning by 5.82%. Despite this progress, the overall performance of VLMs on spatial reasoning remains less than satisfactory. We hope that our study can set the stage for future research in multimodal spatial reasoning and encourage further investigations into refining the reasoning techniques, moving VLMs closer to human-level proficiency in understanding and reasoning over real-world environments.

Limitations

The TopViewRS dataset primarily evaluates model performance in entity recognition, localization, and spatial reasoning over 2D top-view maps. However, it does not yet include task-oriented planning with spatial awareness, which involves more complex sequential decision-making and dynamic interactions.

Further, our dataset assumes one correct answer per question, but exploring scenarios with multiple correct answers or no correct answers could further challenge systems and provide valuable insights.

We also advocate for further research to explore how spatial awareness in models impacts downstream tasks such as navigation instruction generation (Li et al., 2024a) and task completion by language agents in real-world environments (Parashar et al., 2023).

Moreover, our study is currently limited to 2D top-view maps, whereas spatial reasoning can encompass a variety of modalities and perspectives, such as 3D point clouds.

From the perspective of the models, the rapid progress in VLMs makes it hard to include all new releases such as Idefics 2 (Laurençon et al., 2024). Additionally, multimodal in-context learning (MICL) remains underexplored and is only supported by VLMs trained with interleaved image-text data (Baldassini et al., 2024). Although not universal across all VLMs, MICL has been effective in handling out-of-distribution tasks (Zhang et al., 2024b), which could also be interesting in TopViewRS, especially with semantic maps as visual inputs. In future work, we aim to extend our analysis to include more modalities, evaluate a broader range of models and their capabilities, and investigate additional downstream tasks involving spatial awareness.

Ethics Statement

Our research strictly follows ethical guidelines, focusing on data privacy, bias mitigation, and societal impact. During the dataset construction, we carefully check the licenses of the software we use and follow it strictly. The human participants in our study are recruited from our university with bachelor's degree and are guaranteed compensation above the local minimum average. They have consented to the use of their annotations in our research. We do not see any potential risk of our project.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? *Preprint*, arXiv:2404.15736.

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2021. Weakly supervised relative spatial reasoning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1908–1918.

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*.

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.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Junmo Cho, Jaesik Yoon, and Sungjin Ahn. 2024. Spatially-aware transformers for embodied agents. In *The Twelfth International Conference on Learning Representations*.

Maurice Diesendruck, Jianzhe Lin, Shima Imani, Gayathri Mahalingam, Mingyang Xu, and Jie Zhao. 2024. Learning how to ask: Cycle-consistency refines prompts in multimodal foundation models. *Preprint*, arXiv:2402.08756.

Tianai Dong, Alberto Testoni, Luciana Benotti, and Raffaella Bernardi. 2021. Visually grounded follow-up questions: a dataset of spatial questions which require dialogue history. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 22–31, Online. Association for Computational Linguistics.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

Russell Epstein, E Z Patai, Joshua Julian, and Hugo Spiers. 2017. The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, 20:1504–1513.

Christian Freksa. 1991. *Qualitative Spatial Reasoning*, pages 361–372. Springer Netherlands, Dordrecht.

Gemini Team Google. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Meera Hahn, Jacob Krantz, Dhruv Batra, Devi Parikh, James Rehg, Stefan Lee, and Peter Anderson. 2020. Where are you? localization from embodied dialog. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 806–822, Online. Association for Computational Linguistics.

Meera Hahn and James M. Rehg. 2022. Transformer-based localization from embodied dialog with large-scale pre-training. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 295–301, Online only. Association for Computational Linguistics.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *Preprint*, arXiv:2307.12981.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. 2024. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. *Preprint*, arXiv:2402.12259.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint*, arXiv:2306.16527.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

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

Chengzu Li, Chao Zhang, Simone Teufel, Rama Sanand Doddipatla, and Svetlana Stoyanchev. 2024a. Semantic map-based generation of navigation instructions. *Preprint*, arXiv:2403.19603.

Hao Li, Jinfa Huang, Peng Jin, Guoli Song, Qi Wu, and Jie Chen. 2023b. Weakly-supervised 3d spatial reasoning for text-based visual question answering. *IEEE Transactions on Image Processing*, 32:3367–3382.

Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, Hao Tian, Enze Xie, Jiangwei Xie, Li Chen, Tianyu Li, Yang Li, Yulu Gao, Xiaosong Jia, Si Liu, Jianping Shi, Dahua Lin, and Yu Qiao. 2024b. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2151–2170.

Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, Hao Tian, Enze Xie, Jiangwei Xie, Li Chen, Tianyu Li, Yang Li, Yulu Gao, Xiaosong Jia, Si Liu, Jianping Shi, Dahua Lin, and Yu Qiao. 2024c. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(4):2151–2170.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023c. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, Wei Wang, and Min Zhang. 2023d. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *Preprint*, arXiv:2311.07536.

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

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

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. 2023c. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. 2023. Evaluating cognitive maps and planning in large language models with cogeval. In *Advances in Neural Information Processing Systems*, volume 36, pages 69736–69751. Curran Associates, Inc.

Laksh Nanwani, Anmol Agarwal, Kanishk Jain, Raghav Prabhakar, Aaron Monis, Aditya Mathur, Krishna Murthy Jatavallabhula, A. H. Abdul Hafez, Vineet Gandhi, and K. Madhava Krishna. 2023. Instance-level semantic maps for vision language navigation. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

OpenAI. 2022. Chatgpt blog post.

OpenAI. 2023. GPT-4V(ision) Technical Work and Authors.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Priyam Parashar, Vidhi Jain, Xiaohan Zhang, Jay Vakil, Sam Powers, Yonatan Bisk, and Chris Paxton. 2023. Slap: Spatial-language attention policies. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 3571–3596. PMLR.

Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. 2023. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023a. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, and Xi Li. 2023b. Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8690–8699.

Navid Rajabi and Jana Kosecka. 2023. Towards grounded visual spatial reasoning in multi-modal vision language models. *Preprint*, arXiv:2308.09778.

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

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When do we not need larger vision models? *Preprint*, arXiv:2403.13043.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. 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.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Damien Teney, Lingqiao Liu, and Anton Van Den Hengel. 2017. Graph-structured representations for visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3233–3241.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Preprint*, arXiv:2404.02905.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

David Unger, Nikhil Gosala, Varun Ravi Kumar, Shubhankar Borse, Abhinav Valada, and Senthil Yogamani. 2023. Multi-camera bird's eye view perception for autonomous driving. *Preprint*, arXiv:2309.09080.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. *Preprint*, arXiv:2311.13165.

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

Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. 2024. Evaluating spatial understanding of large language models. *Transactions on Machine Learning Research*.

Cheng Yan, Guansong Pang, Lei Wang, Jile Jiao, Xuetao Feng, Chunhua Shen, and Jingjing Li. 2021. Bv-person: A large-scale dataset for bird-view person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10943–10952.

Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. Coupling large language models with logic programming for robust and general reasoning from text. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5186–5219, Toronto, Canada. Association for Computational Linguistics.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *Preprint*, arXiv:2306.13549.

Chao Zhang, Mohan Li, Ignas Budvytis, and Stephan Liwicki. 2024a. Dialoc: An iterative approach to embodied dialog localization. *Preprint*, arXiv:2403.06846.

Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. 2024b. On the out-of-distribution generalization of multimodal large language models. *Preprint*, arXiv:2402.06599.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v(ision) is a generalist web agent, if grounded. *Preprint*, arXiv:2401.01614.

Ruiqi Zhong, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. 2021. Are larger pretrained language models uniformly better? comparing performance at the instance level. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3813–3827, Online. Association for Computational Linguistics.

A Additional Related Work

In addition to Section 2 which provides a brief overview of previous work most relevant to our work, for completeness we also provide additional related work focused on unimodal spatial reasoning from text only.

