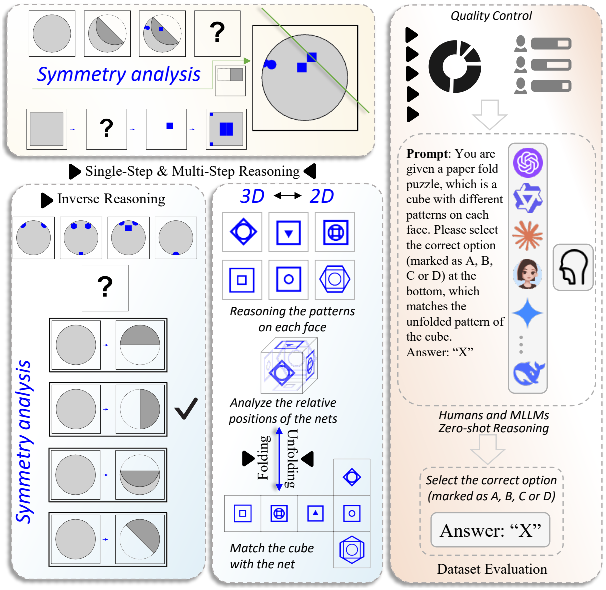
Paper Folding Puzzles: Can Multimodal Large Language Models Perform Spatial Reasoning?

Dibin Zhou[1], Yantao Xu[1], Zongming Huang[1], Zengwei Yan[1], Wenhao Liu[1], Yongwei Miao[1], Jianfeng Ren[2], Fuchang Liu[1]*

[1] School of Information Science and Technology, Hangzhou Normal University
[2] The Digital Port Technologies Lab, School of Computer Science, University of Nottingham Ningbo China
{20080094, 2022210402040, 2022211703035, 1096856083, 20110039, ywmiao, liufc}@hznu.edu.cn,
jianfeng.ren@nottingham.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) largely lag human-level performance on abstract visual reasoning (AVR), which requires models to infer latent rules from visual question sets and generalize them to novel scenarios. Most AVR benchmarks are constrained to narrow and repetitive 2D patterns, involving relatively simple spatial relationships and assessing limited dimensions of reasoning ability. Drawing inspiration from real-world paper folding challenges, we propose Paper Folding Puzzles (PFP), a rigorously designed benchmark specifically developed to assess spatial reasoning capabilities. It comprises 150K visual question-answering samples across five diverse tasks, ranging from basic 2D geometric reasoning to 3D spatial understanding. The developed benchmark dataset can be employed to assess core spatial reasoning abilities essential to human cognition, encompassing fundamental symmetry reasoning and 3D spatial comprehension. Furthermore, we conduct a comprehensive evaluation of 18 leading MLLMs (both closed- and open-source variants) on the PFP benchmark to assess their spatial reasoning capabilities. Our findings show that most MLLMs achieve near-chance performance on FPF, exhibiting substantial performance gaps ($> 30\%$) relative to human baselines across all tasks. This highlights a critical research gap in improving spatial reasoning capabilities of MLLMs. The dataset and code are available at https://github.com/hznuer/PFP_bench.


Symmetry analysis
?
Quality Control
Single-Step & Multi-Step Reasoning
Inverse Reasoning
3D ↔ 2D
Reasoning the patterns on each face
Analyze the relative positions of the nets
Folding
Unfolding
Match the cube with the net
Symmetry analysis
Humans and MLLMs Zero-shot Reasoning
Select the correct option (marked as A, B, C or D)
Answer: "X"
Dataset Evaluation

Prompt: You are given a paper fold puzzle, which is a cube with different patterns on each face. Please select the correct option (marked as A, B, C or D) at the bottom, which matches the unfolded pattern of the cube. Answer: "X"

Figure 1: Overview of task structure in the PFP dataset. The benchmark includes five sub-tasks through folding and unfolding tasks.

Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) (OpenAI 2023; Google 2025; Anthropic 2024) have demonstrated impressive visual reasoning capabilities. Abstract Visual Reasoning (AVR) (Zhang et al. 2019; Hill et al. 2019) poses a distinct challenge by requiring models to infer abstract patterns and structural rules with minimal reliance on contextual grounding. However, most existing AVR benchmarks are constrained to overly simplified 2D settings with limited geometric diversity (Małkiński and Mańdziuk 2023, 2025), offering only narrow assessments of spatial cognition. In contrast, Paper Folding Puzzles (PFP) bridge abstract visual reasoning and real-world spatial cognition through familiar paper-folding tasks. Unlike prior benchmarks limited to simple 2D patterns, PFP introduces diverse and geometrically rich challenges that test generalization and interpretability. As MLLMs evolve, this benchmark offers a compact yet rigorous testbed for evaluating spatial reasoning, with potential applications in robotics, design, and education.

Existing datasets such as RAVEN (Zhang et al. 2019), G-set (Mańdziuk and Żychowski 2019), VAP (Hill et al. 2019), DOPT (Webb et al. 2020), ARC (Chollet 2019), LEGO (Tang et al. 2025) and iVISPAR (Mayer et al. 2025) fall into the geometric category, where the visual elements primarily consist of well-defined shapes and structured layouts. In contrast, datasets like Bongard-LOGO (Nie et al. 2020) and SVRT (Fleuret et al. 2011) represent the abstract type, featuring more irregular or symbolic visual forms. MARVEL (Jiang et al. 2024) encompasses both types, combining structured geometry with abstract visual patterns. From the perspective of pattern dimensionality, most of the aforementioned datasets are limited to 2D geometric pat-

*Corresponding author.

terns, with the exception of MARVEL, LEGO, and iVISPAR, which incorporate 3D geometric structures to support richer spatial reasoning.

