VisWordBench: Bridging the Gap in Cross-modal Reasoning for Multimodal Large Language Models

Tianshu Zhang[1*], Junzhe Chen[1*], Yean Cheng[2], Demin Zhu[3], Haoze Zheng[4], and Lijie Wen[1†]

[1] Tsinghua University
[2] Z.ai
[3] Central University of Finance and Economics
[4] Xinjiang University
[*]Equal contribution. [†]Corresponding author.

Abstract. Although recent multimodal large language models (MLLMs) have advanced rapidly in vision–language reasoning, their capability boundaries, particularly in cross-modal integration and reasoning, remain underexplored. Existing benchmarks primarily focus on evaluating unimodal or loosely coupled multimodal abilities, leaving a gap in assessing complex cross-modal reasoning. To address this gap, we introduce VisWordBench, a benchmark comprising 2,625 English and 2,000 Chinese visual word puzzles with detailed human annotations. The benchmark is designed to evaluate not only fundamental perceptual understanding and world knowledge but also deep cross-modal reasoning skills that require consistently integrating visual and linguistic information. Through evaluations on VisWordBench, we find that current MLLMs exhibit under-diversified hypothesis search during reasoning. We therefore propose a data construction pipeline and a training-free inference-time steering strategy that promotes more diverse hypothesis exploration during CoT reasoning. Experimental results show consistent improvements in reasoning quality and hypothesis diversity, supporting the utility of VisWordBench for studying multimodal reasoning. The benchmark is available at: `https://zenodo.org/records/20957955`.

1 Introduction

Recent studies have substantially improved the reasoning and understanding abilities of multimodal large language models (MLLMs). Building on advances in large language models, MLLMs can now integrate visual and textual information more coherently, enabling perception, comprehension, and complex reasoning across modalities [61, 81].

Despite these advances, multimodal reasoning still faces several distinctive challenges. In cross-modal reasoning scenarios, models must continuously integrate visual information into text-centered semantic contexts, yet their ability to sustain this integration remains limited. Existing benchmark systems also have

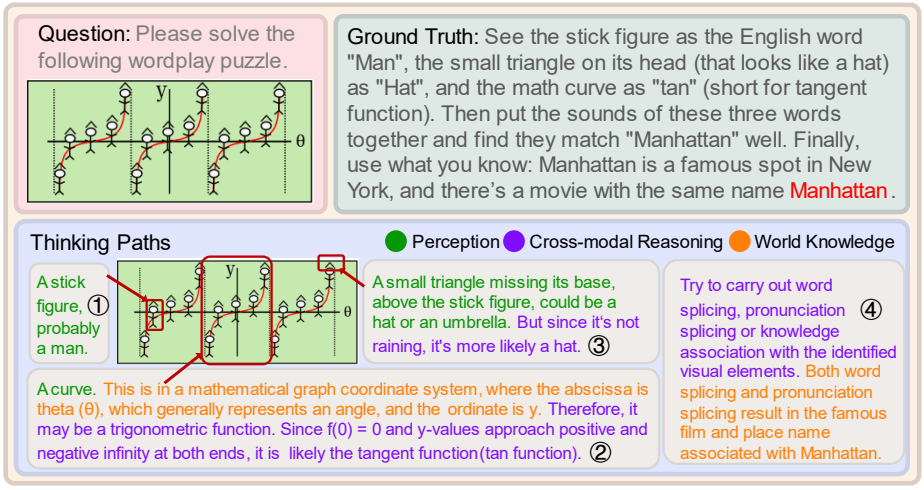


Fig. 1: Example of VisWordBench. Solving the puzzles requires world knowledge, visual perception, and the ability to integrate visual information during CoT reasoning.

important gaps. General VQA benchmarks [9,30,42] cover basic image–text correspondence, typically involve relatively simple question–answer formats, and therefore fail to assess deeper cross-modal reasoning in complex semantic contexts. One category of benchmarks focuses on pure textual reasoning, using tasks such as doctoral-level scientific questions to increase reasoning difficulty. While these tasks evaluate logical inference within text, they fail to measure reasoning that involves both visual and linguistic modalities. Another category centers on purely visual or abstract reasoning tasks, such as puzzles and games like 2048 [8, 47, 50, 55, 57, 68]. Although easy to synthesize, such benchmarks largely ignore textual reasoning, world knowledge utilization, and cross-modal information integration. Consequently, there remains a need for a benchmark that systematically evaluates cross-modal reasoning.

To address these challenges, we introduce VisWordBench, a minimal yet highly compositional cross-modal reasoning testbed designed to function as a litmus test for multimodal intelligence. Rather than simulating a single application scenario, VisWordBench distills the core cognitive requirements underlying real-world multimodal tasks into compact visual word puzzles that demand perception, world knowledge utilization, and cross-modal reasoning including iterative hypothesis verification. The benchmark comprises 2,625 English and 2,000 Chinese visual word puzzles, spanning a wide range of difficulty levels. As mentioned in [21], this type of problem is OOD for MLLMs, hence ideal for revealing limitations that are not apparent on existing benchmarks.

As illustrated in Figure 1, each puzzle in our benchmark is a compact visual scene encoding a multi-step wordplay task. In the example, the model must first perceive visual details: identify the stick figure ("man"), recognize the triangle

Table 1: Comparison between VisWordBench and existing multimodal benchmarks. A check mark (✓) indicates that the benchmark covers the corresponding capability, while a cross (✗) means it does not.

Benchmark	#Samples	Question Types	Diverse Visual Perception	World Knowledge Utilization	Cross-modal Reasoning	Numbers
ARC-AGI [10]	1,000	Reasoning Puzzles	✗	✗	✗	1,000
TET [16]	490	Perception puzzles	✓	✗	✗	490
VISUALPUZZLES [59]	1,168	Reasoning Puzzles	✓	✗	✓	1,168
PRISM-Bench [55]	1,044	Reasoning Puzzles	✓	✗	✗	1,044
VisuLogic [76]	1,000	Visual Logic	✓	✗	✗	1,000
VisuRiddles [80]	–	Visual Riddles	✓	✗	✗	–
A-OKVQA [58]	24,903	General VQA	✓	✓	✗	24,903
WorldQA [91]	1,007	Video QA	✓	✓	✗	1,007
MathVista [45]	6,141	Math	✓	✗	✓	6,141
MME-CC [89]	1,173	Cognitive Questions	✓	✗	✓	1,173
VisWordBench (Ours)	**4,625**	**puzzles**	✓	✓	✓	**4,625**

on its head as a "hat" and interpret the plotted curve as the tangent function "tan". The right side of the figure shows that the model needs to subsequently draw on perception, cross-modal reasoning, and world knowledge to fuse spatial layout, mathematical symbolism, and linguistic cues into the final answer "Manhattan". Notably, solving these puzzles requires iterative hypothesis generation and verification, closely mirroring exploratory reasoning in downstream tasks.

Extensive evaluations on VisWordBench reveal a substantial gap between models and human performance in cross-modal reasoning tasks. To better understand this gap, we further conduct a structural analysis of model reasoning traces by modeling chains of thought as canonical tree-structured exploration processes with revisit edges, and introducing quantitative metrics for branching depth, branching distribution, and revisit behavior. This analysis shows that the gap is primarily driven by under-diversified hypothesis search within the reasoning process, i.e., the models' tendency to insufficiently explore alternative hypotheses during inference, often resulting in repetitive local reasoning rather than high-level re-exploration. To mitigate this limitation, we propose two key innovations: (1) a novel data construction framework that inherently embeds multi-level semantic relationships and reasoning dependencies, encouraging more diversified hypothesis generation during reasoning; and (2) a training-free inference enhancement method that guides models to perform deeper and more diversified hypothesis search within the reasoning chain.

