Beyond Final Answers: CRYSTAL Benchmark for Transparent Multimodal Reasoning Evaluation

Wayner Barrios and SouYoung Jin

Dartmouth College, Hanover, NH, USA
{wayner.j.barrios.quiroga.gr, souyoung.jin}@dartmouth.edu

Abstract. Modern multimodal large language models (MLLMs) achieve impressive results on vision-language benchmarks, yet existing evaluations judge only final answers, making shortcuts indistinguishable from genuine understanding. We introduce **CRYSTAL** (***C****lear* ***R****easoning via* ***Y****ielded* ***S****teps,* ***T****raceability and* ***L****ogic*), a diagnostic benchmark with 6,372 instances that evaluates multimodal reasoning through verifiable intermediate steps. We propose two complementary metrics: *Match F1*, which scores step-level precision and recall via semantic similarity matching, and *Ordered Match F1*, which further penalizes disordered reasoning chains. References are constructed through a Delphi-inspired pipeline where four independent MLLMs generate trajectories, aggregated via semantic clustering and validated through human quality gates. Evaluation of 20 MLLMs, including commercial frontier systems not used during benchmark construction, reveals systematic failures invisible to accuracy: universal cherry-picking (precision far exceeds recall), non-monotonic scaling trade-offs, and disordered reasoning, where even the strongest reasoners preserve under 60% of matched steps in correct order. Beyond evaluation, we propose the **Causal Process Reward (CPR)**, a multiplicative reward that couples answer correctness with step-level alignment, and **CPR-Curriculum**, which progressively increases reasoning difficulty during training. CPR-Curriculum achieves +32% Match F1 via GRPO where additive reward strategies fail, improving reasoning without manual step annotation.

Keywords: Multimodal reasoning · Vision-language models · Benchmark · Process evaluation · Reinforcement learning

1 Introduction

Modern multimodal large language models (MLLMs) have achieved impressive performance on vision-language benchmarks by integrating pretrained visual encoders with large language models [24, 42, 46]. Datasets such as MathVista [24] consolidate diverse mathematical reasoning tasks, while RealWorldQA [46] challenges models with spatial understanding in real-world images. Yet a critical limitation persists: *these benchmarks judge performance by final answers alone.*

Input image

Q: Which of the 3 objects is the smallest?
A. Right B. Left C. Middle **Ground-Truth: C**

Previous Benchmarks *(Answer-Only)*

Model Answer: **C** ✓ **Accuracy: 100%** *(reasoning invisible)*

CRYSTAL *(Answer + Step-Level Evaluation)*

Model's Predicted Steps	Match	
Three gaming consoles on desk	✓	**Contradiction!**
Middle is ***larger*** *than others*	×	Claims *larger*
Most powerful based on size	×	but answers
Final Answer: C	✓	*smallest*

Match F1: 0.15 **Accuracy: 100%**

Fig. 1: The lucky guess problem. LLaVA-v1.6-7B answers correctly (C) but contradicts itself by claiming the middle console is *larger* while selecting it as smallest. Previous benchmarks score 100%; CRYSTAL compares the model's predicted steps against reference reasoning steps via Match F1 (0.15), exposing flawed reasoning.

Without observing intermediate steps, shortcuts become indistinguishable from genuine understanding [55]. Recent theoretical analysis shows that answer-centric evaluation structurally incentivizes hallucination by penalizing models that signal uncertainty [19]. This motivates evaluation frameworks that assess *how* models reason, not merely whether answers are correct.

Consider the example in Figure 1 from RealWorldQA, which asks "*Which of the 3 objects is the smallest?*" given an image of three Xbox consoles. A model answers correctly (C: middle console), but its reasoning reveals a critical contradiction: it states the middle console is *larger* than the others while claiming it is the smallest. Traditional benchmarks award full credit to this "lucky guess" despite fundamentally flawed reasoning (Match F1: 0.15). When reasoning trajectories remain unobserved, systematic errors in perception or logic go undetected, and models exploit evaluation shortcuts rather than develop genuine understanding. Recent work addresses this by separating rationale generation from answer inference [5, 55]. However, elicited steps typically lack structured checkpoints for machine-verifiable evaluation [48], and rarely expose *where* in the reasoning chain a failure originates [55].

We introduce **CRYSTAL** (***C****lear* ***R****easoning via* ***Y****ielded* ***S****teps,* ***T****raceability and* ***L****ogic*), a diagnostic benchmark with verifiable intermediate checkpoints.[1] Each instance contains reasoning steps scored via two novel metrics: *Match F1*, which evaluates step-level precision and recall through semantic similarity matching, and *Ordered Match F1*, which further penalizes disordered reasoning chains. By exposing the intermediate steps, CRYSTAL makes visible *where* a model's

[1] Benchmark and data: `https://huggingface.co/datasets/waybarrios/CRYSTAL`; code: `https://github.com/SEE-AI-Lab/crystal`.

reasoning breaks down, whether at perception or inference, rather than collapsing every failure into a single binary outcome.

We construct references through a Delphi-inspired pipeline [30]: four independent MLLMs from different families generate trajectories, which are aggregated via semantic clustering and validated by a fifth model plus human quality gates (Section 3.1).

Beyond evaluation, CRYSTAL's step-level references enable a new training paradigm. Standard reinforcement learning rewards combine accuracy and reasoning as independent additive terms, allowing models to maximize accuracy through guessing while ignoring reasoning quality. We propose the **Causal Process Reward (CPR)**, which multiplicatively couples both objectives: the model receives full reward only when it answers correctly *and* produces faithful reasoning steps. We further introduce **CPR-Curriculum**, which stabilizes learning by progressively increasing reasoning difficulty during training.

Our contributions: **(i)** CRYSTAL, a diagnostic benchmark with 6,372 instances containing verifiable intermediate reasoning steps for fine-grained evaluation; **(ii)** Match F1 and Ordered Match F1, two complementary metrics that measure step-level reasoning quality via semantic similarity matching and penalize disordered reasoning chains, respectively; **(iii)** CPR and CPR-Curriculum, multiplicative reward strategies that improve both accuracy and Match F1 via GRPO [37] without manual step annotation; **(iv)** evaluation of 20 MLLMs revealing universal cherry-picking behavior, persisting even in commercial frontier systems not used during benchmark construction.

