This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SpatiaLQA: A Benchmark for Evaluating Spatial Logical Reasoning in Vision-Language Models

Yuechen Xie[1,3*], Xiaoyan Zhang[1*], Yicheng Shan[2*], Zhu Hao[3],
Rui Tang[3†], Rong Wei[3], Mingli Song[1], Yuanyu Wan[1], Jie Song[1†]

[1]Zhejiang University, [2]The University of Sydney, [3]Manycore Tech Inc.

Abstract

Vision-Language Models (VLMs) have been increasingly applied in real-world scenarios due to their outstanding understanding and reasoning capabilities. Although VLMs have already demonstrated impressive capabilities in common visual question answering and logical reasoning, they still lack the ability to make reasonable decisions in complex real-world environments. We define this ability as spatial logical reasoning, which not only requires understanding the spatial relationships among objects in complex scenes, but also the logical dependencies between steps in multi-step tasks. To bridge this gap, we introduce Spatial Logical Question Answering (SpatiaLQA), a benchmark designed to evaluate the spatial logical reasoning capabilities of VLMs. SpatiaLQA consists of 9,605 question answer pairs derived from 241 real-world indoor scenes. We conduct extensive experiments on 41 mainstream VLMs, and the results show that even the most advanced models still struggle with spatial logical reasoning. To address this issue, we propose a method called recursive scene graph assisted reasoning, which leverages visual foundation models to progressively decompose complex scenes into task-relevant scene graphs, thereby enhancing the spatial logical reasoning ability of VLMs, outperforming all previous methods. Code and dataset are available at https://github.com/xieyc99/SpatiaLQA.

1. Introduction

Vision-Language Models (VLMs) [23, 37, 38, 44, 64, 88, 89, 97, 98] have recently been increasingly applied to interpret and reason about complex real-world scenes, achieving remarkable progress across various domains such as Visual Question Answering (VQA) [5, 13, 18, 55], task planning [57, 84, 92], image captioning [52, 79, 83], scene understanding [10, 15, 24, 25, 40, 46, 50, 73, 95], scene segmentation [33–36, 67], and medical diagnosis [11, 12]. As shown in the first two examples of Fig. 1, state-of-the-art VLMs such as GPT-4o [28] perform well on common VQA [27, 42, 66] and logical reasoning [61, 90] tasks. However, in the third example, it fails to remove the obstacles above the bottom book before picking it up, indicating that its performance on tasks requiring both spatial understanding and logical reasoning remains unsatisfactory. We refer to this important task as *spatial logical reasoning*. Such tasks not only require models to possess strong spatial understanding [9, 17] but also demand the ability to reason through a sequence of logically consistent steps [72].

Moreover, although spatial logical reasoning shares similarities with Embodied Question Answering (EQA) [39, 57, 85, 92], as both require an understanding of spatial relations and reasoning, its research value is distinct and irreplaceable. The primary focus of EQA is to evaluate whether an agent can translate abstract language instructions into physically executable action sequences under the constraints of real-world dynamics and control strategies. These action sequences are typically selected from a predefined and limited set of motor primitives (*e.g.* move forward, turn left), forming a closed output space. In contrast, spatial logical reasoning does not involve any execution component. Instead, it emphasizes whether the model can deduce a logically consistent and spatially coherent multi-step reasoning process purely at the visual-semantic level, where the answers belong to an open vocabulary space. This open-ended nature demands higher levels of cognitive and linguistic abstraction, reflecting a model's intrinsic reasoning and compositional understanding rather than its ability to map instructions to preset fixed actions. Therefore, spatial logical reasoning serves as the cognitive basis for EQA, without relying on direct physical interactions. Advancing spatial logical reasoning is not only essential for improving performance in embodied tasks, but also for enhancing the overall reasoning capacity of VLMs across diverse real-world domains. Unfortunately, existing benchmarks fail to systematically reflect the performance of VLMs in this as-

*Equal Contribution, †Corresponding Author.

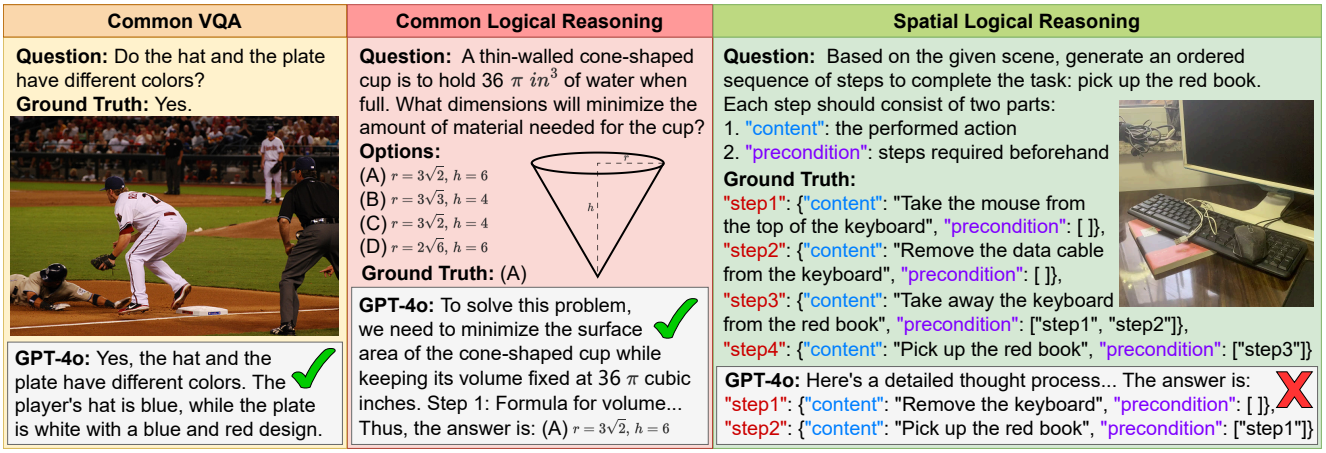

Figure 1. Common VQA [27] typically involve recognizing visual content and factual knowledge, while common logical reasoning [90] focuses on abstract, symbolic problem-solving. Spatial logical reasoning, in contrast, requires integrating both spatial understanding and multi-step logical reasoning to accomplish tasks in real-world scenes.

pect, leaving a critical gap that constrains their safe and effective deployment in real-world scenarios [63, 93, 94].

