

VERIFY: A Benchmark of Visual Explanation and Reasoning for Investigating Multimodal Reasoning Fidelity

Jing Bi¹, JunJia Guo¹, Guangyu Sun², Susan Liang¹, Luchuan Song¹, Yolo Y. Tang¹, Jinxi He^{1,4*}, Jiarui Wu^{1,4}, Ali Vosoughi¹, Chen Chen², Jason J. Corso³, Chenliang Xu¹

¹University of Rochester, Rochester, NY, USA

²University of Central Florida, Orlando, FL, USA

³University of Michigan, Ann Arbor, MI, USA

⁴Carnegie Mellon University, Pittsburgh, PA, USA

{jing.bi, jguo40, sliang22, lsong11, yunlong.tang}@rochester.edu

{ali.vosoughi, chenliang.xu}@rochester.edu, {ginh, jiaruiwu}@andrew.cmu.edu

guangyu@ucf.edu, chen.chen@crcv.ucf.edu, jjcorso@umich.edu

Abstract

Visual reasoning is central to human cognition. Although recent Multimodal Large Language Models (MLLMs) have demonstrated impressive performance across language and vision-language tasks, existing benchmarks primarily measure recognition-based skills. To our knowledge, VERIFY is the first to jointly provide: (i) post-graduate exam-style items expressed primarily as diagrams; (ii) human-authored, per-item stepwise reasoning, supporting process-level evaluation beyond final-answer accuracy; and (iii) a public evaluation harness for reproducible, process-aware scoring. This design decouples perception/recognition from higher-order abstraction and deduction, enabling precise diagnosis of where models fail. Across strong MLLMs, we observe substantial headroom under our process metrics and distinct failure modes (e.g., Abstraction $\rightarrow$ Deduction), despite superficially fluent explanations. We release the dataset, annotation guidelines, baselines, and quality audits including inter-annotator agreement, to facilitate transparent adoption.

1 Introduction

Visual reasoning is a fundamental component of human intelligence, shaping how individuals acquire knowledge and interpret the world (Ittelson, 1960; Kosslyn, 1996; Pylyshyn, 2003). As a core mechanism for concept understanding, it has been the subject of extensive research (Małkiński & Mańdziuk, 2025). However, classic models, often trained on small-scale datasets, struggle to generalize beyond narrow domains (Lake et al., 2017; Marcus, 2018; Szegedy et al., 2013; Małkiński & Mańdziuk, 2025). In contrast, recent advances in MLLM (Li et al., 2023; Zhu et al., 2023; Bi et al., 2024; Liu et al., 2023a; Maaz et al., 2023; Tang et al., 2024b; Zhang et al., 2023) have achieved near-human performance in various vision-language tasks (Liang et al., 2024; Caffagni et al., 2024; Tang et al., 2023; Hua et al., 2024a; Tang et al., 2025). However, existing benchmarks predominantly evaluate recognition-based skills, such as object detection, image captioning, and optical character recognition, which primarily rely on perceptual abilities (Zang et al., 2024; Wu & Xie, 2024; Lin et al., 2024; Dong et al., 2024).

Recent studies have started evaluating MLLM performance in more cognitively demanding domains, including reasoning and chart-based visual understanding (Lu et al., 2024; Masry et al., 2022), which require reasoning of numerical and logical operations. However, these benchmarks often conflate reasoning ability with domain knowledge drawn from the underlying LLM, making it hard to isolate and accurately assess reasoning based solely

*Jinxi He and Jiarui Wu completed part of this work at the University of Rochester and are now at Carnegie Mellon University.

Table 1: **Comparison of datasets** with an added column indicating whether *stage-by-stage reasoning checks* are explicitly performed and whether the base items are *human-authored exam-style problems* (“Human annotated”). RP marks the presence of *human-annotated reasoning paths/rationales*. Only **VERIFY** performs explicit *stage reasoning* verification (P/R/A/D). [†]For ARC, H-ARC reports 76.2% (train) and 64.2% (eval). [‡]VCR provides human-labeled rationale *choices* rather than free-form chains. [§]MMMU reports 11,500 items in total, but most evaluations use the public MMMU_val split of 900 examples; we therefore list 900 as the size to reflect standard practice.

Name	Size	Source	RP	Stage reasoning	Human annotated	Modality	Audience	Human	Year
SVRT (Fleuret et al., 2011)	23	Synthesized	×	×	×	Images	Adults (UC)	–	2011
Raven’s Prog. Matrices (RAPM) (Raven et al., 1989)	48	Psychometric exam	×	×	✓	Images	Adults (above-average IQ)	–	1989
RAVEN (Zhang et al., 2019)	14,000	Synthesized	×	×	×	Images	Psychology students	84%	2019
PGM (RPM) (Barrett et al., 2018)	1,44M (train/val/test)	Synthesized	×	×	×	Images	–	–	2018
VAEC (Webb et al.) (Webb et al., 2020)	~228k	Synthesized	×	×	×	Images	–	–	2020
ARC (Chollet, 2019)	600	Synthesized (programmatic)	×	×	×	Images	Crowdworkers	76.2% [†]	2019
MARVEL (Jiang et al., 2024)	770	Web-curated	×	×	×	Images	Puzzle solvers	68.9%	2024
IQTest (Lu et al., 2024)	228	Online puzzle platforms	×	×	×	Images+Text	General (online)	–	2023
MM-IQ (Cai et al., 2025)	2,710	Mixed (exam/web)	×	×	×	Images	General	51.3%	2025
VQA-X (Park et al., 2018)	–	VQA v2 subset	✓	×	×	Image+Text	General	–	2018
VCR (Zellers et al., 2019)	290,000+	Movies (web)	✓ [‡]	×	×	Image+Text	General	–	2019
ScienceQA (Lu et al., 2022)	21,208	K-12 curricula	✓	×	×	Image+Text	K-12	90.2%	2023
MMMU (Yue et al., 2024a)	900 (val) [§]	Exams, textbooks	×	×	×	Image+Text	College / professional	–	2023
GPQA (Diamond) (Rein et al., 2023)	198	Expert-curated	✓	×	✓	Text-only	Graduate-level	–	2023
VERIFY (ours)	600	National Exam	✓	✓	✓	Image+Text	UG & Grad	48.4%	2025

on visual information. Additionally, most evaluations emphasize accuracy as the primary metric while neglecting the decision-making quality and the depth of the reasoning process.