2 Related Work

Answer-Centric Benchmarks. Recent evaluation suites remain largely answer-centric despite broader scope. *MathVista* and *MathVision* probe mathematical reasoning with 3,040 competition-level problems across 16 disciplines [24, 42], *MMMU* targets expert-level multi-discipline understanding [52], while *SUGAR-CREPE* and *VisMin* isolate compositional failures through minimal-pair contrastive examples [1, 15]. *MMEvalPro* calibrates multimodal benchmarks with triplet-based evaluation to reduce lucky guessing [17], and *LIME* demonstrates that carefully curated subsets can match full-benchmark signal at lower cost [57]. Despite their breadth, these benchmarks score only final answers, limiting visibility into intermediate reasoning.

Process Evaluation & Decoupling. Complementary work evaluates reasoning processes through intermediate steps. *Multimodal-CoT* separates rationale generation from answer inference to reduce hallucination [55], *Visual CoT* and *Visual Sketchpad* provide step-annotated rationales and diagrammatic chains of thought [16,36], while *MME-CoT* benchmarks CoT quality and robustness [18]. *CoMT* requires multimodal outputs (creation, deletion, update, selection) [5], and *MINERVA* evaluates complex video reasoning with LLM judges for multi-step temporal inference [29]. Decoupling approaches like *Prism* and *ViperGPT* separate perception from symbolic reasoning [33, 40], and *MPBench* emphasizes

Table 1: Key statistics of CRYSTAL.

Statistic	Value	Source	Count
Total examples	6,372	MathVision [42]	3,039 (47.7%)
Avg. steps	11.6	ScienceQA-IMG [25]	2,017 (31.7%)
Step range	3–42	RealWorldQA [46]	765 (12.0%)
Easy / Med / Hard	48.5 / 44.1 / 7.4%	MMVP [41]	299 (4.7%)
Annotation	Multi-agent + human	PlotQA [28]	252 (3.9%)

trajectory-level scoring [48]. These motivate machine-checkable, step-level verification of reasoning chains.

Evaluation Methodology & Hallucination. Binary answer-centric evaluation structurally incentivizes hallucination by penalizing abstention over guessing, favoring confident incorrect predictions over calibrated uncertainty [19]. This misaligns evaluation with trustworthy deployment. Our work addresses this by evaluating reasoning processes: Match F1 rewards semantically justified steps while penalizing spurious reasoning (low precision) and incomplete coverage (low recall), aligning incentives with transparent, verifiable inference.

Robustness & Reliability. Trustworthy reasoning requires robustness and calibration. *MultiTrust* covers truthfulness, safety, and fairness across 32 tasks [54]. MLLMs remain vulnerable to adversarial attacks despite instruction tuning [6], while calibration methods like *Self-Calibrated Tuning* and *CaRot* improve OOD detection [31, 50]. Real-world benchmarks reveal persistent gaps in long-horizon multimodality [53], and compositional probes expose sensitivity to minimal changes [1, 15]. Our design emphasizes deterministic step graders and trajectory coherence at the process level.

3 CRYSTAL Dataset & Benchmark

We introduce **CRYSTAL** (***C**lear **R**easoning via **Y**ielded **S**teps, **T**raceability and **L**ogic*), a diagnostic benchmark that evaluates multimodal reasoning *step by step*. Unlike traditional VQA benchmarks that only assess final answer correctness, CRYSTAL provides each question with a sequence of natural language reference reasoning steps that capture the intermediate inferences required to arrive at the correct answer. These references are generated through a multi-agent framework (Section 3.1) that aggregates outputs from independent MLLMs via semantic clustering, ensuring diverse yet high-quality reasoning paths. CRYSTAL spans 6,372 questions covering visual perception, compositional reasoning, spatial relations, counting, and logical inference. Evaluation compares predicted steps against references using semantic similarity matching (Match F1) alongside answer accuracy, enabling fine-grained analysis of *where* and *why* models fail. Table 1 summarizes key statistics.

Figure 2 illustrates representative examples from CRYSTAL spanning diverse reasoning scenarios. Each instance combines visual inputs with natural language

RealWorldQA

Q: In which direction is the dog traveling?
GT: Left to Right

Reference Steps (6):
1. Person walking dog across street.
2. Dog's head turned toward right.
3. Body orientation indicates rightward movement.
4. Motion blur suggests active movement.
5. Leash extends from left to right.
6. Dog traveling left to right.

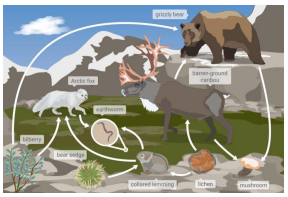
MMVP

Q: Can you see the key "Z" in the image?
GT: No

Reference Steps (10):
1. The image shows a close-up of keyboard keys.
2. Foreground keys include a key labeled '9' with 'Pg Up' text.
3. Another visible key is labeled '6' with a right arrow.
4. Other keys in view carry navigation labels rather than letters.

[*6 more steps*]

ScienceQA



Q: Which organism contains matter from lichen?
GT: Mushroom

Reference Steps (15):
1. Food web shows energy flow.
2. Arrows = matter transfer direction.
3. Lichen at bottom level.
4. Arrow from lichen to mushroom.
5. Mushroom consumes lichen matter.
6. Mushroom contains lichen matter.

[*9 more steps*]

Fig. 2: CRYSTAL spans diverse multimodal reasoning scenarios. Three examples: **(Left)** RealWorldQA (spatial understanding), **(Middle)** MMVP (fine-grained perception), **(Right)** ScienceQA (multi-hop logic). Parentheses give the reference step count per example.

questions and verifiable step-by-step reasoning paths, enabling fine-grained evaluation of where models succeed (correct perception, valid inference) versus where they fail (hallucinated objects, logical errors).

3.1 Multi-Agent Reasoning Step Generation

Each input consists of a question Q, an image I, and the correct answer A. Our goal is to produce an ordered sequence of minimal reasoning steps necessary to derive A from Q and I. The pipeline comprises four phases with two iterative quality gates inspired by the Delphi method [30].

Phase 1: Independent generation. Four open-source MLLMs from different architecture families (Qwen2.5-VL-72B [43], InternVL3-76B [56], Gemma3-27B [12], and Llama-4-Maverick [27]) independently generate candidate reasoning steps from the triplet (Q, I, A), using distinct random seeds to reduce correlated errors [3, 21].