Despite the advancements, current AVR research still falls short in the following aspects: 1) Limited research has focused on spatial reasoning grounded in geometric symmetry, as opposed to simple 2D geometric attributes. 2) While prior research has explored certain aspects of 3D spatial reasoning, the cognitive process of inferring relationships between 2D layouts and their folded 3D counterparts—essential to tasks like mental rotation and physical assembly—has received limited attention in existing benchmarks. 3) The ability to construct 3D structures from 2D representations—or conversely, to unfold 3D objects into 2D layouts—is a fundamental component of human spatial cognition. This skill not only plays a vital role in the early development of geometric thinking in children, but also serves as a core cognitive foundation in fields such as elementary geometry (Rows 1917) and computer-aided design. However, despite recent advances in MLLMs, their ability to perform such 2D–3D spatial transformations remains underexplored and poorly understood. Existing benchmarks rarely capture the nuanced reasoning required for this class of tasks, leaving a gap between model capabilities and the spatial reasoning demands found in real-world applications.

To enhance the research on 3D spatial reasoning, we develop the PFP dataset, which comprises a diverse collection of over 150,000 carefully curated visual question–answer pairs across five distinct tasks, organized into three main reasoning categories (Fig. 1). These five groups encompass fundamental evaluations of spatial reasoning skills, including tasks such as single-step and multi-step sequential reasoning, reverse reasoning, 3D folding and 2D unfolding reasoning. Paper Folding Puzzles extends beyond existing datasets such as RAVEN, MARVEL, LEGO, and iVISPAR by supporting not only isolated 2D or 3D abstract visual reasoning, but also spatial reasoning that involves transformations between 2D and 3D structures.

We evaluate a set of state-of-the-art MLLMs on the proposed PFP dataset to assess their capabilities in abstract and spatial reasoning. The evaluated models include six leading closed-source systems: Doubao1.6-Flash (Doubao Team 2025), Gemini-2.5-Flash (Gemini Team 2025), Sonnet-3.7-Thinking (Claude Team 2025), GPT-4o, GPT-4o-mini (OpenAI 2023), and o4-mini (OpenAI 2025). In addition, we include a diverse collection of open-source models, such as GLM-4.1V-Thinking-Flash-10B (GLM-V Team 2025), Qwen2-VL-[7B/72B] (Wang, Bai et al. 2024), and Qwen2.5-VL-[7B/32B/72B] (Bai, Chen et al. 2025). The experimental results reveal the substantial challenges posed by our benchmark in both abstract pattern recognition and 2D–3D spatial reasoning. Our benchmark evaluation also reveals a substantial gap between current MLLMs and human-level spatial reasoning: humans outperform all six tested closed-source MLLMs by over 30%, while twelve leading open-source models exhibit even weaker performance, with most nearing random baselines. These results highlight the need for more structured spatial reasoning and understanding capabilities in future MLLMs.

In summary, Paper Folding Puzzles offers a thorough assessment of MLLMs' sequential reasoning and spatial comprehension abilities. The main contributions of our work are as follows:

- **Evaluation for single-step and multi-step sequential reasoning.** Built upon a step-by-step construction process, Paper Folding Puzzles is the first benchmark explicitly designed to evaluate both single-step and multi-step sequential reasoning grounded in symmetry. Each task in this benchmark requires reasoning about up to 3 folds and 4 cutout shapes.
- **Evaluation of bidirectional 2D–3D spatial understanding.** Our benchmark offers a diverse suite of tasks designed to systematically evaluate MLLMs' spatial reasoning capabilities across both 2D and 3D domains in both directions. These include folding 2D nets into 3D cubes, unfolding 3D cubes into flat nets, and inferring folding procedures from final cutout shapes. Together, these tasks enable a comprehensive evaluation of both forward and inverse spatial reasoning.
- **Dataset Synthesis and Extensive Experiments.** Our benchmark reveals significant deficiencies even in advanced MLLMs. We constructed 150,000 puzzles and evaluated state-of-the-art models on a 3,000 test subset, comparing their results with human performance to uncover key insights into their strengths, weaknesses, and future directions for enhancing AVR reasoning.

Related Work

AVR Benchmarks. The domain of AVR (Zhang et al. 2019) focuses on tasks that require abstracting and analogizing visual patterns across different contexts. These tasks typically involve recognizing relational structures and transformation rules that govern simple 2D shapes and their visual properties (Małkiński and Mańdziuk 2025). Existing AVR benchmarks vary significantly across multiple dimensions (Małkiński and Mańdziuk 2023). Some focus on clearly defined geometric shapes (Zhang et al. 2019; Mańdziuk and Żychowski 2019; Hill et al. 2019; Webb et al. 2020; Chollet 2019), while others employ more abstract visual forms (Nie et al. 2020; Fleuret et al. 2011). The underlying rules in AVR benchmarks can be either explicit or abstract. For example, PGM (Barrett et al. 2018) employs five well-defined rules such as progression, XOR, OR, and AND, while Bongard problems (Nie et al. 2020) rely on abstract rules that distinguish between two groups of visual panels. AVR tasks include classification, generation, and description, typically framed as either completion (inferring missing elements) or discrimination (identifying anomalies). Datasets like RAVEN (Zhang et al. 2019), PGM (Barrett et al. 2018), and VAP (Hill et al. 2019) emphasize domain transfer, requiring models to generalize learned rules. Others, such as VAEC and DOPT (Webb et al. 2020), focus on extrapolation across systematic spatial variations (e.g., position, size). These benchmarks provide a controlled setting to assess MLLMs' abstract reasoning abilities.

Multimodal Large Language Models. The integration of rich visual representations from vision encoders (Radford et al. 2021) with the strong reasoning capabilities of large language models (LLMs) (Touvron et al. 2023; Chiang et al. 2023) has led to the development of MLLMs (Li et al. 2022; Dai et al. 2023; OpenAI 2023; Liu et al. 2023). These models have been successfully applied to a wide range of vision-language tasks, such as image captioning (Young et al. 2014; Agrawal et al. 2019), VQA (Antol et al. 2015; Goyal et al. 2017; Manmadhan and Kovoor 2020), VCR (Zellers et al. 2019; Xie et al. 2019), and Physical Reasoning (Bakhtin et al. 2019; Riochet et al. 2022), achieving notable performance in zero-shot (Li et al. 2022), few-shot (Alayrac et al. 2022; Zhao et al. 2024), and chain-of-thought reasoning abilities (Huang et al. 2023). To better understand the scope of MLLMs' capabilities, prior studies have investigated their performance on geometric (Kazemi et al. 2023) and mathematical reasoning tasks (Wang et al. 2025). Closely related to our work are the studies by Jiang et al. (Jiang et al. 2024) and Tang et al. (Tang et al. 2025), which assess MLLMs' multidimensional abstract reasoning and 3D spatial understanding. However, these evaluations are either limited in scale or fail to consider spatial reasoning across 2D and 3D representations. In this work, we address this gap by conducting a large-scale evaluation that provides comprehensive insights into the abstract reasoning abilities of MLLMs, with a particular focus on symmetry in 2D space and the bidirectional reasoning between 2D and 3D representations.