Spatial Reasoning on Text. Spatial reasoning has been investigated with the advancement of LLMs (Yamada et al., 2024). Various benchmarks have been proposed to evaluate models' spatial reasoning abilities, including relative spatial relation recognition (Weston et al., 2016; Mirzaee et al., 2021; Shi et al., 2022), natural language navigation (Yamada et al., 2024), and planning (Momennejad et al., 2023). Mirzaee and Kordjamshidi (2022) suggest that introducing synthetic data of spatial reasoning when pre-training helps to improve the spatial awareness of the model. Yang et al. (2023) justify the feasibility of using a logical form as an intermediate representation to improve the spatial reasoning ability in easy scenarios. Instead of describing the spatial relations with natural language, Wu et al. (2024) feed the model with a 2D square grid similar to ASCII-art format and prove that visualising the reasoning procedure explicitly helps to improve the model's ability in multi-hop spatial reasoning. Constrained by language descriptions, most datasets focus on reasoning over symbols within simple scenarios (*e.g. grid-based navigation*) and are synthetically generated. However, real-life scenarios are often more complex and rich in physical semantics. This raises concerns about the models' actual spatial reasoning abilities compared to their proficiency in understanding linguistic patterns.

B Further Details on Dataset Construction

The TopViewRS is derived from Matterport3D (Chang et al., 2017) and is supposed to be used for non-commercial academic use only, under the Term of Use (Matterport End User Licence Agreement For Academic Use of Model Data).

In addition to the main content in Section 4, we provide further details with regard to TopViewRS dataset construction in what follows.

B.1 Top-View Map Construction

To ensure high-quality top-view map representations, we exclude the 3D environments with low coverage of mesh grids. We also prefer environments that are single-floor, in order to avoid the obstruction of objects from different floors. After manually going through 90 building-scale 3D environments from Matterport3D (Chang et al., 2017), we select a total of 7 scenes: `17DRP5sb8fy`, `2azQ1b91cZZ`, `2t7WUuJeko7`, `5LpN3gDmAk7`, `EU6Fwq7SyZv`, `8WUmhLawc2A`, `i5noydFURQK`.

Photo-Realistic Top-View Map. We extract realistic top-view maps using MeshLab by placing an orthographic camera on the top of the 3D scenes and taking a camera shot.

Semantic Top-View Map. We construct them using the Habitat simulation environment (Manolis Savva* et al., 2019; Szot et al., 2021). For each building floor, Matterport3D contains the 2D and 3D semantic segmentation human annotations, which can be retrieved to identify the type of objects as well as the rooms. The 3D coordinates of the entity's (object and room) center (x_i, y_i, h_i) and the size of the entity's bounding box (w_x, w_y, w_h) can also be retrieved as part of the circumstantial information. This information is then used for the construction of the semantic top-view map.

When we obtain the object information for the purpose of constructing a top-view semantic map, we design certain rules to exclude specific types of objects from all 40 object annotation categories of Matterport3D. We believe these objects could either 1) be less meaningful in terms of semantics or 2) take up a large area in the semantic map, which obstructs other objects beneath. The filtered objects include: `'misc'`, `'ceiling'`, `'objects'`, `'floor'`, `'wall'`, `'void'`, `'curtain'`, `'column'`, `'beam'`, `'board panel'`.

We also filter out the objects based on their heights h_{obj} and sizes w_{obj} compared to the rooms' heights h_{room} and sizes w_{room}. We only keep the objects if they satisfy the following relations:

$$0.9 \times (h_{room} - \frac{1}{2} w_{room}) \leq h_{obj} - \frac{1}{2} w_{obj}$$

$$1.1 \times (h_{room} + \frac{1}{2} w_{room}) \geq h_{obj} + \frac{1}{2} w_{obj}$$

After having all the object annotations, we use the `get_topdown_map` API of the Habitat simulator to get the top-down map of the scene, which describes the navigable area and the overall shape of the environment, but without any object annotations. Based on this map, we then draw the bounding boxes with different colors to represent the objects in the environments. Considering that the objects on the top may obstruct the bottom objects in the top-view map, to mimic this characteristic, we create the semantic top-view map based on the heights of the objects, where lower objects are drawn first. Table 4 shows the mapping between the RGB values and object types used for the creation of a semantic top-view map in our work.