To address this important yet unexplored issue, we introduce Spatial Logical Question Answering (SpatiaLQA) in this work, a benchmark dataset consisting of 9,605 image–text question answer (QA) pairs collected from 241 indoor scenes spanning 13 scene categories. Considering the difficulty of acquiring such data, particularly since scenes involving complex logical relationships need to be carefully arranged, the entire annotation process is divided into three stages: manual annotation, subgraph extraction augmentation, and graph expansion augmentation. Specifically, we first manually annotated 2,401 real indoor scene images, assigning one QA pair to each image. We then applied subgraph extraction augmentation to these 2,401 samples, which derives subsets of the original answer steps based on their logical dependencies, and obtained 2,251 new QA pairs. Finally, we performed graph expansion augmentation on the combined 4,652 samples, where heuristic methods were used to append several logically consistent steps to the original answers, resulting in 4,953 new QA pairs.

In addition, we systematically evaluated 41 representative VLMs. Specifically, the evaluation process consists of three steps. First, we use GPT-4o to perform step-level matching between the model's predictions and the ground-truth annotations. Next, the Hungarian algorithm [58] is applied to generate the optimal one-to-one step matching that achieves the maximum number of pairs. Finally, based on the matched results, we calculate precision and recall for both the content and preconditions. Extensive experiments show that most models perform poorly in spatial logical reasoning, especially in complex tasks that require many steps.

To improve the spatial logical reasoning capabilities of VLMs, we propose a method called *recursive scene graph assisted reasoning*. Specifically, our method consists of three steps: (1) We first use Depth Anything V2 [81] and SAM [31, 47, 48] to obtain the depth map and segmentation map of the scene image; (2) Based on the original image and these perception results, we take the object specified in the task as the initial *source object* and perform the first round of scene graph generation using the VLM. This process identifies the objects in direct contact with the source object, referred to as *target objects*, along with their relative spatial relationships. Then construct a scene graph with the source and target objects as nodes and the spatial relationships as edges. This scene graph serves as the input for the next iteration, where the previous target objects are regarded as new source objects, and the process repeats until reaching the maximum iteration number; (3) Finally, the generated scene graph and the prompt are jointly fed into the VLM to produce the final answer. By leveraging domain-specialized visual foundation models, our method incrementally decomposes the complex visual scenes into task-relevant scene graphs. This hierarchical perception process allows the VLM to focus on the spatial environment surrounding the target objects, thereby facilitating more accurate multi-step reasoning.

In this work, we make four primary contributions: (1) We identify and define spatial logical reasoning as a critical yet underexplored capability of VLMs, highlighting its importance for reasoning across interdependent spatial and logical steps in real-world scenarios. (2) We introduce SpatiaLQA, a large-scale benchmark consisting of 9,605 image–text QA pairs across 241 indoor scenes spanning 13 scene categories, to comprehensively evaluate spatial logical reasoning. (3) We conduct a systematic evaluation of 41 representative VLMs using GPT-4o and the Hungarian algorithm, revealing that most models struggle with spatial logical reasoning, particularly in complex tasks that require many steps. (4) We propose a novel method, recur-

Benchmarks	Modality	SU	LR	Real Scene	Answer Type	Multi-step	Precondition	LLM/VLM Scoring	Size
CLEVR [30]	I	✓	✗	✗	Open	✗	✗	✗	853.6K
GQA [27]	I	✓	✗	✓	Open	✗	✗	✗	>1M
MMBench [49]	I	✓	✗	✓	MC	✗	✗	✓	3.2K
Spatial-MM [65]	I	✓	✗	✓	MC	✗	✗	✗	2.3K
SpatialRGPT-Bench [17]	I	✓	✗	✓	MC	✗	✗	✗	1.5K
Open3DVQA [91]	I	✓	✗	✗	Open	✗	✗	✓	9.0K
MathVista [51]	I	✗	✓	✗	MC/Open	✗	✗	✗	6.1K
MMMU [90]	I	✗	✓	✗	MC/Open	✗	✗	✗	11.5K
GeoQA [14]	I	✗	✓	✗	MC	✗	✗	✗	5.0K
ChartQA [56]	I	✗	✓	✗	Open	✗	✗	✗	32.7K
MT-EQA [86]	I	✓	✓	✗	MC	✗	✗	✗	20.0K
MP3D-EQA [77]	I/P	✓	✓	✓	MC	✗	✗	✗	1.1K
OpenEQA [53]	I/V	✓	✓	✓	Open	✗	✗	✓	2.1K
EmbSpatial-Bench [21]	I	✓	✓	✓	MC	✗	✗	✗	3.6K
EmbodiedBench [82]	I	✓	✓	✗	MC	✓	✗	✗	1.1K
SpatiaLQA (Ours)	I	✓	✓	✓	Open	✓	✓	✓	9.6K

Table 1. Comparison between SpatiaLQA and other benchmarks (VQA, logical reasoning and EQA). 'SU' and 'LR' denote spatial understanding and long-range reasoning, respectively. 'I', 'P' and 'V' denote image, point cloud and video, respectively. 'Open' and 'MC' stand for open-vocabulary and multiple-choice respectively. 'Multi-step' specifies whether the answers involve multiple steps. 'Precondition' indicates whether each step in the answer is annotated with its preconditions (i.e., which steps must be completed beforehand).

sive scene graph assisted reasoning, which utilizes visual foundation models to decompose complex scenes into task-specific scene graphs, improving the spatial logical reasoning capability of VLMs.

2. Related Work

We present the main differences between SpatiaLQA and some representative benchmarks in Tab. 1, and provide detailed analyses of how SpatiaLQA differs from VQA, logical reasoning, and EQA in Sec. 2.1. Sec. 2.2 presents the current state of VLMs.

2.1. Related Benchmarks

Visual Question Answering. Common VQA [5, 54, 55] primarily focused on recognizing image content and handling short-range reasoning tasks, such as object or attribute recognition [27], counting [30], and simple spatial relation reasoning [53]. In addition, their efforts largely remained at the level of single-step factual question answering. In contrast, SpatiaLQA emphasizes spatial logical reasoning, where the model must perform long-range reasoning and infer a sequence of dependent operations based on spatial relations, rather than simply outputting a factual answer.

Logical Reasoning. Logical reasoning tasks, such as mathematical reasoning [14, 51, 90], assess a model's ability to establish causal, implicational, and consistency relationships within complex information. However, these tasks are mostly confined to abstract textual or symbolic spaces, where the reasoning process is decoupled from real-world spatial information and visual structures. In contrast, SpatiaLQA focuses on spatial logical reasoning, which requires models to jointly understand spatial relationships and causal logic within realistic scenes. Fundamentally, spatial logical reasoning bridges spatial understanding and logical reasoning, representing a more challenging and practically significant form of reasoning.