Phase 2: Semantic clustering and ordering. We embed all candidate steps using a sentence encoder $f(\cdot)$ and compute pairwise cosine similarity. Steps with similarity $\geq \tau$ form edges in an undirected graph; connected components define semantic clusters $\{\mathcal{C}_k\}$, each grouping paraphrastic formulations of the same

reasoning step. For each cluster, we select the representative minimizing average within-cluster dissimilarity:

$$x_k^\star \;=\; \arg\min_{x \in \mathcal{C}_k} \frac{1}{|\mathcal{C}_k|} \sum_{y \in \mathcal{C}_k} \big(1 - \mathrm{sim}(x, y)\big). \tag{1}$$

Representatives are ordered into a coherent reasoning chain. On the first Delphi round, ordering follows the original question logic. On each subsequent round t, the ordering is updated by minimizing edit distance to the previous round's sequence, stabilizing step order across refinements. This clustering effectively implements self-consistency voting [45] at the step level.

Phase 3: Automated validation. A fifth MLLM (Molmo-72B [9]) validates logical soundness, sequence coherence, visual grounding in I, and consistency with answer A. Failed examples re-enter Phase 1 with fresh seeds (*Iteration Loop 1*): new candidate steps are generated, re-clustered, and re-validated, mirroring Delphi consensus building where contributors revise proposals after aggregated feedback [30].

Phase 4: Human quality gate. A trained annotator verifies that perceptual claims are visible in I, reasoning transitions are logically sound, and executing the steps yields A. Rejected examples restart from Phase 1 (*Iteration Loop 2*); fewer than 5% require re-iteration. Together, both loops form a closed refinement cycle that mirrors multi-annotator label aggregation [7, 34]. The supplementary material visualizes the overall pipeline workflow and provides prompt templates, annotator guidelines, and the annotation interface.

3.2 Evaluation Metrics

We evaluate models via two metrics: *Match F1* measures step-level reasoning quality through semantic similarity matching, while *Accuracy* measures final answer correctness.

Match F1. Given predicted steps $\mathcal{P}_i = \{p_{ij}\}$ and reference steps $\mathcal{G}_i = \{g_{ik}\}$ for example i, we compute pairwise cosine similarity using sentence encoder $f(\cdot)$:

$$S_{jk} = \cos\big(f(p_{ij}), f(g_{ik})\big). \tag{2}$$

Steps are matched via greedy 1:1 assignment over pairs with $S_{jk} \geq \tau$ (threshold), producing a set of matched pairs $\mathcal{A}_i$. We define true positives $\mathrm{TP}_i = |\mathcal{A}_i|$, false positives $\mathrm{FP}_i = |\mathcal{P}_i| - \mathrm{TP}_i$, and false negatives $\mathrm{FN}_i = |\mathcal{G}_i| - \mathrm{TP}_i$. Precision and recall are:

$$\mathrm{Prec}_i = \frac{\mathrm{TP}_i}{\max\{|\mathcal{P}_i|, 1\}}, \quad \mathrm{Rec}_i = \frac{\mathrm{TP}_i}{\max\{|\mathcal{G}_i|, 1\}}. \tag{3}$$

Per-example F1 combines precision and recall via harmonic mean (0 if both denominators are non-empty yet $\mathrm{TP}_i = 0$; 1 if $|\mathcal{P}_i| = |\mathcal{G}_i| = 0$). Dataset-level **Match F1** is the macro-average over all N evaluation examples:

$$\text{Match-F1} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{F1}_i. \tag{4}$$

We use `all-distilroberta-v1` [35] with $\tau = 0.35$ (selected via ablation in Section 4.4).

Ordered Match F1. We extend Match F1 with the Longest Increasing Subsequence (LIS) ratio [10] to penalize disordered chains. Given $m = |\mathcal{A}_i|$ matched pairs, we sort them by reference index and extract the corresponding prediction indices $(j_1, \ldots, j_m)$. The LIS ratio measures the fraction of these indices that form a non-decreasing subsequence:

$$\text{LIS-ratio}_i = \frac{\text{LIS}(j_1, \ldots, j_m)}{m}. \tag{5}$$

Ordered Match F1 combines content quality with ordering:

$$\text{Ordered-F1}_i = \text{F1}_i \cdot \big((1 - \alpha) + \alpha \cdot \text{LIS-ratio}_i\big), \tag{6}$$

with $\alpha = 0.3$. When $\alpha = 0$, this reduces to standard Match F1.

Accuracy. Answer correctness uses type-adapted comparison: tolerance-based matching for numerics, exact matching for categoricals, and LLM-as-judge for free-form text, macro-averaged over all examples.

3.3 Causal Process Reward

Standard reward formulations for reinforcement learning from reasoning combine accuracy and reasoning quality additively (e.g., $R = w_f R_{\text{fmt}} + w_a R_{\text{acc}} + w_r R_{\text{reason}}$, where w_f, w_a, w_r weight format compliance, answer accuracy, and reasoning quality respectively; see the supplementary material), allowing the accuracy term to dominate and the model to maximize reward by guessing correctly while omitting reasoning steps. We propose the **Causal Process Reward (CPR)**, which uses a multiplicative interaction that *causally* links answer correctness to step-level alignment:

$$R_{\text{CPR}} = \begin{cases} a_w + s_w \cdot \text{F1}_{\text{step}} & \text{if answer correct} \\ s_w \cdot \text{F1}_{\text{step}} \cdot \lambda & \text{otherwise} \end{cases} \tag{7}$$

where a_w is the answer bonus, s_w is the step bonus (weighting reasoning quality), F1_{step} is the step-level Match F1 between predicted and reference steps, and $\lambda = 0.3$ penalizes wrong answers. Under this formulation, a correct guess without reasoning receives only a_w (no step bonus), while good reasoning with a wrong answer is heavily discounted, ensuring neither objective can be independently maximized. We extend CPR with **CPR-Curriculum**, a two-phase training strategy inspired by staged reward shaping [8]. Phase 1 trains with only format and accuracy rewards (no reasoning signal) to establish stable answer generation. Phase 2 initializes from the Phase 1 checkpoint and introduces the full CPR reward at a lower learning rate, preserving the accuracy foundation while adding reasoning supervision. Within Phase 2, we apply progressive difficulty scheduling: training begins with examples having fewer reference steps (simpler reasoning chains) and gradually introduces examples with more steps

(complex multi-hop reasoning), preventing early-stage collapse before the model learns to generate structured reasoning. Since accuracy and reasoning can produce conflicting gradients, both strategies use PCGrad [51] to project conflicting gradient components, preventing one objective from undermining the other during optimization.