Paper Folding Puzzles

Attributes and Rules

For 2D folding problems, we employ twelve types of symmetric folds, including horizontal, vertical, and both forward and backward diagonal mirror folds, covering 2-way, 3-way, and 4-way symmetries. Additionally, four asymmetric folds, such as corner folds with randomly sampled axes or positions, are included. Base shapes are uniformly generated across fold types and further augmented with 1–4 cutout elements (`circle`, `rectangle`, `square`, `hexagon`) according to predefined probabilities. Cutout placements are sampled from `corner`, `edge`, and `center` positions to construct the final folded patterns. The distractor options are primarily generated by altering the types of symmetry axes used in the 1- to 3-fold processes, as well as modifying the number, size, and shape of the cutout elements, leading to variations in the resulting patterns' size, shape, orientation, and quantity. For 3D folding and unfolding tasks, we design problems based on face composition transformations that incorporate visual patterns and spatial configurations. Each cube has 24 viewpoints, 11 unfoldable nets, and 30 surface pattern variations. In folding tasks, distractors are generated by altering viewpoints or swapping patterns to change face adjacency and opposition. Problems are categorized as easy or hard based on the consistency of reference face (e.g., front) features. For unfolding tasks, we use 11 cube nets with 30 distinct surface patterns. Distractors are constructed by modifying net shapes or face patterns, and difficulty is determined by whether the 2D nets across choices share the same layout.

2D/3D Spatial Reasoning

In the context of the PFP dataset, 2D spatial reasoning refers to the mental manipulation of two-dimensional shapes, including understanding how objects transform through reflection, translation, rotation, and composition within the 2D plane. Unlike other reasoning tasks, it requires not only pattern recognition and the comprehension of sequential folding procedures, but also the ability to infer mirror symmetry and predict the mirrored outcomes of cuts after unfolding. Such multi-step sequential reasoning further demands short-term memory, temporal ordering, and mental simulation–all of which present unique challenges for current MLLMs. In the context of the PFP dataset, 3D spatial reasoning involves the mental construction and deconstruction of three-dimensional structures, understanding spatial relationships between 2D and 3D representations, and reasoning about object transformations in 3D space–such as rotation, perspective changes, and folding from 2D nets into 3D forms. This requires MLLMs to perform tasks such as matching a 2D net to its corresponding 3D object, reasoning about perspectives and visible surfaces, and understanding adjacency relationships between object faces. By evaluating these abilities, the PFP benchmark highlights critical limitations in current MLLMs and provides insight into the next steps needed for advancing spatially-grounded multimodal intelligence.

Task Definition

To thoroughly evaluate spatial reasoning in MLLMs, we define five categories of tasks. As illustrated in Fig. 1, the PFP dataset comprises: (1) symmetry-based single-step and multi-step spatial reasoning tasks, (2) their corresponding inverse reasoning variants, and (3) tasks involving bidirectional reasoning between 2D and 3D structures, including folding and unfolding procedures.

Tpye 1: Single-Step Reasoning. This task focuses on single-step spatial reasoning and includes four subtasks: 1-Fold, 2-Fold, 3-Fold, and Others. Each subtask represents a different level or type of folding complexity.

(1) 1-Fold: This category includes problems involving a single folding operation, such as horizontal, vertical, diagonal folds. It also covers symmetric variants along parallel axes and corner folds. These folds typically result in two-layer structures. (2) 2-Fold: This category involves two sequential folding operations selected from the same set of directions. The resulting configurations produce two to four layers, requiring more complex spatial reasoning than the 1-Fold category. (3) 3-Fold: Problems in this category are constructed using three consecutive folding operations, leading to intricate structures with two to eight layers and increased reasoning difficulty. (4) Others: For simulation efficiency, corner folds—typically lacking inter-layer interaction or sequential dependencies—are merged into a single equivalent operation. As a result, 1-fold cases may include up to four simultaneous corner folds. These cases introduce substantial geometric irregularity and spatial ambiguity; we isolate them to form a distinct subtask for focused evaluation.

Tpye 2: Multi-Step Reasoning. This task is built upon the 3-Fold subset of Type 1 (Single-Step Reasoning), but reformulates the question type. Instead of predicting the final state, this task requires identifying the step that does not occur during the transition from the initial to the target state.

Tpye 3: Inverse Reasoning. This task is designed as an inverse reasoning task based on the 1-Fold, 2-Fold, and Others subsets from Type 1. Instead of predicting the outcome of folding operations, this task presents a cut-out pattern and requires humans or MLLMs to infer the folding steps that could have produced it. To control task difficulty, only one-step or two-step folding operations are used.

Tpye 4: 3D Folding Reasoning. In this task, a 2D net of a cube with patterns on its faces is provided, and the model is required to identify the corresponding 3D cube after folding. To control for difficulty, we categorize the task into two subsets (easy and hard) depending on whether the viewpoint of the target cube remains fixed or varies. Viewpoint variation is generally recognized as adding complexity to spatial reasoning.

