VERDICT: Training-Free Step-Wise Verification of Multimodal Reasoning via Disagreement-Aware Consensus

Rohit Sinha*[1], Kunal Tilaganji*[1], Tanuja Ganu[2], Nagarajan Natarajan[2], Amit Sharma[2], and Vineeth Balasubramanian[1,2]

[1] Indian Institute of Technology, Hyderabad
https://www.iith.ac.in/
[2] Microsoft Research
https://www.microsoft.com/en-us/research/lab/microsoft-research-india/

Abstract. Multimodal large language models often generate reasoning chains containing subtle errors that lead to incorrect answers. Current verification approaches have notable limitations. Existing approaches either require expensive labelled supervision with inconsistent cross-task performance or aggregate scores from multiple sources by simple aggregations, missing a key insight: when these scores disagree, that disagreement itself carries important information about whether a reasoning step is truly valid or not. We formalise this as a coupled scoring problem among disparate, frozen verifiers, interpretable as a coordination game with a unique closed-form equilibrium where agreement signals valid steps while disagreement reveals instability. Towards this end, we propose a training-free domain-agnostic step-wise verification approach we call **VERDICT**: **VER**ification via **D**isagreement-**I**nformed **C**oupled **T**hresholding. To our knowledge, **VERDICT** is the first training-free verifier that makes the structure of cross-modal disagreement explicit and actionable. It computes consensus scores through a closed-form solution, enabling both disagreement-aware filtering and stability-conscious ranking of reasoning steps. Evaluated across six benchmarks, **VERDICT** consistently improves over the base model by up to +5.95%, and performs competitively with domain-specific critics that demand extensive supervision, demonstrating that cross-modal agreement provides robust verification signals without task-specific adaptation.

Keywords: Multimodal Reasoning · Step-wise Verification · Multi-Agent Consensus · Training-Free Verification

1 Introduction

Multimodal large language models (MLLMs) [1, 18] have demonstrated impressive multi-step reasoning capabilities over images and text. By decomposing complex questions into chains of intermediate steps, these models can tackle compositional visual reasoning tasks that require both perceptual grounding and logical inference [4, 37]. Yet their reasoning chains often contain subtle errors,

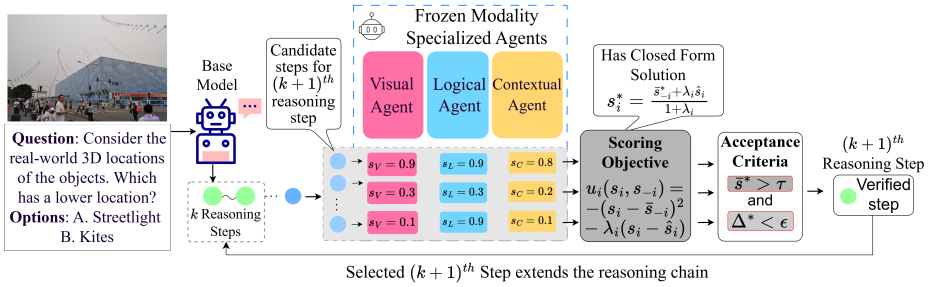

Fig. 1: An illustrative overview of **VERDICT**.Given a multimodal question and partial reasoning trace, the base model generates candidate continuations that are independently scored by three frozen, modality-specialized agents. A closed-form consensus computation transforms the raw scores into disagreement-aware consensus scores; a dual acceptance criterion then filters by mean confidence and consensus dispersion, selecting the most stable step to extend the reasoning chain.

unsupported visual claims, logical gaps, and hallucinated spatial relationships that are difficult to detect precisely because each step *appears* locally plausible [14, 17]. These errors propagate: a single flawed step early in the chain can cascade through subsequent reasoning, ultimately leading to an incorrect final answer. This motivates *step-level* verification, catching errors where they arise rather than after they have shaped the trajectory of the reasoning chain.

Existing verification approaches fall into two broad categories, each with notable limitations. **Domain Specific critics**, including process reward models (PRMs) [16, 19, 29] and their multimodal extensions such as VisualPRM [31], Sherlock [6], LLaVA-Critic [33], and MM-Verify [26], score individual reasoning steps using trained models. However, these require carefully labeled supervision, often obtained through expensive Monte Carlo rollouts or human annotations [19, 32]. More critically, they show inconsistent cross-task performance: while they may improve accuracy on one benchmark, they frequently degrade on others. This limitation is especially acute in data-scarce domains, where the labeled supervision these critics require is prohibitively expensive or simply unavailable.**Training-free aggregation methods** like Weaver [24] and reward model ensembles [5,9] combine scores from multiple sources, but they miss a critical insight. Consider two scoring patterns: unanimous confidence $(0.7, 0.7, 0.7)$ and conflicting evidence $(0.9, 0.3, 0.9)$. Simple averaging treats both as equivalent, yet they reflect fundamentally different verification states. This problem is compounded by well-documented capability tensions within MLLMs themselves: visual grounding can disrupt reasoning [11, 36], visual instruction tuning degrades language performance [25], and extended reasoning chains cause forgetting of earlier context [27]. These tensions mean that different evaluation perspectives can legitimately disagree about the same reasoning step and that disagreement carries diagnostic value.

