Jigsaw-Puzzles: From Seeing to Understanding to Reasoning in Vision-Language Models

Zesen Lyu[1,2], Dandan Zhang[2], Wei Ye[1], Fangdi Li[2,3], Zhihang Jiang[1,2], Yao Yang[2]*
[1]Hangzhou Institute for Advanced Study, UCAS
[2]Zhejiang Lab, [3]Zhejiang University
{lvzesen23, yewei23, jiangzhihang23}@mails.ucas.ac.cn,
{danae.zdd, yangyao}@zhejianglab.com, FdbendyLi@zju.edu.cn

Abstract

Spatial reasoning is a core component of human cognition, enabling individuals to perceive, comprehend, and interact with the physical world. It relies on a nuanced understanding of spatial structures and inter-object relationships, serving as the foundation for complex reasoning and decision-making. To investigate whether current vision-language models (VLMs) exhibit similar capability, we introduce Jigsaw-Puzzles, a novel benchmark consisting of 1,100 carefully curated real-world images with high spatial complexity. Based on this dataset, we design five tasks to rigorously evaluate VLMs' spatial perception, structural understanding, and reasoning capabilities, while deliberately minimizing reliance on domain-specific knowledge to better isolate and assess the general spatial reasoning capability. We conduct a comprehensive evaluation across 24 state-of-the-art VLMs. The results show that even the strongest model, Gemini-2.5-Pro, achieves only 77.14% overall accuracy and performs particularly poorly on the *Order Generation* task, with only 30.00% accuracy, far below the 90%+ performance achieved by human participants. This persistent gap underscores the need for continued progress, positioning Jigsaw-Puzzles as a challenging and diagnostic benchmark for advancing spatial reasoning research in VLMs. Our project page is at https://zesen01.github.io/jigsaw-puzzles.

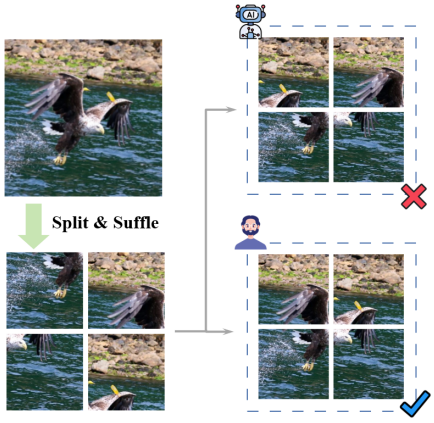

Figure 1: Jigsaw-Puzzles example. While human participants effortlessly reconstruct the original spatial layout, all tested VLMs fail to recover the correct order.

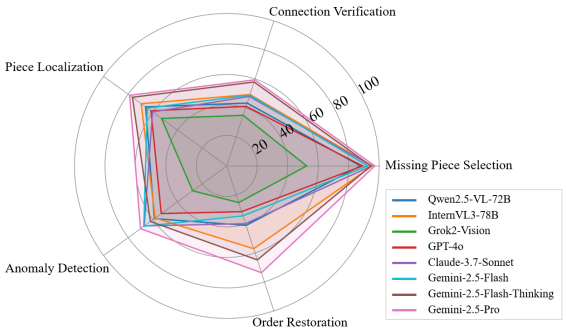

Figure 2: Evaluation of VLMs on Jigsaw-Puzzles. The plot reports the accuracy of 8 representative VLMs on 5 tasks.

1 Introduction

The road to artificial general intelligence (AGI) demands more than language or vision alone: it requires models to possess a human-like spatial reasoning capability by constructing structured representations of the physical world (Lake et al., 2017). Spatial reasoning refers not just to the perception of visual input, but to the capability to comprehend spatial arrangements, model structural relations, and infer the geometry and layout of a scene. These capabilities are fundamental to human cognition and develop naturally through everyday perception and interaction (Ishikawa and Newcombe, 2021). In contrast, current VLMs, while highly capable in tasks such as image captioning (Young et al., 2014; Lin et al., 2014; Sharma et al., 2018), visual question answering (Krishna et al., 2017; Singh et al., 2019; Marino et al., 2019), and image-text retrieval (Schuhmann et al., 2021; Thapliyal et al., 2022;

*Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25992–26003
November 4-9, 2025 ©2025 Association for Computational Linguistics

Bitton-Guetta et al., 2023), consistently struggle with tasks requiring spatial reasoning (Stogiannidis et al., 2025). We show an example in Figure 1 and report the performance of some tested VLMs on Jigsaw-Puzzles in Figure 2. This gap underscores a critical limitation: while current VLMs have made substantial progress in basic visual understanding, they continue to struggle with structured spatial reasoning, which is essential for grounded understanding in real-world scenarios. Bridging this gap is essential for progressing towards generalizable human-like spatial cognition and ultimately AGI.

However, existing benchmarks have yet to provide a comprehensive evaluation of spatial reasoning capability in VLMs under complex, real-world visual scenarios. Some works (Fu et al., 2024; Li et al., 2024; Liu et al., 2024; Yue et al., 2024) focus primarily on foundational visual understanding by systematically evaluating perception, comprehension, and basic visual reasoning, revealing notable limitations in these areas. Although a few recent efforts (Pothiraj et al., 2025; Stogiannidis et al., 2025; Ren et al., 2025; Tang et al., 2025) have attempted to evaluate the spatial reasoning capability of VLMs, they often rely on overly synthetic settings, task-specific constraints, or domain-dependent priors (Song et al., 2025) (See Appendix A for examples), limiting the capability to capture generalizable spatial reasoning under natural visual conditions. A truly effective evaluation of human-like spatial reasoning capability should model the task as a multi-stage cognitive process—beginning with perception, advancing through structural understanding, and culminating in high-level reasoning. Such reasoning must be grounded in the visual richness and ambiguity of real images, requiring the integration of spatial structural modeling and goal-directed reasoning (Chen et al., 2024). Yet, this critical dimension of spatial cognition remains largely overlooked in existing benchmarks, underscoring the need for new benchmarks that move beyond narrow task formulations and embrace the full complexity of spatial reasoning.

To overcome the limitations of existing benchmarks in evaluating spatial reasoning capability of VLMs under real-world conditions, we introduce Jigsaw-Puzzles, a novel benchmark inspired by the cognitive mechanisms underlying human puzzle-solving. Puzzle-solving naturally reflects the multi-stage cognitive process (Fissler et al., 2018), making it a compelling testbed for spatial reasoning in VLMs. Unlike earlier studies that applied puzzle-solving primarily to visual representation learning (Noroozi and Favaro, 2016; Wang et al., 2022; Shen et al., 2024), our benchmark employs it to evaluate the spatial reasoning ability of VLMs under realistic visual conditions.