Experimental results show that both components improve reasoning performance across multiple MLLMs. These findings establish VisWordBench as a diagnostic benchmark for cross-modal reasoning and indicate that diversified hypothesis search is a useful target for improving MLLM reasoning.

Our contributions are summarized as follows:

• We introduce VisWordBench, a new benchmark containing 2,625 English and 2,000 Chinese visual word puzzles with detailed human annotations. It evaluates models' abilities in perception, world knowledge, and cross-modal reasoning. As a compact yet cognitively demanding setting, VisWordBench serves as a diagnostic litmus test for multimodal intelligence.

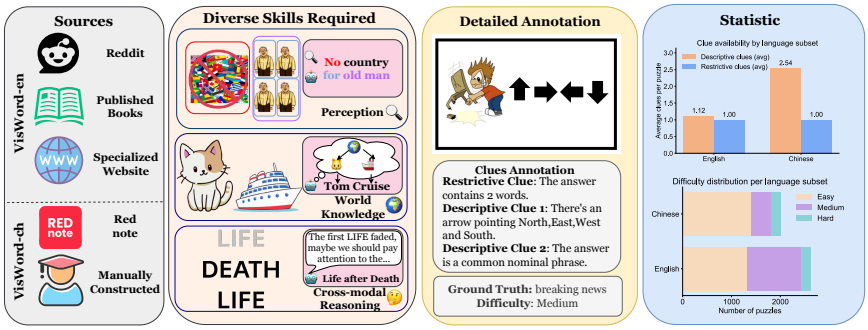


Fig. 2: Overview of VisWordBench, a bilingual benchmark of 2,625 English and 2,000 Chinese visual word puzzles targeting perception, world knowledge, and cross-modal reasoning, with dense restrictive and descriptive clue annotations.

• Extensive evaluations on both open-source and proprietary MLLMs show a large performance gap between humans and models. To further diagnose this gap, we model reasoning traces as canonical tree-structured exploration processes and introduce quantitative structural metrics, which reveal under-diversified hypothesis search within reasoning chains as a key bottleneck.

• To address this, we develop a data synthesis pipeline and an inference steering method that promote more diversified hypothesis generation and search during reasoning without additional training, achieving consistent improvements over strong baselines.

2 Related work

2.1 Multimodal Large Language Models

Following the development of LLMs, MLLM research has shifted from perception-focused modeling toward visual instruction tuning and human-aligned multimodal instruction following. Early systems [12, 36, 39, 40] and scaled successors [4,38,40,79] mainly align a visual encoder with a pretrained LLM using large-scale conversational multimodal data, improving general visual understanding and open-ended interaction. However, these models still largely treat vision as an input interface to the LLM and provide limited explicit support for complex multimodal reasoning. Recent reasoning-oriented models, such as OpenAI o1 and DeepSeek-R1 [13], train long CoT behavior and optimize it with reinforcement learning. Multimodal successors [14,66,74,87] apply similar strategies to vision–language inputs, combining long-form CoT and reflection for stronger multimodal reasoning.

2.2 MLLM Evaluation Benchmarks

With the rapid development of MLLMs, multimodal benchmarks have evolved from perception toward knowledge- and reasoning-centric evaluation. Early work targeted visual recognition [16,32], grounding [64,72,83], and reference resolution [17,23,33,44,46]; broader suites then added captioning [6,23,35], visual QA [1,2, 25,26,65], general multimodal comprehension [19,41,71,75], and world knowledge [3, 37, 63, 93]. Recent efforts probe higher-order visual reasoning by introducing STEM [45, 56, 67, 69, 78, 84], chart and document understanding [31, 48, 49, 60], spatial composition [15, 28, 43, 82], multi-step reasoning [24, 29, 52, 54, 70], and abstract tasks [51,62], alongside broader generalist probes [22,27,86]. In parallel, puzzle-based benchmarks split into two tracks: those isolating visual reasoning with minimal ancillary skills [11, 18, 57, 59, 77, 90], and those coupling puzzles with broader competencies [5, 34, 53, 85, 88, 92].

Despite the breadth of prior efforts, most existing benchmarks remain fragmented, targeting isolated abilities with a notable absence of rigorous cross-modal reasoning evaluation. As summarized in Table 1, there is still no integrative evaluation that simultaneously measures diverse visual perception, world knowledge, and cross-modal reasoning. Our proposed benchmark is designed precisely to fill this gap. In particular, recent visual-reasoning benchmarks such as VisuLogic and VisuRiddles emphasize abstract or fine-grained visual reasoning, CodePercept [20] focuses on code-grounded STEM perception, and knowledge-centric benchmarks such as WorldQA stress multimodal world knowledge. These settings are valuable but leave different parts of the reasoning chain isolated. VisWordBench instead requires a model to first ground visual clues, map them into linguistic units, retrieve relevant world knowledge when needed, and verify candidate answers in a single cross-modal chain. This positioning makes VisWordBench complementary to prior benchmarks and suitable for diagnosing whether MLLMs can sustain visual–linguistic integration during exploratory reasoning.

3 Dataset Construction

Our benchmark, VisWordBench, consists of a total of 4,625 puzzles, including 2,625 in English and 2,000 in Chinese, denoted as VisWordBench-en and VisWordBench-ch respectively. The English puzzles primarily originate from published books, websites dedicated to word puzzles, and social media platforms such as Reddit, with all samples being manually collected. The Chinese puzzles are sourced mainly from RedNote, with the majority being manually collected and a small subset being artificially generated. An overview of the data sources, key characteristics, representative examples, and basic statistics is provided in Figure 2. We implemented a rigorous deduplication process, ensuring uniqueness by both answer-based and image-based deduplication. To further reduce contamination risk, we prioritized recent sources when possible, retained source metadata for auditability, and removed near-duplicate visual or answer-equivalent samples across training and test splits. Each sample is annotated

Table 2: Representative models' performances on our VisWordBench.

Models	VisWordBench-en				VisWordBench-ch			
	Easy	Medium	Hard	All	Easy	Medium	Hard	All
Qwen2.5-VL-7B	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00
GLM-4.1v-Thinking-9B	4.67	0.88	0.00	2.71	1.99	0.50	1.73	1.65
Kimi-vl-thinking-16B-A3B	3.12	1.31	0.00	2.02	5.87	4.02	4.10	5.35
GLM-4.5v-106B-A12B	23.66	7.40	2.66	15.34	4.47	1.51	2.89	3.70
doubao-1.5-thinking-vision-pro-250428	27.95	8.32	1.44	17.65	9.02	4.03	1.73	7.30
doubao-seed-1.6-vision-250815	36.01	14.60	3.03	24.51	9.52	4.79	5.64	8.20
Gemini-2.5-pro	57.12	40.22	9.57	46.28	16.97	5.29	8.21	13.80
GPT-5	**72.62**	**54.99**	**22.73**	**61.37**	**24.29**	**13.35**	**14.36**	**23.15**
Human	**89.07**	**61.23**	**9.50**	**71.32**	**99.58**	**98.74**	**5.50**	**90.25**

with its correct answer, and we explicitly handle puzzles that may have multiple valid solutions: valid alternative answers discovered during validation are added to the accepted-answer set, while overly ambiguous puzzles are removed. We calibrated difficulty based on human performance. For VisWordBench-en, we identify 1322 easy samples, 1103 medium-level samples, and 200 hard samples, and for VisWordBench-ch, we identify 1408 easy samples, 397 medium-level samples, and 195 hard samples.