4 Experiments

We evaluate 20 MLLMs (16 open-source, 1B–38B; 4 commercial) on CRYSTAL. Our experiments address: **(i)** step-level performance beyond accuracy, **(ii)** cherry-picking in commercial frontier models, **(iii)** reasoning step ordering, and **(iv)** step-level rewards for training.

4.1 Experimental Setup

Dataset. CRYSTAL contains 6,372 questions from five benchmarks (MathVision [42] 47.7%, ScienceQA-IMG [25] 31.7%, RealWorldQA [46] 12.0%, MMVP [41] 4.7%, PlotQA [28] 3.9%), averaging 11.6 reasoning steps per question (Table 1). Questions are stratified by complexity: easy (48.5%, 8.6 steps), medium (44.1%, 13.3 steps), hard (7.4%, 20.5 steps), using model-agnostic features (step count, question length, linguistic markers).
Baselines. We evaluate 20 MLLMs spanning open-source and commercial systems. *Open-source* (16 models): Qwen2.5-VL (3B, 7B, 32B) [43], Qwen3-VL (2B, 8B, 32B) [2], InternVL3.5 (1B to 38B) [44], Gemma3 (4B, 12B) [12], Llama 3.2-11B [26], LLaVA-v1.6 (7B) [23], and MiniCPM-v2.6 (8B) [49]. *Commercial* (4 models): GPT-5, GPT-5-mini, GPT-5.2 Instant [32], and Gemini 2.5 Flash [13]. All models use official endpoints or pretrained checkpoints without task-specific fine-tuning. Crucially, commercial models were *not* used in the CRYSTAL reference generation pipeline, providing an unbiased assessment of whether observed patterns generalize to frontier systems.
Metrics. *Accuracy* measures final answer correctness via fuzzy matching; *Match F1* evaluates step-level reasoning through semantic similarity matching (all-distilroberta-v1, $\tau = 0.35$). Match F1 combines precision (fraction of predicted steps aligned with references) and recall (fraction of reference steps covered).
GRPO Training Data. For the GRPO training experiment [37], we use 30,312 examples from ScienceQA-IMG [25] (6,218 samples) and TextVQA [38] (24,094 samples), generated via the pipeline from Section 3.1. The supplementary material provides the complete pipeline workflow diagram, prompt templates, benchmarking settings, GRPO training configuration, and qualitative examples.

4.2 Main Results

Table 2 presents evaluation results across 20 MLLMs, revealing systematic gaps between answer correctness and reasoning transparency.

Table 2: Evaluation of 20 MLLMs on CRYSTAL. Match F1 (all-distilroberta-v1, $\tau = 0.35$): step-level reasoning quality; P/R: step precision/recall; LIS: fraction of matched steps in correct order; Ord. F1 ($\alpha = 0.3$): order-penalized Match F1. Top three per column in **green**, blue, yellow.

Model	Params	Accuracy	Match F1	P	R	Steps	LIS	Ord. F1
Commercial MLLMs								
GPT-5 [32]	n/a	**57.99%**	0.612	0.925	0.479	5.29	0.636	0.539
GPT-5-mini [32]	n/a	55.59%	**0.773**	**0.978**	0.669	7.57	0.560	**0.670**
GPT-5.2 Instant [32]	n/a	47.35%	0.564	0.974	0.416	4.64	0.648	0.501
Gemini 2.5 Flash [13]	n/a	53.95%	0.673	0.701	**0.765**	**17.10**	0.584	0.579
Qwen Family								
Qwen3-VL-8B [2]	8B	57.66%	0.659	0.827	0.590	7.37	0.624	0.572
Qwen3-VL-32B [2]	32B	49.22%	0.718	0.819	0.704	10.56	0.581	0.617
Qwen2.5-VL-32B [43]	32B	47.63%	0.653	0.943	0.524	5.86	0.619	0.572
Qwen3-VL-2B [2]	2B	34.15%	0.595	0.726	0.535	5.94	0.669	0.515
Qwen2.5-VL-7B [43]	7B	30.43%	0.475	0.765	0.365	4.07	0.717	0.422
Qwen2.5-VL-3B [43]	3B	39.85%	0.480	0.898	0.347	3.73	0.723	0.434
InternVL Family								
InternVL3.5-38B [44]	38B	51.21%	0.612	0.892	0.498	5.76	0.643	0.538
InternVL3.5-8B [44]	8B	51.98%	0.530	0.882	0.416	4.96	0.692	0.469
InternVL3.5-4B [44]	4B	37.61%	0.432	0.895	0.325	3.75	0.775	0.387
InternVL3.5-2B [44]	2B	33.02%	0.469	0.725	0.371	3.92	0.731	0.415
InternVL3.5-1B [44]	1B	30.13%	0.330	0.616	0.243	2.51	0.807	0.297
Other Open-Source								
Gemma3-12B [12]	12B	33.83%	0.605	0.838	0.499	5.51	0.673	0.534
Gemma3-4B [12]	4B	28.65%	0.618	0.878	0.506	5.72	0.668	0.547
Llama 3.2-11B [26]	11B	24.83%	0.471	0.713	0.379	4.19	0.726	0.415
LLaVA-v1.6-7B [23]	7B	24.66%	0.512	0.961	0.370	3.94	0.675	0.459
MiniCPM-v2.6-8B [49]	8B	25.54%	0.215	0.709	0.134	1.31	0.854	0.186

Finding 1: Cherry-picking is near-universal. 19 of 20 models exhibit precision substantially exceeding recall (ratios 1.2× to 5.3×), including commercial systems absent from the reference pipeline: GPT-5 (P/R = 0.925/0.479), GPT-5-mini (0.978/0.669), and GPT-5.2 Instant (0.974/0.416). The sole exception is Gemini 2.5 Flash, which generates 17.10 steps per question (vs. 11.6 reference average) with recall (0.765) exceeding precision (0.701). The persistence of cherry-picking across scales (1B to frontier) and 7 model families suggests a fundamental misalignment between training objectives and reasoning transparency [19]. Even GPT-5 (57.99% accuracy) recovers only 47.9% of reference steps. Importantly, commercial models were not part of the reference generation pipeline, yet GPT-5-mini achieves the highest F1 (0.773) and Gemini the highest recall across all 20 models, confirming that Match F1 captures genuine reasoning quality without benchmark construction bias. Low recall reflects genuine omissions rather than over-complete references: Gemini 2.5 Flash attains 0.765 recall (coverage is achievable), while models generating few steps (LLaVA-v1.6: 3.94, GPT-5.2 Instant: 4.64) still reach >0.96 precision, aligning with references yet omitting intermediate reasoning.