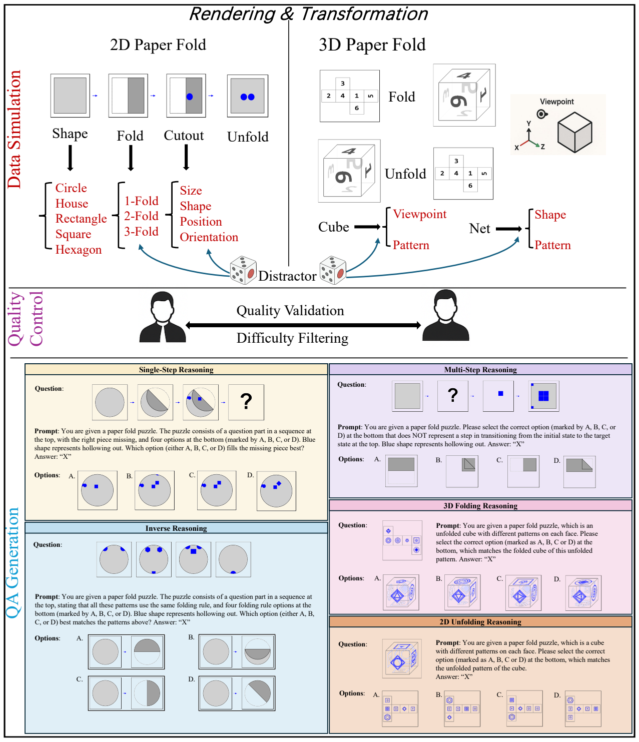
Tpye 5: 2D Unfolding Reasoning. This task reverses the direction: given a 3D cube with distinct face patterns, the model must identify the corresponding 2D unfolded net. We further divide this task into easy and hard subsets based on whether the shapes of the candidate nets differ. Variations in 2D net shapes typically imply distinct unfolding paths, which are generally considered to significantly increase the reasoning difficulty.

Representative examples and prompts for the five task types are illustrated in Fig. 2.

Construction of PFP Dataset

As shown in Fig. 2, the proposed data construction pipeline consists of three major components: data simulation, question–answer generation, and quality control. This pipeline ensures the scalability, accuracy, and reliability of our data.

Construction Procedures. To construct the dataset, we developed a custom paper-folding simulator capable of modeling both 2D and 3D folding processes. For 2D tasks, the tool simulates 1-, 2-, and 3-fold operations over 22,500 base shapes (`circle`, `house`, `rectangle`, `square`, `hexagon`), each combined with 1–4 same-sized cutouts (`circle`, `rectangle`, `square`, `hexagon`). Cutout positions are sampled as `corner` (20%), `edge` (30%), or `center` (50%), and the number of cutouts follows a distribution: 1 (25%), 2 (30%), 3 (40%), 4 (5%). This process yields a total of 90,000 unique 2D reasoning problems. For 3D reasoning, we simulate the unfolding (30,000) and reassembly of cubes (30,000) problems. we designed 11 cube nets, 30 surface pattern variations, and 24 viewpoints to introduce substantial structural diversity and visual confusion. This setup presents a demanding challenge that requires precise spatial reasoning and accurate pattern matching. To ensure data quality and reliability, we adopt a multi-stage human-in-the-loop review process. First, we conduct quality validation to filter out samples with overly small cutouts, irregular shapes, out-of-bound positions, or folded results exceeding paper boundaries. Then, we apply difficulty filtering to exclude QA instances that are overly challenging for humans. The dataset is intentionally curated to prioritize examples that are easy for humans but remain difficult for current MLLMs, highlighting their limitations in spatial reasoning.



Single-Step Reasoning — **Prompt**: You are given a paper fold puzzle. The puzzle consists of a question part in a sequence at the top, with the right piece missing, and four options at the bottom (marked by A, B, C, or D). Blue shape represents hollowing out. Which option (either A, B, C, or D) fills the missing piece best? Answer: “X”

Multi-Step Reasoning — **Prompt**: You are given a paper fold puzzle. Please select the correct option (marked by A, B, C, or D) at the bottom that does NOT represent a step in transitioning from the initial state to the target state at the top. Blue shape represents hollowing out. Answer: “X”

Inverse Reasoning — **Prompt**: You are given a paper fold puzzle. The puzzle consists of a question part in a sequence at the top, stating that all these patterns use the same folding rule, and four folding rule options at the bottom (marked by A, B, C, or D). Blue shape represents hollowing out. Which option (either A, B, C, or D) best matches the patterns above? Answer: “X”

3D Folding Reasoning — **Prompt**: You are given a paper fold puzzle, which is an unfolded cube with different patterns on each face. Please select the correct option (marked as A, B, C or D) at the bottom, which matches the folded cube of this unfolded pattern. Answer: “X”

2D Unfolding Reasoning — **Prompt**: You are given a paper fold puzzle, which is a cube with different patterns on each face. Please select the correct option (marked as A, B, C or D) at the bottom, which matches the unfolded pattern of the cube. Answer: “X”

Figure 2: Data curation pipeline.

Evaluation Protocol. We evaluate MLLMs’ spatial reasoning on the Paper Folding Puzzles dataset using a zero-shot protocol. We generated 141,000 training questions, 6,000 validation questions, and test sets consisting of 3,000 questions. Each model is given standardized input, which includes visual stimuli such as folding sequences, 2D nets, or 3D cubes, along with a task-specific textual prompt. Examples of these prompts are illustrated in Fig. 2.

Dataset Statistics. As illustrated in Fig. 3, the dataset consists of 150,000 questions uniformly distributed across five task types, each contributing 20% of the total. Within each task type, questions are further divided into 2–4 subtypes based on folding complexity, except for Type 2. Due to the multi-step nature of Type 2 reasoning, which typically requires three folds, we did not introduce additional folding steps, as such complexity poses challenges even for human participants. In terms of spatial characteristics, 60% of the tasks involve 2D reasoning and 40% involve 3D reasoning. Additionally, from the perspective of bidirectional spatial reasoning, 60% of the tasks involve forward reasoning (Types 1, 2, and 4), while the remaining 40% focus on backward reasoning (Types 3 and 5). Furthermore, the four options in our single-choice questions are approximately uniformly distributed.