To support future research, we annotate each VisWordBench sample with clues. Each puzzle has two types. The first, the restrictive clue, specifies the number of words in the answer and letters per word, or characters for Chinese puzzles. These clues aim to improve verification awareness during CoT. However, they are rule-generated and interact poorly with tokenization, which may mislead MLLMs. Their rigid format cannot capture higher-dimensional answer properties, leading to incomplete or incorrect inferences. Therefore, we seek clues that better guide reasoning and provide richer contextual hints.

To address this limitation, we introduce descriptive clues that provide additional, high-level guidance for the answer. These clues may indicate the answer category, such as a country name, or highlight key puzzle features, such as vertical letter arrangements. They give models a more informative path toward the correct solution without directly revealing the answer.

In pilot experiments, Gemini-2.5-pro performed well on VisWordBench-en and was used to generate clues for English puzzles, followed by human review. Its performance on VisWordBench-ch was insufficient, so Chinese clues were manually created. These descriptive clues were annotated by eight fluent annotators, achieving a Fleiss' Kappa of 0.81.

For each puzzle, we provide one restrictive clue and one to three descriptive clues. VisWordBench-en contains 1.12 descriptive clues per puzzle on average, while VisWordBench-ch contains 2.54. Further details can be found in Appendix A.

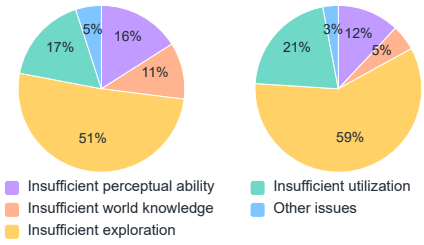


Table 3: Comparison across models on four metrics we proposed for evaluating exploration in the reasoning process.

Model	nABD	BDE	SRR	ARR
GLM-4.1v-thinking	0.79	0.19	0.43	0.31
GLM-4.5v	0.75	0.21	0.39	0.25
Gemini-2.5-pro	0.66	0.28	0.21	0.11

Fig. 3: (Left) Error analysis of GLM-4.1v-thinking on VisWordBench-en and doubao-1.5-thinking-vision-pro-250428 on VisWordBench-ch.

4 Benchmark Setup and Analysis

This section describes the benchmark setup and analyzes model performance and failure patterns on VisWordBench.

4.1 Problem: Deficiencies in Reasoning

The overall performance of eight representative large vision-language models is shown in Table 2. The results lead to three observations. First, model performance on VisWordBench remains far below human performance, particularly on VisWordBench-ch. Human annotators achieve over 90% accuracy, whereas the best-performing model reaches only 23.15%. Models perform relatively better on VisWordBench-en, but still fall short of human annotators. Second, open-source models lag behind closed-source models: the best-performing open-source model, GLM-4.5v, reaches only about one seventh of the leading closed-source model's performance on VisWordBench-ch. Third, larger models generally perform better on this task. For example, GLM-4.1v-Thinking-9B and GLM-4.5v-106B-A12B, trained with similar methodology, show a several-fold performance difference. We further evaluated recent strong systems to test whether this conclusion remains stable. Qwen3.5-9B obtains 5.64% on VisWordBench-en and 1.60% on VisWordBench-ch, while its thinking variant improves the Chinese score to 2.20% and changes the error profile through more propose–verify cycles. Larger systems are stronger but still far from human performance: Qwen3.5-397B reaches 41.07% on VisWordBench-en and 19.80% on VisWordBench-ch, and Claude-Opus-4.6 reaches 46.42% and 21.40%, respectively. These results indicate that explicit thinking and scaling help, but neither closes the cross-modal reasoning gap.

4.2 Error Analysis

We manually analyzed a subset of error cases and grouped the primary failure causes into three categories: **insufficient reasoning ability**, **insufficient**

(a) An ideal reasoning process with multi-dimensional exploration.

(b) A common failure pattern with under-diversified hypothesis search.

Fig. 4: Reasoning patterns modeled as tree diagrams. The left shows an ideal trace, while the right illustrates a common failure. Red arrows denote revisiting.

world knowledge, and **insufficient perceptual ability**. Reasoning failures were further subdivided into three types: **under-diversified hypothesis search**, **insufficient utilization**, and other issues. Under-diversified hypothesis search refers to the model's tendency to repeatedly propose similar or redundant hypotheses during the reasoning chain, preventing it from arriving at the correct solution. Insufficient utilization refers to cases where the model reaches the correct answer but fails to validate it and eventually outputs an incorrect conclusion. Other issues include contradictions within the reasoning process, gradually weakened instruction-following, and similar problems. Insufficient world knowledge refers to the model's lack of familiarity with the background knowledge required to solve the puzzle, such as not understanding relative slang. Insufficient perceptual ability refers to errors in visual perception, such as overlooking small objects in the image or misjudging spatial relationships. We combined human annotation with LLM-based labeling to estimate the frequency of each error type, as shown in Figure 3.

The results show that insufficient reasoning ability is the dominant source of errors, whereas world knowledge and perceptual limitations play secondary roles. Among reasoning failures, Under-diversified hypothesis search accounts for over half of the cases, making it the primary bottleneck. This observation is also consistent with the accuracy distribution across different difficulty levels. As shown in Table 2, the models perform worse on medium-difficulty tasks (where human annotators require some thinking to provide an answer) than on difficult tasks (where human annotators are unable to provide an answer). The hard split is therefore the human-difficulty tail rather than an indication of leakage: these items are calibrated to require rare cultural associations, subtle wordplay, or atypical visual-to-phonetic mappings. This pattern suggests that models often become trapped in a narrow reasoning subspace rather than systematically exploring alternatives, highlighting diversified hypothesis search as a key limitation.

To complement human annotation, we analyze the structure of MLLM reasoning traces in our benchmark to identify exploration failures in CoT reasoning. Each trace is modeled as a canonical reasoning tree with revisit edges, as shown in Table 3, Figure 4 . We decompose traces into atomic steps and map them

to three node types: observation, interpretation, and answer. Observation and interpretation nodes may form multiple layers, enabling deep reasoning beyond a shallow pipeline. Semantically equivalent observation/interpretation nodes are merged into canonical states to avoid over-fragmentation, while answer nodes remain distinct. Revisit edges record returns to prior states.

Based on this structure, we quantify hypothesis search quality with four metrics. Let $\mathcal{B}$ denote the set of branching events in the reasoning tree, and let $d(b)$ be the depth of branching event $b \in \mathcal{B}$. Let $B_k = |\{b \in \mathcal{B} : d(b) = k\}|$ be the number of branching events at depth k, and let $K = \{k : B_k > 0\}$ be the set of active branching depths.