To address this gap, we introduce **VERIFY** (Visual Explanation and Reasoning for Investigating Multimodal Reasoning Fidelity), a novel benchmark designed to rigorously evaluate the reasoning fidelity of state-of-the-art MLLMs. **VERIFY** comprises a diverse and challenging set of visual reasoning tasks that require both generalization and abstraction, as shown in Figure 2.

Compared with prior visual-reasoning suites such as SVRT, RAVEN, PGM/RPM, VAEC, ARC, MARVEL, and MM-IQ (see Table 1), **VERIFY** is the largest benchmark that jointly provides a unique **intersection** of (i) graduate-level, exam-sourced *diagram-centric* visual reasoning, (ii) multimodal inputs where the diagram is indispensable, and (iii) *expert-verified, stage-wise* reasoning paths (Perception/Recognition/Abstraction/Deduction) for process-aware diagnosis.

Prior Rationale-rich datasets such as VCR, VQA-X/ACT-X, A-OKVQA, ScienceQA, and e-SNLI-VE provide brief process supervision but are not designed as graduate-level, cross-disciplinary exams (Zellers et al., 2019; Park et al., 2018; Schwenk et al., 2022; Lu et al., 2022; Do et al., 2020). GQA uses structured programs, not natural-language reasoning (Hudson & Manning, 2019). GPQA and GPQA-Diamond target graduate-level difficulty but are text-only and much smaller (Rein et al., 2023; 2024). In contrast, **VERIFY** is a curated, test-only suite with dense, multi-step, expert-verified reasoning paths, following the high-signal evaluation philosophy of RAPM, ARC (Table 1). Recent multimodal benchmarks address related but distinct goals. MMMU-Pro (Yue et al., 2024b) focuses on knowledge-intensive, shortcut-resistant evaluation but lacks diagram-centric, process-annotated reasoning. MM-Star (Chen et al., 2024b) emphasizes vision indispensability, not process supervision. MathVerse (Zhang et al., 2024a) targets visual mathematics with step-level evaluation but is domain-specific. MM-IQ (Cai et al.,

Question

Choose the most appropriate option to fill in the question mark so that the figures follow a certain pattern.

THIS

HELLO

HAIR

?

FAIR

BOOK

EYE

BYE

A
B
C
D

Answer: A

Human Reasoning Path

Since the composition of the figures differs and there is no obvious attribute rule, priority should be given to considering numerical rules. The first figure in the question stem contains a single curve, so we consider the number of curves. Upon observation, it is found that each figure in the question stem has only one curve, hence the ? (blank) figure should also have only one curve. Only option A meets this criterion.

Pattern: Line Property

OPENAI-o1 Correctness: ✗ Recognition: 0, Abstraction: 2, Deduction: 1

A concise way to see why “FAIR” fits is to note that in each of the given words, if you drop the first letter you still (more or less) get a recognizable English word:

- THIS → HIS
- HELLO → ELLO (“ello” is at least a colloquial form)
- HAIR → AIR

Of the answer choices, only “FAIR” continues that pattern (FAIR → AIR). The others (“BOOK,” “EYE,” “BYE”) do not reduce to a standard English word by removing just the first letter. Hence “FAIR” is the best fit.

This answer reveals **Semantic Dominance Bias** in model reasoning. To overcome this, letters must be perceived as shapes rather than symbols. Like the well-known “How many ‘r’s in strawberry?” puzzle, the model struggles recognizing letters visually.

Figure 1: This example illustrates that current MLLMs predominantly rely on simple visual cues (such as letters) for reasoning, often overlooking more complex patterns involving shapes or line properties. **VERIFY** provides human-annotated reasoning paths, enabling deeper evaluation of the reasons and circumstances under which models fail.

2025) offers broad AVR/IQ-style coverage but lacks expert-verified, stage-wise traces. In contrast, VERIFY uniquely provides process-annotated, exam-derived, diagram-centric reasoning diagnostics absent from prior suites. Moreover, a preregistered cross-regional study with participants from five geographic regions, all completing the assessment in English, finds that regional coefficients in multivariable logistic regression have 91% confidence intervals overlapping zero, providing no evidence of cultural bias.

Furthermore, drawing on foundational studies in human visual reasoning (Fukunaga, 1990; Duda et al., 2000; Russell & Norvig, 2020; Barsalou, 1999), we propose novel metrics that decompose reasoning into four stages, as illustrated in Figure 3. These metrics address limitations in current automated evaluation methods (Tyen et al., 2023), which often fail to capture nuanced differences between models. By comparing with human-annotated reasoning paths, VERIFY provides deeper insights into why models may misinterpret specific visual elements.

Leveraging this benchmark, we systematically evaluate contemporary MLLMs to obtain a comprehensive analysis of their visual reasoning capabilities and failure modes. Our findings show that even the strongest model we test (Qwen3-VL-32B-Instruct) reaches only 31.7% accuracy (190/600), far below a human baseline of 48.4% and only modestly better than random guessing (25%). Qualitative analysis reveals a critical insight: *models adept at perceiving visual details often fail at abstraction or deduction* (Figure 1), frequently latching onto superficial cues or partial patterns instead of constructing globally consistent explanations. Process-level evaluation further shows that, for advanced visual reasoning models, the reasoning path is more informative than the final answer: models frequently explore multiple reasoning trajectories during problem-solving, but these trajectories often diverge from human reasoning.