Experimental Results

Experimental Settings

Model Selection We compared various MLLMs that represent the state-of-the-art for both closed-source and open-

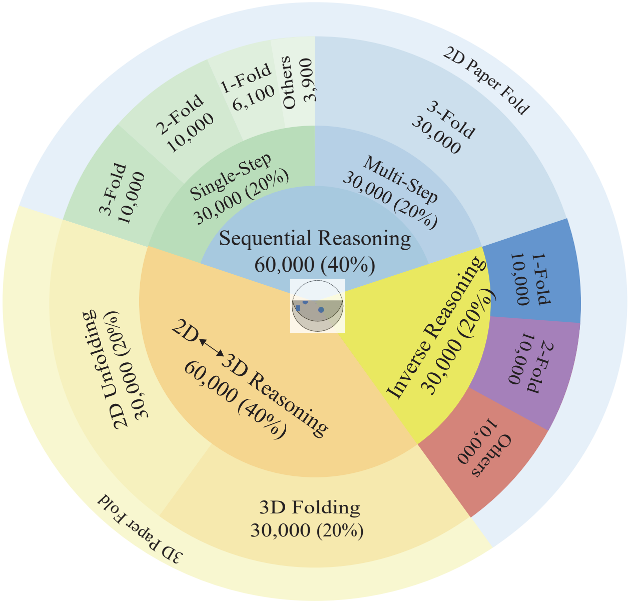


Figure 3: Problem statistics in Paper Folding Puzzles.

source models, covering a diverse range of architectures, sizes, and training processes.

We selected closed-source models based on their leaderboard rankings (LMArena 2025; LLMStats 2025), and open-source models based on response stability and validity. The final set includes models used in recent spatial reasoning studies (Tang et al. 2025; Jiang et al. 2024; Ahrabian et al. 2024).

Closed-source MLLMs. We include API-based MLLMs including Doubao1.6-Flash (Doubao Team 2025), Gemini-2.5-Flash (Gemini Team 2025), Sonnet-3.7-Thinking (Claude Team 2025), GPT-4o (20241120), GPT-4o-mini (OpenAI 2023), and o4-mini (OpenAI 2025).

Open-source MLLMs. We evaluate MLLMs from 4.5B to 72B: GLM-4.1V-Thinking-Flash-10B (GLM-V Team 2025), Qwen2-VL-[7B/72B] (Wang, Bai et al. 2024), Qwen2.5-VL-[7B/32B/72B] (Bai, Chen et al. 2025), Emu3-8B (Emu3 Team 2024), Pixtral-12B (Agrawal et al. 2024), Idefics3-8B (Laurençon et al. 2024), InternVL2.5-8B (Chen, Wang et al. 2025), MiniCPM-V2.6-8B (Yao, Yu et al. 2024), LLaVA-1.5-13B (Liu et al. 2023) and DeepSeek-VL2-4.5B (Wu, Chen et al. 2024).

Human Evaluation. To estimate the upper-bound performance on the Paper Folding Puzzles benchmark, we recruited 20 human participants aged between 20 and 40 years, all of whom are graduates of universities with science and engineering majors, and none of whom had received prior training in spatial reasoning. Each participant was asked to solve 30 questions from every task, ensuring comprehensive coverage of all pattern types and task configurations. The overall accuracy across these 20 participants is reported as the human performance baseline.

Evaluation Metrics. For all visual questions in Paper Folding Puzzles, we adopt accuracy (%) as the evaluation metric. Most models were evaluated three times per task, with the best overall performance reported. Exceptions include Sonnet-3.7-Thinking, Gemini-2.5-Flash, and o4-mini, which were excluded from repeated runs due to excessive runtime (over 5 hours) and high costs (over $8) per task.

Single-Step Reasoning

Table 1 presents the performance of MLLMs on single-step sequence reasoning tasks with increasing folding complexity. Doubao1.6-Flash and GLM-4.1V-Thinking-Flash-10B achieve the highest accuracy (41.67%), followed by Qwen2-VL-72B (34.67%). Task difficulty increases progressively from 1-Fold to 3-Fold and Others, reflecting the greater number of folding steps and the increasing complexity of local symmetry. These patterns are consistent with human cognitive expectations.

Top-performing models show accuracy trends similar to human performance but still encounter two main challenges. First, perception errors occur when models misinterpret subtle differences in cutout size, orientation, shape, or number. These variations are typically easy for humans to discern. Second, corner-fold tasks (categorized under ”Others”) lead to substantial reasoning errors due to the lack of global symmetry, posing challenges for both MLLMs and humans; however, human accuracy remains approximately 30% higher.

Overall, although current MLLMs demonstrate a degree of reasoning ability, they remain limited in fine-grained visual discrimination and in handling spatial reasoning tasks that involve symmetry. These limitations underscore the perceptual challenges identified by the AVR benchmark.

Multi-Step Reasoning

In this task, MLLMs perform even more poorly due to the interdependence among answer choices, which requires short-term memory and stronger reasoning abilities such as ordering and elimination. For humans, multi-step reasoning often reveals additional clues through the structure of the options. Once the correct step order is determined, incorrect choices can be efficiently ruled out. As shown in Table 1, human accuracy significantly exceeds that of all tested MLLMs. Specifically, human performance is more than twice as high as that of the best-performing model, o4-mini (41.00%), while all other models score below 35%.

Error analysis of the top-performing MLLMs reveals that most mistakes occur when identifying the third folding step. This suggests that these models struggle to understand the sequential dependencies across the four options. Unlike humans, they struggle to reason through the steps in a hierarchical manner, first identifying the initial folds and then applying elimination to determine the final step.

Inverse Reasoning

Inverse reasoning tasks challenge models to reconstruct the original 1- to 2-step folding sequence based on a given 2D cutout pattern and select the correct answer accordingly. Unlike sequential reasoning, which emphasizes alignment with a process, this task focuses on pattern induction from observed outcomes, a skill in which humans are excellent.