Normalized Average Branching Depth (nABD). Let D denote the maximum depth of the reasoning tree for a given sample. We define the normalized average branching depth as

$$\text{nABD} = \frac{1}{D} \cdot \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} d(b) = \frac{\sum_{k \in K} k\, B_k}{D \sum_{k \in K} B_k}$$

A higher nABD indicates that branching tends to occur later (closer to the leaves) relative to the overall depth of the reasoning tree.

Branch Depth Entropy (BDE). BDE measures how evenly branching is distributed across depths. Define

$$\text{BDE} = -\frac{\sum_{k \in K} \frac{B_k}{\sum_{j \in K} B_j} \log \frac{B_k}{\sum_{j \in K} B_j}}{\log |K|}$$

A higher BDE indicates that branching is distributed across multiple depths, while a lower BDE indicates that branching is concentrated at a single stage.

State Revisit Rate (SRR). Let V_{state} be the set of canonical nodes, and let $r(v)$ be the revisit count of state node $v \in V_{\text{state}}$. We define

$$\text{SRR} = \frac{\sum_{v \in V_{\text{state}}} r(v)}{|V_{\text{state}}|}.$$

A higher SRR indicates more frequent returns to previously explored intermediate reasoning states.

Answer Revisit Rate (ARR). Let V_{ans} be the set of answer proposal nodes, and let R_{ans} be the number of answer revisit events, i.e., the number of answer proposals whose string has already appeared earlier in the same trace.We define $\text{ARR} = \frac{R_{\text{ans}}}{|V_{\text{ans}}|}$. A higher ARR indicates stronger repetition in answer proposals.

For degenerate cases ($D = 0$ or $|\mathcal{B}| = 0$ for nABD; $|K| \leq 1$ for BDE), we set the corresponding metric to 0.

Our results 3 reveal a consistent pattern of structurally under-diversified hypothesis search. Branch-depth entropy is generally low, indicating that hypothesis branching tends to concentrate at a narrow stage of reasoning rather than being distributed across the full decision hierarchy. Moreover, branching occurs predominantly at later depths, suggesting that diversification happens mainly at the answer-generation stage. In other words, models tend to enumerate candidate outputs without adequately exploring alternative high-level hypotheses earlier in the reasoning process. We also observe relatively high intermediate-state revisit rates, showing that long reasoning traces frequently loop over similar mid-level hypotheses instead of reconsidering upstream assumptions and reorienting the reasoning trajectory from different conceptual starting points. These structural metrics complement, rather than replace, semantic error analysis: in concrete failures, the collapse often occurs at visual-to-phonetic mapping, idiom retrieval, or final answer verification. This aligns with our qualitative observations: MLLMs often follow nearly identical high-level reasoning paths and vary primarily in their final conclusions, rather than engaging in genuinely diversified hypothesis search.

5 Method

The benchmark results and error analysis indicate that under-diversified hypothesis search is a major limitation in the CoT process. We address this issue with a synthetic data generation procedure and a training-free inference-time intervention that encourages more diversified hypothesis search.

5.1 Exploratory Prompting

We first query a powerful MLLM, denoted $\mathcal{M}_{\text{MLLM}}$, with a structured prompt $\pi_{\text{explore}}(q)$: $\tau_0 = \mathcal{M}_{\text{MLLM}}(\pi_{\text{explore}}(q))$, which asks the model to: 1) propose multiple incorrect answers $\{a_t^-\}$, 2) analyze failures, and 3) propose and verify a correct answer a^+. While this approach increases the hypothesis search diversity of τ_0, it reduces answer accuracy, suggesting intrinsic reasoning deficiencies that prompting alone cannot fix, thereby lowering the number of usable samples.

Multi-Stage Aggregation and Refinement To mitigate this limitation, we design a multi-stage procedure that explicitly aggregates diverse reasoning branches with controlled redundancy, which can be formalized as:

(i) **Multi-rollout collection.** For each question q, we perform n rollouts: $\mathcal{T}(q) = \{\tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(n)}\}$, each sampled from $\mathcal{M}_{\text{MLLM}}$ uniformly.

(ii) **Segmentation into units.** We employ an LLM $\mathcal{M}_{\text{seg}}$ to segment each trajectory: $\tau^{(i)} = \{u_j^{(i)}\}_{j=1}^{k_i}$, where $u_j^{(i)} = (\text{proposal}, \text{verification})$.

(iii) **Answer-level de-duplication.** We use an auxiliary large language model to evaluate semantic equivalence among candidate answers. Semantically redundant answers are removed because they add little to diversified hypothesis search, yielding the unique set $\mathcal{U}(q)$.

(iv) **Diversified hypothesis search set construction.** We construct a compact hypothesis search set $\mathcal{U}(q) = \{u_i\}_{i=1}^{m}$, where each unit represents a plausible hypothesis together with its corresponding reasoning step. We concatenate these units to form a diversified hypothesis-search trajectory $\tau_{\text{search}}(q)$. Note that the terminal answer a_m is not required to be correct, as the goal is to preserve hypothesis diversity rather than enforce early convergence.

(v) **Ground-truth–guided closure (without leakage).** Given the ground-truth answer $a^\star$, we query $\mathcal{M}_{\text{MLLM}}$ with prompt $\pi_{\text{close}}(q, a^\star)$ to generate a closing rationale $r^\star$ that resolves the task under the correct hypothesis. We then employ a rewriting model $\mathcal{M}_{\text{rewrite}}$ to remove any explicit leakage, yielding $\tilde{r}^\star$. The final trajectory becomes $\tau_{\text{final}}(q) = \tau_{\text{search}}(q) \cup \{(\tilde{r}^\star, a^\star)\}$.

(vi) **Global polishing.** Finally, we apply a global fluency model $\mathcal{M}_{\text{polish}}$ to refine $\tau_{\text{final}}(q)$, ensuring coherent transitions across hypotheses and natural language flow.

Empirical Observations Empirically, this pipeline yields higher-quality CoT trajectories with substantially more diversified hypothesis search than standard distillation: $Q(\tau_{\text{final}}) \gg Q(\tau_0)$, as supported by the results in Section 6. The structured aggregation over $\mathcal{T}(q)$ enables more diverse reasoning coverage while maintaining logical coherence.

5.2 Steering to Incentivize Diversified Hypothesis Search

We first identify samples whose primary error source is under-diversified hypothesis search during reasoning. We refer to these as hypothesis-underdiversified samples, which consist of the original model inputs paired with their corresponding incorrect outputs. For each such case, we further construct hypothesis-diversified counterparts using the synthetic data generation method introduced in the previous subsection. In these counterparts, the same model inputs are paired with synthesized outputs that reflect a more diversified hypothesis search process and lead to correct reasoning outcomes. Together, these two sets of samples form the foundation for our intervention design.

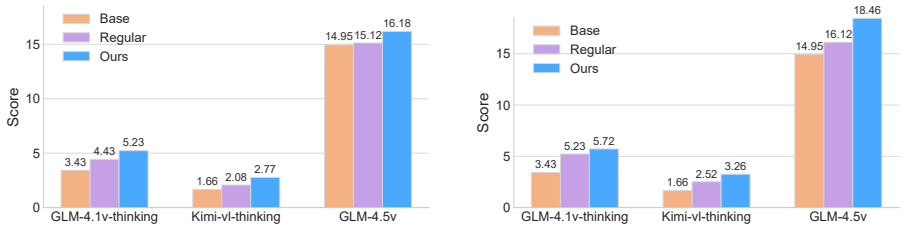
Inspired by previous work [7], we adopt an inference-time steering approach to enhance the exploration capability of MLLMs during the CoT process.