RGB Values	Label
[31, 119, 180]	void
[174, 199, 232]	wall
[255, 127, 14]	floor
[255, 187, 120]	chair
[44, 160, 44]	door
[152, 223, 138]	table
[214, 39, 40]	picture
[255, 152, 150]	cabinet
[148, 103, 189]	cushion
[197, 176, 213]	window
[140, 86, 75]	sofa
[196, 156, 148]	bed
[227, 119, 194]	curtain
[247, 182, 210]	chest_of_drawers
[51, 105, 30]	plant
[199, 199, 199]	sink
[188, 189, 34]	stairs
[219, 219, 141]	ceiling
[23, 190, 207]	toilet
[158, 218, 229]	stool
[57, 59, 121]	towel
[82, 84, 163]	mirror
[107, 110, 207]	tv_monitor
[156, 158, 222]	shower
[99, 121, 57]	column
[140, 162, 82]	bathtub
[181, 207, 107]	counter
[206, 219, 156]	fireplace
[140, 109, 49]	lighting
[189, 158, 57]	beam
[231, 186, 82]	railing
[231, 203, 148]	shelving
[132, 60, 57]	blinds
[173, 73, 74]	gym_equipment
[214, 97, 107]	seating
[231, 150, 156]	board_panel
[123, 65, 115]	furniture
[165, 81, 148]	appliances
[206, 109, 189]	clothes
[222, 158, 214]	objects

Table 4: RGB values and corresponding labels.

After having the top-view maps of the whole floor, we crop them into smaller rooms according to the region boundaries obtained from the Habitat simulator.

B.2 Structured Question Framework Design

In order to minimize human labor and standardize the collection pipeline, we adopt the template-based question generation method following the practice of Liu et al. (2023a); we design 15 different templates in total to construct the sub-tasks for each task. In particular, we consider benchmarking different perspectives of the model's ability within each task in a fine-grained manner when designing the templates. The question templates are also multi-scale in terms of objects or rooms with full or partial top-view maps for Top-View Recognition, Top-View Localization and Static Spatial Reasoning. For Dynamic Spatial Reasoning, the designed questions evaluate the recognition and reasoning from the scale of single navigation points (Dynamic Action Counting and Spatial Localization) to the whole path (Dynamic Relative Spatial Reasoning).

Below we provide the designed templates for all 9 sub-tasks, with some examples shown in Figure 1. In what follows, we also introduce the logic for selecting the correct answer and other wrong choices when constructing the multiple-choice questions.

B.2.1 Top-View Recognition

Table 5 shows the templates we use for the Top-View Recognition task. Considering that some objects and rooms may be hard to recognize from the top view, in addition to the set of filtered objects, we also remove some objects (`'picture'`, `'mirror'`, `'window'`, `'blinds'`, `'towel'`, `'furniture'`, `'door'`, `'tv_monitor'`, `'cabinet'`) and rooms (`'hallway'`, `'entryway/foyer/lobby'`, `'tv'`) when we use the templates to generate questions.

B.2.2 Top-View Localization

Table 6 shows the templates for the Top-View Localization task. For the objects, we adopt the same set as for Top-View Recognition. Concerning rooms, we define a set of rooms that are easy and natural to recognize for humans, spanning: `'office'`, `'workout/gym/exercise'`, `'kitchen'`, `'bedroom'`, `'dining room'`, `'bar'`, `'balcony'`, `'toilet'`, `'bathroom'`,

Object Recognition	
Template 1	Which of the following objects are in the room?
Template 2	Which of the following objects are not in the room?
Scene Recognition	
Template 1	What room is this?
Template 2	What types of rooms are included in the top-view map below?

Table 5: Templates for Object and Scene Recognition sub-tasks.

Object Localization	
Template 1	Where is the <object> in the top-down map?
Scene Localization	
Template 1	Where is the <room> in the top-down map?
Template 2	What objects does <room> have?

Table 6: Templates for Object and Scene Localization sub-tasks.

Scene Counting	
Template 1	How many <room> are there in the map?
Relative Spatial Relation	
Template 1	What's <object1>'s relative spatial relation to <object2>?
Template 2	What's <room1>'s relative spatial relation to <room2>?

Table 7: Templates for Scene Counting and Relative Spatial Relation sub-tasks.

Dynamic Action Counting	
Template 1	How many turning <action> are there along the path?
Dynamic Relative Spatial Reasoning	
Template 1	Which direction does the navigation path head for?
Dynamic Spatial Localization	
Template 1	What rooms does the navigation path pass by?
Template 2	At which room does the navigation path <action>?
Template 3	At which object does the navigation path <action>?