Embodied Question Answering. EQA focuses on the feasibility and effectiveness of actions in interactive environments, such as navigation [53, 71, 96] and manipulation [20, 39, 57] tasks. The primary focus of EQA is to evaluate whether an agent can translate abstract language instructions into physically executable action sequences under the constraints of real-world dynamics and control strategies. These action sequences are typically selected from a predefined and limited set of motor primitives, forming a closed output space. In contrast, SpatiaLQA emphasizes whether the model can deduce a logically consistent and spatially coherent multi-step reasoning process purely at the visual-semantic level, where the answers belong to an open vocabulary space. This reasoning ability, namely spatial logical reasoning, forms the cognitive foundation for embodied tasks without relying on physical interaction.

2.2. Vision-Language Models

Recently, VLMs have achieved remarkable success in tasks such as image captioning [52, 79, 83], visual question answering [5, 54], and open-ended multimodal dialogue [29, 80], driven by large-scale pretraining [41, 74], instruction tuning [16, 32], and external tool augmentation [62]. Despite their strong generalization and reasoning capabilities, their performance in complex scenarios that require reasoning across multiple interdependent steps has not been systematically studied. SpatiaLQA directly addresses this gap

Prompt Template

Given a scene image and a task, please provide a concise and accurate list of steps to execute the task in order.
Here is an example of answer:"step1": {"content": "Take away the stapler from the top of the book", "precondition":[ ]}, "step2": {"content": "Remove the keys from the top of the book", "precondition":[ ]}, "step3": {"content": "Remove the toliet paper from the top of the laptop", "precondition":[ ]}, "step4": {"content": "Remove the book from the top of the laptop", "precondition":["step1", "step2"]}, "step5": {"content": "Pick up the laptop", "precondition":["step3", "step4"]}
Please note the following: 1. The example above is only for illustrating the answer format and is not related to the question you need to answer. **2.** All objects in the image are within your reach, so no movement is necessary. **3.** Each step can only perform one action on a single object. **4.** Each step in the answer consists of two parts: the action to be performed in this step ("content") and the steps that must be completed beforehand ("precondition").
Now answer the following question based on the input image: **{Question}**. You only need to return an answer wrapped with "<ans></ans>" tags.

Lounge

Question: Clean the table with the cloth.
Answer:
"step1": {"content": "Remove the cup from the top of the duster cloth", "precondition":[ ]}, "step2": {"content": "Remove the bottle from the top of the duster cloth", "precondition":[ ]}, "step3": {"content": "Pick up the duster cloth", "precondition": ["step1", "step2"]}, "step4": {"content": "Wipe the table with the duster cloth", "precondition": ["step3"]}

Kitchen

Question: Pick up the fruit box.
Answer:
"step1": {"content": "Remove the bowl from the top of the plastic box", "precondition":[ ]}, "step2": {"content": "Remove the cup from the top of the plastic box", "precondition":[ ]}, "step3": {"content": "Remove the plastic storage box from the front of the fruit box", "precondition": ["step1", "step2"]}, "step4": {"content": "Take out the fruit box", "precondition": ["step3"]}

Garage

Question: Put the bottle in the trunk.
Answer:
"step1": {"content": "Pick up the beverage from the top of the trunk", "precondition": [ ]}, "step2": {"content": "Remove the phone from the top of the trunk", "precondition": [ ]}, "step3": {"content": "Open the trunk", "precondition": ["step1", "step2"]}, "step4": {"content": "Put the bottle in the trunk", "precondition": ["step3"]}, "step5": {"content": "Close the trunk", "precondition": ["step4"]}

Office

Question: Pick up the yellow box.
Answer:
"step1": {"content": "Remove the candy from the black leather bag", "precondition": [ ]}, "step2": {"content": "Remove the black leather bag from the top of the yellow box", "precondition": ["step1"]}, "step3": {"content": "Remove the hairpin from the top of the yellow box", "precondition": [ ]}, "step4": {"content": "Pick up the yellow box", "precondition": ["step2", "step3"]}

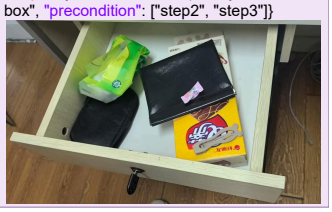

Figure 2. Prompt template and examples of several indoor scenes.

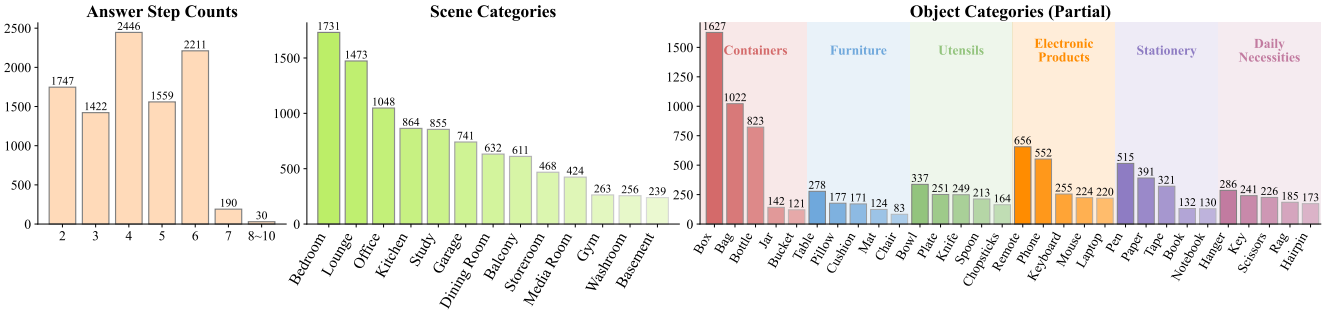

Figure 3. The distributions of answer step counts, scene categories, and partial object categories in SpatiaLQA. The x-axes of the three plots represent the number of answer steps, indoor scene categories, and object categories, while the y-axes indicate the number of samples.

by requiring VLMs to perform ordered and dependency-aware spatial logical reasoning under structured scene constraints. It further provides a systematic analysis of model capabilities in two key aspects: the generation of step content and the inference of preconditions, which enables a comprehensive evaluation of the VLMs' reasoning consistency and spatial understanding, thereby supporting their reliable and safe deployment in real-world scenarios.

3. SpatiaLQA

To evaluate the spatial logical reasoning capabilities of VLMs, we first define the concept of spatial logical reasoning and introduce the SpatiaLQA dataset in Sec. 3.1. Then, we describe the dataset collection process and evaluation metrics in Sec. 3.2 and Sec. 3.3, respectively. In Sec. 3.4, we conduct experiments using SpatiaLQA to determine whether VLMs can effectively perform spatial logical reasoning. Finally, in Sec. 3.5, we analyze the alignment between our metrics and human evaluations, as well as the potential reasons for the poor performance of VLMs.

