LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning?

Kexian Tang[1,3*], Junyao Gao[2,3*], Yanhong Zeng[3†], Haodong Duan[3†], Yanan Sun[3], Zhening Xing[3], Wenran Liu[3], Kai Chen[3], and Kaifeng Lyu[1‡]

[1] Tsinghua University, Beijing, China
[2] Tongji University, Shanghai, China
[3] Shanghai AI Laboratory, Shanghai, China
tangkx25@mails.tsinghua.edu.cn, klyu@mail.tsinghua.edu.cn

Abstract. Many real-world applications of spatial intelligence, such as robotic control, autonomous driving, and automated assembly, require spatial reasoning across multiple sequential steps. However, the extent to which current Multimodal Large Language Models (MLLMs) possess this capability remains largely unexplored. Inspired by LEGO construction, a recreational activity that critically relies on multi-step spatial reasoning, we introduce **LEGO-Puzzles**: a benchmark designed to systematically evaluate the spatial reasoning capabilities of MLLMs from basic spatial understanding to multi-step planning. LEGO-Puzzles contains two task sets. The **Elementary** set covers 11 visual question-answering (VQA) tasks with 1,100 carefully curated samples to test elementary spatial reasoning skills that are cruical for LEGO assembly. The **Planning** set directly requires the model to generate a step-by-step plan for assembling a target LEGO structure, where the tasks are organized into subsets with different planning horizons ranging up to 8. Our evaluation of 29 state-of-the-art MLLMs shows that even the strongest models struggle with elementary reasoning tasks in LEGO construction, falling at least 20% behind human performance. The planning accuracy also quickly drops to 0% as the number of planning steps increases, whereas our human participants solve all the tasks perfectly. Switching the output format from multiple choice to image generation degrades model performance even further, leading to zero accuracy even for planning 3 steps. Overall, LEGO-Puzzles reveals critical limitations in current MLLMs' spatial reasoning capabilities and highlights the need for substantial advances.

Keywords: MLLMs · Multi-Step Spatial Reasoning · Benchmark

1 Introduction

Multimodal Large Language Models (MLLMs) have made remarkable progress and have shown promising capabilities in spatial intelligence [13]. The most elementary capability of spatial intelligence is the ability to perceive and understand

[*] Equal contribution. [†] Project lead. [‡] Corresponding author.

spatial relationships among 3D objects from a static scene, which we refer to as **spatial understanding** in this work. This includes many Visual Question Answering (VQA) tasks, such as identifying object attributes (e.g., shape, color, size), recognizing spatial relations (e.g., left/right, above/below), and counting objects that satisfy certain conditions [17, 22, 23, 40].

Many other tasks not only require the spatial understanding of the given scene, but also critically rely on the ability to imagine the consequences of hypothetical actions applied to the scene. That is, one has to sequentially apply the spatial understanding to two or more real or imaginary scenes to solve these tasks, a capability which we refer to as **sequential spatial reasoning** in this work. Notably, a series of benchmarks have been introduced to evaluate the ability of imagining the scene after changing the camera perspective [21, 27, 32, 34].

However, only a few benchmarks [26, 35, 43] evaluate the sequential spatial reasoning capability of MLLMs beyond one step of action, even though many real-world applications require planning for multiple steps of actions. For example, to assemble a target structure, a robotic arm may perceive the current and target states and plan a long sequence of actions to execute. In home maintenance scenarios, users may send a photo of a broken item to a chatbot for help, and the chatbot needs to plan a sequence of actions for the user to follow step by step. Providing a systematic evaluation of the spatial intelligence of MLLMs from single-step to multiple-step reasoning is hence the focus of this work.

LEGO-Puzzles. In this work, we take inspiration from a common recreational activity, *LEGO construction*, to assess the spatial intelligence of MLLMs. Constructing LEGO models naturally requires reasoning over sequential assembly steps, making it a convenient testbed for evaluating spatial reasoning abilities. Moreover, LEGO construction has been widely used in cognitive science as a reliable indicator of spatial intelligence. Studies have shown that LEGO construction is strongly associated with core spatial skills such as *mental rotation* and *visuo-spatial working memory*, as well as with mathematical performance [16, 28, 29].

We introduce **LEGO-Puzzles**, a benchmark designed to systematically evaluate the spatial reasoning capabilities of MLLMs with progressively increasing levels of spatial and sequential complexity, ranging from basic spatial understanding to multi-step planning. Based on open-source LEGO projects with step-by-step assembly instructions, we curate a diverse set of tasks organized into two sets, the **Elementary** set and the **Planning** set (see Fig. 1).

- The **Elementary** set contains a diverse collection of over 1,100 carefully curated visual question-answering (VQA) pairs spanning 11 distinct tasks, grouped into three levels. The first level assesses MLLMs' **basic spatial understanding** capabilities, including recognition of height relationships, rotation angles, etc. The second and third levels test **single-step** and **multi-step sequential reasoning** abilities of MLLMs based on LEGO assembly sequences to examine models' sequential reasoning ability.
- The **Planning** set is especially designed to assess whether a model's capability of multi-step spatial reasoning is strong enough to generate an explicit