Table 8: Templates for Dynamic Action Counting, Dynamic Relative Spatial Reasoning, and Dynamic Spatial Localization sub-tasks

`'living room'`, `'stairs'`.

B.2.3 Static Spatial Reasoning

Table 7 lists the templates for the Static Spatial Reasoning task. For rooms, we restrict the regions within the same range as in Top-View Localization. Concerning objects, we focus on the objects that are common and large enough to recognize in daily life, which includes: `'chair'`, `'table'`, `'cushion'`, `'sofa'`, `'bed'`, `'chest_of_drawers'`, `'sink'`, `'toilet'`, `'bathtub'`, `'stool'`, `'plant'`, `'stairs'`, `'shower'`, `'fireplace'`, `'gym_equipment'`, `'seating'`.

B.2.4 Dynamic Spatial Reasoning

For Dynamic Action Counting, we define that a valid turn should involve more than a 30-degree rotation. For Dynamic Relative Spatial Reasoning, the direction is also defined by the relative spatial relation between the starting point and ending point, where the spatial description is determined by 30-degree intervals.

Multiple-Choice Question-Answer Pairs. For the answer to the questions, because we have all the spatial information and semantic annotation of the objects in the scene, we write a set of rules with code for each type of question in order to automatically obtain the golden answer according to the simulation environments. For all the wrong choices in the multiple-choice settings, they are randomly chosen from other possible candidates of the same kind (e.g. objects, rooms, numbers, etc.). After having all the options for multiple-choice questions, we randomize the order of the options to make the correct choices evenly distributed among possible options A, B, C, and D.

B.3 Alignment with Human Judgments

In our preliminary quality control, we realized that semantic annotations of environments may sometimes be inaccurate. Moreover, even though we exclude some unreasonable objects, the top view of certain objects can sometimes be challenging to recognize, even for humans. To address these issues, we have implemented a second stage in our dataset creation process: alignment and verification based on human judgments.

When validating the automatically collected data, the human participants are supposed to check the correctness of the question-answer pair and choose one of the following four actions according to their own judgments: **1)** skip the instance if it cannot be repaired and/or looks strange, **2)** modify the pair by replacing the options or the entities in the question in order to make it answerable by humans, **3)** correct the answer if it is wrong, **4)** keep the data if it is answerable by humans and correct. In order to ensure the quality of the dataset, we communicated to the human participants that they are supposed to be cautious when 'accepting' a data point/instance. On a practical level, the participants may either discard this data point or modify the options of this data to make the correct choice more distinguishable by humans. This helps to exclude the data points where different human judges may diverge and thus ensure the alignment between the dataset and general human judgments. We assure that the alignment process does not include any information with regard to personal identification or offensive content.

In our experiments, we also provide the corresponding rules of how we obtain the answer for the model with textual description in the prompt (see Appendix C.2).

B.4 Dataset Statistics

We provide further insight into different portions of the TopViewRS dataset with regard to the object and room distribution in Figure 4, whereas statistics over different sub-tasks are provided in Table 9.

The visualization demonstrates that the objects or regions that are hard to recognize (*e.g. gym equipment, utility room, etc.*) have fewer occurrences in the dataset compared to those which are easier to identify and typically more common (*e.g. bed, table, bedroom, etc.*). *Bed, chair* and *table* are the top-3 most frequently mentioned objects and *bedroom, dining room* and *living room* are the most common regions in the dataset. Among all the spatial descriptions, the diagonal spatial relations (*e.g. top right, up left*) are more frequently referred to as the correct choice as relative spatial descriptions in Static Spatial Reasoning while being less frequently used as absolute spatial descriptions in Top-View Localization.

Regarding the dataset size per each sub-task, object-level recognition and localization take a large portion of data in the Top-View Recognition and Localization tasks. For Static Spatial Reasoning, reasoning over relative spatial relations takes the main part of the data. Dynamic Spatial Localization has the largest number of data instances overall. The numbers are different with realistic maps and semantic maps for each task. The disparity stems from the second stage of dataset creation, where the human annotators have excluded more data points associated with more complex, photo-realistic maps due to various possible reasons.