Consider an MLLM with parameters θ. Let $\mathbf{X} = \{x_1, \ldots, x_p\}$ denote text tokens and $\mathbf{V} = \{v_1, \ldots, v_q\}$ denote visual tokens. We build a single stream $\mathbf{H}^{(0)} = [\mathbf{V} \,\|\, \mathbf{X}]$ and pass it through L Transformer layers with N heads. Define the layer map:

$$\Phi^{(l)}(\mathbf{H}) \;=\; \sum_{n=1}^{N} \mathrm{Attn}_n^{(l)}(\mathbf{H})\, \mathbf{W}_o^{(l)},$$

so the residual update is: $\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \Phi^{(l)}(\mathbf{H}^{(l)})$.

Fused token and activations. Let the fused index be $s = p + q + 1$. For paired runs $i = 1, \ldots, N$,



(a) Steering. **(b)** Training.

Fig. 5: Performance comparison between the original model and the model after (left) steering and (right) training. "Base" denotes performance before applying the corresponding procedure. "Regular" (steering) / "Answer-only" (training) indicates using data obtained through conventional distillation settings, while "Ours" uses data generated by our synthetic data pipeline.

$$\mathbf{a}_{i,n}^{(l)} = \big(\mathrm{Attn}_n^{(l)}(\mathbf{H}^{(l)})\big)_{s,:}, \quad \mathbf{a}_{i,n}^{\prime(l)} = \big(\mathrm{Attn}_n^{(l)}(\mathbf{H}^{\prime(l)})\big)_{s,:}.$$

The per-head activation shift is: $\mathbf{V}_n^{(l)} = \frac{1}{N}\sum_{i=1}^{N}\big(\mathbf{a}_{i,n}^{(l)} - \mathbf{a}_{i,n}^{\prime(l)}\big)$.

Intervention. We modify the block as:

$$\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \sum_{n=1}^{N}\Big(\mathrm{Attn}_n^{(l)}(\mathbf{H}^{(l)}) + \alpha\,\mathbf{V}_n^{(l)}\Big)\mathbf{W}_o^{(l)},$$

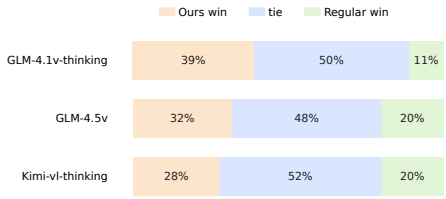
which preserves the architecture while nudging the fused-token behavior toward broader exploration.

6 Experiments

This section evaluates the proposed data synthesis and inference-steering methods.

6.1 Training on Synthetic Data

We first evaluate the synthetic data pipeline by training directly on synthetic data for VisWordBench-en. We randomly select 1000 samples as the training set and use the remaining 1625 as the test set. For GLM-4.1v-thinking and Kimi-vl-thinking-2506, we perform full-parameter fine-tuning, whereas for GLM-4.5v, we apply LoRA fine-tuning with rank 8. The results are shown in Figure 5b. Compared with the regular setting, our pipeline improves performance by 2.29%, 1.60%, and 3.51% for the three models, respectively. The gains are larger for the larger model, suggesting that stronger models may benefit more from diversified hypothesis-search trajectories.



Backbone	Setting	nABD	BDE	SRR	ARR
GLM-4.1v-thinking	+Regular	0.81	0.18	0.45	0.33
	+Ours	**0.78**	**0.22**	**0.41**	**0.32**
GLM-4.5v	+Regular	0.77	0.20	0.42	0.26
	+Ours	**0.73**	**0.24**	**0.38**	**0.21**

Fig. 6: (Left) Hypothesis search diversification after steering compared with the original model. "Regular" denotes steering using data obtained through conventional settings. (Right) Comparison on the proposed structural metrics.

6.2 Inference Steering

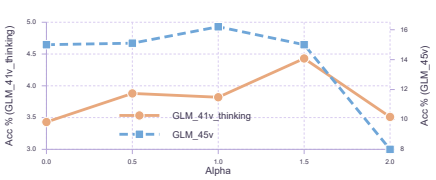
During the implementation, we incorporated the obtained intervention vectors as weights into the model to adapt to mainstream inference frameworks, thereby enhancing efficiency. For GLM-4.1v-thinking and Kimi-vl-thinking-2506, we set α to 1, while for GLM-4.5v, we set α to 1.5. As shown in Figure 5a, inference steering improves performance by 1.80%, 1.11%, and 2.23%, respectively. These relative gains are comparable to those obtained by fine-tuning. To assess how much of the performance improvement can be attributed to enhanced hypothesis search diversity, we replaced the hypothesis-diversified set with standard distillation data and observed only marginal gains. This suggests that the observed improvement does not simply arise from better supervision signals, but primarily from promoting more diversified hypothesis search during reasoning. These results further validate the effectiveness of our inference-steering approach and reinforce our earlier conclusion that under-diversified hypothesis search in the CoT process is a key contributor to the model's suboptimal performance.

6.3 Analysis and Ablations

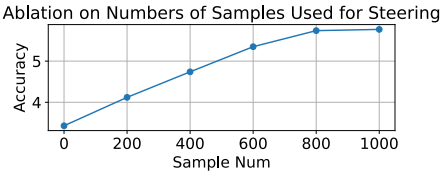
Improvement in Hypothesis Search Diversity We combined manual and LLM-based annotation to assess responses from the original and steered models, focusing on their relative hypothesis search diversity during the CoT process.

Results in Figure 6 show that inference steering on our data improves hypothesis search diversity. Models explore more diverse hypotheses during reasoning, indicating that the method mitigates CoT search collapse. A representative case appears in Appendix B. In this case, the base model fails by repeatedly performing ineffective self-reflection without generating alternatives beyond superficial spelling cues. In contrast, steered models explore multiple plausible hypotheses and reach the correct answer.

Ablation on α During the inference-steering stage, we use a hyperparameter, the intervention intensity α. Figure 7a presents the ablation results. When α is too low, the intervention is too weak to be effective. When α is too high,



(a) Effect of hyperparameters α on the accuracy for GLM-4.1v-thinking and GLM-4.5v models.



(b) Steering sample size ablation ($\alpha = 1$).

Fig. 7: Ablation studies on GLM-4.1v-thinking. (Left) Effect of hyperparameters α. (Right) Steering sample size ablation with $\alpha = 1$, evaluated on VisWordBench-en.

the intervention becomes overly aggressive and interferes with the model's core abilities, reducing performance.

Ablation on Sample Nums We also conduct an ablation study on the number of samples used to estimate the steering vectors. As shown in Figure 7b, the method is not very sensitive to this number and achieves its best performance with 600–800 samples, indicating that it remains effective in low-data settings.

Generalization to Broad Multimodal Reasoning Tasks A natural question is whether VisWordBench are overly specialized and whether our inference steering reflects transferable multimodal reasoning. We evaluate OOD generalization on MathVista [45] and LogicVista [73]. Using a strict zero-shot transfer setup with the same intervention vector derived on VisWordBench, we observe improvements on both benchmarks, as shown in Figure 8, suggesting that the acquired vectors promote more diversified hypothesis search. We also test GLM-4.1v-thinking on three additional reasoning benchmarks: steering improves VisuLogic from 24.7% to 25.9%, VisuRiddles from 24.0% to 26.7%, and CrosswordBench WCR_all from 6.04% to 7.58%. These results support the broader relevance of VisWordBench to general cross-modal reasoning.