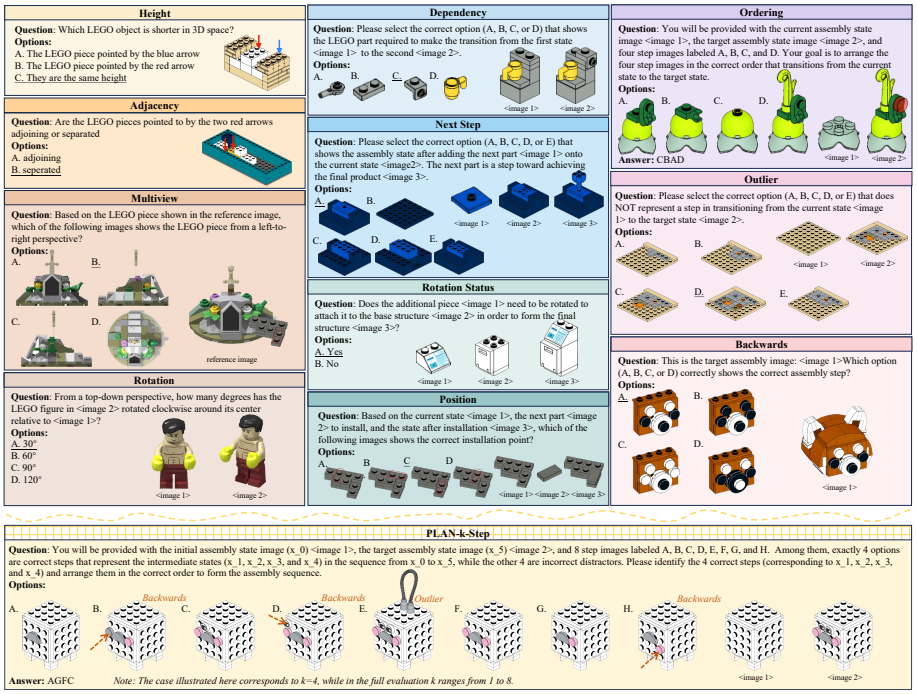

Fig. 1: Task examples of LEGO-Puzzles. The examples above the wavy line correspond to the *Elementary Set*, with columns (from left to right) representing Spatial Understanding, Single-Step Sequential Reasoning, and Multi-Step Sequential Reasoning. The examples below the wavy line correspond to the *Planning Set*. *Note: The questions are slightly simplified for clarity and brevity.*

multi-step plan for assembling a target LEGO structure. The task is organized into subsets with different numbers of planning steps k, which ranges up to 8. While the primary evaluation is conducted in a multiple-choice format, the task also evaluates MLLMs' ability to generate images of the planned intermediate states in multiple turns, mimicking real-world scenarios where a chatbot guides users through a sequence of actions using visual instructions.

Evaluation Results. We conduct comprehensive evaluations of 29 state-of-the-art MLLMs, including proprietary models such as GPT-5 and Gemini-2.5-Pro, as well as leading open-source alternatives [9, 38, 42]. The evaluation results reveal significant limitations in the spatial reasoning capabilities of MLLMs. Even the strongest models lag behind humans by at least 20% on the tasks in the elementary set. Among open-source models, only a few achieve performance notably above random guessing. This human-model gap becomes more pronounced in the planning tasks, where the multiple-choice accuracy quickly drops below 50% once the number of steps exceeds 3, and eventually approaches 0% as the number of steps further increases. In contrast, human participants achieve perfect

accuracy on all these planning tasks. When switching the output format from multiple choice to image generation, MLLMs can usually generate images with the correct appearance, but fail to place LEGO pieces in the correct positions. As a result, the accuracy drops to zero even for planning 3 steps.

Our Contributions. Our main contributions are as follows:

- **Comprehensive evaluation framework.** Our benchmark includes a diverse set of tasks with progressively increasing spatial and sequential complexity, ranging from basic spatial understanding to multi-step planning.
- **Natural but challenging planning tasks.** Our planning tasks are naturally constructed from open-source human-designed LEGO projects but can pose significant challenges for frontier MLLMs to solve.
- **Evaluation of multi-turn image generation.** Beyond VQA evaluation, our planning tasks also evaluate the ability of MLLMs to generate planned intermediate states as images in multiple turns. All the frontier MLLMs fail completely even for 3 planning steps.

See also appendix for detailed comparison with existing benchmarks.

2 LEGO-Puzzles

In this section, we introduce LEGO-Puzzles, a benchmark designed to evaluate the multi-step spatial reasoning capability of MLLMs through a progressive and comprehensive evaluation framework. The benchmark contains two sets: (1) the **Elementary set**, which includes a diverse set of LEGO-related VQA tasks spanning spatial understanding, single-step and multi-step sequential reasoning (Sec. 2.1); (2) the **Planning set**, which is specifically designed to evaluate whether a model's reasoning capability is strong enough to generate an explicit multi-step plan for assembling a target LEGO structure (Sec. 2.2). Our data curation process is described in Sec. 2.3.

2.1 Elementary Set

The Elementary Set is designed to test fundamental saptial reasoning skills that are crucial for LEGO assembly. To enable a more comprehensive and progressively structured evaluation, we define three levels of tasks. This framework is grounded in insights from cognitive psychology and human developmental stages in acquiring spatial intelligence [13, 30]. Using LEGO building as a concrete and intuitive example, we observe that humans typically develop spatial reasoning abilities in stages: from basic *spatial understanding (Level 1)*, to reasoning through *individual assembly steps (Level 2)*, and ultimately to reasoning across *multiple sequential steps (Level 3)*. Based on this developmental trajectory, our benchmark is divided into three levels, as illustrated in Fig. 1.

Level 1: Spatial Understanding. This level focuses on the ability to *understand the spatial relationships* between each LEGO piece and how the pieces appear and relate when viewed from different perspectives in 3D space: (1) *Height:*

Distinguish the relative heights of LEGO objects. (2) *Adjacency:* Determine whether LEGO objects are adjacent or separated. (3) *Rotation:* Calculate the angle of rotation between a LEGO object and its corresponding rotated version. (4) *Multiview:* Predict the current LEGO status from different viewpoints.

Level 2: Single-Step Sequential Reasoning. Building upon spatial understanding, this level is designed based on single-step actions, mirroring how humans typically progress from spatial perception to carrying out one assembly step at a time during the building process: (5) *Rotation Status:* Determine whether a LEGO piece requires rotation before installation. In contrast to *Rotation* Task in Level 1, this task focuses on reasoning from an assembly perspective, with finer granularity. (6) *Position:* Identify the correct assembly position to place the next LEGO piece. (7) *Next-Step:* Predict the next LEGO status based on the current status and the new pieces. (8) *Dependency:* Identify which pieces are necessary to transition from the current to the next assembly stage.

Level 3: Multi-Step Sequential Reasoning. The final level is built upon the single-step reasoning skills from *Level 2* and require planning over multiple sequences (involving up to 7 intermediate stages). It assesses the ability to reason across multiple steps in the assembly process: (9) *Backwards:* Given the target assembly stage, determine the correct intermediate step in the LEGO build process. (10) *Ordering:* Determine the correct assembly order of the provided final LEGO images. (11) *Outlier:* Detect the LEGO status that does not belong to the provided assembly sequence.

Task Summary. The Elementary set consists of over 1,100 visual question-answering (VQA) pairs derived from 407 LEGO building instructions, encompassing 11 tasks across spatial understanding, single-step and multi-step sequential reasoning. Each task contains 100 samples to ensure balanced evaluation across categories.

2.2 Planning Set

To further investigate the long-horizon multi-step sequential reasoning abilities of MLLMs, we construct the Planning Set, which goes beyond elementary spatial reasoning by requiring step-by-step plan. Within this set, we design two tasks: **Plan-*k*-Step-VQA** and **Plan-*k*-Step-Generation**, which evaluate the model's ability to generate a multi-step plan in VQA and image generation settings, respectively.