Finding 2: Accuracy and reasoning fidelity diverge. GPT-5 leads accuracy (57.99%) but ranks 8th in F1 (0.612), while GPT-5-mini leads F1 (0.773) at lower accuracy (55.59%). Cross-family comparisons reveal architectural choices dominate scale: Gemma3-4B (0.618 F1) outperforms InternVL3.5-38B (0.612 F1) despite 9.5× fewer parameters [22, 39]. Models like LLaVA-v1.6 (0.961 precision, 3.94 steps) and GPT-5.2 Instant (0.974 precision, 4.64 steps) achieve near-perfect step accuracy by generating fewer, safer steps, masking reasoning failures through confident omissions [4, 47]. This divergence is enabled by the structure of the evaluation: 36.4% of CRYSTAL questions are multiple-choice or binary (ScienceQA-IMG, MMVP), where models can exploit statistical regularities in answer distributions and surface-level visual cues to select correct answers without complete reasoning [11, 14]. This motivates step-level rewards that explicitly incentivize reasoning coverage (Section 4.5).

Finding 3: Scaling exhibits non-monotonic reasoning trade-offs. Within model families, scaling parameters does not uniformly improve both accuracy and reasoning quality. Qwen3-VL-32B achieves 0.718 F1 (2nd overall) but lower accuracy (49.22%) than Qwen3-VL-8B (57.66%, 0.659 F1), indicating that the larger model produces more thorough reasoning chains at the cost of answer correctness. Conversely, InternVL3.5-4B yields lower F1 (0.432) than InternVL3.5-2B (0.469) despite higher accuracy (37.61% vs. 33.02%), suggesting that scale can improve answer extraction while suppressing reasoning coverage. These non-monotonic patterns demonstrate that accuracy and reasoning quality respond differently to parameter scaling [39], reinforcing the necessity of joint evaluation through metrics like Match F1.

4.3 Reasoning Order Analysis

Match F1 measures *what* steps are present but not *whether they appear in a logically coherent sequence*. Using Ordered Match F1 (Eq. 6), which penalizes out-of-sequence steps via the LIS ratio (Eq. 5), we evaluate whether models organize reasoning chains in the correct order. The last two columns of Table 2 report LIS and Ordered F1 across all 20 MLLMs.

Finding 4: No model preserves coherent reasoning order. Among the top reasoners by Match F1 (> 0.66), no model preserves more than 60% of matched steps in the correct relative order. GPT-5-mini achieves the highest Ordered F1 (0.670) but with LIS $= 0.560$, meaning 44% of its matched steps appear out of sequence. Qwen3-VL-32B (LIS $= 0.581$) and Gemini 2.5 Flash (LIS $= 0.584$) show similar ordering degradation despite strong Match F1. Notably, LIS alone can be misleading: smaller models such as InternVL3.5-1B (LIS $= 0.807$) and MiniCPM-v2.6 (LIS $= 0.854$) achieve high ratios simply because generating 2 to 3 matched steps is trivially ordered regardless of reasoning quality. Ordered Match F1 accounts for this by weighting LIS with Match F1, ensuring that high ordering scores require substantive step coverage.

Finding 5: Ordering is orthogonal to cherry-picking. Comparing Findings 1 to 3 with ordering results reveals that cherry-picking (high precision, low

recall) and disordered reasoning are independent failure modes. GPT-5.2 Instant (precision 0.974) cherry-picks aggressively yet maintains moderate order (LIS = 0.648), while Gemini 2.5 Flash, the only model with recall exceeding precision, achieves comparable ordering (0.584). This suggests that current MLLMs retrieve relevant reasoning steps but fail to organize them into coherent chains regardless of their coverage strategy, pointing to a fundamental gap in sequential reasoning that training objectives do not address.

4.4 Ablation Studies

We conduct 100 experiments testing 4 sentence encoders across 5 thresholds ($\tau \in \{0.30, 0.35, 0.40, 0.45, 0.50\}$) on 5 baselines. Figure 3 shows encoder choice dominates threshold selection: DistilRoBERTa-v1 [35] achieves 0.520 F1 versus 0.471–0.479 for competitors, with 4–8pp advantages transferring across all model families. Threshold variation yields only 2–3pp swings, confirming that evaluation reliability depends on encoder quality independent of the model being evaluated. Critically, model rankings remain stable across all 20 encoder-threshold configurations, indicating that the matching captures consistent semantic relationships rather than artifacts of a particular similarity function. This cross-encoder stability serves as an empirical proxy for human alignment: if the matching were unreliable, different encoders would produce contradictory rankings. We adopt all-distilroberta-v1 with $\tau = 0.35$, and validate the matching directly against human judgment below.

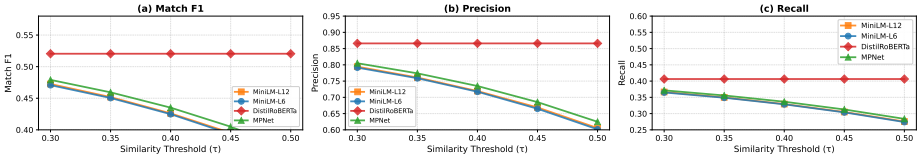


Fig. 3: Ablation: encoder and threshold. 4 encoders across 5 thresholds, averaged over all models. (a) `all-distilroberta-v1` achieves the highest Match F1 (+4.9pp at $\tau = 0.35$). (b–c) higher thresholds reduce precision and recall for most encoders, while DistilRoBERTa stays stable. We select $\tau = 0.35$.