C Experiments: Additional Information

C.1 Inference Parameters

We adopt most of the inference parameters for each model from the implementations of VLMEvalKit (OpenCompass Contributors, 2023). Table 10 shows the configuration of the inference process for different models. If not specified in Table 10, we use the default configuration in Huggingface.

C.2 Prompts

Table 11 and 12 show the prompt templates of each task used in the main experiments (Table 1) with realistic and semantic top-view maps as visual input, respectively. Table 13 and 14 show the prompt templates used for Chain-of-Thought reasoning using realistic and semantic top-view maps (Table 3).

Within the prompt templates, `<QUESTION>` and `<OPTIONS>` are replaced with the question and option list $O = \{o_0, o_1, o_2, o_3\}$ (*e.g. "A. bed; B. chair; C. table; D. cushion"*). For semantic top-view maps, `<MAPPING>` is replaced with the RGB-object mapping, as shown below.

```
(196, 156, 148) -> bed
(44, 160, 44) -> door
...
```

In the task of Dynamic Spatial Reasoning, `<TASK-SPECIFIC INSTRUCTION>` contains the rules of how we obtain the answer from the simulator for the sub-task Dynamic Action Counting, which is described as follows.

```
Suppose you are a navigation agent tracing
the path. Your job is to assess whether
there's a turn at each intermediate point
```

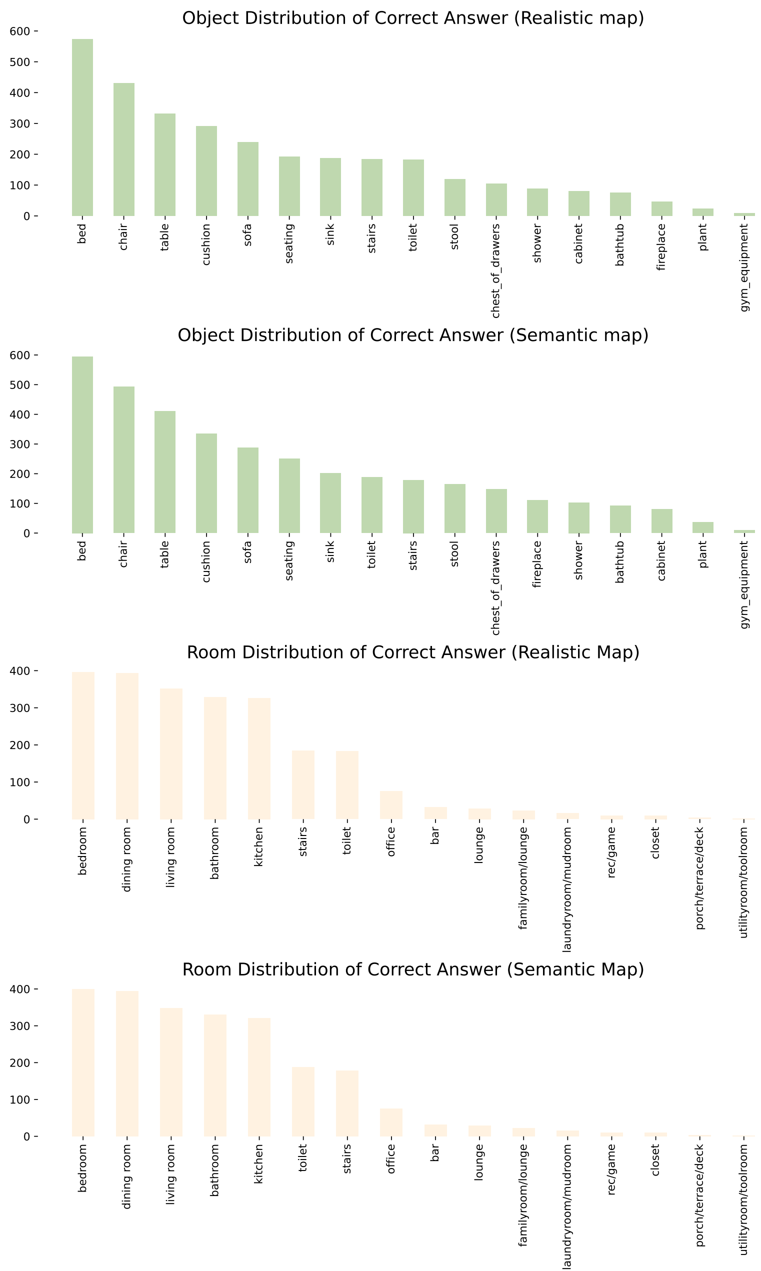


Figure 4: Additional statistics of the TopViewRS dataset.