	Models	Single-Step Reasoning					Multi-Step Reasoning					
		1-Fold	2-Fold	3-Fold	Others	Overall	Circle	Hexagon	House	Rectangle	Square	Overall
Closed-Source	Doubao1.6-Flash	40.98	45.50	43.50	28.21	41.67	24.17	25.83	31.67	29.17	27.50	27.67
	GPT-4o-mini	30.33	29.00	25.50	19.23	26.83	28.00	30.50	25.50	28.00	26.50	27.70
	GPT-4o	31.97	31.00	28.50	26.92	29.83	25.83	29.17	35.83	24.17	27.50	28.50
	o4-mini	34.43	35.50	31.00	26.92	31.96	40.83	43.33	34.17	33.33	53.33	41.00
	Gemini-2.5-Flash	26.23	31.5	28.50	32.05	29.5	33.33	41.67	36.67	28.33	29.17	33.83
	Sonnet-3.7-Thinking	31.97	33.00	31.00	25.64	31.17	39.17	30.00	35.00	28.33	40.83	34.67
Open-Source	GLM-4.1V-10B	37.70	45.50	48.00	21.79	41.67	22.50	28.33	35.83	20.00	28.33	27.00
	Qwen2.5-VL-7B	27.05	26.50	19.00	16.67	22.83	19.17	23.33	22.50	20.83	24.17	22.00
	Qwen2.5-VL-32B	28.69	31.5	29.00	24.36	29.17	28.33	31.67	14.17	27.50	25.00	25.33
	Qwen2.5-VL-72B	33.61	35.00	31.00	21.79	31.67	21.67	34.17	26.67	27.50	19.17	25.83
	Qwen2-VL-72B	30.33	42.50	34.00	23.08	34.67	21.67	28.33	34.17	27.50	18.33	26.00
	Emu3-8B	24.59	25.50	28.50	28.21	26.70	24.17	24.17	26.67	22.50	21.83	23.87
	Pixtral-12B	22.13	26.50	27.50	30.77	26.50	30.83	25.83	30.83	30.83	33.33	30.33
	Idefics3-8B	31.97	27.00	28.00	14.10	26.67	-/-	-/-	-/-	-/-	-/-	-/-
	InternVL2.5-8B	18.85	20.00	28.00	19.23	22.33	18.33	25.00	28.33	30.00	28.33	26.00
	MiniCPM-V2.6-8B	24.59	25.50	17.00	17.95	21.50	25.00	25.83	28.57	27.50	26.67	26.67
	LLaVA-1.5-13B	27.78	28.50	26.00	28.21	27.50	20.00	30.83	29.17	23.33	28.33	26.33
	DeepSeek-VL2-4.5B	24.59	26.50	25.00	28.21	25.83	16.67	20.00	20.83	24.17	25.00	21.33
	Human	85.25	80.50	67.50	60.30	72.64	92.50	90.83	83.33	83.33	82.50	86.50

Table 1: Evaluation on Single-Step and Multi-Step Sequential Reasoning. Light gray indicates the best performance for each task among all models in the test sets, respectively.

Humans achieve the highest overall accuracy on this task (87.67%), while closed-source MLLMs generally outperform open-source ones. The best-performing model, o4-mini, reaches 37.83%, followed by Gemini-2.5-Flash and Sonnet-3.7-Thinking, both exceeding 30%, indicating relatively stronger inductive capabilities, as shown in Table 2.

Error analysis of the top performing MLLMs indicates that a majority of their mistakes originate from misidentifying the direction and location of the symmetry axis during the first fold. Humans rarely make errors at this early stage. In terms of difficulty, humans find 2-Fold tasks significantly more challenging than 1-Fold ones, whereas MLLMs exhibit relatively consistent performance across both. However, in the Others subtask involving corner folds, human performance remains stable, while most MLLMs show a significant drop. This result aligns with the trends observed in the Single-Step Reasoning task and indicates that current MLLMs still face challenges in understanding moderately complex geometric symmetries and performing inductive reasoning based on spatial relationships.

3D Folding Reasoning

As shown in Table 2, o4-mini continues to outperform other MLLMs but still lags behind human performance by more than 50%. Viewpoint variation has little effect on most MLLMs, except for Qwen2-VL-72B and Qwen2.5-VL-72B. In contrast, humans are more sensitive to viewpoint changes. The hard group, which includes viewpoint variation, shows approximately a 5% drop in accuracy compared to the easy group. This suggests that 3D spatial memory plays a significant role in solving these problems.

Overall, aside from o4-mini, most MLLMs perform close to random guessing, indicating a lack of effective 3D spatial reasoning capabilities. Error analysis of o4-mini shows that it correctly identifies basic geometric shapes such as circles, triangles, squares, hexagons, and crosses. However, it struggles to reason about 3D relationships between faces, particularly in distinguishing adjacent faces from opposite ones. This is a task that humans tend to solve with ease.

2D Unfolding Reasoning

From Table 2, we observe that the performance gap between humans and MLLMs is the largest among all task types. Both open-source and closed-source MLLMs achieve accuracies below 30%. In contrast, human performance improves on the 2D unfolding task compared to 3D folding. This may be due to the fact that 2D nets lie on a flat plane, making them easier to memorize and match. Additionally, humans find it easier to apply elimination strategies based on 2D structures. Consequently, human performance does not significantly decline in the hard group with 2D net shape variations compared to the easy group without such variations.

The comparison between the 2D unfolding and 3D folding tasks highlights the current limitations of state-of-the-art MLLMs in spatial reasoning on the 3D AVR benchmark. To improve performance, future models will need enhanced mechanisms for memorizing spatial positions and reasoning about relationships between adjacent and opposite faces. Moreover, incorporating flexible elimination strategies similar to those employed by humans based on key visual features could further boost reasoning accuracy.

Key Research Findings

We include evaluation results for our tasks and summarize key findings as follows. Several models failed to produce valid or task-relevant responses on certain tasks and