Human agreement. A benchmark metric is only as trustworthy as its agreement with human semantic judgment. We sample 100 (predicted step, reference step) pairs from three models spanning the performance spectrum: GPT-5 [32] (57.99% acc.), Qwen2.5-VL-7B [43] (30.43%), and Llama 3.2-11B [26] (24.83%). To stress-test the decision boundary, we stratify pairs by cosine similarity: 33 low (<0.20), 33 borderline (0.20–0.50), and 34 high (>0.50). A human annotator labels each pair as semantically equivalent or not, *blinded* to the encoder's score and decision; this adversarial sampling over-represents hard cases, yielding a conservative estimate of agreement.

Table 3: Human agreement with embedding-based step matching. Agreement between the encoder's decisions and a blinded human annotator on 100 step pairs stratified by cosine similarity. Agreement is perfect below the operating threshold ($\tau = 0.35$); disagreements stay in the ambiguous borderline zone.

Similarity Band	Pairs	Agreement	Note
< 0.20	33	100.0%	No false matches
$[0.20, 0.35)$	14	100.0%	No false matches
$[0.35, 0.50)$	19	68.4%	Borderline zone
$[0.50, 0.70)$	25	64.0%	Borderline zone
≥ 0.70	9	88.9%	Clear matches
Overall	**100**	**84.0%**	$\kappa = 0.534$

Below the operating threshold ($\tau = 0.35$), encoder and annotator agree on *every* pair (47/47): the encoder **never produces a false match on semantically unrelated steps**, the most critical property for a benchmark metric, since false matches would inflate scores and mask reasoning deficiencies. Overall agreement is 84% ($\kappa = 0.534$), consistent with inter-annotator rates for comparable tasks [35]; all 16 disagreements are false positives, cases where the encoder proposes a match the annotator declines, and they concentrate in the borderline band (15 at $0.35 \leq \text{sim} < 0.70$, one at $\text{sim} \geq 0.70$) where semantic equivalence is genuinely ambiguous. The encoder produces no false negatives, and because this stratified sampling deliberately over-represents borderline pairs, effective agreement in normal evaluation exceeds 84%.

Order metric selection. We compare two candidates for measuring step ordering in Ordered Match F1: Kendall's Tau (normalized to $[0, 1]$) [20] and LIS ratio (Eq. 5). Figure 4 reports both metrics on a representative subset of six models deliberately spanning strong (GPT-5-mini, Qwen3-VL-32B), moderate (Gemini 2.5 Flash, GPT-5), and weak (InternVL3.5-1B, MiniCPM-v2.6-8B) reasoning profiles to stress-test discrimination across the full performance spectrum.

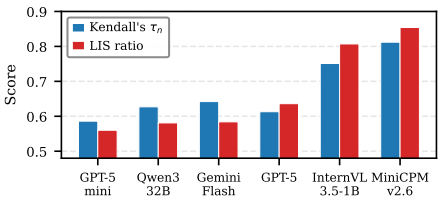


Fig. 4: Order metric comparison on a representative subset of 6 models. Both metrics rise for weak models generating few steps, but LIS ratio provides wider inter-group discrimination (0.56–0.85 vs. 0.58–0.81), clearly exposing trivially ordered few-step outputs.

Kendall's τ_n shows limited discrimination: GPT-5-mini (7.57 steps, $\tau_n = 0.586$) and InternVL3.5-1B (2.51 steps, $\tau_n = 0.751$) differ by only 0.165 points. LIS ratio exposes this contrast more clearly (0.560 vs. 0.807, gap of 0.247), reflecting that models generating few matched steps achieve trivially high ordering. We adopt LIS for its wider dynamic range and direct interpretability as the fraction of steps readable in correct order.

4.5 Reinforcement Learning with Step-Level Rewards

A key question is whether CRYSTAL can guide model improvement through training, not just evaluation. We apply our Causal Process Reward (CPR) and CPR-Curriculum (Section 3.3) as GRPO [37] reward strategies to two architecturally distinct models under *identical* reward weights, learning rate, and data: Qwen2.5-VL-3B-Instruct [43] and InternVL3.5-4B [44], the latter pairing an InternViT encoder with a Qwen3-based backbone (4×A100, DeepSpeed ZeRO-3; full hyperparameters in the supplementary material). Both baselines already appear in Table 2, so the comparison rests on pre-existing references. Table 4 shows CPR-Curriculum simultaneously improves accuracy *and* Match F1 on both: Qwen gains +7.67pp accuracy and +0.153 F1, while InternVL gains +8.15pp and +0.401 F1, its recall more than doubling (0.325→0.811) as it expands from 3.75 to 9.49 steps, exceeding even Qwen3-VL-32B's coverage (0.704 recall, Table 2) at 8× fewer parameters. Both show a small LIS drop (longer chains add ordering noise) yet higher Ordered F1, so content gains outweigh the ordering cost. With no architecture-specific tuning, CPR-Curriculum is not a Qwen-specific artifact but a reward formulation that transfers across multimodal architectures, improving reasoning without manual step annotation (qualitative examples in the supplementary material).

Table 4: GRPO with step-level rewards. CPR-Curriculum improves both accuracy and Match F1 over the base model on two architecturally distinct backbones, using identical reward weights and protocol. All metrics use all-distilroberta-v1, $\tau = 0.35$.

Configuration	Acc (%)	Match F1	Prec	Rec	Steps	LIS	Ord. F1
Qwen2.5-VL-3B							
Baseline	39.85	0.480	0.898	0.347	3.73	0.723	0.434
CPR-Curriculum	**47.52**	**0.633**	**0.963**	**0.493**	**5.10**	0.626	**0.560**
Δ	+7.67	+0.153	+0.065	+0.146	+1.37	−0.097	+0.126
InternVL3.5-4B							
Baseline	37.61	0.432	0.895	0.325	3.75	0.775	0.387
CPR-Curriculum	**45.76**	**0.833**	**0.903**	**0.811**	**9.49**	0.562	**0.719**
Δ	+8.15	+0.401	+0.008	+0.486	+5.74	−0.213	+0.332

Reward Strategy Comparison. We compare four reward strategies (Table 5) to isolate the effect of step-level supervision. The *Composite* reward combines