Benchmark	Understanding	Reasoning	High Visual Complexity	Great Scalability	Automated Construction
Capture (Pothiraj et al., 2025)	✓	✓	×	×	×
Mind the Gap (Stogiannidis et al., 2025)	✓	✓	×	×	×
VGRP (Ren et al., 2025)	✓	✓	×	×	×
LEGO-Puzzles (Tang et al., 2025)	✓	✓	×	✓	✓
Jigsaw-Puzzles (Ours)	✓	✓	✓	✓	✓

Table 1: Comparison of spatial reasoning benchmarks.

In total, Jigsaw-Puzzles comprises 1,100 carefully curated real-world images and features five different tasks. First, we begin with the *Missing Piece Selection* task to evaluate VLMs' basic spatial understanding capability, which serves as the essential foundation for spatial reasoning. Building on this foundation, we introduce four reasoning-centric tasks: *Piece Localization*, *Connection Verification*, *Anomaly Detection*, and *Order Restoration*. These tasks are designed to assess various facets of spatial reasoning, including adjacency modeling, local structural consistency, spatial localization, geometric transformation understanding, and multi-step spatial reasoning.

Compared to existing benchmarks for evaluating spatial reasoning capability in VLMs, Jigsaw-Puzzles offers three key advantages, as summarized in Table 1: (1) Higher visual complexity. Jigsaw-Puzzles uses real-world images with diverse and rich visual elements, and significantly outperforms benchmarks based on synthetic images (Ren et al., 2025; Stogiannidis et al., 2025; Tang et al., 2025) and simple visual scenes (Pothiraj et al., 2025). This enables more realistic and challenging spatial reasoning evaluation. (2) Greater scalability. Any natural image that satisfies the construction rules can be directly used to generate puzzle tasks, without the need to manually synthesize target images. (3) Fully automated construction pipeline. All Jigsaw-Puzzles tasks are generated automatically without manual annotation, with each question paired with a unique deterministic answer. This feature enables low-cost dataset construction and facilitates continuous expansion and refinement.

In summary, we introduce Jigsaw-Puzzles, a novel benchmark for systematically evaluating the human-like spatial reasoning capability of VLMs in realistic visual settings. Our main contributions are as follows:

A new benchmark for spatial reasoning. We

introduce Jigsaw-Puzzles, a puzzle-inspired benchmark constructed through a fully automated pipeline that improves existing benchmarks in visual complexity and scalability, while enabling structured evaluation of spatial reasoning in VLMs.

Comprehensive evaluation and analysis. We evaluate 24 state-of-the-art VLMs on Jigsaw-Puzzles and conduct detailed analysis. Our findings expose consistent limitations in current VLMs and provide insights to guide future improvements in spatial reasoning capability.

Open-sourced dataset and construction tools. We release the full dataset along with the automated generation scripts to support the evaluation and continued advancement of spatial reasoning in VLMs under real-world visual scenarios, as well as to facilitate future benchmark expansion.

2 Related Work

General VLMs Evaluation Benchmarks. With the rapid progress of VLMs, systematically evaluating their diverse capabilities has become a key challenge. Although many benchmarks have been introduced, most focus primarily on visual understanding. MME (Fu et al., 2024) evaluates instruction following, perception, and basic reasoning across 14 subtasks, revealing persistent issues like object hallucination and limited spatial understanding. SEED-Bench (Li et al., 2024) includes 19,000 multiple-choice questions across 12 dimensions and shows continued struggles with text recognition and temporal reasoning. MMBench (Liu et al., 2024) offers fine-grained bilingual evaluations, enhancing the robustness of multilingual assessment. MMMU (Yue et al., 2024) provides 11,500 questions across 183 subfields and 30 image types to test expert-level reasoning, yet even advanced models like Gemini display notable knowledge gaps. While these benchmarks have advanced the evaluation of perceptual and semantic understanding, none systematically assess spatial reasoning—the core aspect of human cognition. This highlights the pressing need for more challenging and diagnostic benchmarks specifically targeting spatial reasoning capability in VLMs.

Spatial Reasoning Evaluation Benchmarks in VLMs. Several recent benchmarks have aimed to evaluate the spatial reasoning capability of VLMs. Capture (Pothiraj et al., 2025) assesses occluded object counting, revealing that VLMs struggle to form coherent spatial representations under occlusion. Mind the Gap (Stogiannidis et al., 2025) evaluates spatial relations, navigation, and mental rotation, showing that VLMs often perform near chance level, indicating limited spatial cognition. VGRP (Ren et al., 2025) introduces 20 visual grid puzzles across varying difficulty levels to assess visual perception, rule-following, and logical reasoning. LEGO-Puzzles (Tang et al., 2025) provides 1,100 visual QA pairs over 11 subtasks to measure basic and multi-step spatial reasoning. Results consistently show that current VLMs struggle with perceptual complexity, rotation reasoning, and sequential reasoning. Despite these efforts, most benchmarks rely on simplified scenarios, failing to reflect the complexity of real-world spatial environments, thereby limiting their generalizability. More diagnostic benchmarks grounded in natural visual settings are needed to advance human-level spatial reasoning in VLMs.

3 Jigsaw-Puzzles

In this section, we introduce Jigsaw-Puzzles, a scalable and comprehensive benchmark designed to evaluate the spatial reasoning capability of VLMs in realistic visual environments. Specifically, Section 3.1 outlines the motivation and definition of each task, while Section 3.2 describes the dataset construction process, including image selection and the automated generation of question–answer pairs.

3.1 Task Definition

To systematically evaluate the spatial reasoning capability of VLMs, we design tasks around the core cognitive stages underlying human spatial reasoning—seeing, understanding, and reasoning. Inspired by the human process of solving jigsaw puzzles, our benchmark simulates how individuals integrate fragmented visual information into a coherent whole: beginning with the perception of local visual cues, followed by the comprehension of spatial relationships, and culminating in multi-step spatial reasoning to reconstruct the original scene. This sequence naturally reflects the progression from low-level perception to high-level spatial reasoning. Accordingly, the tasks span spatial understanding, single-step, and multi-step spatial reasoning, collectively providing a comprehensive evaluation across different levels of spatial reasoning. Figure 3 shows examples of each task in Jigsaw-Puzzles.