7 Conclusion

In this work, we present VisWordBench, a bilingual visual word-puzzle benchmark for evaluating MLLMs' perception, world knowledge, and cross-modal reasoning. Evaluation on eight representative MLLMs reveals a substantial performance gap between models and human annotators. Our structural analysis shows that this gap largely stems from under-diversified hypothesis search during CoT. To address this, we propose a data generation pipeline that encourages diversified hypothesis search and an inference-steering approach that promotes it without additional training. Experiments show consistent gains from both components.

The additional analyses clarify the scope of these findings. Compared with visual-only reasoning benchmarks and knowledge-centric VQA, VisWordBench

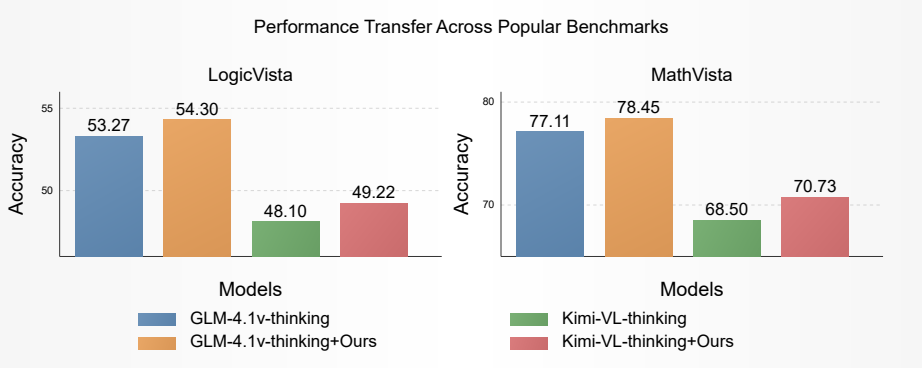


Fig. 8: We evaluate the generalizability of VisWordBench and our method on LogicVista and MathVista under the same settings as Sections 5 and 6.

requires models to connect visual grounding, linguistic mapping, world knowledge retrieval, and answer verification within one reasoning chain. We reduce ambiguity by retaining valid alternative answers and filtering overly subjective puzzles, and mitigate contamination risk through image- and answer-level deduplication with source metadata. OOD transfer results on MathVista and LogicVista, together with VisuLogic, VisuRiddles, and CrosswordBench, suggest that the learned steering signal is not puzzle-specific, but is associated with a broader propose–verify reasoning pattern.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2024YFB3309702), Beijing Natural Science Foundation (QY25044), the National Nature Science Foundation of China (No. 62021002), Tsinghua BNRist and Beijing Key Laboratory of Industrial Big Data System and Application.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Batra, D., Parikh, D.: Vqa: Visual question answering (2016)
2. Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., Sanghai, S.: Gqa: Training generalized multi-query transformer models from multi-head checkpoints (2023)
3. Anjum, K., Arshad, M.A., Hayawi, K., Polyzos, E., Tariq, A., Serhani, M.A., Batool, L., Lund, B., Mannuru, N.R., Bevara, R.V.K., Mahbub, T., Akram, M.Z., Shahriar, S.: Domain specific benchmarks for evaluating multimodal large language models (2025)
4. Bai, J., Ye, C., et al.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2407.10322 (2024)

5. Barrett, D.G.T., Hill, F., Santoro, A., Morcos, A.S., Lillicrap, T.: Measuring abstract reasoning in neural networks (2018)
6. Chen, J., Zhang, T., Huang, S., Niu, Y., Sun, C., Zhang, R., Zhou, G., Wen, L., Hu, X.: Omnidpo: A preference optimization framework to address omni-modal hallucination (2025)
7. Chen, J., Zhang, T., Huang, S., Niu, Y., Zhang, L., Wen, L., Hu, X.: Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models (2025)
8. Chen, L., Gao, H., Liu, T., Huang, Z., Sung, F., Zhou, X., Wu, Y., Chang, B.: G1: Bootstrapping perception and reasoning abilities of vision-language model via reinforcement learning (2025)
9. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? (2024)
10. Chollet, F., Knoop, M., Kamradt, G., Landers, B.: Arc prize 2024: Technical report (2024)
11. Chollet, F., Knoop, M., Kamradt, G., Landers, B., Pinkard, H.: Arc-agi-2: A new challenge for frontier ai reasoning systems (2025)
12. Dai, W., Li, J., Li, D., Hoi, S.C.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
13. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025)
14. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. ArXiv **abs/2503.21776** (2025), https://api.semanticscholar.org/CorpusID:277349899
15. Feng, Z., Kang, Z., Wang, Q., Du, Z., Yan, J., Shi, S., Yuan, C., Liang, H., Deng, Y., Li, Q., Yang, R., An, A., Zheng, L., Wang, W., Chen, S., Xu, S., Liang, Y.,

Yang, J., Guo, B.: Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes (2025)
16. Gao, H., Huang, Z., Xu, L., Tang, J., Li, X., Liu, Y., Li, H., Hu, T., Lin, M., Yang, X., Wu, G., Bi, B., Chen, H., Zhang, W.: Pixels, patterns, but no poetry: To see the world like humans (2025)
17. Ghosh, S., Evuru, C.K.R., Kumar, S., Tyagi, U., Nieto, O., Jin, Z., Manocha, D.: Visual description grounding reduces hallucinations and boosts reasoning in lvlms (2025)
18. Gritsevskiy, A., Panickssery, A., Kirtland, A., Kauffman, D., Gundlach, H., Gritsevskaya, I., Cavanagh, J., Chiang, J., Roux, L.L., Hung, M.: Rebus: A robust evaluation benchmark of understanding symbols (2024)
19. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models (2024)
20. Guan, T., Yang, Z., Wan, J., et al.: Codepercept: Code-grounded visual stem perception for mllms (2026)
21. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., Chen, J., Huang, J., Lei, K., Yuan, L., Luo, L., Liu, P., Ye, Q., Qian, R., Yan, S., Zhao, S., Peng, S., Li, S., Yuan, S., Wu, S., Cheng, T., Liu, W., Wang, W., Zeng, X., Liu, X., Qin, X., Ding, X., Xiao, X., Zhang, X., Zhang, X., Xiong, X., Peng, Y., Chen, Y., Li, Y., Hu, Y., Lin, Y., Hu, Y., Zhang, Y., Wu, Y., Li, Y., Liu, Y., Ling, Y., Qin, Y., Wang, Z., He, Z., Zhang, A., Yi, B., Liao, B., Huang, C., Zhang, C., Deng, C., Deng, C., Lin, C., Yuan, C., Li, C., Gou, C., Lou, C., Wei, C., Liu, C., Li, C., Zhu, D., Zhong, D., Li, F., Zhang, F., Wu, G., Li, G., Xiao, G., Lin, H., Yang, H., Wang, H., Ji, H., Hao, H., Shen, H., Li, H., Li, J., Wu, J., Zhu, J., Jiao, J., Feng, J., Chen, J., Duan, J., Liu, J., Zeng, J., Tang, J., Sun, J., Chen, J., Long, J., Feng, J., Zhan, J., Fang, J., Lu, J., Hua, K., Liu, K., Shen, K., Zhang, K., Shen, K., Wang, K., Pan, K., Zhang, K., Li, K., Li, L., Li, L., Shi, L., Han, L., Xiang, L., Chen, L., Chen, L., Li, L., Yan, L., Chi, L., Liu, L., Du, M., Wang, M., Pan, N., Chen, P., Chen, P., Wu, P., Yuan, Q., Shuai, Q., Tao, Q., Zheng, R., Zhang, R., Zhang, R., Wang, R., Yang, R., Zhao, R., Xu, S., Liang, S., Yan, S., Zhong, S., Cao, S., Wu, S., Liu, S., Chang, S., Cai, S., Ao, T., Yang, T., Zhang, T., Zhong, W., Jia, W., Weng, W., Yu, W., Huang, W., Zhu, W., Yang, W., Wang, W., Long, X., Yin, X., Li, X., Zhu, X., Jia, X., Zhang, X., Liu, X., Zhang, X., Yang, X., Luo, X., Chen, X., Zhong, X., Xiao, X., Li, X., Wu, Y., Wen, Y., Du, Y., Zhang, Y., Ye, Y., Wu, Y., Liu, Y., Yue, Y., Zhou, Y., Yuan, Y., Xu, Y., Yang, Y., Zhang, Y., Fang, Y., Li, Y., Ren, Y., Xiong, Y., Hong, Z., Wang, Z., Sun, Z., Wang, Z., Cai, Z., Zha, Z., An, Z., Zhao, Z., Xu, Z., Chen, Z., Wu, Z., Zheng, Z., Wang, Z., Huang, Z., Zhu, Z., Song, Z.: Seed1.5-vl technical report (2025)
22. Hao, Y., Gu, J., Wang, H.W., Li, L., Yang, Z., Wang, L., Cheng, Y.: Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark (2025)
23. Hong, W., Cheng, Y., Yang, Z., Wang, W., Wang, L., Gu, X., Huang, S., Dong, Y., Tang, J.: Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models (2025)
24. Hong, Y., Yi, L., Tenenbaum, J.B., Torralba, A., Gan, C.: Ptr: A benchmark for part-based conceptual, relational, and physical reasoning (2021)