Task: Plan-*k*-Step-VQA. *PLAN-k-Step-VQA* integrates the multi-step tasks *Backwrds*, *Ordering* and *Outlier* from the Elementary set to model the realistic reasoning planning and execution across multiple sequences under noisy conditions. *PLAN-k-Step-VQA* first extends the original *Ordering* setting, which was restricted to a fixed *PLAN-4-Step*, into a scalable sequence length k setting ranging from 1 to 8. Next, we add an additional k erroneous options from the

tasks *Backwrds* and *Outlier* on each k-length sequence, resulting in a total of $2k$ options. These noisy conditions increase the task's complexity and make the examination of MLLMs' multi-step reasoning ability more realistic and challenging. Specifically, we construct 20 test cases for every sequence length k. Each test case includes the current LEGO object (x_0), the target LEGO object (x_{k+1}), the correct intermediate LEGO objects ($x_1^c, x_2^c, \ldots, x_k^c$), and the erroneous LEGO objects ($x_1^e, x_2^e, \ldots, x_k^e$), along with the corresponding text instructions. The model is required to select the k correct intermediate steps and order them according to the assembly sequence.

Task: Plan-*k*-Step-Generation. To further investigate whether MLLMs can transfer their sptial reasoning abilities to image generation and perform realistic full-stage planning, we design a more challenging image generation task: *PLAN-k-Step-Generation*. It mirrors how humans naturally perform LEGO assembly: given an initial assembly state and a final target state, an intelligent system should generate the intermediate states step by step to demonstrate how the build is completed. Motivated by this idea, we design an interactive, step-by-step generative framework, where the model produces one intermediate state at each turn. Given the initial state x_0 and the final target state x_{k+1}:

- Turn 1: The model receives x_0 and x_{k+1} and is asked to generate the next assembly state x_1.
- Turn 2: The model is then asked to generate x_2, conditioned on its previously generated output x_1.
- Turn (t): At each subsequent turn, the model generates x_{t+1} based on the entire history of generated states.

This process continues until the model produce x_k, yielding a complete generative trajectory. We set $k = 2, 3, 4, 5$. We do not evaluate longer horizons because even at $k = 3$ the strongest models already fail completely (Sec. 4.2).

2.3 Dataset Curation

As illustrated in Fig. 2, our pipeline consists of three key steps: data collection, question-answer generation, and quality control. This design ensures the scalability, accuracy, and reliability of our data.

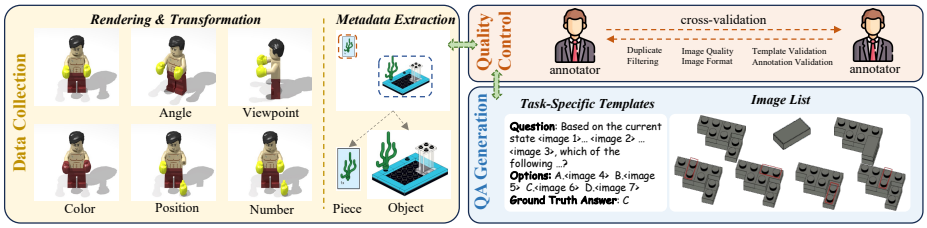

Fig. 2: Data curation pipeline. Our pipeline comprises: data collection, question-answer generation, and quality control.

Data Collection. Data collection consists of three stages. *(1). LEGO Project Collection.* We collect a diverse set of open-source LEGO source files from the Internet, each containing detailed step-by-step building instructions and part lists. To ensure suitable task complexity, we filter for projects with moderate final size. Extremely large builds exhibit high structural complexity, and the small visual changes introduced by added pieces make it difficult for models to detect step-wise differences. Conversely, overly small builds are filtered out for low spatial complexity and insufficient steps for multi-step reasoning. We also ensure category diversity, covering animals, furniture, vehicles, and more, to increase task variety. *(2). Rendering and Transformation.* We render each project into PDF format using the public software Studio[4], keeping the camera viewpoint fixed across steps to maintain spatial and temporal consistency. The tool allows flexible editing of the source files, enabling us to modify part attributes such as type, quantity, color, and position as needed for task construction. For example, in the *Rotation* and *Multiview* task, we apply POV-Ray style rendering and adjust lighting to simulate different viewing angles. In the *Backward* task, we introduce deliberate errors in part attributes to generate incorrect assembly states. *(3).Metadata Extraction and Unification.* We employ PDF-Extract[5] to extract structured information from the rendered PDFs, including individual LEGO pieces and assembled objects. All visual assets are processed under a unified naming convention and organized for downstream question-answer generation across all defined task types.

Question-Answer Generation. To support scalable and structured data construction, we manually design task-specific templates tailored to each task. Each example includes an instruction defining the model's role, a question referencing input images using tokens like `<image x>`, and a ground-truth answer. For example, in the *Position* task, the model is provided with the current assembly state, the part to install, the resulting state, and several candidate placements. This standardized template design allows flexibility in the number of input images required by each task, making it suitable for both single-step and multi-step reasoning scenarios. More details are provided in the appendix.

Quality Control. To ensure data quality and reliability, we implement a multi-stage human-in-the-loop review process. *1)Duplicate filtering.* We perform duplication checks by computing similarity scores across rendered images, identifying visually redundant or recolored samples. These are further reviewed manually, and duplicates are removed to reduce redundancy and improve evaluation quality. *2)Image Quality and Format Check.* We manually verify all rendered images to ensure consistency in camera perspective, correctness of part attributes (e.g., color, shape, quantity), and adherence to naming conventions. Any files with errors are either corrected or discarded. *3)Template and annotation validation.* Each question-answer pair is verified by three trained annotators. Reviewers confirm that the images referenced by tokens (`<image 1>`, `<image 2>`, ...) appear

[4] `https://www.bricklink.com/v3/studio/download.page`
[5] `https://github.com/opendatalab/PDF-Extract-Kit`

in the correct order. Samples with unresolved disagreements are either revised or removed.

Given the vast number of high-quality open-source LEGO projects available, our pipeline is inherently scalable and can be extended in an semi-automated manner to support larger and more diverse benchmarks in the future.