	Models	Inverse Reasoning				3D Folding			2D Unfolding		
		1-Fold	2-Fold	Others	Overall	Easy	Hard	Overall	Easy	Hard	Overall
Closed-Source	Doubao1.6-Flash	27.00	28.00	15.50	23.50	28.33	25.00	26.67	25.00	22.00	23.50
	GPT-4o-mini	28.00	30.50	25.50	28.00	24.00	26.67	25.33	26.33	23.00	24.67
	GPT-4o	29.50	29.50	21.00	26.67	29.67	23.33	26.50	29.00	19.33	24.17
	o4-mini	44.50	34.00	35.00	37.83	30.67	30.67	30.33	29.00	29.33	29.17
	Gemini-2.5-Flash	29.00	42.00	24.50	31.83	30.67	27.00	28.83	23.33	25.33	24.33
	Sonnet-3.7-Thinking	40.00	35.50	21.00	32.17	25.33	28.67	27.00	22.00	27.33	24.67
Open-Source	GLM-4.1V-10B	26.50	28.50	21.00	25.33	26.33	26.33	26.33	23.67	22.67	23.17
	Qwen2.5-VL-7B	23.00	26.50	17.50	22.33	24.00	28.67	26.33	23.67	22.67	23.17
	Qwen2.5-VL-32B	24.00	31.50	24.00	26.50	27.67	25.00	26.33	31.67	23.67	25.83
	Qwen2.5-VL-72B	23.50	27.00	32.50	25.83	31.33	20.00	25.67	24.67	26.00	25.33
	Qwen2-VL-72B	24.50	23.50	21.00	23.00	23.67	18.67	21.17	23.33	21.00	22.17
	Idefics3-8B	27.00	28.50	21.00	25.50	28.00	23.34	25.67	27.67	27.33	27.33
	InternVL2.5-8B	15.50	17.00	21.00	17.83	29.34	24.67	27.00	31.67	21.00	26.33
	MiniCPM-V2.6-8B	24.00	22.00	20.05	22.02	26.00	26.34	26.17	24.34	23.67	24.00
	LLaVA-1.5-13B	26.50	21.00	32.50	26.67	23.33	28.67	25.67	22.34	24.67	23.50
	DeepSeek-VL2-4.5B	16.00	16.50	21.50	18.00	24.67	27.67	26.17	24.00	20.00	22.00
	Human	92.00	79.00	92.00	87.67	85.00	80.33	82.67	84.00	83.33	83.67

Table 2: Evaluation on Inverse, 3D Folding and 2D Unfolding Reasoning. Light gray indicates the best performance for each task among all models in the test sets, respectively.

were therefore excluded from the corresponding performance comparisons.

1) MLLMs significantly lag human performance in spatial reasoning. Experimental results reveal a clear and substantial performance gap between humans and MLLMs, with human accuracy surpassing that of the best MLLM by at least 30% and exceeding it by more than 50% in some tasks. While both groups exhibit declining accuracy with increased folding complexity and symmetry variation, humans consistently outperform MLLMs, especially in recognizing symmetry and interpreting intricate folding patterns. Humans rely on holistic and intuitive pattern recognition, whereas MLLMs tend to follow rigid, component-wise analysis strategies that struggle with subtle visual distinctions and processes of elimination. For instance, in the Type 1 Single-Step Reasoning task, humans perform relatively worse than in the other four tasks, likely because detecting fine-grained differences in size, position, and shape after one to three folds is more difficult than identifying global features or applying elimination strategies. These findings underscore fundamental differences in spatial reasoning: MLLMs lack the holistic comparison skills and elimination-based strategies that are essential to human cognition.

2) Closed-source models outperform open-source models. Both open-source and closed-source MLLMs exhibit limited performance across all reasoning tasks, as shown in Table 1 and Table 2, with most models performing only slightly better than random guessing. Even the leading closed-source model, o4-mini, achieves only modest accuracies of 31.96%, 41.00%, 37.83%, 30.33%, and 29.17% across the five task types. Overall, closed-source models tend to outperform open-source models, particularly on some 2D tasks, while the performance gap is less pronounced in the two 3D tasks. Among open-source models, the Qwen series demonstrates relatively consistent results, with larger variants generally outperforming smaller ones. Nonetheless, open-source MLLMs continue to struggle with core challenges such as task comprehension, step-wise reasoning, and fine-grained visual recognition. Common failure cases include misinterpreting folding operations, providing vague or imprecise descriptions, and failing to comply with output format constraints, revealing a significant gap in deep task understanding and structured response generation compared to their closed-source counterparts.

Conclusion and Future Work

We present Paper Folding Puzzles, a large-scale benchmark designed to evaluate five types of 2D to 3D spatial reasoning abilities in MLLMs through a structured set of paper folding tasks. Unlike prior AVR benchmarks focused on abstract 2D patterns, our dataset combines 2D symmetry and 3D face-relation reasoning within a real-world-inspired paper folding domain. Experimental results across 18 representative MLLMs and human participants reveal a substantial performance gap, highlighting key limitations in current models' geometric understanding and spatial reasoning.

However, the benchmark remains an early-stage effort. Future work can expand the dataset in both scale and diversity by incorporating more complex fold operations, additional geometric patterns, and human–computer interaction settings. In parallel, evaluation protocols can be further enriched, for instance by examining MLLMs' performance under varied prompting strategies, increasing reasoning complexity with multi-problem sequences, or introducing dynamic task composition.

Acknowledgments

This work was partially supported by the Joint Funds of the National Natural Science Foundation of China under Grant No. U24A20249, and the Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ23F020002.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel Object Captioning at Scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8947–8956.

Agrawal, P.; Antoniak, S.; Hanna, E. B.; Bout, B.; Chaplot, D.; Chudnovsky, J.; Costa, D.; Monicault, B. D.; and et al. 2024. Pixtral 12B. arXiv:2410.07073.

Ahrabian, K.; Sourati, Z.; Sun, K.; Zhang, J.; Jiang, Y.; Morstatter, F.; and Pujara, J. 2024. The Curious Case of Nonverbal Abstract Reasoning with Multi-Modal Large Language Models. In *First Conference on Language Modeling*. Philadelphia, Pennsylvania, USA: colmweb.org.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; and et al. 2022. Flamingo: A Visual Language Model for Few-shot Learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 23716–23736. Red Hook, NY, USA: Curran Associates Inc.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2025-07-25.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2425–2433. Santiago, Chile: IEEE Computer Society.

Bai, S.; Chen, K.; et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bakhtin, A.; van der Maaten, L.; Johnson, J.; Gustafson, L.; and Girshick, R. 2019. PHYRE: A New Benchmark for Physical Reasoning. arXiv:1908.05656.

Barrett, D. G. T.; Hill, F.; Santoro, A.; Morcos, A. S.; and Lillicrap, T. 2018. Measuring Abstract Reasoning in Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 511–520. Stockholm, Sweden: PMLR.