- **VERIFY:** We introduce a novel benchmark that treats visual reasoning as a multi-stage cognitive process, providing 600 graduate-level, diagram-centric items and human-annotated, stage-wise reasoning paths. This enables both answer-based and process-based evaluation of MLLMs under a single, standardized protocol.
- **Beyond Accuracy:** We propose new metrics and analysis tools that score models at the level of reasoning stages (Perception, Recognition, Abstraction, Deduction), report binomial and Wilson confidence intervals, and provide text-only controls and image-perturbation ablations. These tools support a nuanced and comprehensive evaluation of visual reasoning beyond traditional accuracy.
- **Comprehensive Evaluation:** We benchmark state-of-the-art MLLMs, spanning both proprietary and open-source. Despite this breadth, the best model (Qwen3-VL-32B-Instruct) attains only 31.7% accuracy (190/600) versus a human baseline of 48.4%, revealing substantial headroom and offering detailed insights into models’ strengths, weaknesses, and process-level failure modes.

2 Related Work

Visual Reasoning. Visual reasoning entails understanding and manipulating visual information to infer logical relationships (Agrawal et al., 2016; Goyal et al., 2017). Early benchmarks such as VQAv1 and VQAv2 (Agrawal et al., 2016; Goyal et al., 2017) centered on object-based question answering, while CLEVR (Johnson et al., 2017) and GQA (Hudson & Manning, 2019) introduced compositional, programmatic queries for structured reasoning. Expanding beyond static images, video datasets like CoPhy and Space (Baradel et al., 2020; Janny et al., 2022; Duan et al., 2021) focus on physical and causal inference in dynamic scenes. Our work is closely related to *abstract* visual reasoning, which does not rely on external domain knowledge. Datasets such as RAVEN, SVRT, CVR, Bongard-HOI, and Bongard-LOGO (Zhang et al., 2019; Fleuret et al., 2011; Zerroug et al., 2022; Jiang et al., 2023; Nie et al., 2020) explore relational patterns among shapes and visual elements. However, as noted by Van der Maas et al. (2021), many existing datasets have rigid configurations and limited reasoning paradigms, restricting their ability to generalize.

Reasoning Benchmarks. Most widely used reasoning benchmarks focus on text-based question answering, mathematics, or code inference (Cobbe et al., 2021; Hendrycks et al., 2021; Srivastava et al., 2022; Suzgun et al., 2022), where step-by-step solutions can be

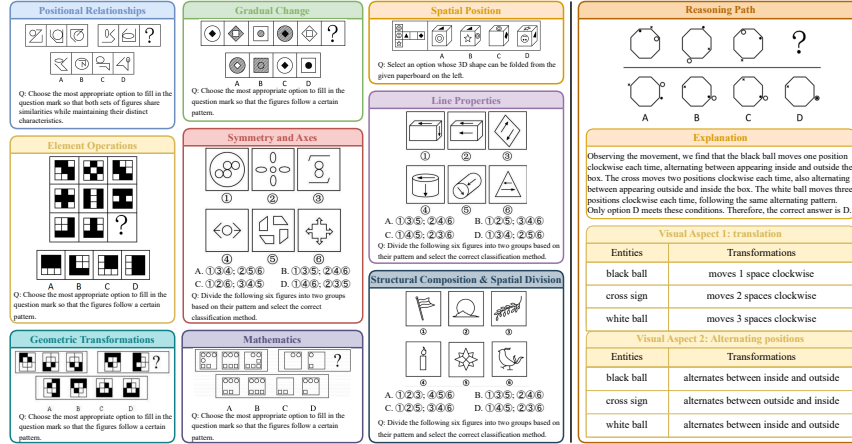


Figure 2: Categories from VERIFY cover a range of patterns, from logical operations to 3D geometry and mathematics. The right panel presents a human reasoning path, demonstrating how visual transformations and rotations lead to the answer.

programmatically verified. In the *visual* domain, benchmarks such as CLEVR (Johnson et al., 2017), GQA (Hudson & Manning, 2019), CLEVRER (Yi et al., 2019), Bongard (Nie et al., 2020), RAVEN (Zhang et al., 2019), MMComposition (Hua et al., 2024b), and VidComposition (Tang et al., 2024a) introduce compositional reasoning tasks over images or videos. Recent works (Jiang et al., 2025) extend these efforts by incorporating multi-modal chain-of-thought (CoT) prompting, often blending external knowledge with visual cues. Among the most relevant is MM-IQ (Cai et al., 2025), which focuses on abstract visual puzzles but does not provide annotations to clarify reasoning transitions.

3 Data

We construct our items in the spirit of Raven’s Progressive Matrices: non-verbal, diagrammatic pattern-completion tasks designed to probe domain-general abstract reasoning rather than curriculum-specific knowledge. Source problems are adapted from eight years of high-stakes, postgraduate-level national civil-service reasoning exams. From this raw pool, we remove items that (i) require external factual knowledge, (ii) allow multiple plausible answers, (iii) are underspecified or ambiguous, or (iv) are not primarily diagram-driven. Only items with a unique, self-contained solution—dependent solely on the visual stimulus and general logic—are retained. This results in a benchmark with real-world difficulty, controlled ambiguity, and minimal risk of domain or cultural bias. Further details on curation and filtering are provided in the appendix.

Leakage and contamination audit. Because the source material is drawn from large-scale examinations, we audit VERIFY for contamination. Reverse image search over all 600 diagrams (Google Lens and TinEye) returned no exact match; only 8 items (1.3%) had a visually similar look-alike, none carrying reasoning paths or appearing in a known MLLM benchmark, and all images are MD5-deduplicated. More importantly, what VERIFY scores—the stage-level Recognition/Abstraction/Deduction paths—was authored by our annotators and never published online, so the evaluation target cannot have been memorized. The results are also inconsistent with leakage: the best model reaches only 31.7% against a 48.4% human baseline, and frontier proprietary systems do not lead.