3 Evaluation on the Elementary Set

3.1 Experimental Setting

Benchmark Models. For the Elementary Set, we extensively evaluate 29 models, covering a diverse range of architectures, sizes, and training processes. For open-source models, we evaluate Qwen3-VL-[8B/32B]-Instruct [4], Qwen2.5-VL-[7B/72B] [5], Qwen2-VL-[7B/72B] [38], InternVL3.5-[8B/38B] [39], InternVL3-[8/78B] [45], InternVL2.5-[8B/78B] [8], MiniCPM-V2.6 [44], VILA1.5-13B [24], Idefics3-8B [19], DeepSeek-VL2-[Tiny/Small] [42], Pixtral-12B [2], LLaVA-OneVision-7B [20], and EMU3 [41]. For proprietary models, we evaluate GPT-5 [33], GPT-o3 [31], GPT-4o , GPT-4o-mini [1], Claude-3.5-Sonnet [3], Gemini-2.5-Pro [11], Gemini-2.0-Flash, Gemini-1.5-Flash [37], and Gemini-1.5-Pro. Moreover, all evaluations are conducted in a zero-shot setting for a fair comparison.
Baselines. We provide two baselines for comparison:

- *Random*: the accuracy of random selection, assuming equal probability for all options.
- ↑ *Random*: the *p-value–based critical value*, which indicates the minimum accuracy required to statistically surpass random guessing at a given significance level ($p = 0.05$).

Evaluation Metrics. For all questions in the Elementary set, we adopt *exact match accuracy* (%) as the evaluation metric. For multiple-choice questions, we follow VLMEvalKit [12], applying rule-based matching as a first step, and resort to LLM-based choice extraction and matching using ChatGPT [1] when heuristic matching fails. For the *Ordering* task, we adopt the same strategy: rule-based extraction of the predicted step sequence from the model's response, followed by LLM-based extraction and matching when necessary.

3.2 Results

We report results on the Elementary Set in Tab. 1 and Tab. 2. We summarize key findings as below.

Clear Gap Between Proprietary and Open-Source Models. As shown in Tab. 1, there is a clear performance gap between proprietary and open-source MLLMs. GPT-5 and Gemini-2.5-Pro exhibits strong spatial reasoning capabilities, achieving overall accuracies of 68.8% and 72.0%, respectively, while most open-source MLLMs perform only marginally better than ↑*Random*.

Table 1: Full evaluation results of 29 MLLMs on the Elementary set. Dark Gray and Light Gray indicates the best performance for each task among all models and open-source models respectively. We highlight the top 3 models by overall performace using Dark Green, Medium Green, and Light Green, respectively.

Models	Spatial Understanding				Single-Step Reasoning				Multi-Step Reasoning			Overall
	Height	Adjacency	Rotation	Multiview	Next-Step	Dependency	Rotation Stat.	Position	Backwards	Ordering	Outlier	
Proprietary												
GPT-5	62.0	69.0	72.0	67.0	83.0	92.0	55.0	59.0	73.0	85.0	75.0	72.0
GPT-o3	66.0	71.0	64.0	66.0	83.0	92.0	53.0	54.0	66.0	80.0	65.0	69.1
GPT-4o	49.0	66.0	41.0	51.0	65.0	87.0	51.0	51.0	53.0	72.0	49.0	57.7
GPT-4o-mini	31.0	53.0	26.0	51.0	27.0	71.0	57.0	32.0	50.0	7.0	27.0	39.3
Claude-3.5-Sonnet	39.0	60.0	42.0	48.0	61.0	78.0	58.0	37.0	49.0	54.0	64.0	53.6
Gemini-2.5-Pro	57.0	64.0	65.0	65.0	79.0	88.0	64.0	67.0	59.0	82.0	67.0	68.8
Gemini-2.0-Flash	35.0	70.0	49.0	45.0	69.0	81.0	54.0	46.0	56.0	46.0	43.0	54.0
Gemini-1.5-Flash	29.0	58.0	28.0	45.0	57.0	77.0	57.0	32.0	28.0	20.0	51.0	43.8
Gemini-1.5-Pro	35.0	58.0	38.0	56.0	59.0	84.0	61.0	39.0	35.0	44.0	59.0	51.6
Open-source												
Qwen3-VL-32B-Instruct	38.0	59.0	41.0	58.0	62.0	91.0	65.0	58.0	73.0	34.0	48.0	57.0
Qwen3-VL-8B-Instruct	35.0	67.0	30.0	42.0	62.0	81.0	58.0	40.0	41.0	22.0	30.0	46.2
Qwen2.5-VL-72B	30.0	61.0	27.0	27.0	55.0	72.0	58.0	47.0	60.0	33.0	43.0	46.6
Qwen2.5-VL-7B	35.0	60.0	22.0	27.0	26.0	60.0	49.0	25.0	24.0	5.0	13.0	31.5
Qwen2-VL-72B	40.0	62.0	37.0	51.0	57.0	79.0	49.0	43.0	34.0	26.0	31.0	46.3
Qwen2-VL-7B	31.0	57.0	30.0	40.0	44.0	70.0	48.0	26.0	13.0	9.0	28.0	36.0
InternVL3.5-VL-38B	42.0	51.0	42.0	51.0	61.0	74.0	56.0	32.0	53.0	5.0	20.0	44.3
InternVL3.5-VL-8B	43.0	54.0	26.0	36.0	41.0	52.0	57.0	31.0	32.0	6.0	33.0	37.4
InternVL3-VL-78B	49.0	65.0	37.0	54.0	60.0	80.0	61.0	34.0	25.0	20.0	40.0	47.7
InternVL3-VL-8B	39.0	58.0	24.0	42.0	30.0	57.0	53.0	25.0	30.0	3.0	33.0	35.8
InternVL2.5-78B	41.0	62.0	32.0	47.0	60.0	79.0	58.0	32.0	40.0	15.0	37.0	45.7
InternVL2.5-8B	35.0	53.0	23.0	37.0	38.0	48.0	64.0	25.0	35.0	0.0	29.0	35.2
MiniCPM-V2.6	26.0	56.0	22.0	44.0	34.0	50.0	51.0	29.0	23.0	0.0	19.0	32.2
VILA1.5-13B	26.0	55.0	26.0	35.0	17.0	34.0	48.0	26.0	12.0	4.0	22.0	27.7
Idefics3-8B	29.0	51.0	23.0	23.0	18.0	20.0	47.0	30.0	24.0	4.0	24.0	26.6
DeepSeek-VL2-Small	31.0	52.0	36.0	41.0	38.0	57.0	59.0	28.0	41.0	3.0	26.0	37.5
DeepSeek-VL2-Tiny	32.0	52.0	36.0	24.0	27.0	25.0	47.0	27.0	26.0	4.0	16.0	28.7
Pixtral-12B	31.0	68.0	24.0	24.0	21.0	38.0	53.0	21.0	24.0	3.0	37.0	31.3
LLaVA-OneVision-7B	42.0	59.0	21.0	41.0	30.0	50.0	59.0	26.0	20.0	0.0	22.0	33.6
EMU3	31.0	52.0	24.0	25.0	17.0	25.0	47.0	25.0	24.0	0.0	20.0	26.4
Baseline												
Random Guessing	33.0	50.0	25.0	25.0	20.0	25.0	50.0	25.0	25.0	4.2	20.0	27.5
↑ Random ($p < 0.05$)	42.0	59.0	33.0	33.0	28.0	33.0	59.0	33.0	33.0	9.0	28.0	35.5