Task 1: Missing Piece Selection. The task evaluates VLMs' spatial understanding capability. Given

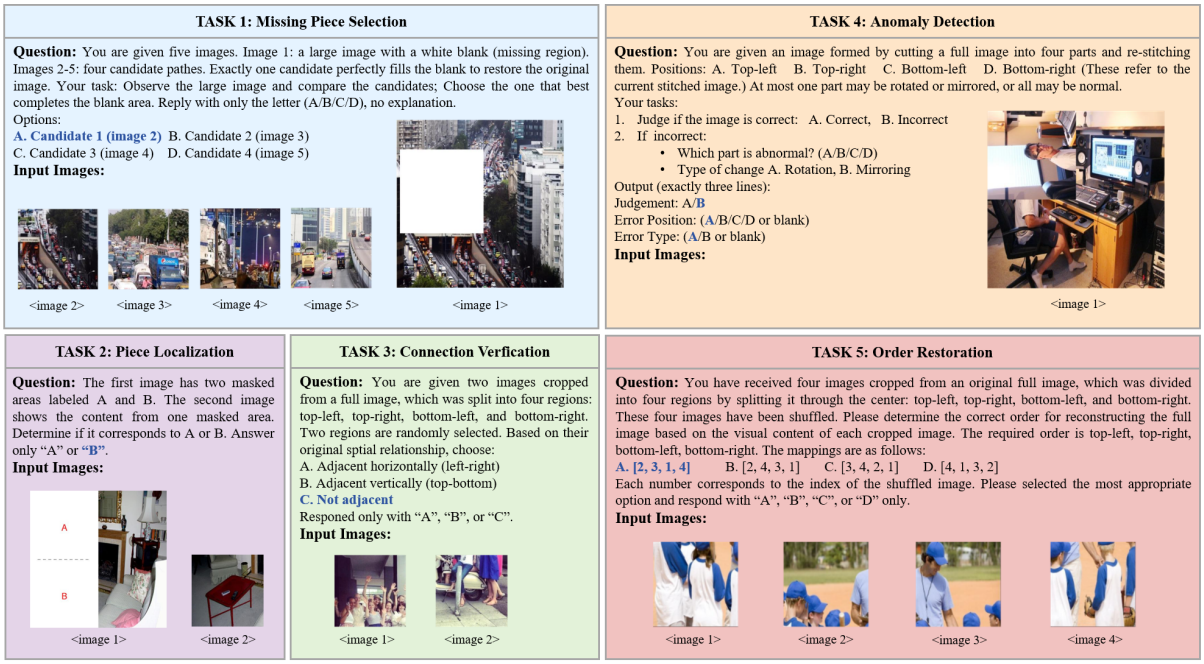

Figure 3: Task examples of Jigsaw-Puzzles. Note: the questions above are slightly simplified for clarity and brevity, and the **blue** option indicates the correct answer.

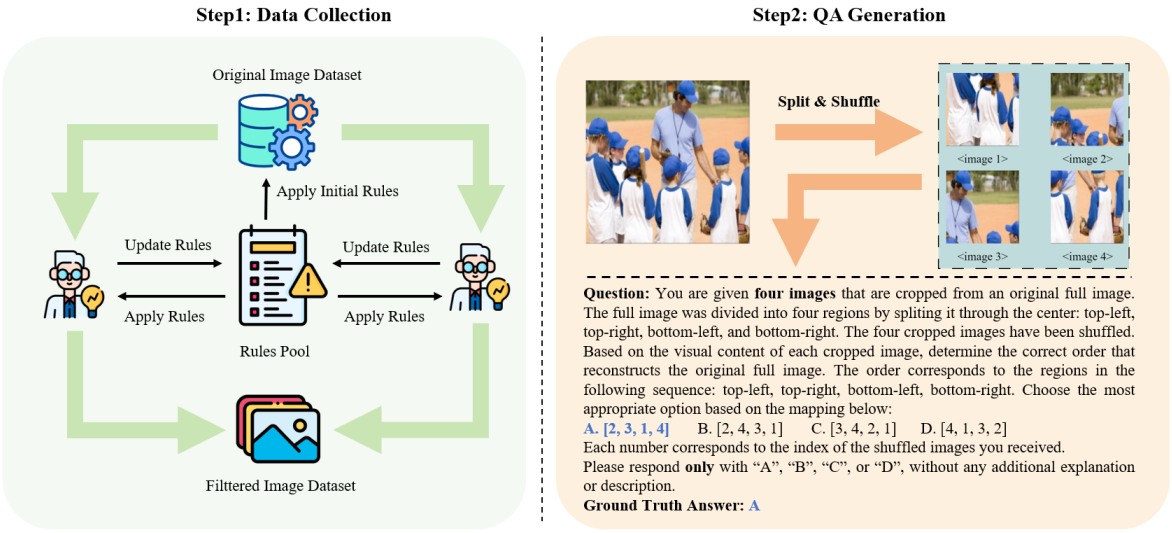

Figure 4: Dataset curation pipeline. Step 1 filters candidate images through expert-defined rules to build a spatial reasoning dataset. Step 2 uses automated templates to generate task-specific QA pairs from the curated images.

an image with a missing region, VLMs need to select the correct patch from four candidates. We define two difficulty levels: Easy, where distractors are randomly chosen, and Hard, where distractors are selected using CLIP (Radford et al., 2021) similarity to closely resemble the ground-truth patch, increasing the task's difficulty. *Task 2: Piece Localization*. This task evaluates spatial localization as a representative single-step spatial reasoning capability. Given a partially masked image and one masked patch, VLMs must identify the patch's original position. Difficulty is controlled by grid size and number of masked regions: Easy (2×2 with two masks), Hard (3×3 with four masks), increasing spatial complexity. *Task 3: Connection Verification*. This task evaluates adjacency reasoning, which also falls under single-step spatial reasoning. The full image is divided into 2×2 grids, and two patches are randomly selected. VLMs are asked to determine their spatial relationship in

the original image (e.g., above-below, left-right, or non-adjacent). *Task 4: Anomaly Detection*. This task targets local spatial transformation detection, a process that inherently involves single-step spatial reasoning. One region in 2×2 grids is randomly rotated, mirrored, or left unchanged. The model must detect the change, locate the region, and identify the transformation. *Task 5: Order Restoration*. This task integrates the capabilities assessed in the previous tasks and serves as a complex multi-step spatial reasoning challenge. A complete image is split into four shuffled patches. VLMs must identify the correct order to reconstruct the original spatial layout.