Comparison to adjacent benchmarks. Prior multimodal evaluation resources often satisfy only one (or at most two) of the following desiderata: (i) strong diagram-centric abstract reasoning, (ii) rigorous human stepwise rationales suitable for process evaluation, and (iii) real-world exam-sourced difficulty with careful ambiguity control. VERIFY is constructed to occupy this missing intersection by pairing exam-derived diagram reasoning with stepwise traces and a process-aware evaluation harness.

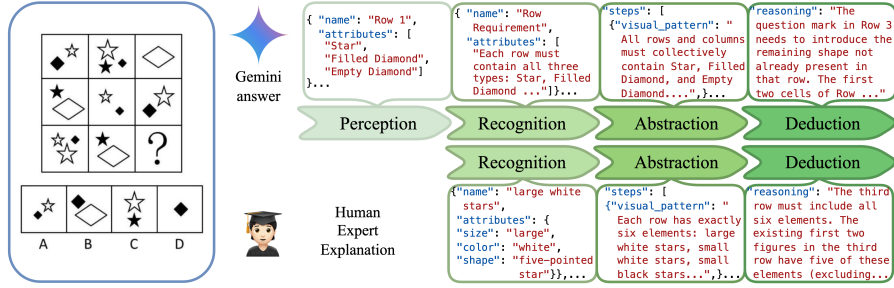


Figure 3: We divide the reasoning into four key stages inspired by human visual reasoning: perception, recognition, abstraction, and deduction. Unlike general visual tasks, where perception involves detecting raw visual features, humans often have implicit perception because the provided visual elements are already structured for useful components. Even for shown complex problems, a model with strong visual abilities—like Gemini—can effectively analyze patterns and logical structures to determine the correct answer.

MMMU-Pro (Yue et al., 2024b) improves robustness in multi-discipline multimodal tasks by filtering text-only questions and augmenting evaluation, but focuses on college-level knowledge and lacks diagrammatic pattern reasoning with expert stepwise traces. In contrast, VERIFY targets post-graduate difficulty with detailed human rationales. MMStar (Chen et al., 2024b) emphasizes vision-indispensable selection and leakage-aware evaluation, but does not provide process supervision for stage-level reasoning. MathVerse (Zhang et al., 2024a) covers diagram-based visual mathematics with step-level evaluation, but is domain-specific and does not address broader post-graduate exam-style reasoning. MM-IQ (Cai et al., 2025) offers a broad AVR/IQ-style suite for abstraction testing; VERIFY complements it with expert-verified stepwise traces (P/R/A/D) and process-aware scoring to diagnose multimodal reasoning failures beyond final-answer accuracy.

Core Novelty

VERIFY is the first benchmark that *combines* (1) real-world, exam-sourced *postgraduate* diagram-centric abstract reasoning with careful ambiguity control; (2) expert-verified, step-by-step traces suitable for process supervision and evaluation; and (3) a public, process-aware evaluation harness enabling stage-level, diagnostic scoring and error localization beyond accuracy.

3.1 Reasoning Path

A distinctive feature of our dataset is the inclusion of human-annotated reasoning paths that explicitly document each step leading to the correct answer. Unlike previous work (Zhang et al., 2024b; Cai et al., 2025), which typically relies on generated reasoning paths with only the final answer verified by reviewers, our approach ensures a higher standard of reliability. To guarantee both accuracy and clarity, each reasoning path is independently annotated by three experts, with any discrepancies resolved through consensus.

Solving visual reasoning problems involves a structured cognitive process. Humans first extract meaningful feature representations through perception and feature extraction (Fukunaga, 1990). These features are then matched against learned patterns to recognize objects and scenes (Duda et al., 2000). Once the elements are identified, they are abstracted into symbolic models that capture relationships and context. Finally, deductive inference is applied to these abstractions to draw conclusions and guide decision-making (Russell & Norvig, 2020). To provide a structured framework for evaluation, we define four key stages in the visual reasoning process, as illustrated in Figure 3: 1. **Perception**: Processing visual input to detect features such as shapes, colors, and orientations; 2. **Recognition**: Selecting the task-relevant visual aspects and entities—answering *which elements matter?* (e.g., “the relevant elements are the black dot, the white dot, and the cross”); 3. **Abstraction**: Converting those aspects into a rule or relation—answering *what rule relates them?* (e.g., “the black dot moves one step clockwise”, or “the count alternates by parity”); 4. **Deduction**: Applying logical reasoning to infer missing details or predict outcomes based on the abstractions.

Table 2: Taxonomy and distribution of visual reasoning patterns in VERIFY, sorted by frequency. Mathematics (MA) is the most prevalent category at 24.3%, followed by Element Operations (EO, 12.6%) and Positional Relationships (PR, 12.5%). The remaining categories are more evenly distributed, indicating broad coverage of diverse visual reasoning patterns.

Code	%	Definition
MA	24.3	<i>Mathematics</i> : Focuses on numerical aspects, including element counts and repetitive patterns.
EO	12.6	<i>Element Operations</i> : Captures interactions between shapes, such as layering, merging, splitting, and reassembly.
PR	12.5	<i>Positional Relationships</i> : Focuses on how elements relate in terms of position, alignment, and directional shifts.
GT	10.2	<i>Geometric Transformations</i> : Transformations like rotation, translation, and reflection, occurring in sequences.
SA	9.2	<i>Symmetry and Axes</i> : Identifies various types of symmetry, including axial, central, and rotational.
LP	9.2	<i>Line Properties</i> : Analyzes stroke characteristics such as type, angle, continuity, and curvature.
SC	8.7	<i>Structural Composition</i> : Examines enclosures, layering, connections, and shading to reveal deeper structural patterns.
GC	8.5	<i>Gradual Change</i> : Tracks progressive changes in shape, size, or position, forming logical patterns.
SP	5.1	<i>Spatial Position</i> : Describes element arrangement and positional shifts via movement, rotation, or progression.