Table 2: Comparing top-performing MLLMs with human proficiency on LEGO-Puzzles-Lite. The best results are marked in **bold**. The top 3 overall performances are highlighted in Dark Green, Medium Green, and Light Green, respectively.

Models	Spatial Understanding				Single-Step Reasoning				Multi-Step Reasoning			Overall
	Height	Adjacency	Rotation	Multiview	Next-Step	Dependency	Rotation Stat.	Position	Backwards	Ordering	Outlier	
LEGO-Puzzles-Lite												
Human proficiency	**70.0**	**95.0**	**95.0**	**100.0**	**90.0**	**100.0**	**100.0**	**95.0**	**95.0**	**95.0**	**95.0**	93.6
GPT-5	55.0	70.0	80.0	85.0	75.0	95.0	55.0	60.0	70.0	85.0	60.0	71.8
GPT-o3	65.0	70.0	75.0	70.0	75.0	95.0	55.0	55.0	80.0	75.0	65.0	70.9
Gemini-2.5-Pro	40.0	70.0	70.0	55.0	85.0	90.0	50.0	50.0	55.0	80.0	80.0	65.9
Qwen3-VL-32B-Instruct	45.0	55.0	30.0	50.0	70.0	95.0	65.0	75.0	80.0	50.0	50.0	60.5
InternVL3-78B	50.0	65.0	35.0	50.0	65.0	75.0	60.0	40.0	20.0	30.0	40.0	48.2
Qwen2.5-VL-72B	25.0	70.0	25.0	35.0	65.0	70.0	65.0	45.0	55.0	20.0	55.0	48.2

Human Outperforms the Best MLLMs by Over 20%. To study the performance gap between human and MLLMs on the Elementary set of LEGO-Puzzles, we randomly select 20 questions from each task to create **LEGO-Puzzles-Lite**, resulting in a total of 220 QA pairs, and use this subset for investigation. We invited 30 human experts to solve these questions, ensuring a diverse and representative evaluation of human-level reasoning. As shown in

Tab. 2, human experts consistently achieve significantly higher overall performance (93.6%). In contrast, current MLLMs fall short, with even the most advanced models, Gemini-2.5-pro and GPT-5, trailing over 20% behind human performance across all tasks.

Hard Cases in the Elementary Set. As shown in Tab. 1, most models (22/29) perform worse than ↑*Random* in the *Height* task. Part of the cause is that we observe that MLLMs tend to answer questions based on the object's height in its 2D projection rather than in its true 3D perspective. See appendix for some examples. This is consistent with the observation in 3DSRBench [27], where MLLMs exhibit a similar tendency on real-world images. Our construction of the *Height* task contains many images that can lead to different answers between 2D and 3D perspectives, which further exacerbates the problem.

Another task that MLLMs struggle with is the *Rotation* task, where 16 out of 29 models fall below ↑*Random*. This shows that current MLLMs may not perceive and distinguish object orientation changes reliably, which could

Beyond basic spatial understanding, increasing the reasoning complexity further degrades the performance of MLLMs. For example, when the task changes from *Next-Step* to *Ordering*, the accuracy drops significantly for most models.

MLLMs Perform Poor in Multi-Step Reasoning. LEGO-Puzzles also reveals substantial limitations in MLLMs' sequential reasoning capabilities, particularly when multiple reasoning steps are required. As shown in Tab. 1, almost half of the models perform below ↑ *Random* in the *Ordering* task, with some models (e.g., InternVL2.5-8B, LLaVA-OneVision-7B) failing completely. Similar trends are observed in the *Backwards* task, where 12 out of 20 open-source models perform below ↑ *Random*. To further explore the upper bound of MLLMs in multi-step reasoning, we conduct the more challenging and length-controllable task in Sec. 4.

4 Evaluation on the Planning Set

4.1 VQA Results

To further investigate the long-horizon multi-step sequential reasoning abilities of MLLMs, we design a scalable and challenging task ***PLAN-k-Step-VQA***, which controls the number of reasoning steps k to model the realistic reasoning planning and execution across multiple sequences under noisy conditions (defined in Sec. 2.2). We select the four top-performing models in Sec. 3: GPT-5, Gemini-2.5-Pro, Qwen3-VL-32B-Instruct, and InternVL3-78B.

Evaluation Metrics. We use *exact match accuracy* as in the Elementary set, and additionally include two metrics: (1) *set match accuracy*(%), defined as the proportion of predictions in which the predicted steps contain exactly the same elements as the ground truth, regardless of order. (2) *overlap ratio* (%), which measures the fraction of predicted steps that also appear in the ground-truth sequence without considering order. The results are shown in Fig. 3.

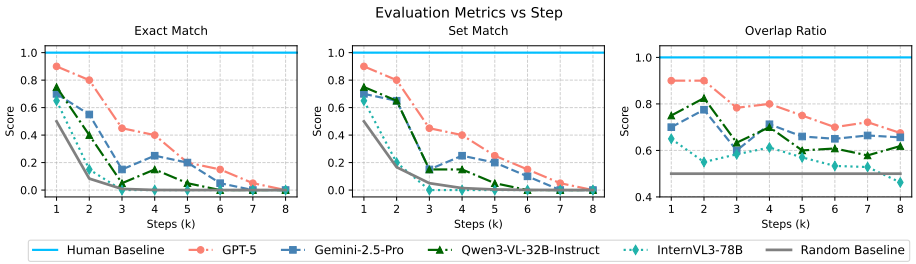

Fig. 3: Evaluation on *PLAN-k-Step* of different MLLMs across varying numbers of steps.