Overall, the five puzzle-inspired tasks in Jigsaw-Puzzles cover a broad spectrum of spatial reasoning challenges—from basic spatial understanding to single-step and multi-step spatial reasoning—enabling a comprehensive evaluation of spatial reasoning in VLMs.

3.2 Dataset Curation

As illustrated in Figure 4, our dataset curation pipeline consists of two main stages: data collection and QA generation.

Data Collection. We integrate data collection and quality control into a unified process. Starting from the CC3M (Sharma et al., 2018) dataset, we apply task-specific filtering criteria—including minimum resolution and aspect ratio constraints—to construct an initial image pool of approximately 10,000 candidate images. Two human experts iteratively review the image pool while incrementally refining a shared set of filtering rules. Based on these evolving rules, they collaboratively filter the initial dataset to obtain the final set of high-quality, structurally diverse images. See Appendix B for the rules pool. To enhance generalizability, we emphasize semantic and structural diversity throughout the dataset.

QA Generation. To support scalable and consistent QA pairs generation, each task type is associated with a specific construction template. QA pairs are automatically generated using these templates. Figure 3 illustrates simplified examples of the templates, full versions are provided in Appendix B.

4 Evaluation on Jigsaw-Puzzles

4.1 Experimental Setting

Benchmark Models. We evaluate 24 VLMs on Jigsaw-Puzzles, covering a diverse range of model scales and training paradigms. For open-source models, we evaluate Qwen2-VL-72B (Wang et al., 2024a), QvQ-72B-Preview (Qwen, 2024), Qwen2.5-VL-[7B/32B/72B] (Bai et al., 2025), InternVL3-[8B/14B/38B/78B] (Zhu et al., 2025), Kimi-VL-A3B-[Instruct/Thinking] (Du et al., 2025), Phi-4-multimodal-instruct (Abouelenin et al., 2025), Aya-Vision-[8B/32B] (Dash et al., 2025), and Mistral-Small-3.1-24B-Instruct (Mistral, 2025). For proprietary models, we evaluate Claude-[3.5/3.7]-Sonnet (Anthropic, 2024), Gemini-[2.0/2.5]-Flash, Gemini-2.5-Flash-Thinking, Gemini-2.5-Pro (Anil et al., 2023), GPT-4o, GPT-4o-mini (Achiam et al., 2023), and Grok-2-Vision (Grok, 2024). Notably, QvQ-72B-Preview, Kimi-VL-A3B-Thinking, Gemini-2.5-Flash-Thinking, and Gemini-2.5-Pro are categorized as reasoning-enhanced models. All models, supporting multi-image input, are evaluated in a zero-shot setting with hardware scaled to their parameter size, see details in Appendix C.

Evaluation Metric. Since each QA pair in Jigsaw-Puzzles has a single correct answer, we use exact match accuracy (%) as the primary metric to evaluate VLMs' performance on each task.

Baselines. We provide two baselines for comparison: (1) Random, which assumes equal probability across all options and calculates expected accuracy accordingly. (2) p-value-based critical value, which reports the minimum accuracy required to outperform random guessing at a significance level of p=0.05.

Human Performance. To evaluate human performance, we construct a subset called Jigsaw-Puzzles-Lite by sampling 220 images from the full dataset. Three human participants complete all tasks on this subset under the same conditions as VLMs—without access to any external tools or the internet. Their performance serves as an empirical upper bound for spatial reasoning capability.

4.2 Main Results

Table 2, 3 report the performance of 24 VLMs on Jigsaw-Puzzles. Building on these results, we conduct a comprehensive and systematic analysis. We summarize several key findings as below.

Spatial Reasoning Remains a Challenge for

Models	Missing Piece Selection		Piece Localization		Connection	Anomaly	Order	Overall
	Easy	Hard	Easy	Hard	Verification	Detection	Restoration	
Baseline								
Random Guessing	25.00	25.00	50.00	25.00	33.33	28.13	25.00	30.21
↑ Random ($p < 0.05$)	27.30	27.30	52.50	27.30	35.70	30.50	27.30	32.56
Proprietary Models								
Grok2-Vision	64.55	52.45	53.00	41.00	34.91	27.73	25.27	42.70
GPT-4o-mini	96.45	83.64	59.45	37.82	44.36	57.91	33.18	58.97
GPT-4o	95.00	89.18	61.55	53.45	41.09	53.18	31.55	60.71
Claude-3.5-Sonnet	99.73	94.55	62.45	41.09	45.64	67.45	35.00	63.70
Claude-3.7-Sonnet	99.55	95.09	60.27	44.55	47.91	67.00	39.82	64.88
Gemini-2.0-Flash	92.09	85.64	63.55	54.00	44.91	68.73	34.27	63.31
Gemini-2.5-Flash	98.82	92.45	64.55	54.55	48.82	67.36	34.45	65.86
Gemini-2.5-Flash-Thinking	**99.55**	**94.73**	**76.64**	**51.27**	**57.91**	**62.00**	**64.82**	**72.42**
Gemini-2.5-Pro	**99.91**	**97.18**	**78.82**	**61.09**	**59.36**	**70.00**	**73.64**	**77.14**
Open-source Models								
Kimi-VL-A3B-Instruct	67.91	52.55	51.45	29.82	37.91	21.82	32.82	42.04
Kimi-VL-A3B-Thinking	**84.64**	**58.09**	**56.36**	**32.64**	**28.00**	**30.91**	**20.36**	**44.43**
Phi-4-multimodal-instruct	63.45	51.64	60.91	37.45	36.64	43.36	27.64	45.87
Qwen2.5-VL-7B	87.18	63.27	54.27	36.18	38.55	45.00	28.09	50.36
Aya-Vision-8B	26.82	27.27	49.64	26.55	35.00	12.91	24.73	28.99
InternVL3-8B	98.09	85.45	53.91	35.45	44.91	46.82	34.36	57.00
InternVL3-14B	99.73	88.73	59.64	40.18	51.73	49.09	40.91	61.43
Mistral-Small-3.1-24B-Instruct	26.91	28.55	54.27	31.27	38.27	52.36	26.45	36.87
Qwen2.5-VL-32B	97.27	76.82	62.09	40.36	50.09	61.55	33.64	60.26
Aya-Vision-32B	24.27	25.27	51.45	28.36	37.18	41.45	25.00	33.28
InternVL3-38B	99.00	91.36	63.45	42.64	56.73	30.73	54.64	62.65
Qwen2-VL-72B	95.55	75.36	55.27	40.64	40.55	42.36	33.55	54.75
QVQ-72B-Preview	**77.82**	**52.18**	**53.09**	**36.73**	**41.82**	**47.73**	**34.64**	**49.14**
Qwen2.5-VL-72B	99.36	87.82	65.91	45.27	43.36	58.82	41.00	63.08
InternVL3-78B	99.73	95.55	69.45	52.27	49.27	58.18	57.09	68.79