3.1. Overview of SpatiaLQA

Spatial logical reasoning refers to the ability of a model to solve complex problems by outputting a series of logically coherent steps through spatial understanding and logical reasoning. This capability is crucial yet fundamentally challenging for VLMs, as the model must integrate spatial understanding with logical reasoning, which involves precise spatial perception and tightly coordinated multi-step causal reasoning to ensure safe and effective operation.

However, existing datasets often focus solely on either spatial understanding or logical reasoning, while neglecting the integrated aspect of spatial logical reasoning described above. To bridge this gap, we introduce SpatiaLQA, a benchmark designed to comprehensively evaluate the spatial logical reasoning capabilities of VLMs. The dataset consists of 9,605 QA pairs collected from 241 scenes across

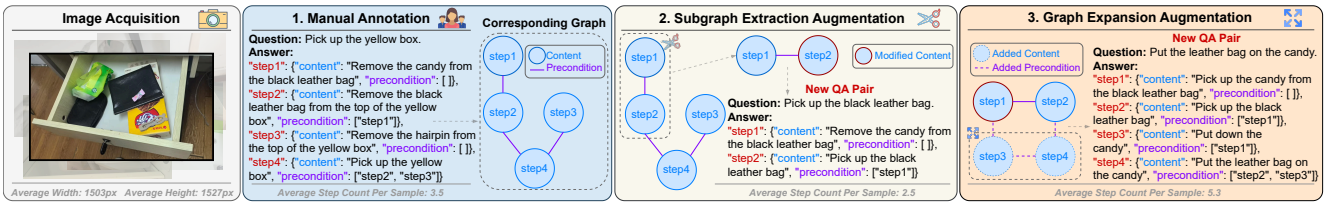

Figure 4. The data collection pipeline for SpatiaLQA. Note that although the graph expansion augmentation in the figure is applied only to the data from subgraph extraction augmentation, we actually also applied graph expansion augmentation to the manually annotated data.

13 real-world indoor scene categories. The prompt template and QA examples are shown in Fig. 2, which includes the required answer format and example outputs that the model should follow. Fig. 3 shows the distribution of answer step counts, scene categories, and partial object categories. It shows that the answer step counts are broadly distributed across the range of 2–10, reflecting a diverse level of task complexity (in general, questions with more answer steps are more challenging). Additionally, the dataset's images come from 13 common indoor scene categories, and the QA pairs involve over a thousand distinct objects (only a subset is shown in the figure), which demonstrates that SpatiaLQA encompasses a rich variety of scenes and objects.

3.2. Dataset Collection Process

We first collected 2,401 real indoor scene images from 241 locations across 13 scene categories, each depicting a complex, multi-step task. Given the challenges in collecting such data, particularly because scenes with complex logical dependencies require deliberate setup, as illustrated in Fig. 4, the collection process consists of three stages: manual annotation, subgraph extraction augmentation, and graph expansion augmentation. The details are as follows:

Manual Annotation. We first employed trained annotators to manually label the 2,401 images, assigning one QA pair per image with answers ranging from 2–8 steps. Notably, we did not annotate implicit preconditions: for instance, if 'step$k$' depends on 'step$j$' and 'step$l$' depends on 'step$k$', we do not mark 'step$j$' as a precondition of 'step$l$'. To ensure data quality, we conducted two rounds of review and correction by professional annotators. Each annotation was represented as an undirected graph, where nodes correspond to step contents and edges indicate logical dependencies between steps, serving as the basis for the following stages.

Subgraph Extraction Augmentation. We applied subgraph extraction augmentation to these 2,401 samples, generating 2,251 new QA pairs. This method derives subgraphs (each containing at least one edge) from the original annotations based on their logical dependencies, forming new QA pairs that share the same image, with the question corresponding to the final step of the subgraph.

Graph Expansion Augmentation. We generated 4,953 new QA pairs using graph expansion augmentation, based on the samples containing 'Remove' and 'Pick up' from the previous two stages. Specifically, assuming the original question and the final step is 'Pick up B', with an intermediate step being 'Remove A', the graph expansion augmentation would change the question to 'Place B on A', modify 'Remove A' to 'Pick up A', and add two additional steps at the end: 'Put down A' and 'Place B on A'. The generated QA pairs share the same image as the original sample.

Note that all data have been manually verified to ensure their reasonableness and feasibility.

3.3. Evaluation Metrics

Although human evaluation is the gold standard for open-vocabulary tasks, it is costly and time-consuming, making automatic metrics preferable for benchmarking. To this end, we first use GPT-4o and the Hungarian algorithm to match the predicted results with the ground-truth annotations, as shown in Fig. 5, and then calculate the recall and precision based on the matching results. Specifically, the evaluation process is divided into three steps: (1) Use GPT-4o to match the predicted steps with the ground truth steps based on the image, i.e., determining whether the semantics of the content in each step are consistent in the image. Given an annotated answer with m steps and a predicted answer with n steps, we use GPT-4o to generate a matching matrix with m rows and n columns, where the values are either 0 or 1. 0 means the two steps differ, while 1 means they are the same. Note that in this step, a predicted (annotated) step can be matched with multiple annotated (predicted) steps. (2) Apply the Hungarian algorithm to filter the matching matrix, removing redundant matches and achieving the maximum one-to-one match between the predicted steps and the annotated steps, resulting in the filtered matching matrix. (3) Using the filtered matching matrix for all samples, we calculated the recall R_c and precision P_c for the all content, as well as the recall R_p and precision P_p for the all preconditions. Finally, we used the F1 score F_c and F_p for content and preconditions as the evaluation metrics.

Please refer to the supplementary material for the prompt used to generate the matching matrix.

3.4. The Evaluation of VLMs

To evaluate the spatial logical reasoning capabilities of VLMs, we conducted experiments on SpatiaLQA with 41

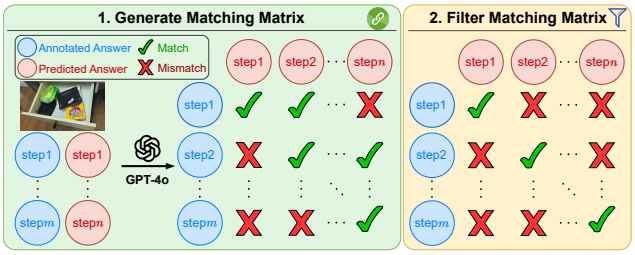

Figure 5. The matching process between the predicted and annotated steps. We first use GPT-4o to match the predicted steps and annotated steps in pairs based on the image (allowing one-to-many matches), resulting in a matching matrix. Then, we apply the Hungarian algorithm to filter the matching matrix, removing redundant matches to achieve the maximum one-to-one matches.

representative VLMs, covering a wide range of types, including those in non-thinking mode and thinking mode, as well as VLMs that only support text–image input and general VLMs that support mixed multimodal inputs. For details on VLM prompts and hyperparameters, please refer to the supplementary materials.