Performance Degradation when k Increases. As shown in Fig. 3, all models exhibit a clear decline in *exact match accuracy* as the number of planning steps k increases. GPT-5 drops from 90% at $k = 1$ to 0% at $k = 8$, while both Qwen3-VL-32B-Instruct and InternVL3-78B reach 0% once $k > 4$. For $k > 3$, the accuracy of all models drops below 50%.

Model Struggles in Selecting the Correct Set of Steps. Comparing the accuracy of *exact match* and *set match*, GPT-5 exhibits almost identical curves. This indicates that errors emerge already at the set-selection stage: the model fails to identify the correct elements and consequently selects an incorrect set of steps, let alone ordering them correctly.

Clear Gap Between Open-Source and Proprietary Models. For *overlap ratio*, even when *exact match accuracy* falls to 0% at $k = 7/8$, the proprietary models still achieve ∼65%, indicating that they often identify a substantial portion of the correct steps. Conversely, the open-source models degrade to near random-guessing levels as $k > 4$.

MLLMs Perform Far Below Human Level. To establish a human baseline, we invited five expert participants to solve the same set of tasks. Humans achieved a perfect score of 100% across all values of k, showing no decline as the number of steps increased. On average, they completed all tasks in about 85 minutes. This demonstrates that, unlike MLLMs, humans can reliably handle long-horizon multi-step reasoning tasks. In contrast, current MLLMs struggle to track and integrate spatial transformations across multiple steps, where accumulated errors lead to inconsistent predictions in longer reasoning chains.

4.2 Image Generation Results

Experiments in Sec. 4.1 show that MLLMs struggle to maintain consistent multi-step reasoning as the number of steps increases, even in a simplified multiple-choice setting. To further investigate their generative ability, we construct a full-stage planning task (defined in Sec. 2.2) that requires models to generate intermediate assembly states step by step.

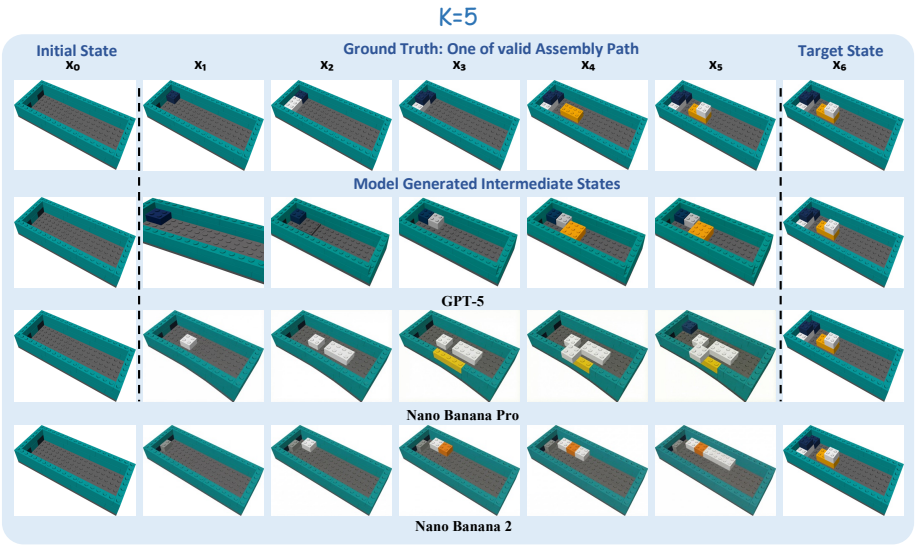

Fig. 4: Qualitative results of generative full-stage planning. The figure provides qualitative examples of the generated intermediate states for both models at $k = 5$.

Benchmark Models. We evaluate three strong current generation models: GPT-5, Gemini-3-pro-image-preview (Nano Banana Pro) and gemini-3.1-flash-image-preview (Nano Banana 2).

Evaluation Protocol. To construct ground-truth planning paths, we first apply topology expansion to enumerate all valid assembly sequences from the initial state to the final state under physical and assembly constraints, where lower parts must be assembled before upper parts. This produces multiple feasible topological trajectories rather than a single fixed path. We then perform multi-turn generation, where the model outputs one intermediate state at each turn conditioned on the previously generated history.

For evaluation, each generated step is matched against all topology-expanded candidates at that step using mean squared error (MSE). After per-step matching, we check whether the matched steps can be concatenated into a complete valid sequence in the topology-expanded set. If the matched steps form a full legal trajectory, the case is counted as correct (score 1); otherwise it is counted as incorrect (score 0). This automatic evaluation process makes it possible to quantitatively assess the correctness of the generated multi-step planning paths, which would be infeasible to evaluate manually due to the large number of possible trajectories.

We set the planning step $k = 2, 3, 4, 5$. We do not evaluate longer horizons because even at $k = 3$ the strongest models already fail completely. Although some generated images show reasonable appearance similarity, none of the models succeed in producing a fully correct generative planning trajectory. We construct 50 cases for this experiment, with 10 cases for each planning horizon k. The

quantitative results and qualitative examples are shown in Fig. 5 and Fig. 4, respectively.

Even the strongest current MLLMs fail to perform full-stage generative planning. The examples in Fig. 4 show that the models cannot produce reasonable intermediate states, with errors often appearing from the very first step. This suggests that current models still lack the ability to perform consistent multi-step reasoning and apply such reasoning to generate correct intermediate images. True generative multi-step planning requires not only spatial understanding, but also consistent visual imagination, temporal reasoning, and implicit physical constraints, which remain challenging for current MLLMs. More analysis is provided in Sec. 5.

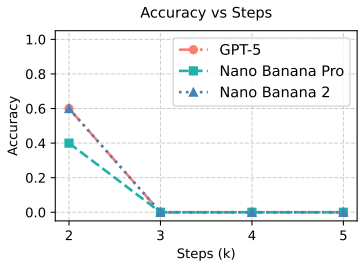

Fig. 5: Quantitative results of generative full-stage planning. The figure shows the quantitative results for GPT-5, Nano Banana Pro and Nano Banana 2 across scaling steps $k = 2, 3, 4, 5$.

5 Discussion

5.1 Correlation with Real-world Spatial Reasoning

Table 3: Pearson correlation coefficients (PCC) and p-values for *height* and *adjacency* tasks.