25. Hu, Q., Wang, A., Song, J., Qiu, D., Liu, Q., Su, J.: Boosting visual knowledge-intensive training for lvlms through causality-driven visual object completion (2025)
26. Huang, J., Zhang, J.: A survey on evaluation of multimodal large language models (2024)
27. Izadi, A., Banayeeanzade, M.A., Askari, F., Rahimiakbar, A., Vahedi, M.M., Hasani, H., Baghshah, M.S.: Visual structures helps visual reasoning: Addressing the binding problem in vlms (2025)
28. Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., Yi, L.: Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models (2025)
29. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning (2025)
30. Kafle, K., Kanan, C.: An analysis of visual question answering algorithms (2017)
31. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images (2016)
32. Kim, J., Ji, H.: Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. p. 6187–6207. Association for Computational Linguistics (2024)
33. Li, H., Qu, H., Zhang, X.: Ddfav: Remote sensing large vision language models dataset and evaluation benchmark (2024)
34. Li, H., Jiang, B., Naehu, A., Song, R., Zhang, J., Tjandrasuwita, M., Ekbote, C., Chen, S.S., Balachandran, A., Dai, W., Chang, R., Liang, P.P.: Puzzleworld: A benchmark for multimodal, open-ended reasoning in puzzlehunts (2025)
35. Li, J., Lu, W., Fei, H., Luo, M., Dai, M., Xia, M., Jin, Y., Gan, Z., Qi, D., Fu, C., Tai, Y., Yang, W., Wang, Y., Wang, C.: A survey on benchmarks of multimodal large language models (2024)
36. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
37. Lin, C., Lyu, H., Xu, X., Luo, J.: Ins-mmbench: A comprehensive benchmark for evaluating lvlms' performance in insurance (2025)
38. Liu, H., Li, C., Lee, Y.J.: Llava-next: Improved reasoning for large vision-language models. arXiv preprint arXiv:2406.08637 (2024)
39. Liu, H., Li, C., Li, Y., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
40. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2024)
41. Liu, S., Ying, K., Zhang, H., Yang, Y., Lin, Y., Zhang, T., Li, C., Qiao, Y., Luo, P., Shao, W., Zhang, K.: Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models (2024)
42. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024)
43. Liu, Z., Gao, K., Liang, S., Xiao, B., Qiao, L., Ma, L., Jiang, T.: Beyond the visible: Benchmarking occlusion perception in multimodal large language models (2025)
44. Lu, J., Rao, J., Chen, K., Guo, X., Zhang, Y., Sun, B., Yang, C., Yang, J.: Evaluation and enhancement of semantic grounding in large vision-language models (2024)

45. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts (2024)
46. Lu, Y., Li, R., Jing, L., Wang, J., Du, X., Guo, Y., Ruozzi, N., Xiang, Y.: Multimodal reference visual grounding (2025)
47. Lyu, Z., Zhang, D., Ye, W., Li, F., Jiang, Z., Yang, Y.: Jigsaw-puzzles: From seeing to understanding to reasoning in vision-language models (2025)
48. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning (2022)
49. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images (2020)
50. Mayer, J., Ballout, M., Jassim, S., Nezami, F.N., Bruni, E.: ivispar – an interactive visual-spatial reasoning benchmark for vlms (2025)
51. Małkiński, M., Mańdziuk, J.: Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. ACM Computing Surveys **57**(7), 1–36 (Feb 2025)
52. Nagar, A., Jaiswal, S., Tan, C.: Zero-shot visual reasoning by vision-language models: Benchmarking and analysis (2024)
53. Pham, T.H., Nguyen, P.V., Hung, D.T., Duong, B.T., Thanh, V.N., Ngo, C., Truong, T.Q., Hy, T.S.: Iqbench: How "smart" are vision-language models? a study with human iq tests (2025)
54. Plaat, A., Wong, A., Verberne, S., Broekens, J., van Stein, N., Back, T.: Multi-step reasoning with large language models, a survey (2025)
55. Qian, Y., Wan, C., Jia, C., Yang, Y., Zhao, Q., Gan, Z.: Prism-bench: A benchmark of puzzle-based visual tasks with cot error detection (2025)
56. Qiao, R., Tan, Q., Dong, G., Wu, M., Sun, C., Song, X., GongQue, Z., Lei, S., Wei, Z., Zhang, M., Qiao, R., Zhang, Y., Zong, X., Xu, Y., Diao, M., Bao, Z., Li, C., Zhang, H.: We-math: Does your large multimodal model achieve human-like mathematical reasoning? (2025)
57. Ren, Y., Tertikas, K., Maiti, S., Han, J., Zhang, T., Süsstrunk, S., Kokkinos, F.: Vgrp-bench: Visual grid reasoning puzzle benchmark for large vision-language models. arXiv preprint arXiv:2503.23064 (2025)
58. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge (2022)
59. Song, Y., Ou, T., Kong, Y., Li, Z., Neubig, G., Yue, X.: Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge (2025)
60. Tang, L., Kim, G., Zhao, X., Lake, T., Ding, W., Yin, F., Singhal, P., Wadhwa, M., Liu, Z.L., Sprague, Z., Namuduri, R., Hu, B., Rodriguez, J.D., Peng, P., Durrett, G.: Chartmuseum: Testing visual reasoning capabilities of large vision-language models (2025)
61. Team, C., Yue, Z., Lin, Z., Song, Y., Wang, W., Ren, S., Gu, S., Li, S., Li, P., Zhao, L., Li, L., Bao, K., Tian, H., Zhang, H., Wang, G., Zhu, D., Cici, He, C., Ye, B., Shen, B., Zhang, Z., Jiang, Z., Zheng, Z., Song, Z., Luo, Z., Yu, Y., Wang, Y., Tian, Y., Tu, Y., Yan, Y., Huang, Y., Wang, X., Xu, X., Song, X., Zhang, X., Yong, X., Zhang, X., Deng, X., Yang, W., Ma, W., Lv, W., Zhuang, W., Liu, W., Deng, S., Liu, S., Chen, S., Yu, S., Liu, S., Wang, S., Ma, R., Wang, Q., Wang, P., Chen, N., Zhu, M., Zhou, K., Zhou, K., Fang, K., Shi, J., Dong, J., Xiao, J., Xu, J., Liu, H., Xu, H., Qu, H., Zhao, H., Lv, H., Wang, G., Zhang, D., Zhang, D., Zhang, D., Ma, C., Liu, C., Cai, C., Xia, B.: Mimo-vl technical report (2025)
62. Teney, D., Wang, P., Cao, J., Liu, L., Shen, C., van den Hengel, A.: V-prom: A benchmark for visual reasoning using visual progressive matrices (2019)