Chen, Z.; Wang, W.; et al. 2025. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv:2412.05271.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/. Accessed: 2025-07-27.

Chollet, F. 2019. On the Measure of Intelligence. arXiv:1911.01547.

Claude Team. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-02-25.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-language Models with Instruction Tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 49250–49267. Red Hook, NY, USA: Curran Associates Inc.

Doubao Team. 2025. Doubao Large Language Model 1.6-flash. https://www.volcengine.com/product/doubao/. Accessed: 2025-07-23.

Emu3 Team. 2024. Emu3: Next-Token Prediction is All You Need. arXiv:2409.18869.

Fleuret, F.; Li, T.; Dubout, C.; Wampler, E. K.; Yantis, S.; and Geman, D. 2011. Comparing Machines and Humans on A Visual Categorization Test. *Proceedings of the National Academy of Sciences*, 108(43): 17621–17625.

Gemini Team. 2025. We're Expanding Our Gemini 2.5 Family of Models. https://blog.google/products/gemini/gemini-2-5-model-family-expands/. Accessed: 2025-06-17.

GLM-V Team. 2025. GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. arXiv:2507.01006.

Google, G. T. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6325–6334.

Hill, F.; Santoro, A.; Barrett, D. G. T.; Morcos, A. S.; and Lillicrap, T. 2019. Learning to Make Analogies by Contrasting Abstract Relational Structure. arXiv:1902.00120.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; and et al. 2023. Language Is Not All You Need: Aligning Perception with Language Models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 72096–72109. Red Hook, NY, USA: Curran Associates Inc.

Jiang, Y.; Zhang, J.; Sun, K.; Sourati, Z.; Ahrabian, K.; Ma, K.; Ilievski, F.; and Pujara, J. 2024. MARVEL: Multidimensional Abstraction and Reasoning through Visual Evaluation and Learning. In *Proceedings of the 38th Advances in Neural Information Processing Systems (NeurIPS)*, 46567–46592. Vancouver, Canada: Curran Associates, Inc.

Kazemi, M.; Alvari, H.; Anand, A.; Wu, J.; Chen, X.; and Soricut, R. 2023. GeomVerse: A Systematic Evaluation of Large Models for Geometric Reasoning. arXiv:2312.12241.

Laurençon, H.; Marafioti, A.; Sanh, V.; and Tronchon, L. 2024. Building and Better Understanding Vision-language Models: Insights and Future Directions. arXiv:2408.12637.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 12888–12900. PMLR.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 34892–34916. Red Hook, NY, USA: Curran Associates Inc.

LLMStats. 2025. LLM Leaderboard. https://llm-stats.com/. Accessed: 2025-07-27.

LMArena. 2025. Leaderboard Overview. https://lmarena.ai/leaderboard. Accessed: 2025-07-27.

Manmadhan, S.; and Kovoor, B. C. 2020. Visual Question Answering: A State-of-the-art Review. *Artificial Intelligence Review*, 53(8): 5705–5745.

Mayer, J.; Ballout, M.; Jassim, S.; Nezami, F. N.; and Bruni, E. 2025. iVISPAR – An Interactive Visual-Spatial Reasoning Benchmark for VLMs. arXiv:2502.03214.

Małkiński, M.; and Mańdziuk, J. 2023. A Review of Emerging Research Directions in Abstract Visual Reasoning. *Information Fusion*, 91: 713–736.

Małkiński, M.; and Mańdziuk, J. 2025. Deep Learning Methods for Abstract Visual Reasoning: A Survey on Raven's Progressive Matrices. *ACM Computing Surveys*, 57(7): 1–36.

Mańdziuk, J.; and Żychowski, A. 2019. DeepIQ: A Human-Inspired AI System for Solving IQ Test Problems. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. Budapest, Hungary: IEEE Computer Society.

Nie, W.; Yu, Z.; Mao, L.; Patel, A. B.; Zhu, Y.; and Anandkumar, A. 2020. Bongard-LOGO: A New Benchmark for Human-Level Concept Learning and Reasoning. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 16468–16480. Vancouver, Canada: Curran Associates, Inc.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI. 2025. Introducing OpenAI o3 and o4-mini. https://openai.com/zh-Hans-CN/index/introducing-o3-and-o4-mini/. Accessed: 2025-04-16.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 8748–8763. PMLR.

Riochet, R.; Castro, M. Y.; Bernard, M.; Lerer, A.; Fergus, R.; Izard, V.; and Dupoux, E. 2022. IntPhys 2019: A Benchmark for Visual Intuitive Physics Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5016–5025.

Rows, S. T., ed. 1917. *Geometric Exercises in Paper Folding*. Chicago, London.: The Open Court Publishing Company.

Tang, K.; Gao, J.; Zeng, Y.; Duan, H.; Sun, Y.; Xing, Z.; Liu, W.; Lyu, K.; and Chen, K. 2025. LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning? arXiv:2503.19990.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; and et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wang, P.; Bai, S.; et al. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

Wang, P.; Li, Z.-Z.; Yin, F.; Ran, D.; and Liu, C.-L. 2025. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 19541–19551. Nashville TN: IEEE Computer Society.

Webb, T. W.; Dulberg, Z.; Frankland, S. M.; Petrov, A. A.; O'Reilly, R. C.; and Cohen, J. D. 2020. Learning Representations that Support Extrapolation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 10136–10146. PMLR.

Wu, Z.; Chen, X.; et al. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv:2412.10302.

Xie, N.; Lai, F.; Doran, D.; and Kadav, A. 2019. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. arXiv:1901.06706.

Yao, Y.; Yu, T.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. arXiv:2408.01800.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6713–6724. Long Beach, CA, USA: IEEE Computer Society.

Zhang, C.; Gao, F.; Jia, B.; Zhu, Y.; and Zhu, S.-C. 2019. RAVEN: A Dataset for Relational and Analogical Visual rEasoNing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2425–2433. Long Beach, CA, USA: IEEE Computer Society.

Zhao, H.; Cai, Z.; Si, S.; Ma, X.; An, K.; Chen, L.; Liu, Z.; Wang, S.; Han, W.; and Chang, B. 2024. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. arXiv:2309.07915.