Table 3: **LLM judge for reasoning pairs under direct comparison.** Although direct comparison yields only **15% accuracy**, the judge still assigns consistently high scores to both strong and weak reasoning paths. This suggests that the model fails to reliably distinguish reasoning quality through direct comparison alone.

r1	r2	deepseek-r1:32b		llama3.3	
		Score _{r+c}	Score _r	Score _{r+c}	Score _r
o1	GTR+P	7.85	6.40	8.05	8.20
o1	GTR+	7.60	6.80	8.15	8.35
o1	GTR	7.80	6.35	7.95	8.20
GTR	GTR+P	8.45	7.15	9.05	9.40
GTR	GTR+	8.70	8.00	9.25	9.85
o1	GTR+P	8.10	7.80	—	—

For humans, the Perception stage is often implicit and may even be bypassed, as some initially relevant visual elements may later prove irrelevant (Treisman & Gelade, 1980; Wolfe & Horowitz, 2017). As a result, Recognition becomes the first explicit step, where annotators identify only the essential visual aspects and structure the reasoning path accordingly (Wardle & Baker, 2020; Ghafari et al., 2022). After annotation, we summarize and identify key visual reasoning patterns for evaluation, as illustrated in Figure 2. To address these limitations, we refined and expanded the set of reasoning patterns, building upon existing categories while introducing clearer structural distinctions as shown in Table 2.

4 Evaluation

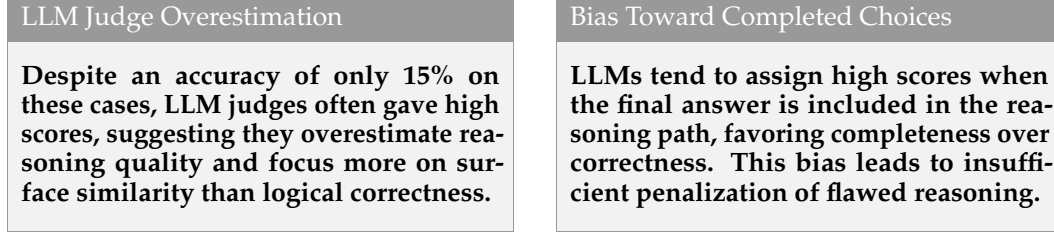
We initially adopt the standard *LLM-as-a-judge* protocol, comparing model-generated reasoning to human references. However, our analysis indicates that plain text judging can overestimate reasoning quality by rewarding superficial textual properties (e.g., verbosity and fluency) rather than logical correctness, consistent with recent findings (Zhou et al., 2024; Chen et al., 2024a). This motivates a more structured evaluation approach. Recent work advocates *structured* pipelines in which an LLM first extracts instance-specific criteria or evidence and scoring is performed component-wise: decomposed criteria-based evaluation improves correlation with expert judgments relative to single-score pointwise judging (Yu et al., 2025), while form-filling and rubric-based schemes further improve human alignment and interpretability (Liu et al., 2023b; Hashemi et al., 2024). Similarly, claim-level factuality metrics decompose outputs into atomic, verifiable units to obtain more reliable assessments (Min et al., 2023; Song et al., 2024). Motivated by these results, we propose a stage-wise framework that decomposes multimodal reasoning into distinct stages and evaluates each stage independently, rather than relying on end-to-end textual similarity.

4.1 Failure Cases of Existing Evaluation Metrics

To understand the limitations of current plain text judging methods, we first tested three leading LLMs—DeepSeek-R1:32B, DeepSeek-R1:70B, and LLaMA3.3-70B—as automated judges. We randomly selected 120(20%) examples from our dataset and used the o1 model’s response as the candidate reasoning path. We compare four types of reasoning paths: 1. **o1**: Model-generated reasoning path (OpenAI-o1). 2. **GTR**: Ground-truth (Human-annotated)

reasoning path. 3. **GTR+**: GTR decomposed into Recognition, Abstraction, and Deduction stages. 4. **GTR+P**: GTR+ with additional human-labeled pattern details. We score model responses on a 0–10 scale using two approaches: 1. **r+c**: the final answer +c(choice). 2. **r**: Only Reasoning steps.

Table 3 shows a striking discrepancy: human evaluators marked only 6 out of 40 answers as correct, yet LLMs consistently assigned scores in the range of 6.35–9.85. This disparity indicates that LLMs tend to overestimate reasoning quality, often relying on superficial similarity rather than a rigorous evaluation of logical correctness. Notably, LLaMA3.3-70B assigns comparably high scores to both GTR and its structured variants (GTR+ and GTR+P), while it awards slightly lower scores to o1.



4.2 Proposed Evaluation Framework

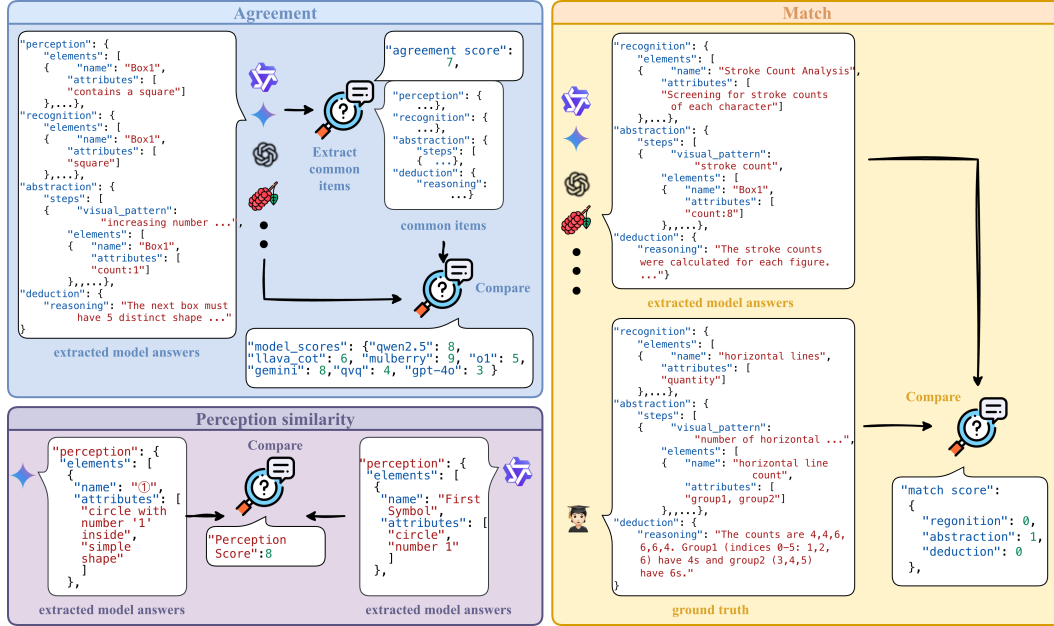


Figure 4: This diagram illustrates proposed metrics: Agreement, which extracts common elements across model and compares their consistency; Match, which assesses alignment between model and ground truth across Recognition, Abstraction, and Deduction; and Similarity, which measures the similarity of extracted perception attributes across models.