In addition, we recruited a human participant to establish human-level performance on SpatiaLQA. We provided the participant with an answer template in Fig. 2 and asked him to sequentially answer all the questions in SpatiaLQA.

The evaluation results are shown in Tab. 2, with the F1 scores being our primary metrics. We first share some observations and comments as follows:

(1) Generally, within the same series, VLMs with larger parameter sizes perform better, newer versions of VLMs perform better (*e.g.*, Qwen2.5-VL-7B-Ins vs. Qwen3-VL-4B-Ins), and VLMs with a thinking mode outperform those with a non-thinking mode (*e.g.*, Kimi-VL-A3B-Ins vs. Kimi-VL-A3B-Thinking). In addition, the performance of proprietary models typically surpasses that of open-source models. These confirm the effectiveness of the benchmarking and the correctness of the evaluation metrics.

(2) Humans achieved excellent performance on the benchmark, with the lowest metric exceeding 90%. However, even the best-performing VLMs show a significant gap compared to humans, particularly in precondition prediction, which indicates that there is a notable performance gap between VLMs and humans in spatial logical reasoning.

(3) The prediction of preconditions generally performs worse than the prediction of content, indicating that VLMs have significant deficiencies in causal reasoning. Even though they may predict the steps roughly, they do not understand the logical relationships between them.

(4) The recall for content and preconditions is generally lower than precision, indicating that VLMs tend to output more certain answers during prediction, avoiding incorrect 'false positives'. Specifically, the model generates very certain step contents or preconditions, but for uncertain steps, it may choose not to predict or skip them, leading to the omission of some steps that should have been identified. According to the statistics, the best-performing GPT-5 produces answers with an average of 3.1 steps, while the annotated answers have an average of 4.2 steps, which further confirms this observation.

	Size	R_c	P_c	F_c	R_p	P_p	F_p
human	-	97.6	97.6	97.6	92.3	92.7	92.5
BLIP2-OPT [38]	4B	24.6	99.9	39.5	0.0	0.0	-
BLIP2-Flan-T5-xl [38]	4B	24.0	74.3	36.3	0.1	0.2	0.1
BLIP2-Flan-T5-xxl [38]	12B	41.3	67.7	51.3	0.0	0.2	0.0
LLaVA1.5-7B [43]	7B	40.3	39.8	40.0	6.4	9.0	7.5
LLaVA1.5-13B [43]	13B	39.1	49.4	42.5	9.1	19.3	12.3
LLaVA1.6-mistral [45]	7B	41.5	47.3	44.2	7.2	13.6	9.4
LLaVA1.6-vicuna-7B [45]	7B	47.0	48.6	47.8	10.0	15.5	12.2
LLaVA1.6-vicuna-13B [45]	13B	42.0	48.5	45.0	11.9	24.9	16.1
Phi3.5-vision-Ins [1]	4B	36.6	38.4	37.5	5.5	10.3	7.2
Gemma3-12B-Ins [68]	12B	41.0	64.6	50.2	9.1	24.9	13.4
Gemma3-27B-Ins [68]	27B	44.1	60.5	51.0	9.9	20.6	13.4
Qwen2.5-VL-3B-Ins [8]	3B	32.5	75.8	45.5	3.3	7.4	4.6
Qwen2.5-VL-7B-Ins [8]	7B	32.5	73.2	45.1	3.3	15.5	5.5
Qwen2.5-VL-32B-Ins [8]	32B	42.0	75.5	54.0	9.2	24.6	13.4
Qwen2.5-VL-72B-Ins [8]	72B	60.0	82.9	69.6	23.9	44.8	31.2
Qwen3-VL-4B-Ins [70]	4B	38.8	78.8	52.0	12.5	46.5	19.6
Qwen3-VL-8B-Ins [70]	8B	44.0	65.7	52.7	13.4	36.2	19.0
Cosmos-Reason1 [6]	7B	48.2	66.9	56.1	11.9	40.3	18.3
InternVL3.5-4B-Ins [75]	4B	42.0	67.8	51.9	11.3	30.5	16.5
InternVL3.5-8B-Ins [75]	8B	43.1	71.5	53.8	13.4	33.6	19.2
InternVL3.5-14B-Ins [75]	14B	49.3	66.8	56.8	14.0	24.9	17.9
GLM-4.1V-9B-Base [26]	9B	45.6	78.1	57.6	12.1	31.2	17.5
GLM-4.1V-9B-Thinking [26]	9B	44.0	82.7	57.5	12.4	37.1	18.6
Kimi-VL-A3B-Ins [69]	16B	28.7	78.2	42.0	1.1	11.4	2.1
Kimi-VL-A3B-Thinking [69]	16B	32.0	92.2	47.5	4.4	42.3	8.0
DeepSeek-VL2-Small [78]	16B	43.4	84.9	57.4	10.5	59.1	17.9
MiniCPM-V-4.5 [87]	9B	50.0	59.4	54.2	10.9	18.6	13.7
SpaceOm [13]	4B	34.1	71.2	46.1	8.7	35.1	13.9
Pixtral [2]	12B	48.1	70.0	57.0	13.7	31.2	19.0
Qwen-VL-Plus [7]	-	38.3	86.0	53.0	9.5	37.5	15.1
Qwen-VL-Max [7]	-	62.0	83.0	70.9	25.6	45.2	32.7
GPT-4o-mini [28]	-	54.3	76.4	63.5	14.1	26.4	18.4
GPT-4o [28]	-	59.0	78.5	67.4	19.0	37.0	25.1
GPT-4.1-mini [59]	-	53.8	87.1	66.5	21.2	51.4	30.0
GPT-4.1 [59]	-	65.2	84.2	73.5	30.2	51.3	38.0
GPT-5-mini [60]	-	64.5	86.6	73.9	32.8	56.2	41.2
GPT-5 [60]	-	67.8	86.4	76.0	39.2	58.5	47.0
Claude-3-7-sonnet [3]	-	47.1	80.2	59.3	19.0	52.3	27.9
Claude-4-sonnet [4]	-	59.4	82.0	68.9	27.8	52.1	36.3
Gemini-2.5-flash [19]	-	62.0	88.6	72.9	29.5	57.3	38.9
Gemini-2.5-pro [19]	-	65.3	86.2	74.3	31.0	53.6	39.3

Table 2. The evaluation results of 41 VLMs. 'Ins' indicates that the model is an 'instruction-tuned' version. Recall and precision are used as reference metrics and are marked in gray. The best and second-best F1 scores (excluding human results) are marked in red and blue, respectively, and we use a dashed line to separate open-source VLMs (above) and proprietary VLMs (below).