Task	Sub-Task	Realistic	Semantic
TVR	Object Recognition	195	198
	Scene Recognition	97	97
TVL	Object Localization	410	470
	Scene Localization	100	97
SSR	Scene Counting	43	48
	Relative Spatial Relation	1,004	1,245
DSR	Dynamic Action Counting	668	668
	Dynamic Spatial Localization	2,436	2,436
	Dynamic Relative Spatial Reasoning	586	586
Total		5,539	5,845

Table 9: Distribution of sub-tasks with realistic and semantic top-view maps.

Idefics 9B&80B	
max_new_tokens	20
LLaVANext 7B&13B&34 B	
temperature	0
num_beams	1
max_new_tokens	20
do_sample	False
top_p	None
XComposer2	
temperature	1
beams	5
max_token	20
repetition_penalty	1
do_sample	False
Qwen-VL	
max_new_tokens	20
GPT4V	
temperature	0
max_tokens	1024
img_size	512
img_detail	low
Gemini	
temperature	0
max_tokens	1024

Table 10: Configurations of inference parameters.

and sum up the total turns for the final outcome.

For other sub-tasks in Dynamic Spatial Reasoning, <TASK-SPECIFIC INSTRUCTION> is replaced with an empty string.

C.3 Additional Experimental Results

Table 15 shows the fine-grained sub-task performance of all the models, which corresponds to Figure 3 in the main paper.

Realistic Top-View Maps
Top-View Recognition, Top-View Localization and Static Spatial Reasoning This is a top-view map of a room. Please respond to the question below by selecting one choice from a list of available options provided. Your response should only include the letter of the chosen option (A, B, C, or D) with no additional explanation. Question: <QUESTION> Options: <OPTIONS>; Answer:
Dynamic Spatial Reasoning This is a top-view map of a room with the navigation path. The path starts from the green triangle (RGB [0, 255, 0]) and ends at the red star (RGB [255, 0, 0]). The direction of the path is denoted by a series of yellow arrows (RGB [255, 255, 0]), with intermediate points highlighted in RGB [25, 255, 255]. <TASK-SPECIFIC INSTRUCTION> Please respond to the question below by selecting one choice from a list of available options provided. Your response should only include the letter of the chosen option (A, B, C, or D) with no additional explanation. Question: <QUESTION> Options: <OPTIONS>; Answer:

Table 11: Prompt templates for main experiments with realistic top-view maps.

Semantic Top-View Maps

Top-View Recognition, Top-View Localization and Static Spatial Reasoning

This is a semantic top-view map of a room. Various objects are depicted by colored bounding boxes, each with its corresponding color, and there may be instances of overlap between them. Below are the RGB color codes associated with each object, presented in the format RGB -> Object:
<MAPPING>
Please respond to the question below by selecting one choice from a list of available options provided. Your response should only include the letter of the chosen option (A, B, C, or D) with no additional explanation.
Question: <QUESTION>
Options: <OPTIONS>;
Answer:

Dynamic Spatial Reasoning

This is a semantic top-view map of a room with the navigation path. In the semantic map, various objects are depicted by colored bounding boxes, each with its corresponding color, and there may be instances of overlap between them. The navigation path starts from the green triangle (RGB [0, 255, 0]) and ends at the red star (RGB [255, 0, 0]). The direction of the path is denoted by a series of yellow arrows (RGB [255, 255, 0]), with intermediate points highlighted in RGB [25, 255, 255]. Below are the RGB color codes associated with each object and symbol, presented in the format RGB -> Object:
<MAPPING>
<TASK-SPECIFIC INSTRUCTION> Please respond to the question below by selecting one choice from a list of available options provided. Your response should only include the letter of the chosen option (A, B, C, or D) with no additional explanation.
Question: <QUESTION>
Options: <OPTIONS>;
Answer:

Table 12: Prompt templates for main experiments with semantic top-view maps.