63. Wang, D., Zhu, D.: Benchmarking the ancient books capability of multimodal large language models. npj Heritage Science **13** (07 2025)
64. Wang, H., Li, X., Huang, Z., Wang, A., Wang, J., Zhang, T., Zheng, J., Bai, S., Kang, Z., Feng, J., Wang, Z., Zhang, Z.: Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology (2025)
65. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., Ding, M., Tang, J.: Lvbench: An extreme long video understanding benchmark (2024)
66. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025)
67. Wang, Z., Sun, J., Zhang, W., Hu, Z., Li, X., Wang, F., Zhao, D.: Benchmarking multimodal mathematical reasoning with explicit visual dependency (2025)
68. Wang, Z., Zhu, J., Tang, B., Li, Z., Xiong, F., Yu, J., Blaschko, M.B.: Jigsaw-r1: A study of rule-based visual reinforcement learning with jigsaw puzzles (2025)
69. Wu, C., Lian, S., Liu, Z., Zhang, L., Yang, L.T., Chen, K.: Dynasolidgeo: A dynamic benchmark for genuine spatial mathematical reasoning of vlms in solid geometry (2025)
70. Wu, Q., Yang, X., Zhou, Y., Fang, C., Song, B., Sun, X., Ji, R.: Grounded chain-of-thought for multimodal large language models (2025)
71. Xia, P., Han, S., Qiu, S., Zhou, Y., Wang, Z., Zheng, W., Chen, Z., Cui, C., Ding, M., Li, L., Wang, L., Yao, H.: Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models (2025)
72. Xiao, L., Yang, X., Lan, X., Wang, Y., Xu, C.: Towards visual grounding: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence p. 1–20 (2026)
73. Xiao, Y., Sun, E., Liu, T., Wang, W.: Logicvista: Multimodal llm logical reasoning benchmark in visual contexts (2024)
74. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step (2024)
75. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models (2025)
76. Xu, W., Wang, J., Wang, W., et al.: Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models (2025)
77. Xu, Y., Li, C., Zhou, H., Wan, X., Zhang, C., Korhonen, A., Vulić, I.: Visual planning: Let's think only with images (2025)
78. Xu, Y., Chen, L., Zhang, L., Wang, W., Jin, Q.: Polychartqa: Benchmarking large vision-language models with multilingual chart question answering (2025)
79. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., Kendre, S.: xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872 (Aug 2024)
80. Yan, H., Liu, X., Wang, H., et al.: Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning (2025)

81. Yang, B., Wen, B., Ding, B., Liu, C., Chu, C., Song, C., Rao, C., Yi, C., Li, D., Zang, D., Yang, F., Zhou, G., Zhang, G., Shen, H., Peng, H., Ding, H., Wang, H., Fan, H., Ju, H., Huang, J., Cao, J., Chen, J., Hua, J., Chen, K., Jiang, K., Tang, K., Gai, K., Wei, M., Wang, Q., Wang, R., Na, S., Zhang, S., Mao, S., Huang, S., Zhang, T., Gao, T., Chen, W., Yuan, W., Wu, X., Hu, X., Lu, X., Zhang, Y.F., Yang, Y., Chen, Y., Lu, Z., Wu, Z., Ling, Z., Yang, Z., Li, Z., Xu, D., Gao, H., Li, H., Wang, J., Ren, L., Hu, Q., Wang, Q., Wang, S., Luo, X., Li, Y., Hu, Y., Zhang, Z.: Kwai keye-vl 1.5 technical report (2025)
82. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces (2025)
83. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions (2016)
84. Yuan, F., Yan, Y., Jiang, Y., Zhao, H., Feng, T., Chen, J., Lou, Y., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: Gsm8k-v: Can vision language models solve grade school math word problems in visual contexts (2025)
85. Zeng, Y., Huang, W., Huang, S., Bao, X., Qi, Y., Zhao, Y., Wang, Q., Chen, L., Chen, Z., Chen, H., Ouyang, W., Zhao, F.: Agentic jigsaw interaction learning for enhancing visual perception and reasoning in vision-language models (2025)
86. Zha, Y., Zhou, K., Wu, Y., Wang, Y., Feng, J., Xu, Z., Hao, S., Liu, Z., Xing, E.P., Hu, Z.: Vision-g1: Towards general vision language reasoning with multi-domain data curation (2025)
87. Zhan, Y., Zhu, Y., Zheng, S., Zhao, H., Yang, F., Tang, M., Wang, J.: Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning (2025)
88. Zhang, H., Guo, H., Guo, S., Cao, M., Huang, W., Liu, J., Zhang, G.: Ing-vp: Mllms cannot play easy vision-based games yet. arXiv preprint arXiv:2410.06555 (2024)
89. Zhang, K., Yang, C., Wen, Z., Yuan, S., Wang, Q., Huang, C., Zhu, G., Wang, H., Lu, H., Wen, J., Jiao, J., Luo, L., Liu, L., Wu, S., Zhu, X., Zhang, X., Zhang, G., Lin, Y., Shi, G., Fu, C., Huang, W.: Mme-cc: A challenging multi-modal evaluation benchmark of cognitive capacity (2025)
90. Zhang, Y., Bai, H., Zhang, R., Gu, J., Zhai, S., Susskind, J., Jaitly, N.: How far are we from intelligent visual deductive reasoning? (2024)
91. Zhang, Y., Zhang, K., Li, B., et al.: Worldqa: Multimodal world knowledge in videos through long-chain reasoning (2024)
92. Zhang, Z., Chen, Z., Zhang, Z., Sun, Y., Tian, Y., Jia, Z., Li, C., Liu, X., Min, X., Zhai, G.: Puzzlebench: A fully dynamic evaluation framework for large multimodal models on puzzle solving (2025)
93. Zhao, Y., Liu, H., Long, Y., Zhang, R., Zhao, C., Cohan, A.: Financemath: Knowledge-intensive math reasoning in finance domains (2024)