4.2.1 Human-Validated Evaluation

A central goal of VERIFY is to evaluate intermediate reasoning *faithfully*, rather than inadvertently rewarding verbosity, stylistic fluency, or a judge model’s preferences. We therefore design our evaluation to be (i) grounded in *human-authored stage-level references* whenever ground truth exists, and (ii) explicitly *validated against expert human judgments* for every component that relies on model-based extraction or comparison. Concretely, we standardize the response format across all evaluated models, extract stage-specific structured representations, and then score them against human annotations. We further report answer accuracy throughout, ensuring that any process-level score is interpreted against ground-truth.

We conduct a targeted alignment study on a random sample of 100 problems. First, we evaluate the reliability of stage segmentation/extraction: two expert annotators independently judge whether the extracted segment for each stage is (i) correctly localized to the intended stage and (ii) semantically faithful to the original explanation. Across 300 stage traces (3 stages $\times$ 100 items), we obtain Cohen’s $\kappa = 0.82$ for localization and $\kappa = 0.78$ for faithfulness; the extractor matches the human majority vote in 91.5% of cases. Second, we validate the *matching/judging* component directly. The same participants rate the semantic similarity of structured reasoning paths for the 100 sampled problems (3 stages each), yielding 3,600 representation pairs. Human similarity scores and model-predicted similarity exhibit strong agreement (Pearson $r = 0.68$). Together, these results demonstrate that our end-to-end evaluator (extraction + matching) closely tracks expert human judgments, providing an empirically grounded basis for stage-level scoring.

4.2.2 Evaluation of Perception Stage

Perception ($\hat{P}$) lacks item-level ground truth, so we report two unsupervised diagnostics: 1. *Common Element Agreement*: Measures how consistently models mention salient visual elements, indicating perceptual reliability. 2. *Inter-Model Perception Similarity*: Computes pairwise similarity of extracted elements, revealing shared biases and failure modes across models. Crucially, neither Agreement nor Similarity is treated as a correctness criterion, and neither awards credit for matching other models. Instead, they summarize perceptual consistency and correlation structure, helping diagnose *how* models fail. All claims of correctness in our benchmark are grounded in (i) answer accuracy and (ii) stage-level comparisons to human-authored ground truth in Recognition, Abstraction, and Deduction.

4.2.3 Evaluation of Recognition, Abstraction, and Deduction Stages

For Recognition, Abstraction, and Deduction, we evaluate correctness via *direct matching to human-authored stage-level ground truth* (Figure 4). We first extract structured representations from model explanations and then compare them to the corresponding human annotations using our stage-wise metrics. This design ensures that credit is assigned only when a model’s intermediate representation aligns with the human reference under an explicit rubric. We consider three candidate LLMs for the evaluation (Qwen3-32B, DeepSeek-R1, and QWQ). Based on the human-alignment study above, QWQ achieves the strongest agreement with expert judgments, and we therefore adopt it as the primary model for stage-level extraction and scoring. Additional implementation details, prompts, rubrics, and qualitative examples are provided in the Appendix.

5 Experiments and Results

Table 4: Performance comparison of MLLMs with or without CoT support.

Model	COT	GT	MA	GC	LP	EO	PR	SA	SC	SP	ALL
Open-Source MLLMs											
Qwen3-VL-32B-Instruct	×	0.311	0.267	0.365	0.418	0.255	0.412	0.300	0.260	0.360	0.317
Qwen3-VL-32B-Thinking	✓	0.295	0.293	0.308	0.327	0.345	0.490	0.167	0.240	0.413	0.315
Qwen3-VL-8B-Thinking	✓	0.377	0.200	0.288	0.364	0.236	0.314	0.267	0.315	0.333	0.302
Qwen3-VL-4B-Thinking	✓	0.197	0.253	0.327	0.400	0.236	0.353	0.367	0.233	0.373	0.290
InternVL3.5-4B	×	0.180	0.280	0.231	0.364	0.226	0.353	0.333	0.281	0.307	0.282
InternVL3.5-8B-think	✓	0.311	0.360	0.212	0.236	0.145	0.275	0.300	0.226	0.347	0.267
InternVL3.5-4B-think	✓	0.262	0.307	0.288	0.255	0.218	0.216	0.367	0.185	0.333	0.257
QVQ	✓	0.197	0.280	0.288	0.200	0.382	0.235	0.167	0.240	0.293	0.257
Qwen3-VL-8B-Instruct	×	0.180	0.280	0.288	0.273	0.327	0.255	0.167	0.240	0.267	0.255
LLaVA-COT	✓	0.197	0.267	0.269	0.218	0.182	0.216	0.300	0.253	0.360	0.253
Qwen3-VL-4B-Instruct	×	0.279	0.200	0.308	0.309	0.309	0.314	0.233	0.185	0.227	0.248
Mulberry	✓	0.230	0.240	0.212	0.273	0.218	0.157	0.267	0.295	0.240	0.245
Qwen2.5(72B)	×	0.230	0.267	0.212	0.218	0.273	0.235	0.367	0.219	0.213	0.238
InternVL3.5-8B	×	0.262	0.227	0.231	0.200	0.182	0.314	0.300	0.205	0.253	0.233
Proprietary MLLMs											
O1	✓	0.279	0.320	0.308	0.291	0.273	0.275	0.233	0.295	0.373	0.300
Gemini-2.5	✓	0.311	0.293	0.192	0.345	0.291	0.275	0.367	0.219	0.227	0.267
GPT-4o	×	0.246	0.227	0.269	0.200	0.255	0.353	0.267	0.219	0.227	0.243