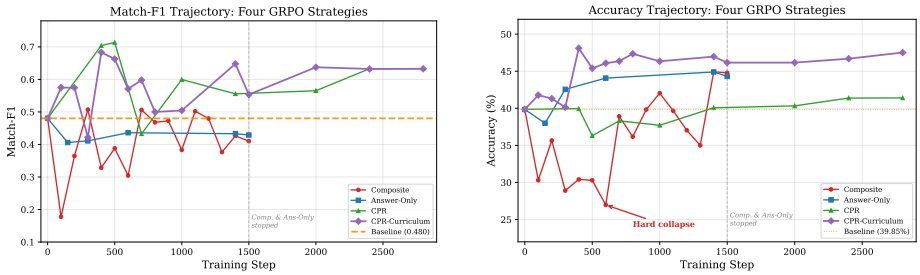


Fig. 5: Training dynamics. *Left:* Match F1; *Right:* Accuracy. Composite and Answer-Only diverge to NaN gradients by step 1,500 (Composite collapses at 600), while CPR variants train stably through 2,800 steps.

format, accuracy, and reasoning terms additively ($R = w_f R_{\text{fmt}} + w_a R_{\text{acc}} + w_r R_{\text{reason}}$), allowing the model to maximize each independently, achieving high accuracy through guessing while producing minimal reasoning. *Answer-Only* removes reasoning ($w_r = 0$) as a control. CPR and CPR-Curriculum use multiplicative coupling (Eq. 7) with initial weights $a_w = 0.65, s_w = 0.35$; we ablate this choice in Table 6.

Table 5: Reward strategy comparison. Best checkpoint per strategy on CRYSTAL test set. CPR strategies use weights $a_w = 0.65, s_w = 0.35$; a full weight ablation is presented in Table 6. Composite and Answer-Only reach comparable accuracy but fail to improve reasoning (F1 $\leq$ 0.43). CPR-based strategies improve both.

Strategy	Acc (%)	Match F1	Prec	Rec	LIS	Ord. F1
Baseline	39.85	0.480	0.898	0.347	0.723	0.434
Composite	<u>44.92</u>	0.426	**0.983**	0.284	0.743	0.392
Answer-Only	44.30	0.429	0.803	0.308	0.700	0.380
CPR	41.40	**0.633**	<u>0.975</u>	<u>0.489</u>	0.631	**0.560**
CPR-Curriculum	**47.52**	**0.633**	0.963	**0.493**	0.626	**0.560**

Answer-Only and Composite reach similar accuracy (44.30% and 44.92%) while reasoning stays flat (F1: 0.429 and 0.426), confirming the reasoning term is ignored under additive optimization. Composite collapses at step 600 (recall: 0.284); both strategies diverged to NaN gradients at step 1,500, requiring early termination (Figure 5). CPR's multiplicative coupling resolves this: the model must produce both correct answers and aligned reasoning, achieving F1 = 0.633 (+32%) and the highest Ordered F1 (0.560), while training stably through 2,800 steps. CPR-Curriculum further improves accuracy while maintaining comparable F1 and ordering quality [18, 55].

Reward Weight Sensitivity. We vary a_w and s_w in Eq. 7 across six configurations on a 15% held-out validation split (4,546 samples; Table 6). Beyond downstream performance, we report two diagnostics: the *discrimination gap* (mean reward difference between correct and incorrect outputs) and the *reward–reasoning correlation* r(F1) (Pearson correlation between reward and Match F1). These reveal a fundamental trade-off: higher s_w increases r(F1) but reduces discrimination. Both extremes are unstable: too answer-heavy ($a_w = 0.80$) causes KL spikes from overconfident updates, while too step-heavy ($s_w \geq 0.50$) elevates gradient norms from insufficient discrimination. Our configuration ($a_w = 0.65$, $s_w = 0.35$) achieves the highest accuracy (82.70%), the lowest $\mathrm{KL}_{\max}$ (0.372), and the highest r(F1) among stable configs (0.448), confirming it as the optimal balance.

Table 6: Reward weight sensitivity. Six configurations on a 15% held-out split (4,546 samples). Both extremes are unstable ($\nabla\uparrow$); $\mathrm{KL}_{\max}$ is U-shaped with its minimum at $a_w = 0.65$, which also gives the highest accuracy and r(F1) in the stable range.

		Val Performance			**Reward Signal**		**Stability**	
a_w	s_w	Acc (%)	M-F1	Rec	Discrim.	r(F1)	$\mathrm{KL}_{\max}$	∇
0.80	0.20	81.79	0.676	0.553	0.865	0.338	1.274	↑
0.70	0.30	82.61	0.731	0.627	0.797	0.410	0.456	↓
0.65	0.35	**82.70**	0.728	0.634	0.763	0.448	**0.372**	↓
0.50	0.50	79.39	0.730	0.629	0.662	0.563	1.218	↑
0.35	0.65	78.53	**0.736**	0.642	0.560	0.670	1.684	↑
0.20	0.80	74.81	0.721	**0.647**	0.459	**0.752**	1.521	↑

↑ = increasing gradient norms (unstable). **Bold** = best per column; underline = best within viable range (gap ≥ 0.70). Shaded = selected.

5 Conclusion

We introduced CRYSTAL, a benchmark that evaluates multimodal reasoning step by step via Match F1 and Ordered Match F1. Across 20 MLLMs, we find universal cherry-picking, non-monotonic scaling trade-offs, and disordered reasoning. CPR-Curriculum achieves +32% Match F1 via GRPO where additive strategies fail. We publicly release the CRYSTAL benchmark, reference steps, and evaluation code to support reproducible research.

Limitations. References may not capture all valid paths. The fixed encoder and threshold are robust (Section 4.4) but domain-specific alternatives may help. We validate CPR on two architectures (Qwen2.5-VL-3B, InternVL3.5-4B) but only at the 3–4B scale; behavior at larger scale remains open, and Ordered Match F1 does not yet model causal step dependencies.

Acknowledgments

This work was supported by startup funds provided by Dartmouth College. The authors also acknowledge support from the National Science Foundation under CAREER Award No. 2541968.