3.5. Analysis and Discussions

In this section, we delve into two key questions: (1) the alignment between our metric and human judgment; (2) the underlying causes of VLMs' poor performance.

Human Alignment of VLM-based Evaluation. We employ GPT-4o when matching predicted steps with annotated

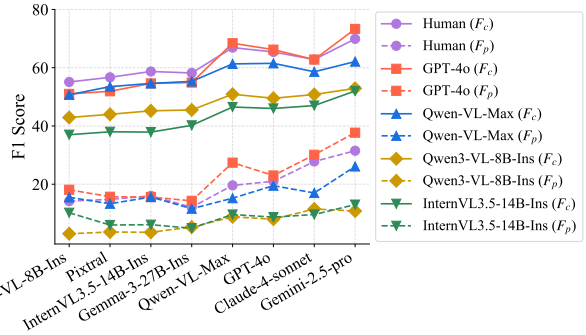

Figure 6. Evaluation results of human and different scoring VLMs. The x-axis represents the eight representative VLMs being evaluated. Each point with the same marker shape denotes the F1 scores obtained by the same scoring VLM or by human evaluators. The solid lines indicate the F1 scores for content, while the dashed lines represent the F1 scores for preconditions.

ones. In the following analysis, we investigate the consistency between the *scoring* VLM (used to generate the matching matrix) and the human evaluators. To this end, we selected eight representative VLMs as the evaluated models and four VLMs as the scoring VLMs. All evaluations were conducted on 300 randomly sampled instances from SpatiaLQA. Fig. 6 presents the evaluation results obtained from human evaluators and different scoring VLMs, while Tab. 3 compares the outcomes between the scoring VLMs and human evaluators. The results show that evaluation scores vary significantly depending on which VLM is used as the scoring model. Notably, proprietary models (Qwen-VL-Max and GPT-4o) yield results that are more consistent with human evaluations, likely because they learn more stable semantic similarity judgment patterns from larger, higher-quality data, bringing their judgments closer to human intuition. Specifically, proprietary models generally exhibit higher correlation coefficients and lower mean absolute errors (around 3 percentage points), whereas open-source models show mean absolute errors exceeding 10 percentage points. Furthermore, GPT-4o achieves the highest correlation and lowest mean absolute error. Therefore, to maintain consistency with human judgment, we adopt GPT-4o as the scoring VLM.

The Underlying Causes of Poor Performance. To answer this question, we selected four representative VLMs and analyzed their performance more comprehensively from three dimensions: the number of annotated answer steps, annotation source, and scene category. As shown in Fig. 7, model performance generally decreases as the number of annotated answer steps increases. Moreover, model performance shows clear patterns across different annotation sources: VLMs perform best on data generated by subgraph extraction augmentation, followed by manual annotations, and worst on graph expansion augmentation. This trend arises because the samples generated through subgraph extraction

Scoring VLMs	ρ_c	ρ_p	s	Δ
Qwen3-VL-8B-Ins	0.96	0.89	0.98	13.4
InternVL3.5-14B-Ins	0.96	0.78	0.99	14.9
Qwen-VL-Max	0.97	0.86	0.99	3.5
GPT-4o	0.99	0.96	0.99	3.0

Table 3. Comparison between scoring VLM evaluation results and human evaluation results. ρ_c (ρ_p) represents the Pearson correlation coefficient between the content (precondition) F1 scores obtained by VLMs and those obtained by humans. s (Δ) denotes the cosine similarity (mean absolute error) between the F1 scores obtained by VLMs and those obtained by humans.

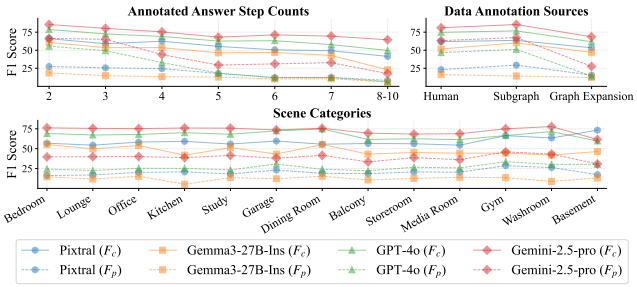

Figure 7. Dimension-wise analysis. Each dot represents the F1 score of a VLM under specific conditions, including different numbers of answer steps, annotation methods, and scene categories. 'Human', 'Subgraph', and 'Graph Expansion' correspond to data from 'manual annotations', 'subgraph extraction augmentation', and 'graph expansion augmentation', respectively.

augmentation have fewer answer steps (simpler problems), whereas the samples generated through graph expansion augmentation contain more steps (more complex problems). However, VLMs exhibit relatively consistent performance across various scene categories.

These observations suggest that VLMs tend to perform worse on tasks requiring more steps, as such tasks demand longer and more stable reasoning processes. VLMs' failures on these tasks result in overall poor performance.

4. Recursive Scene Graph Assisted Reasoning

To address the poor performance of VLMs on complex tasks, we introduce Recursive Scene Graph Assisted Reasoning (RSGAR) in Sec. 4.1 and present its effectiveness and ablation studies in Sec. 4.2. Please refer to the supplementary material for details on hyperparameters and costs.

4.1. Method

As shown in Fig. 8, RSGAR consists of the following three steps: (1) We first employ Depth Anything V2 and SAM to obtain the depth and segmentation maps of the scene image. (2) Using these perception results together with the original image, we designate the task-specified target as the initial source object and perform the first round of scene graph generation with the VLM. During this process, the model identifies the objects in direct contact with the source object, referred to as target objects, and determines their spa-

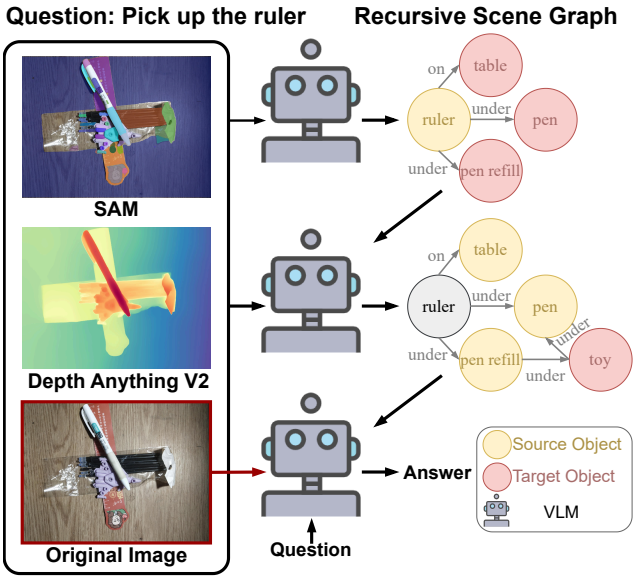

Figure 8. The overview of RSGAR. Scene graph generation and question answering are performed by the same VLM.

tial relationships. A scene graph is then constructed with the source and target objects as nodes and their spatial relationships as edges. This graph serves as input for the next iteration, where the previous target objects are treated as new source objects. The process continues until a predefined maximum iteration T is reached. (3) Finally, the generated scene graph is combined with the task prompt and fed into the VLM to produce the final answer.