Table 2: Full Evaluation Results of 24 VLMs on Jigsaw-Puzzles. VLMs are grouped into proprietary and open-source categories. Dark Green and Light Green indicate the top-1 and top-2 performance within each group, respectively. Results of reasoning-enhanced are marked in **bold**. We also highlight the top three models based on their overall performance, using Dark Blue, Medium Blue, and Light Blue, respectively.

Models	Missing Piece Selection		Piece Localization		Connection	Anomaly	Order	Overall
	Easy	Hard	Easy	Hard	Verification	Detection	Restoration	
Human Performance	99.55	100.00	95.45	91.36	93.18	97.27	97.73	96.36
Proprietary Models								
Claude-3.7-Sonnet	100.00	95.45	55.45	47.27	42.73	68.18	38.64	63.96
Gemini-2.5-Flash	98.18	93.18	58.18	55.45	42.27	66.82	32.27	63.76
Gemini-2.5-Flash-Thinking	**99.55**	**95.91**	**71.82**	**51.82**	**55.00**	**60.91**	**57.27**	**70.33**
Gemini-2.5-Pro	**100.00**	**96.36**	**77.73**	**56.82**	**57.27**	**71.36**	**70.91**	**75.78**
Open-source Models								
Qwen2.5-VL-72B	99.09	86.82	67.73	42.27	40.00	57.73	33.64	61.04
InternVL3-78B	99.55	95.45	70.45	53.64	44.55	61.36	56.82	68.83

Table 3: Comparing Top-Performing VLMs with Human Performance on Jigsaw-Puzzles-Lite. The human performance is highlighted in Dark Green. Results of reasoning-enhanced are marked in **bold**. The top three overall performance are highlighted in Dark Blue, Medium Blue, and Light Blue, respectively.

VLMs. As shown in Table 3, human participants consistently outperform VLMs, achieving an overall accuracy of 96.36%. By comparison, current VLMs perform considerably worse, with even the strongest models—Gemini-2.5-Pro—lagging more than 20 percentage points behind human accuracy across all tasks. The persistent gap between humans and VLMs highlights the demanding nature of Jigsaw-Puzzles and affirms its utility as a robust benchmark for spatial reasoning evaluation.

Significant Gap Between Open-Source and Proprietary VLMs. As shown in Tables 2, proprietary VLMs consistently outperform open-source VLMs on Jigsaw-Puzzles. Among them, non-reasoning-enhanced proprietary VLMs typically exceed 60% overall accuracy, whereas most open-source VLMs fall short—only InternVL3-[14B/38B/78B] and Qwen2.5-VL-72B surpass this threshold. Reasoning-enhanced proprietary models, such as Gemini-2.5-Flash-Thinking (72.42%)

and Gemini-2.5-Pro (77.14%), further widen this gap. These results reveal a persistent disparity in spatial reasoning performance, suggesting that proprietary VLMs benefit from advantages in model architecture, training strategy, and access to large-scale data. Meanwhile, this finding highlights substantial room for improvement in open-source VLMs toward achieving more robust and generalizable spatial reasoning.

Model Performance in Different Tasks. In the *Missing Piece Selection* task, which primarily targets spatial understanding, most proprietary VLMs perform well under both Easy and Hard settings, demonstrating strong perceptual capability. Although open source models generally perform poorly by comparison, certain models, such as the InternVL3 series and Qwen2.5-VL-72B, achieve perceptual understanding on par with proprietary VLMs. Notably, both the Aya-Vision series and the Mistral-Small-3.1-24B-Instruct models perform poorly across all settings, even at the 32B scale, accuracy remains near random, revealing severe deficits in spatial understanding and instruction following. In single-step spatial reasoning tasks—*Piece Localization*, *Connection Verification*, and *Anomaly Detection*—most VLMs surpass the p-value-based critical value, indicating emerging competence in basic spatial reasoning. However, strong performance remains concentrated in only a few models, particularly reasoning-enhanced proprietary models and the latest open-source InternVL3 series. This disparity becomes even more evident in the multi-step spatial reasoning task—*Order Restoration*, indicating that most VLMs struggle with complex spatial reasoning.

In conclusion, Jigsaw-Puzzles effectively distinguishes VLMs across a spectrum of spatial reasoning capability—from basic understanding to complex multi-step reasoning. As shown by the results in Table 2, substantial room for improvement remains, particularly in multi-step spatial reasoning.

Foundational Spatial Understanding Shapes Reasoning Performance. We analyze task similarity on Jigsaw-Puzzles by computing Pearson correlation coefficients between each task and all others, as proposed by Zhang et al. (2025), as shown in Figure 5. The results show that performance on the *Missing Piece Selection* task—a proxy for spatial understanding, is strongly correlated with performance on spatial reasoning tasks. In contrast, VLMs with weaker spatial understanding often struggle with reasoning tasks, with some performing worse than random on reasoning-intensive tasks. This pattern reflects the human cognitive progression from perception to understanding to reasoning, underscoring the foundational role of spatial understanding in enabling higher-level spatial reasoning in VLMs.

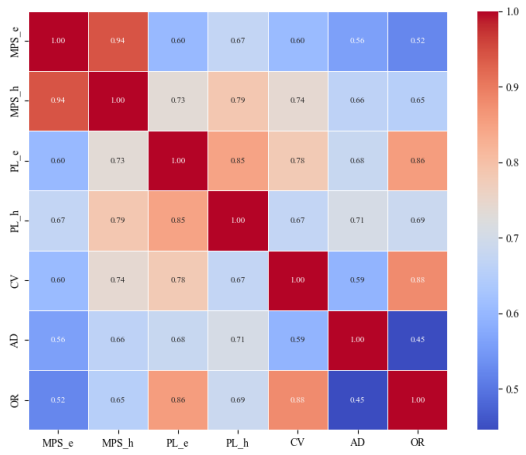

Figure 5: Task Similarity Heatmap. The heatmap illustrates the pairwise correlation between tasks in our benchmark, measured using Pearson correlation coefficients. Task names are abbreviated using the initials of each word (e.g., *Missing Piece Selection* → MPS). The suffixes _e and _h indicate the Easy and Hard settings, respectively.