In this section, we evaluate the visual reasoning abilities of MLLMs with or without CoT. During the initial phase of our evaluation, we tested a range of open-source and smaller models on a randomly selected set of 40 questions. The results revealed that these models struggled, exhibiting both low accuracy and flawed reasoning. Recognizing these limitations, we refined our approach by selecting the most advanced open-source models specifically designed for reasoning: LLaVA-CoT (11B), Mulberry (8B), and QVQ (72B) (Team, 2024), alongside Qwen2.5-72B, a cutting-edge vision model. For proprietary models, we included GPT-4o, as well as two of the most recent leading models: OpenAI-o1 (OpenAI, 2024)

Table 5: Per-pattern diagnostics. *Top*: human–MLLM accuracy gap, in points. *Bottom*: failure-stage distribution (%) over 4,958 incorrect responses from 17 models; codes follow Table 2.

	MA	EO	PR	GT	SA	LP	SC	GC	SP	ALL
Human – best MLLM	+9.0	+12.8	−9.0	+9.3	+12.3	+0.2	+20.5	+16.5	+8.7	+16.7
Recognition failure	74.2	68.1	65.2	65.8	67.5	76.0	72.0	73.9	66.3	—
Abstraction failure	18.6	22.8	26.2	22.5	27.1	15.2	20.9	21.0	22.1	—
Deduction failure	4.3	4.7	6.1	8.4	3.5	3.3	3.3	3.1	8.4	—
Other	3.0	4.4	2.5	3.3	1.9	5.5	3.8	2.0	3.2	—

Perception similarity and agreement between models. Figure 6 presents the similarity matrix, where each model achieves maximal self-alignment (diagonal score of 10) and models are clearly grouped by family. Variants within the same family—such as the Qwen3-VL 4B/8B/32B Instruct vs. Thinking models and the Internvl3.5 4B/8B vs. think variants—show higher mutual similarity than with unrelated systems, indicating that our metric captures family-level inductive biases. Cross-family structure is also meaningful: for example, LLaVA-CoT and Mulberry (4.55), and Gemini and GPT-4o (4.28), form particularly aligned pairs, while QVQ remains relatively isolated, reflecting a distinct way of processing visual inputs. Figure 7 visualizes the average agreement on common objects, where each model is represented by a colored segment and a larger segment indicates a stronger tendency to agree with the others. Proprietary models such as GPT-4o and o1 occupy longer portions of the bar, reflecting higher consistency in identifying shared visual elements, while models like Qwen2.5 correspond to shorter segments, showing moderate levels of agreement.

5.3 Where Does Reasoning Break Down?

The scatter plots above are qualitative, so we quantify both transitions over 7,317 model-item stage records from 17 models. Contrary to the impression the plots give, *Abstraction and Deduction are the more tightly coupled pair*: the median Spearman correlation is $\rho(A, D)=0.696$ against $\rho(R, A)=0.433$, a paired gap of $\Delta\rho = -0.271$ (95% bootstrap CI $[-0.282, -0.219]$, excluding zero), and no model (0/17) shows the reverse ordering. Pearson correlations agree (0.688 vs. 0.439).

Recognition→Abstraction is the dominant bottleneck. Thresholding stage success at 7/10, models that recognize the relevant visual aspects still fail to abstract the governing rule in 65.0% of cases ($N=2,371$), whereas models that form a correct abstraction fail at deduction in only 45.9% of cases ($N=1,666$)—a 19.1-point conditional-failure gap that holds at thresholds 6/10 and 8/10 and reverses for just 1/17 models. The stages are nonetheless separable rather than one latent quality axis: mean stage scores on correct versus incorrect answers are 5.36/4.76 (R), 5.44/4.34 (A), and 5.93/3.49 (D), a gap that widens down the pipeline.

The failure stage depends on pattern type. Failure-stage distributions differ significantly across pattern categories ($\chi^2(24)=82.0$, $p<10^{-4}$, Cramér’s $V=0.074$, a reliable but modest effect). As Table 5 shows, recognition failures peak on LP, MA, and GC; abstraction failures on SA, PR, and EO, where symmetry and positional relations are hard to convert into explicit rules; deduction failures are rarer and concentrate in GT, SP, and PR.

Humans and models fail differently. Humans outperform the best MLLM in 8 of 9 categories and by 16.7 points overall, but the gap is far from uniform. Models are near parity on LP (+0.2) and ahead on PR (−9.0), yet fall well behind on SC, GC, EO, and SA. The categories that separate them most demand holistic abstraction over the whole figure—structure, symmetry, gradual change—which humans grasp intuitively, whereas PR rewards the local, pattern-matchable rules models handle best. VERIFY therefore surfaces distinct reasoning profiles rather than a single difficulty ranking.

6 Conclusion

We introduced VERIFY, a benchmark for evaluating visual reasoning. Our experiments show that current models often struggle to combine perceptual grounding and logical reasoning, excelling at one but not both. While chain-of-thought reasoning improves accuracy, models still face challenges moving from abstract understanding to deduction.