4.2. Experiments

Effectiveness. We adopt five baselines: PhysAgent [18], Chain of Thought (CoT) [76], and three variants of vanilla reasoning with additional inputs of segmentation map (SAM), depth map (Depth Anything V2), or both together. PhysAgent is one of the most advanced methods for enhancing VLMs' physical commonsense and understanding of the real world. We use GPT-4o as the VLM for all methods, and the number of iterations T in RSGAR is set to 5. The results of all methods on SpatiaLQA are shown in Tab. 4a, which shows that RSGAR achieves the best performance among all methods, while CoT attains the second-best result due to its ability to enable more stable reasoning. Other baselines even underperform the vanilla reasoning, indicating that simply incorporating physical priors or visual cues does not contribute to the spatial logical reasoning of VLMs.

The Reason for Effectiveness. To answer this question, we further analyzed RSGAR from the perspective of the number of steps in the annotated answers. As shown in Tab. 4b, although RSGAR shows a slight decrease in performance on samples with fewer answer steps, it significantly improves the performance of VLMs on samples with more answer steps. This suggests that RSGAR can improve VLM performance on complex tasks by explicitly representing the relationships among key objects in the original scene.

	F_c	F_p
GPT-4o	67.4	25.1
+ depth	64.1	19.0
+ seg	50.3	14.7
+ seg&depth	50.5	14.5
PhysAgent	64.7	22.2
CoT	<u>67.6</u>	<u>27.0</u>
RSGAR	**69.8**	**28.1**

(a) RSGAR vs. baselines.

Step Counts	F_c	F_p
2	76.4 (-1.3)	53.9 (-0.7)
3	71.6 (-1.0)	46.4 (-1.9)
4	73.6 (+4.5)	34.8 (+1.9)
5	65.6 (+2.8)	20.0 (+1.6)
6	62.8 (-0.4)	14.3 (+2.5)
7	58.6 (+0.8)	15.2 (+3.0)
8-10	50.2(+0.5)	8.9(+2.4)

(b) F_c/F_p across answer step counts.

Table 4. (a) The results of various methods. **Bold** and <u>underlined</u> scores denote the best and second-best results. '+ depth', '+ seg', and '+ depth&seg' represent vanilla reasoning enhanced with depth maps, segmentation maps, and both, respectively. (b) RSGAR's F_c and F_p across answer step counts. The values in parentheses represent the changes relative to vanilla reasoning.

T	F_c	F_p
1	68.5	27.4
3	68.8	28.1
5	*69.8*	*28.1*
7	70.6	28.7

(a) Iteration number T.

	F_c	F_p
w/o seg&depth	66.5	26.3
w/o seg	66.9	26.5
w/o depth	68.8	27.8
w/ seg&depth	*69.8*	*28.1*

(b) Segmentation/Depth map.

Table 5. Ablation studies on GPT-4o. The default settings are indicated in *italics*. 'seg' and 'depth' denote the segmentation map and depth map, respectively.

Ablation Studies. We set $T = 5$ by default and provide both the depth map and segmentation map during scene graph generation, as indicated by the *italics* in Tab. 5. Tab. 5a shows that the performance of the VLM improves as T increases. This is because a larger T allows the final scene graph to include more information, giving the VLM a more comprehensive understanding of the scene when answering questions. Tab. 5b shows that the performance of RSGAR declines when either the depth map or segmentation map is not used. This is because each of them provides distinct visual information, and both are essential for generating an accurate scene graph.

5. Conclusion

In summary, we introduce SpatiaLQA, a benchmark comprising 9,605 samples from 241 real-world indoor scenes, designed to systematically evaluate the spatial logical reasoning abilities of VLMs. Moreover, we develop an automated evaluation strategy based on GPT-4o, which achieves a high level of consistency with human. By evaluating 41 VLMs, we reveal significant deficiencies in their spatial logical reasoning. Therefore, we propose recursive scene graph assisted reasoning, which effectively enhances GPT-4o's performance on complex tasks. We will further explore how VLMs can be applied to more complex scenarios, such as long-horizon planning and dynamic environments, where models must reason and adapt to changing observations.

6. Acknowledgment

This work received substantial support from the Pioneer R&D Program of Zhejiang (2026SDXT005) and the National Natural Science Foundation of China (62576305).

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv*, 2024. 6

[2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 6

[3] Claude Anthropic. Claude 3.7 sonnet and claude code, 2025. 6

[4] Claude Anthropic. Introducing claude 4, 2025. 6

[5] 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. 1, 3

[6] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 6

[7] 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*, 1(2):3, 2023. 6

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[9] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE, 2025. 1

[10] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21819–21830, 2024. 1

[11] Chengan Che, Chao Wang, Xinyue Chen, Sophia Tsoka, and Luis C. Garcia-Peraza-Herrera. A stitch in time: Learning procedural workflow via self-supervised plackett-luce ranking. In *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. To appear. 1

[12] Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C. Garcia-Peraza-Herrera. LEMON: A Large Endoscopic MONocular Dataset and Foundation Model for Perception in Surgical Settings. In *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. To appear. 1

[13] 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. 1, 6

[14] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021. 3

[15] Kaixuan Chen, Jie Song, Shunyu Liu, Na Yu, Zunlei Feng, Gengshi Han, and Mingli Song. Distribution knowledge embedding for graph pooling. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):7898–7908, 2023. 1

[16] Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *arXiv preprint arXiv:2402.12501*, 2024. 3

[17] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 1, 3

[18] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025. 1, 8, 4

[19] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[20] Yuhong Deng, Di Guo, Xiaofeng Guo, Naifu Zhang, Huaping Liu, and Fuchun Sun. Mqa: Answering the question via robotic manipulation. *arXiv preprint arXiv:2003.04641*, 2020. 3

[21] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024. 3

[22] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 1

[23] Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fak-

eness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025. 1

[24] Le Han, Kaixuan Chen, Minchen Ye, and Nenggan Zheng. Hi-motion: Hierarchical intention guided conditional motion synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9628–9637, 2025. 1

[25] Le Han, Kaixuan Chen, Lei Zhao, Yangbo Jiang, Pengfei Wang, and Nenggan Zheng. Cross-domain animal pose estimation with skeleton anomaly-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1

[26] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 6

[27] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 2, 3

[28] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 6

[29] Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*, 2024. 3

[30] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[32] Dohwan Ko, Sihyeon Kim, Yumin Suh, Minseo Yoon, Manmohan Chandraker, Hyunwoo J Kim, et al. St-vlm: Kinematic instruction tuning for spatio-temporal reasoning in vision-language models. *arXiv preprint arXiv:2503.19355*, 2025. 3

[33] Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring efficient open-vocabulary segmentation in the remote sensing. *arXiv preprint arXiv:2509.12040*, 2025. 1

[34] Bingyu Li, Tao Huo, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring the underwater world segmentation without extra training. *arXiv preprint arXiv:2511.07923*, 2025.