Task	PCC	P-value
Height	0.93	0.00723
Adjacency	0.98	0.00046

While LEGO-Puzzles is built on rendered data, it aims to evaluate fundamental spatial reasoning capabilities that are also essential in real-world scenarios. To assess its generalizability beyond synthetic environments, we compare model performance on LEGO-Puzzles with 3DSRBench [27], a benchmark based on natural images. Both datasets contain conceptually similar tasks—specifically, the *Height* task in LEGO-Puzzles aligns with *Height* in 3DSRBench, and *Adjacency* in LEGO-Puzzles corresponds to the *Location* task in 3DSRBench.

We evaluate all proprietary models tested in LEGO-Puzzles on the corresponding tasks in 3DSRBench and compute the Pearson correlation coefficient [6] to measure consistency in performance across the two datasets. As shown in Tab. 3, the results reveal strong positive correlations: 0.93 for *Height* and 0.98 for *Adjacency*, both statistically significant ($p < 0.01$).

These findings suggest that LEGO-Puzzles not only offers high scalability and precise control in synthetic settings but also captures spatial reasoning patterns that generalize well to natural images. This validates its utility as a proxy for evaluating real-world spatial understanding.

5.2 MLLMs Struggle with Elementary Tasks in Image Generation

The results in Sec. 4.2 show that even at the very first step, the generated images often contain errors, such as misaligned parts, incorrect orientations, or missing pieces. To better understand the source of these failures, we conduct

an additional experiment where the model is required to generate only a single turn. We select five tasks from the *Elementary* set of LEGO-Puzzles: *Rotation**, *Multiview**, *Position**, *Dependency**, and *Next-Step**. For each task, we sample 20 questions, resulting in a total of 100 questions. In this experiment, we change the output format of the tasks from multiple choice to image generation. Instead of selecting from predefined options, the model is required to directly generate the visual output.

Models and Evaluation. We evaluate the open-source models Emu2 [36], GILL [18], and Anole [10], as well as the proprietary models GPT-5, GPT-4o, Gemini-3-pro-image-preview (Nano Banana Pro) and Gemini-2.0-Flash, all of which support long-range sequence input and image output. For evaluation, traditional metrics such as FID [15], CLIPScore [14], and X-IQE [7] mainly assess image fidelity or cross-modal alignment, often relying on pre-trained model priors or fixed scoring heuristics. They struggle to capture the fine-grained spatial accuracy required in LEGO assembly tasks, where even small errors such as misaligned parts or incorrect orientations can invalidate the result. Furthermore, many recent multimodal evaluation metrics depend on GPT-based models [25], introducing uncontrollable bias into the evaluation process. Therefore, we enlist 5 human experts to assess model performance across two dimensions: appearance similarity and instruction following. Each aspect is rated on a scale from 0 to 3. Detailed scoring guidelines are provided in the appendix.

Table 4: Evaluation on *generation*. We conduct human-based evaluation to assess the "Appearance" (App) and "Instruction Following" (IF) scores of GPT-5, GPT-4o, Nano Banana Pro, Gemini-2.0-Flash, Emu2, GILL, and Anole, using a scoring scale from 0 to 3 for both dimensions.

Task \MLLM	GPT-5		Nano Banana Pro		GPT-4o		Gemini-2.0-Flash		Emu2		GILL		Anole	
	App	IF	App	IF	App	IF	App	IF	App	IF	App	IF	App	IF
Rotation*	2.45	1.75	2.60	1.70	2.35	1.75	2.05	1.45	1.70	0.00	0.00	0.00	0.10	0.00
Multiview*	1.90	2.10	2.45	2.55	2.30	2.00	2.40	1.65	1.65	0.25	0.00	0.00	0.05	0.00
Position*	2.75	1.60	3.00	2.30	2.30	1.65	2.85	1.30	0.50	0.00	0.00	0.00	0.00	0.00
Dependency*	1.85	2.20	2.05	1.65	2.05	2.00	1.70	0.95	0.55	0.00	0.00	0.00	0.00	0.00
Next-Step*	2.30	1.75	2.50	1.75	2.25	1.45	1.75	0.05	0.05	0.00	0.00	0.00	0.00	0.00
Overall	2.25	1.88	2.52	1.99	2.25	1.77	2.15	1.08	0.89	0.05	0.00	0.00	0.03	0.00

MLLMs struggle to perform single-turn image generation, not alone full-stage multi-step planning. Tab. 4 presents the human evaluation results across five generation tasks. Overall, proprietary models outperform open-source ones in both appearance consistency (App) and instruction adherence (IF). Among them, Nano Banana Pro achieves the highest overall scores (App: 2.52, IF: 1.99), followed by GPT-5 (App: 2.25, IF: 1.88). However, both models show clear room for improvement in instruction following. GPT-4o and Gemini-2.0-Flash perform slightly worse than GPT-5 and Nano Banana Pro, with lower scores in both dimensions, especially in instruction following (IF: 1.77 and 1.08, respectively). Among open-source models, Emu2 shows some ability to preserve visual appearance (App: 0.89) but fails almost entirely in instruction following (IF: 0.05), treating the task more as image replication than reasoning-based generation. GILL and Anole perform the worst, with near-zero scores across all tasks

and frequently irrelevant outputs. The qualitative results are shown in Fig. 6. More cases are provided in the appendix.

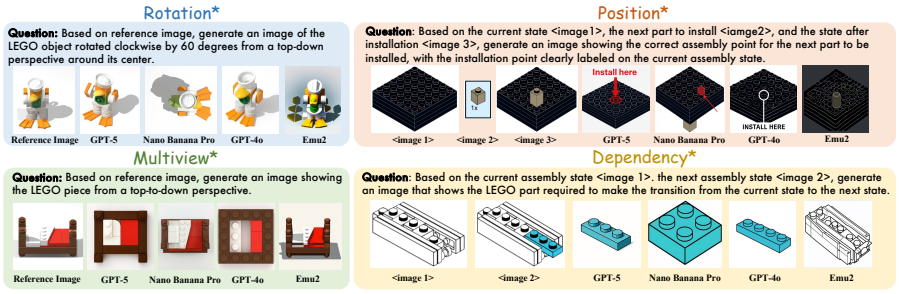

Fig. 6: Qualitative visual results for image generation tasks. Note: The questions above are slightly simplified for clarity and brevity.

Overall, these results show that current models, especially open-source ones, still struggle with spatial reasoning-based image generation in LEGO assembly tasks. Even the strongest models, while producing outputs with high appearance consistency, often fail to fully adhere to the instructions, indicating a gap between visual understanding and precise generation.