Acknowledgment

This research was funded, in part, by the U.S. Government under ARPA-H contract 1AY2AX000062 and the Center of Excellence in Data Science and Artificial Intelligence, an Empire State Development-designated Center of Excellence. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. URL <https://arxiv.org/abs/1505.00468>.
- Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics, 2020. URL <https://arxiv.org/abs/1909.12000>.
- David G. T. Barrett, Felix Hill, Adam Santoro, Ari S. Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 511–520, 2018.
- Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4): 577–660, 1999.
- Jing Bi, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, and Chenliang Xu. Eagle: Egocentric aggregated language-video engine. *arXiv preprint arXiv:2409.17523*, 2024.
- Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*, 2024.
- Huanqia Cai, Yijun Yang, and Winston Hu. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models, 2025.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474. URL <https://aclanthology.org/2024.emnlp-main.474/>.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024b.
- François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve: Corrected visual-textual entailment with natural language explanations, 2020.
- Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption, 2024. URL <https://arxiv.org/abs/2405.19092>.

- Jiafei Duan, Samson Yu Bai Jian, and Cheston Tan. Space: A simulator for physical interactions and causal learning in 3d environments, 2021. URL <https://arxiv.org/abs/2108.06180>.
- Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, New York, NY, 2nd edition, 2000.
- François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.
- Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Boston, MA, 2nd edition, 1990.
- Tara Ghafari, Ali Rahimpour Jounghani, and Hossein Esteky. Where and when matter in visual recognition. *Attention, Perception, & Psychophysics*, 85:404–417, 2022. doi: 10.3758/s13414-022-02607-y.
- Google DeepMind. Gemini 2.0 flash thinking, 2024. URL <https://deepmind.google/technologies/gemini/flash-thinking/>.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL <https://arxiv.org/abs/1612.00837>.
- Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.745. URL <https://aclanthology.org/2024.acl-long.745/>.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*, 2024a.
- Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024b.
- Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- William H Ittelson. Visual space perception. 1960.
- Steeven Janny, Fabien Baradel, Natalia Neverova, Madiha Nadri, Greg Mori, and Christian Wolf. Filtered-cophy: Unsupervised learning of counterfactual physics in pixel space, 2022. URL <https://arxiv.org/abs/2202.00368>.
- Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mmcot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.
- Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions, 2023. URL <https://arxiv.org/abs/2205.13803>.

- Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning, 2024. URL <https://arxiv.org/abs/2404.13591>.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.
- Stephen M Kosslyn. *Image and brain: The resolution of the imagery debate*. MIT press, 1996.
- Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.
- Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodal large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.
- Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL <https://aclanthology.org/2023.emnlp-main.153/>.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafford, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.
- Mikołaj Małkiński and Jacek Mańdziuk. Deep learning methods for abstract visual reasoning: A survey on raven’s progressive matrices. *ACM Computing Surveys*, 57(7):1–36, 2025.
- Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL <https://arxiv.org/abs/2203.10244>.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in*

- Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL <https://aclanthology.org/2023.emnlp-main.741/>.
- Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. *Advances in Neural Information Processing Systems*, 33:16468–16480, 2020.
- OpenAI. Learning to reason with llms, 2024. URL <https://openai.com/index/learning-to-reason-with-llms/>.
- Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Zenon W Pylyshyn. *Seeing and visualizing: It’s not what you think*. MIT press, 2003.
- John C. Raven, John H. Court, and John Raven. *Standard Progressive Matrices*. Oxford Psychologists Press, Oxford, UK, 1989.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa-diamond subset. <https://arxiv.org/abs/2311.12022>, 2024. Hardest 198-question subset of GPQA.
- Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, London, England, 4th edition, 2020.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022*. Springer, 2022. doi: 10.1007/978-3-031-20074-8_9.
- Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.552. URL <https://aclanthology.org/2024.findings-emnlp.552/>.
- AaroHi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023.
- Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, et al. Vidcomposition: Can mllms analyze compositions in compiled videos? *arXiv preprint arXiv:2411.10979*, 2024a.

- Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. *arXiv preprint arXiv:2403.16276*, 2024b.
- Yunlong Tang, Junjia Guo, Pinxin Liu, Zhiyuan Wang, Hang Hua, Jia-Xing Zhong, Yunzhong Xiao, Chao Huang, Luchuan Song, Susan Liang, et al. Generative ai for cel-animation: A survey. *arXiv preprint arXiv:2501.06250*, 2025.
- Qwen Team. Qvq: To see the world with wisdom, December 2024. URL <https://qwenlm.github.io/blog/qvq-72b-preview/>.
- Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.
- Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. *arXiv preprint arXiv:2311.08516*, 2023.
- Han LJ Van der Maas, Lukas Snoek, and Claire E Stevenson. How much intelligence is there in artificial intelligence? a 2020 update. *Intelligence*, 87:101548, 2021.
- Susan G. Wardle and Chris I. Baker. Recent advances in understanding object recognition in the human brain: deep neural networks, temporal dynamics, and context. *F1000Research*, 9:F1000 Faculty Rev–590, 2020. doi: 10.12688/f1000research.22296.1.
- Taylor W. Webb, Zachary Dulberg, Steven M. Frankland, Alexander A. Petrov, Randall C. O’Reilly, and Jonathan D. Cohen. Learning representations that support extrapolation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, 2020.
- Jeremy M. Wolfe and Todd S. Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1:0058, 2017. doi: 10.1038/s41562-017-0058.
- Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.
- Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.
- Fangyi Yu, Nabeel Seedat, Drahomira Herrmannova, Frank Schilder, and Jonathan Richard Schwarz. Beyond pointwise scores: Decomposed criteria-based evaluation of LLM responses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1931–1954, Suzhou (China), November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-industry.136. URL <https://aclanthology.org/2025.emnlp-industry.136/>.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 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 (CVPR)*, 2024a.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2024b.
- Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, pp. 1–19, 2024.

- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *arXiv preprint arXiv:2206.05379*, 2022.
- Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5317–5327, 2019.
- Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024a.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024b. URL <https://arxiv.org/abs/2302.00923>.
- Hongli Zhou, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, and Tiejun Zhao. Mitigating the bias of large language model evaluation. In Sun Maosong, Liang Jiye, Han Xianpei, Liu Zhiyuan, and He Yulan (eds.), *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pp. 1310–1319, Taiyuan, China, July 2024. Chinese Information Processing Society of China. URL <https://aclanthology.org/2024.cc1-1.101/>.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.