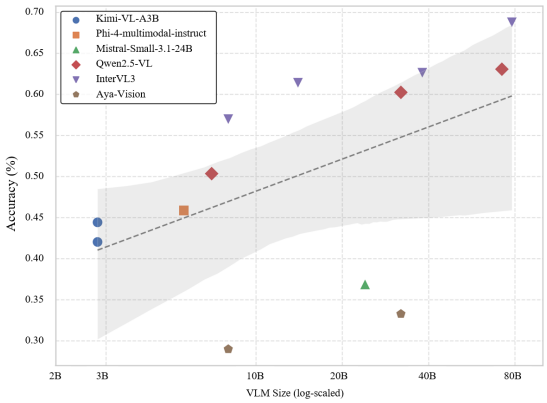

Figure 6: Relationship between VLM size and performance on Jigsaw-Puzzles. Each point represents a VLM, with its accuracy plotted against log-scaled parameter size. A clear positive correlation is observed both across and within model families, indicating that larger models tend to exhibit stronger performance.

Spatial Reasoning Scales with VLM size. We analyze the relationship between VLM size and overall performance on Jigsaw-Puzzles. As shown in Figure 6, our results reveal that VLM accuracy consistently increases with model size, both across all models and within specific fami-

Strong Capacity of Self-Correction

```python
import json
# The user provided four images (implicitly indexed 1 to 4 based on presentation order)
# Image 1: Fist, mic stand right, ceiling
# Image 2: Red shirts midsection, mic stands middle
# Image 3: Leggings lower body, drum kit, mic stand left lower
# Image 4: Singer upper body, mic stand left upper
# Analysis based on visual continuity:
# Connection 1: Microphone stand on the right connects Image 1 (top) and Image 2 (bottom).
# This means: Top-Right = 1, Bottom-Right = 2.
# Connection 2: Microphone stand on the left connects Image 4 (top) and Image 3 (bottom).
# This means: Top-Left = 4, Bottom-Left = 3.
# Connection 3: The singer in Image 4 is raising her right arm, which forms the fist in Image 1.
# This confirms Image 4 is left of Image 1. So Top-Left = 4, Top-Right = 1.
# Connection 4: The positioning of the people in Image 3 (left) and Image 2 (right) seems consistent with the stands. So Bottom-Left = 3, Bottom-Right = 2.
# Derived Order [Top-Left, Top-Right, Bottom-Left, Bottom-Right]:
# [4, 1, 3, 2]   The Initial Incorrect Answer Generated By Gemini-2.5-Pro
# Check the given options:
# A: [3, 4, 2, 1]
# B: [1, 4, 2, 3]   Ground Truth Answer
# C: [3, 1, 2, 4]
# D: [3, 2, 1, 4]
# The derived order [4, 1, 3, 2] is not present in the options.   The Correct Answer Obtained After Self-Correction
# Let's re-evaluate, prioritizing the clearest connections (mic stands).
# Vertical Connection Right: 1 above 2.
# Vertical Connection Left: 4 above 3.
# Now consider the options based on these vertical pairs:
# Option A: [3, 4, 2, 1] - Fails both vertical connections (3 !above 2, 4 !above 1)
# Option B: [1, 4, 2, 3] - TL=1, TR=4, BL=2, BR=3. Vertical TL->BL is 1->2 (Correct). Vertical TR->BR is 4->3 (Correct).
# Option C: [3, 1, 2, 4] - Fails both vertical connections (3 !above 2, 1 !above 4)
# Option D: [3, 2, 1, 4] - Fails both vertical connections (3 !above 1, 2 !above 4)
# Only Option B satisfies the vertical continuity established by the microphone stands.
# Let's check the horizontal continuity for Option B: [1, 4, 2, 3]
# Top: 1 (Fist) left of 4 (Singer). This contradicts the visual of the singer (4) raising her *right* arm (fist should be to her right).
# Bottom: 2 (Red shirts) left of 3 (Leggings). This positioning is plausible.
# Although the arm connection seems inconsistent with Option B, the microphone stand connections strongly support it, and it's the only option satisfying those vertical links. It's possible the arm pose is misleading or interpreted incorrectly, or it belongs to someone else obscured. Given the choices, B is the most structurally sound based on object continuity.
# Final Answer is B.
print(\"B\")
```

B

Figure 7: An example of self-correction. Red shows the initial incorrect answer generated by Gemini-2.5-Pro; Blue indicates the ground-truth answer; Green illustrates the model's self-correction process.

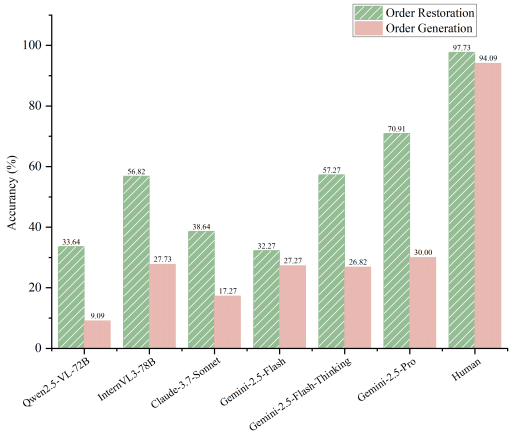

Figure 8: Evaluation of *Order Restoration* and *Order Generation* tasks on Jigsaw-Puzzles-Lite. Without option constraints, VLM accuracy drops significantly—peaking at just 30.00% and falling far short of human performance.

lies (e.g., InternVL3, Qwen2.5-VL). This positive correlation suggests that spatial reasoning capability—like other cognitive competencies (Wang et al., 2024b)—benefits from larger model capacity, which scales with parameter count.