References

1. Awal, R., Ahmadi, S., Zhang, L., Agrawal, A.: Vismin: Visual minimal-change understanding. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. https://arxiv.org/abs/2511.21631 (2025), technical report
3. Breiman, L.: Bagging predictors. Machine Learning **24**(2), 123–140 (1996)
4. Chen, Y., Sikka, K., Cogswell, M., Ji, H., Divakaran, A.: Measuring and improving chain-of-thought reasoning in vision-language models. In: North American Chapter of the Association for Computational Linguistics (NAACL). pp. 192–210 (2024)
5. Cheng, Z., Chen, Q., Zhang, J., Fei, H., Feng, X., Che, W., Li, M., Qin, L.: CoMT: A novel benchmark for chain of multi-modal thought on large vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39 (2025)
6. Cui, X., Aparcedo, A., Jang, Y.K., Lim, S.N.: On the robustness of large multimodal models against image adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
7. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics) **28**(1), 20–28 (1979)
8. DeepSeek-AI: DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature **645**(8081), 633–638 (2025). https://doi.org/10.1038/s41586-025-09422-z
9. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N.A., Hajishirzi, H., Girshick, R., Farhadi, A., Kembhavi, A.: Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 91–104 (2025)

10. Fredman, M.L.: On computing the length of longest increasing subsequences. Discrete Mathematics **11**(1), 29–35 (1975)
11. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R.S., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
12. Gemma Team: Gemma: Open models based on gemini research and technology. https://arxiv.org/abs/2403.08295 (2024), technical report from Google DeepMind
13. Google DeepMind: Gemini 2.5 flash. https://deepmind.google/en/models/gemini/flash/ (2025), accessed 2026-02-15
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
15. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: SUGARCREPE: Fixing hackable benchmarks for vision–language compositionality. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2023)
16. Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N.A., Krishna, R.: Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
17. Huang, J., Chen, L., Guo, T., Zeng, F., Zhao, Y., Wu, B., Yuan, Y., Zhao, H., Guo, Z., Zhang, Y., Yuan, J., Ju, W., Liu, L., Liu, T., Chang, B., Zhang, M.: MMEval-Pro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. In: North American Chapter of the Association for Computational Linguistics (NAACL) (2025)
18. Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., Zhang, B., Fu, C., Gao, P., Li, H.: Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In: International Conference on Machine Learning (ICML) (2025)
19. Kalai, A.T., Nachum, O., Vempala, S.S., Zhang, E.: Why language models hallucinate. arXiv preprint arXiv:2509.04664 (2025)
20. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938)
21. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: Advances in Neural Information Processing Systems (NeurIPS) (1995)
22. Li, Y., Yue, X., Xu, Z., Jiang, F., Niu, L., Lin, B.Y., Ramasubramanian, B., Poovendran, R.: Small models struggle to learn from strong reasoners. In: Findings of the Association for Computational Linguistics (ACL). pp. 25366–25394 (2025). https://doi.org/10.18653/v1/2025.findings-acl.1301
23. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
24. 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. In: International Conference on Learning Representations (ICLR) (2024)
25. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for

science question answering. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
26. Meta AI: Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ (2024), accessed 2025-10-30
27. Meta AI: The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (2025), accessed 2025-10-30
28. Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: Plotqa: Reasoning over scientific plots. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1516–1525 (2020)
29. Nagrani, A., Menon, S., Iscen, A., Buch, S., Mehran, R., Jha, N., Hauth, A., Zhu, Y., Vondrick, C., Sirotenko, M., Schmid, C., Weyand, T.: MINERVA: Evaluating complex video reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
30. Nasa, P., Jain, R., Juneja, D.: Delphi methodology in healthcare research: How to decide its appropriateness. World Journal of Methodology **11**(4), 116–129 (2021). https://doi.org/10.5662/wjm.v11.i4.116
31. Oh, C., Lim, H., Kim, M., Han, D., Yun, S., Choo, J., Hauptmann, A., Cheng, Z.Q., Song, K.: Towards calibrated robust fine-tuning of vision–language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
32. OpenAI: GPT-5 system card. https://openai.com/index/gpt-5-system-card/ (2025), accessed 2026-02-15
33. Qiao, Y., Duan, H., Fang, X., Yang, J., Chen, L., Zhang, S., Wang, J., Lin, D., Chen, K.: Prism: A framework for decoupling and assessing the capabilities of VLMs. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
34. Raykar, V.C., Yu, S., Zhao, L.H., Hermosillo Valadez, G., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. Journal of Machine Learning Research **11**, 1297–1322 (2010)
35. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019)
36. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2024)
37. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024), introduces Group Relative Policy Optimization (GRPO)
38. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards VQA models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8317–8326 (2019)
39. Sun, Y., Wang, H., Li, J., Liu, J., Li, X., Wen, H., Yuan, Y., Zheng, H., Liang, Y., Li, Y., Liu, Y.: An empirical study of LLM reasoning ability under strict output length constraint. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2025)
40. Surís, D., Menon, S., Vondrick, C.: ViperGPT: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)

41. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9568–9578 (2024)
42. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with the math-vision dataset. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2024)
43. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. https://arxiv.org/abs/2409.12191 (2024), technical report
44. 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. arXiv preprint arXiv:2508.18265 (2025)
45. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: International Conference on Learning Representations (ICLR) (2023)
46. xAI: Realworldqa: A benchmark for real-world spatial understanding. https://x.ai/news/grok-1.5v (2024)
47. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: LLaVA-CoT: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
48. Xu, Z., Zhou, P., Ai, J., Zhao, W., Wang, K., Peng, X., Shao, W., Yao, H., Zhang, K.: Mpbench: A comprehensive multimodal reasoning benchmark for process errors identification. In: Findings of the Association for Computational Linguistics (ACL) (2025)
49. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, H., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., Sun, M.: Minicpm-v: A gpt-4v level mllm on your phone. https://arxiv.org/abs/2408.01800 (2024), technical report from OpenBMB
50. Yu, G., Zhu, J., Yao, J., Han, B.: Self-calibrated tuning of vision–language models for out-of-distribution detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
51. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
52. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: 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) (2024)
53. Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., Wang, L., Jin, R., Tan, T.: MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In: International Conference on Learning Representations (ICLR) (2025)
54. Zhang, Y., Huang, Y., Sun, Y., Liu, C., Zhao, Z., Fang, Z., Wang, Y., Chen, H., Yang, X., Wei, X., Su, H., Dong, Y., Zhu, J.: Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2024)
55. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. Transactions on Machine Learning Research (TMLR) (2024)
56. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
57. Zhu, K., Zang, Q., Jia, S., Wu, S., Fang, F., Li, Y., Guo, S., Zheng, T., Guo, J., Li, B., Wu, H., Qu, X., Yang, J., Liu, R., Yue, X., Liu, J., Lin, C., Alinejad-Rokny, H., Yang, M., Ni, S., Huang, W., Zhang, G.: LIME: Less is more for MLLM evaluation. In: Findings of the Association for Computational Linguistics (ACL) (2025)