6 Conclusion

We introduce LEGO-Puzzles, a novel benchmark specifically designed to evaluate spatial understanding, as well as single-step and multi-step sequential reasoning in MLLMs. Inspired by human cognitive patterns in LEGO construction, we create two task sets. The Elementary set includes over 1,100 carefully curated visual question-answering (VQA) samples across 11 distinct tasks, providing diverse scenarios to assess multimodal visual reasoning. The Planning set directly requires the model to generate a step-by-step plan for assembling a target LEGO structure in VQA and image generation settings. We conduct comprehensive experiments with 29 advanced MLLMs, revealing substantial performance gaps compared to humans, particularly in long-horizon planning and the generation of spatially coherent visual outputs. These findings underscore the urgent need to enhance the spatial understanding and sequential reasoning capabilities of multimodal AI.

Acknowledgements

We would like to thank Xinyu Fang, Xiangyu Zhao, and the anonymous reviewers for their insightful comments and feedback that helped improve the quality of this paper. We also thank the broader research community for releasing datasets and evaluation resources that made the experiments in this work possible.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023)
2. Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., De Monicault, B., Garg, S., Gervet, T., et al.: Pixtral 12B. arXiv preprint arXiv:2410.07073 (2024)
3. Anthropic: Claude 3.5 Sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet` (2024), accessed: 2026-06-28
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923 (2025)
6. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson Correlation Coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)
7. Chen, Y., Liu, L., Ding, C.: X-IQE: eXplainable Image Quality Evaluation for Text-to-Image Generation with Visual Large Language Models. arXiv preprint arXiv:2305.10843 (2023)
8. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv preprint arXiv:2412.05271 (2024)
9. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. Science China Information Sciences **67**(12), 220101 (2024)
10. Chern, E., Su, J., Ma, Y., Liu, P.: ANOLE: An Open, Autoregressive, Native Large Multimodal Models for Interleaved Image-Text Generation. arXiv preprint arXiv:2407.06135 (2024)
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
12. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 11198–11201 (2024)
13. Gardner, H.: Frames of Mind: The Theory of Multiple Intelligences. Basic books (2011)
14. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Advances in neural information processing systems **30** (2017)

16. Jirout, J.J., Newcombe, N.S.: Building Blocks for Developing Spatial Skills: Evidence From a Large, Representative U.S. Sample. Psychological science **26**(3), 302–310 (2015)
17. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
18. Koh, J.Y., Fried, D., Salakhutdinov, R.R.: Generating Images with Multimodal Language Models. Advances in Neural Information Processing Systems **36**, 21487–21506 (2023)
19. Laurençon, H., Marafioti, A., Sanh, V., Tronchon, L.: Building and better understanding vision-language models: insights and future directions. arXiv preprint arXiv:2408.12637 (2024)
20. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-OneVision: Easy Visual Task Transfer. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=zKv8qULV6n, accessed: 2026-06-28
21. Li, D., Li, H., Wang, Z., Yan, Y., Zhang, H., Chen, S., Hou, G., Jiang, S., Zhang, W., Shen, Y., et al.: ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models. arXiv preprint arXiv:2505.21500 (2025)
22. Li, Z., Wang, X., Stengel-Eskin, E., Kortylewski, A., Ma, W., Van Durme, B., Yuille, A.L.: Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14963–14973 (2023)
23. Liao, Y.H., Mahmood, R., Fidler, S., Acuna, D.: Reasoning Paths with Reference Objects Elicit Quantitative Spatial Reasoning in Large Vision-Language Models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 17028–17047 (2024)
24. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: VILA: On Pretraining for Visual Language Models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
25. Liu, M., Xu, Z., Lin, Z., Ashby, T., Rimchala, J., Zhang, J., Huang, L.: Holistic Evaluation for Interleaved Text-and-Image Generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 22002–22016 (2024)
26. Ma, L., Wen, J., Lin, M., Xu, R., Liang, X., Lin, B., Ma, J., Wang, Y., Wei, Z., Han, M., et al.: PhyBlock: A Progressive Benchmark for Physical Understanding and Planning via 3D Block Assembly. Advances in Neural Information Processing Systems **38** (2026)
27. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6924–6934 (2025)
28. McDougal, E., Silverstein, P., Treleaven, O., Jerrom, L., Gilligan-Lee, K., Gilmore, C., Farran, E.K.: Assessing the impact of lego® construction training on spatial and mathematical skills. Developmental Science **27**(2), e13432 (2024)
29. McDougal, E., Silverstein, P., Treleaven, O., Jerrom, L., Gilligan-Lee, K.A., Gilmore, C., Farran, E.K.: Associations and indirect effects between lego® construction and mathematics performance. Child development **94**(5), 1381–1397 (2023)

30. Newcombe, N.S., Frick, A.: Early Education for Spatial Intelligence: Why, What, and How. Mind, Brain, and Education **4**(3), 102–111 (2010)
31. OpenAI: Introducing OpenAI o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini` (2025), accessed: 2026-06-28
32. Shiri, F., Guo, X.Y., Far, M.G., Yu, X., Haf, R., Li, Y.F.: An empirical analysis on spatial reasoning capabilities of large multimodal models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21440–21455 (2024), accessed: 2026-06-28
33. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: GPT-5 System Card. arXiv preprint arXiv:2601.03267 (2025)
34. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the Gap: Benchmarking Spatial Reasoning in Vision-Language Models. arXiv preprint arXiv:2503.19707 (2025)
35. Su, Y., Ling, Z., Shi, H., Jiayang, C., Yim, Y., Song, Y.: ActPlan-1K: Benchmarking the Procedural Planning Ability of Visual Language Models in Household Activities. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 14953–14965 (2024)
36. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative Multimodal Models are In-Context Learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14398–14409 (2024)
37. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (2023)
38. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv preprint arXiv:2409.12191 (2024)
39. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv preprint arXiv:2508.18265 (2025)
40. Wang, X., Ma, W., Li, Z., Kortylewski, A., Yuille, A.L.: 3D-Aware Visual Question Answering about Parts, Poses and Occlusions. Advances in Neural Information Processing Systems **36**, 58717–58735 (2023)
41. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-Token Prediction is All You Need. arXiv preprint arXiv:2409.18869 (2024)
42. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv preprint arXiv:2412.10302 (2024)
43. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
44. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: MiniCPM-V: A GPT-4V Level MLLM on Your Phone. arXiv preprint arXiv:2408.01800 (2024)
45. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv preprint arXiv:2504.10479 (2025)