Reasoning-Enhanced Models Show Superior Spatial Reasoning Performance. To assess the spatial reasoning capability of reasoning-enhanced VLMs, we evaluate Gemini-2.5-Flash-Thinking, Gemini-2.5-Pro, Kimi-VL-A3B-Thinking and QvQ-72B-Preview. Except for Gemini-2.5-Pro, each model has a corresponding base version for comparison. As shown in Table 2, these enhanced VLMs consistently achieve higher overall accuracy. For example, Kimi-VL-A3B-Thinking improves from 42.04% to 44.43%, and Gemini-2.5-Flash-Thinking rises from 65.86% to 72.42%. Although QvQ-72B-Preview overall underperforms Qwen2-VL-72B, it achieves better results on spatial reasoning tasks. Notably, Gemini-2.5-Pro achieves the highest overall accuracy (77.14%) among all VLMs tested. Furthermore, the largest improvements occur in the multi-step spatial reasoning task, *Order Restoration*, where reasoning-enhanced VLMs outperform their base counterparts more substantially than in single-step tasks. To explain this, we analyze cases where only Gemini-2.5-Pro answers correctly, with Figure 7 presenting one such example. Gemini-2.5-Pro demonstrates a form of self-correction: when the model's initial prediction is not among the provided options, it will re-evaluate the visual input and revise its judgment. This behavior, facilitated by the reduced answer space under choice constraints, may contribute to the superior performance of reasoning-enhanced models in the *Order Restoration* task.

Further Exploring Multi-Step Spatial Reasoning in VLMs. To further evaluate VLMs' multi-step spatial reasoning beyond the constraints of predefined choices, we introduce the *Order Generation* task based on Jigsaw-Puzzles-Lite. In this setting, VLMs must directly generate the correct sequence of puzzle pieces without relying on answer options, thereby more authentically simulating open-ended spatial reasoning. As shown in Figure 8, current VLMs consistently struggle with this task—Gemini-2.5-Pro, the best-performing model, achieves only 30.00% accuracy, in stark contrast to 94.09% by human participants. This finding reveals that, despite exhibiting strong self-correction behavior under option constraints, existing VLMs face considerable challenges in autonomously constructing coherent spatial reasoning chains. This highlights a significant gap between current VLMs and human-level spatial reasoning in open-ended scenarios.

5 Conclusion

We propose Jigsaw-Puzzles, a novel benchmark for systematically evaluating the spatial reasoning capability of VLMs in real-world visual scenes. Through extensive experiments on 24 representative VLMs, we identify persistent gaps between current VLMs and human-level spatial reasoning—especially in multi-step spatial reasoning tasks. Jigsaw-Puzzles provides a scalable and cognitively grounded benchmark to advance future research on spatial reasoning in VLMs.

Limitations

While Jigsaw-Puzzles provides a structured benchmark tailored for 2D spatial reasoning in static images, it currently does not address 3D perception, temporal sequences, or embodied contexts—each of which represents an important and orthogonal axis of spatial cognition. We view this as a natural next step and encourage future work to extend the benchmark in these directions.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was supported by the Key R&D Program of Zhejiang (2024C01036).

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

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

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, and 1 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*.

Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Patrick Fissler, Olivia Caroline Küster, Daria Laptinskaya, Laura Sophia Loy, Christine AF Von Arnim, and Iris-Tatjana Kolassa. 2018. Jigsaw puzzling taps multiple cognitive abilities and is a potential protective factor for cognitive aging. *Frontiers in aging neuroscience*, 10:408085.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Grok. 2024. Bringing grok to everyone.

Toru Ishikawa and Nora S Newcombe. 2021. Why spatial is special in education, learning, and everyday activities.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seedbench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

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

Mistral. 2025. Mistral small 3.1.

Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. 2025. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. *arXiv preprint arXiv:2504.15485*.

Qwen. 2024. Qvq: To see the world with wisdom.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Yufan Ren, Konstantinos Tertikas, Shalini Maiti, Junlin Han, Tong Zhang, Sabine Süsstrunk, and Filippos Kokkinos. 2025. Vgrp-bench: Visual grid reasoning puzzle benchmark for large vision-language models. *arXiv preprint arXiv:2503.23064*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Liheng Shen, Tetsu Matsukawa, and Einoshin Suzuki. 2024. Satjip: Spatial and augmented temporal jigsaw puzzles for video anomaly detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 27–40. Springer.

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

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. 2025. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. 2025. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*.

Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*.

Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. 2022. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xinglin Wang, Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2024b. Coglm: Tracking cognitive development of large language models. *arXiv preprint arXiv:2408.09150*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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.

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

Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, and Guangtao Zhai. 2025. Redundancy principles for mllms benchmarks. *arXiv preprint arXiv:2501.13953*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Examples of Other Spatial Reasoning Benchmarks

Figure 9 and Figure 10 illustrate two representative question types commonly used to evaluate spatial reasoning capability of VLMs. The tested images are not based on real-world scenes, which limits the capability to evaluate spatial reasoning in VLMs under realistic conditions.

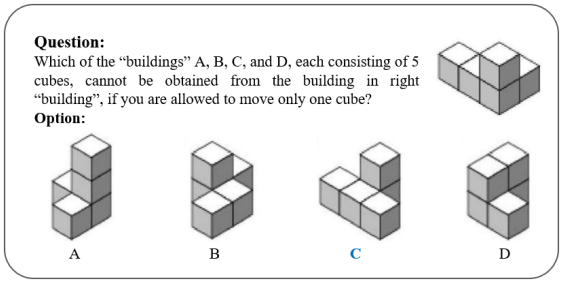

Figure 9: An example of Mind the Gap (Stogiannidis et al., 2025).

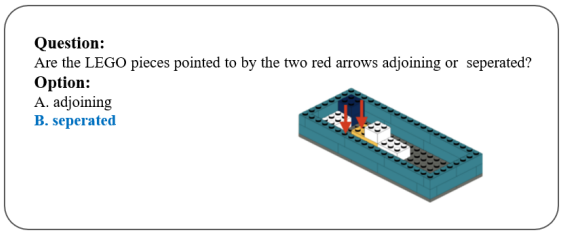

Figure 10: An example of LEGO-Puzzles (Tang et al., 2025).

B Dataset Curation

Rules Pool. Figure 11 shows examples of images that were rejected and accepted based on the filtering rules, the following are the rules defined by experts during the image selection process:

- Removing images containing explicit or violent content.
- Filtering out blurry, low-resolution, or visually ambiguous images.
- Excluding images lacking semantic clarity or spatial structure.
- Discarding images with structural ambiguity (e.g., multiple valid puzzle arrangements).
- Eliminating misaligned images or those with overly small visual elements after cropping, which hinder spatial reasoning.