[35] Bingyu Li, Feiyu Wang, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Maris: Marine open-vocabulary instance segmentation with geometric enhancement and semantic alignment. *arXiv preprint arXiv:2510.15398*, 2025.

[36] Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1308–1317, 2025. 1

[37] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026. 1

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 6

[39] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37: 100428–100534, 2024. 1, 3

[40] Peizheng Li, Zhenghao Zhang, David Holtz, Hang Yu, Yutong Yang, Yuzhi Lai, Rui Song, Andreas Geiger, and Andreas Zell. Spacedrive: Infusing spatial awareness into vlm-based autonomous driving. *arXiv preprint arXiv:2512.10719*, 2, 2025. 1

[41] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024. 3

[42] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in neural information processing systems*, 35: 10560–10571, 2022. 1

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 6

[44] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 1

[45] 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. 6

[46] Sichao Liu, Jianjing Zhang, Robert X Gao, Xi Vincent Wang, and Lihui Wang. Vision-language model-driven scene understanding and robotic object manipulation. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 21–26. IEEE, 2024. 1

[47] Xueyu Liu, Rui Wang, Yexin Lai, Guangze Shi, Feixue Shao, Fang Hao, Jianan Zhang, Jia Shen, Yongfei Wu, and Wen Zheng. Plug-and-play ppo: An adaptive point prompt optimizer making sam greater. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4332–4342, 2025. 2

[48] Xueyu Liu, Xiaoyi Zhang, Guangze Shi, Meilin Liu, Yexin Lai, Yongfei Wu, and Mingqiang Wei. Attack for defense: Adversarial agents for point prompt optimization empowering segment anything model. *arXiv preprint arXiv:2509.18891*, 2025. 2

[49] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 3

[50] Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. Adsqa: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23396–23407, 2025. 1

[51] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3

[52] Duc-Tuan Luu, Viet-Tuan Le, and Duc Minh Vo. Questioning, answering, and captioning for zero-shot detailed image caption. In *Proceedings of the Asian Conference on Computer Vision*, pages 242–259, 2024. 1, 3

[53] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498, 2024. 3

[54] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014. 3

[55] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1, 3

[56] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 3

[57] Aoran Mei, Guo-Niu Zhu, Huaxiang Zhang, and Zhongxue Gan. Replanvlm: Replanning robotic tasks with visual language models. *IEEE Robotics and Automation Letters*, 2024. 1, 3

[58] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957. 2

[59] OpenAI. Introducing gpt-4.1 in the api. Technical report, OpenAI, 2025. Accessed: 2025-05-07. 6

[60] OpenAI. GPT-5 System Card. Technical report, OpenAI, 2025. Accessed: 2025-08-10. 6

[61] Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, et al. Balrog: Benchmarking agentic llm and vlm reasoning on games. *arXiv preprint arXiv:2411.13543*, 2024. 1

[62] Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. Rora-vlm: Robust retrieval-augmented vision language models. *arXiv preprint arXiv:2410.08876*, 2024. 3

[63] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE, 2024. 2

[64] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025. 1

[65] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. *arXiv preprint arXiv:2411.06048*, 2024. 3

[66] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1

[67] Zhiying Song, Kaixuan Chen, Pengfei Wang, Mingli Song, and Nenggan Zheng. Unsupervised action segmentation via multi-scale temporal-interaction enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1

[68] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 6

[69] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 6

[70] Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action. *Qwen Blog. Accessed*, pages 10–04, 2025. 6

[71] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & qa. *arXiv preprint arXiv:1811.00613*, 2018. 3

[72] Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Code2logic: Game-code-driven data synthesis for enhancing vlms general reasoning. *arXiv preprint arXiv:2505.13886*, 2025. 1

[73] Haoming Wang, Qiyao Xue, and Wei Gao. Infinibench: Infinite benchmarking for visual spatial reasoning with customizable scene complexity. *arXiv preprint arXiv:2511.18200*, 2025. 1

[74] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024. 3

[75] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6

[76] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 8

[77] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019. 3

[78] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6

[79] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1, 3

[80] Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. *arXiv preprint arXiv:2502.11903*, 2025. 3

[81] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2

[82] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 3

[83] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36:40924–40943, 2023. 1, 3

[84] Zhutian Yang, Caelan Garrett, Dieter Fox, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Guiding long-horizon task and motion planning with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16847–16853. IEEE, 2025. 1

[85] Fei Yu, Quan Deng, Shengeng Tang, Yuehua Li, and Lechao Cheng. Open-world 3d scene graph generation for retrieval-augmented reasoning, 2025. 1

[86] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019. 3

[87] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. 6

[88] Xinlei Yu, Chengming Xu, Zhangquan Chen, Yudong Zhang, Shilin Lu, Cheng Yang, Jiangning Zhang, Shuicheng Yan, and Xiaobin Hu. Visual document understanding and reasoning: A multi-agent collaboration framework with agent-wise adaptive test-time scaling. *arXiv preprint arXiv:2508.03404*, 2025. 1

[89] Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. *arXiv preprint arXiv:2511.11007*, 2025. 1

[90] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1, 2, 3

[91] Weichen Zhang, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*, 2025. 3

[92] Xiaohan Zhang, Yan Ding, Saeid Amiri, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang. Grounding classical task planners via vision-language models. *arXiv preprint arXiv:2304.08587*, 2023. 1

[93] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025. 2

[94] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024. 2

[95] Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yinjie Lei, Mingkui Tan, and Chuang Gan. Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3761–3771, 2025. 1

[96] Yufeng Zhong, Chengjian Feng, Feng Yan, Fanfan Liu, Liming Zheng, and Lin Ma. Robotrom-nav: A unified framework for embodied navigation integrating perception, planning, and prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6416–6425, 2025. 3

[97] Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. Medeyes: Learning dynamic visual focus for medical progressive diagnosis. *arXiv preprint arXiv:2511.22018*, 2025. 1

[98] Chunzheng Zhu, Yangfang Lin, Jialin Shao, Jianxin Lin, and Yijun Wang. Pathology-aware prototype evolution via llm-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9196–9205, 2025. 1