Realistic Top-View Maps

Static Spatial Reasoning

This is a top-view map of a room. Please respond to the question below by selecting one choice from a list of available options provided. You should explain your reasoning step-by-step by first localizing the entities and then reasoning over the question based on the locations. You should conclude your chosen option (A, B, C, or D) starting with 'The answer is '.
Question: <QUESTION>
Options: <OPTIONS>;
Answer: Let's think step by step.

Table 13: Prompt templates for Chain-of-Thought experiments with realistic top-view maps.

Semantic Top-View Maps

Static Spatial Reasoning

This is a semantic top-view map of a room. Various objects are depicted by colored bounding boxes, each with its corresponding color, and there may be instances of overlap between them. Below are the RGB color codes associated with each object, presented in the format RGB -> Object:
<MAPPING>
Please respond to the question below by selecting one choice from a list of available options provided. You should explain your reasoning step-by-step by first localizing the entities and then reasoning over the question based on the locations. You should conclude your chosen option (A, B, C, or D) starting with 'The answer is '.
Question: <QUESTION>
Options: <OPTIONS>;
Answer: Let's think step by step.

Table 14: Prompt templates for Chain-of-Thought experiments with semantic top-view maps.

		Idefics		**LLaVANext**				**XComposer2**	**Qwen-VL**	**GPT-4V**	**Gemini**
Model Size		9B	80B	vicuna 7B	mistral 7B	vicuna 13B	34B	7B	7B	API	API
Realistic Map											
TVR	Object Recognition	32.31	25.64	66.15	61.03	58.97	65.64	38.97	17.95	68.21	89.23
	Scene Recognition	58.76	28.87	70.10	61.86	67.01	72.16	35.05	45.36	72.16	92.78
TVL	Object Localization	26.83	26.10	40.24	30.24	40.00	50.49	26.34	16.34	40.73	45.21
	Scene Localization	45.00	46.00	50.00	49.00	46.00	53.00	34.00	16.00	69.00	61.00
SSR	Scene Counting	25.58	32.56	16.28	18.60	2.33	16.28	25.58	48.84	76.74	53.49
	Relative Spatial Relations	24.00	25.80	20.02	24.60	21.02	23.01	25.80	13.25	19.82	30.68
DSR	Dynamic Action Counting	31.89	26.80	27.54	26.95	25.30	27.40	27.54	32.34	22.01	26.95
	Dynamic Spatial Localization	42.57	29.27	45.03	23.89	32.88	24.63	22.62	20.11	34.15	35.30
	Dynamic Relative Spatial Reasoning	26.62	23.72	25.77	23.04	17.58	16.21	26.11	18.77	23.72	27.82
Semantic Map											
TVR	Object Recognition	54.55	51.01	92.93	87.37	92.42	98.48	50.00	12.63	100.00	99.49
	Scene Recognition	73.20	76.29	80.41	64.95	79.38	86.60	28.87	34.02	91.75	85.57
TVL	Object Localization	28.51	23.19	22.98	28.94	10.85	34.47	24.47	6.17	41.49	31.28
	Scene Localization	44.33	47.42	37.11	47.42	48.45	57.73	26.80	26.80	58.76	54.64
SSR	Scene Counting	37.50	50.00	12.50	10.42	6.25	4.17	14.58	22.92	37.50	20.83
	Relative Spatial Relations	23.29	27.23	18.96	24.82	17.03	18.96	23.37	14.54	21.12	26.43
DSR	Dynamic Action Counting	36.68	27.69	33.38	26.80	28.89	25.30	27.25	30.39	22.90	26.95
	Dynamic Spatial Localization	39.20	38.79	41.87	25.82	17.98	39.00	19.46	22.95	47.17	33.42
	Dynamic Relative Spatial Reasoning	26.11	24.74	23.72	27.30	17.75	17.58	24.06	18.26	25.26	28.16

Table 15: Fine-grained results of 10 VLMs on different sub-tasks, corresponding to the visualization in Figure 3.