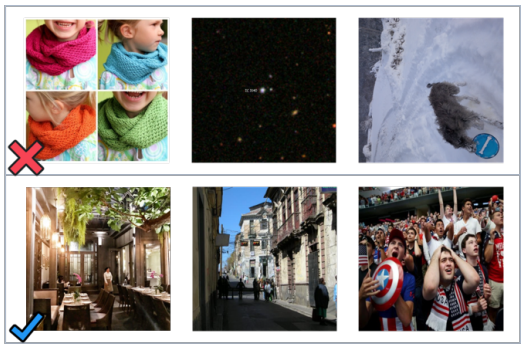

Figure 11: Top: examples of images rejected by expert-defined filtering rules. Bottom: examples of high-quality images that pass the rules.

Task-Specific Template. The following are detailed templates for each task. Note that <image_x> denotes a placeholder for the corresponding image input.

TASK 1: Missing Piece Selection

Question: You will be provided with <image 1>, <image 2>, <image 3>, <image 4>, <image 5>: The first image is a large image with a missing region, represented by a white blank area. The second to fifth images are four candidate patches. Among these four candidate images, exactly one can accurately fill the white blank area in the first image, seamlessly restoring its visual appearance. Your task is: Carefully observe the large image with the missing region; Compare the contents of the four candidate images and select the one that best visually completes the missing area of the first image.
The option mapping is as follows:
A. Candidate 1 (the second input image) B. Candidate 2 (the third input image)
C. Candidate 3 (the fourth input image) D. Candidate 4 (the fifth input image)
Please select the option you believe is correct and output only the corresponding letter (A, B, C, or D), without providing any explanation or analysis.

Figure 12: Template of *Missing Piece Selection*, notably, the templates for the Easy and Hard settings are identical.

TASK 2: Piece Localization (Easy)

Question: <image 1> has two masked areas labeled A and B. <image 2> shows the content from one masked area. Determine if it corresponds to A or B. Answer only "A" or "B".

Figure 13: Template of *Piece Localization (Easy)*.

TASK 2: Piece Localization (Hard)

Question: <image 1> has four masked areas labeled A, B, C, and D. <image 2> shows the content from one masked area. Determine if it corresponds to A, B, C, or D. Answer only "A", "B", "C", or "D".

Figure 14: Template of *Piece Localization (Hard)*.

TASK 3: Connection Verfication

Question: You are given <image 1>, <image 2>, each cropped from an original full image. The full image was divided into four regions: top-left, top-right, bottom-left, and bottom-right by splitting it through the center. Two regions are randomly selected. Based on the spatial relationship of these two regions in the original image, choose the most appropriate option and respond only with the corresponding letter (A, B, or C):
A. The two images are adjacent horizontally (left-right relationship)
B. The two images are adjacent vertically (top-bottom relationship)
C. The two images are not adjacent in the original image
Please respond with only "A", "B", or "C" and no additional text.

Figure 15: Template of *Connection Verification*.

TASK 4: Anomaly Detection

Question: You are given an image that is created by cutting a full image into 4 sub-images and stitching them back together. The positions of the <image 1>, <image 2>, <image 3>, <image 4> in the stitched image are defined as follows:
A. Top-left (the sub-image in the top-left corner)
B. Top-right (the sub-image in the top-right corner)
C. Bottom-left (the sub-image in the bottom-left corner)
D. Bottom-right (the sub-image in the bottom-right corner)
Note: These positions refer to the current stitched image you are looking at. Exactly one of the 4 sub-images may have been rotated or mirrored, or all sub-images may be completely normal. There is at most one abnormal sub-image in each stitched image. Your tasks:
1. Decide whether the stitched image is correct. Options: A. Correct B. Incorrect
2. If incorrect, answer:
   Which sub-image is abnormal? Options: A/B/C/D
   What kind of change has occurred? Options: A. Rotation, B. Mirroring
Output format (exactly 3 lines):
Judgment: A or B
Error Position: A/B/C/D (or leave blank)
Error Type: A/B (or leave blank)
Do not add any explanations or extra text. Example if incorrect (error in bottom-left, rotated):
Judgment: B Error Position: C Error Type: A
Example if correct: Judgment: A Error Position: Error Type: .

Figure 16: Template of *Anomaly Detection*.

TASK 5: Order Restoration

Question: You are given <image 1>, <image 2>, <image 3>, <image 4> that are cropped from an original full image. The full image was divided into four regions by splitting it through the center: top-left, top-right, bottom-left, and bottom-right. The four cropped images have been shuffled. Based on the visual content of each cropped image, determine the correct order that reconstructs the original full image. The order corresponds to the regions in the following sequence: top-left, top-right, bottom-left, bottom-right. Choose the most appropriate option based on the mapping below:
<option list> (A. [2, 3, 1, 4] B. [2, 4, 3, 1] C. [3, 4, 2, 1] D. [4, 1, 3, 2])
Each number corresponds to the index of the shuffled images you received.
Please respond only with "A", "B", "C", or "D", without any additional explanation or description.

Figure 17: Template of *Order Restoration*. Note: <option list> serves as a placeholder for the answer choices. The text in parentheses is an example and should be removed in actual use. One option is the correct answer, while the remaining three are randomly drawn from the other 23 candidates.

TASK 6: Order Generation

Question: You are given <image 1>, <image 2>, <image 3>, <image 4> that are cropped from an original full image. The full image was divided into four regions by splitting it through the center: top-left, top-right, bottom-left, and bottom-right. The four cropped images have been shuffled. Based on the visual content of each cropped image, determine the correct order that reconstructs the original full image. The order corresponds to the regions in the following sequence: top-left, top-right, bottom-left, bottom-right.
Please respond only with a Python-style list of four integers indicating the correct order of the shuffled images, like [2, 3, 1, 4]. Each number corresponds to the index of the shuffled images you received. Please Do not provide any explanation or description.

Figure 18: Template of *Order Generation*.

C Hardware Setup for Evaluating VLMs.

We evaluate open-source VLMs using hardware configurations scaled to model size. Models with fewer than 20B parameters run on a single NVIDIA A100 80GB GPU. Those between 20B and 40B use two NVIDIA A100 80GB GPUs, while models exceeding 40B are evaluated on four NVIDIA A100 80GB GPUs to meet their greater memory and computational demands.