Our key insight is that when a reasoning step is truly valid, disparate judges evaluating it from different perspectives, such as visual grounding, logical con-

sistency, and contextual relevance, should be able to converge. When they *cannot* agree, even after accounting for each other's views, the step is likely unstable. Rather than treating verification as a classification or regression task, we treat it as a *coupled scoring problem* among disparate sources of evidence. We formalize this as a coupled scoring framework among frozen, modality-specialized judges, whose consensus yields disagreement-aware scores. The consensus captures how each judge would revise their belief in light of what others think, producing scores that naturally balance collective confidence against residual disagreement. Unlike variance-based approaches (which ignore belief strength) or simple averaging (which buries disagreement), **VERDICT** is a disagreement-aware consensus framework, which makes the structure of conflict explicit and actionable.

Concretely, three frozen agents, Visual, Logical, and Contextual, disparately score each candidate's reasoning step. A closed-form consensus computation then transforms these raw scores into consensus-adjusted scores, without any iterative optimization or learning. A dual acceptance criterion combines a mean confidence check (ensuring sufficient collective endorsement) with a dispersion check (requiring cross-modal agreement). Among accepted candidates, ranking by consensus confidence selects the most stable continuation. The entire framework operates as a plug-in verifier requiring no modification to the base model, no training data, and no task-specific adaptation. This formulation admits a natural game-theoretic interpretation: each agent is a player in a coordination game, and the closed-form solution corresponds to its unique Nash equilibrium [21], grounding the influence structure in equilibrium theory rather than ad hoc design choices.

Evaluated across six benchmarks spanning spatial reasoning, visual grounding, and multimodal abstraction, **VERDICT** achieves consistent improvements of up to +5.95% over baseline models and performs competitively with or surpasses domain-specific critics that require extensive labeled training data. Crucially, our method never degrades below baseline on any tested benchmark, a property that none of the domain-specific critics we evaluate can claim, as they show significant performance variability across tasks.

It is worth noting that Averaging [5, 30] is symmetric as every judge's deviation counts equally, and the pattern of who disagrees with whom is lost. Variance-based filtering [13] treats a stubborn visual agent and a flexible contextual agent identically. Our disagreement-aware consensus formulation introduces *coupling* where each agent's adjusted score depends on all others, so disagreement from a stubborn agent weighs more heavily: an asymmetry no fixed weighting can recover. This coupling is what distinguishes **VERDICT**. Our contributions are summarized below:

- We formalize step-wise verification as a coupled scoring problem among disparate, frozen verifiers, establishing disagreement structure as a first-class verification signal and proposing disagreement-aware verification as a design principle.
- We introduce **VERDICT**, a consensus verifier with a closed-form solution (the unique equilibrium of a coordination game among modality-specialized

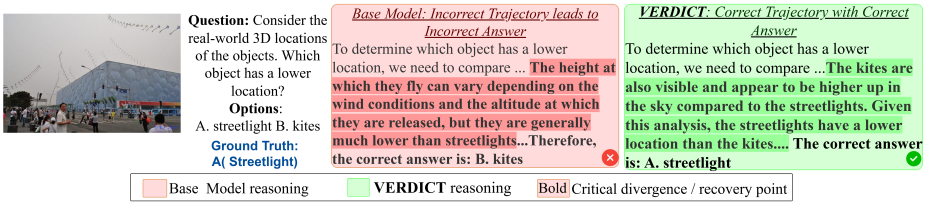

Fig. 2: Qualitative examples showing base model reasoning (orange) and VERDICT-verified reasoning (blue). Bold highlighted text marks the critical sentence where the base model commits to an incorrect reasoning trajectory. VERDICT recovers by selecting an alternative candidate step that corrects the judgment, yielding the correct answer. Full reasoning chains are provided in the supplementary material.

judges) that models cross-modal agreement, enabling both disagreement-aware filtering and stability-conscious ranking of candidate reasoning steps.
- We provide a domain-agnostic, training-free, plug-in implementation that integrates directly into MLLM reasoning without base model modification and demonstrates consistent improvements across six benchmarks spanning spatial reasoning, visual grounding, and multimodal abstraction.
- We conduct comprehensive ablation studies revealing that the framework operates through two complementary mechanisms, rejection and selection, and that consensus scores function optimally as ranking tools rather than binary classifiers.

2 Related Work

Step-level verification has largely developed along two tracks. Process reward models from PRM800K [16] through Math-Shepherd [29] and OmegaPRM [19] to multimodal extensions like MM-PRM [7] and VisualPRM [31] demonstrate that intermediate supervision improves reasoning reliability, but at the cost of expensive human annotation or Monte Carlo rollouts that our approach requires none of. Domain-specific critic models [6, 15, 26, 33, 34] achieve strong results on specific benchmarks yet implicitly collapse heterogeneous evidence into a single score, which, as our experiments show, leads to fragile cross-task generalization; a critic that learns to reward one reasoning style can quietly penalize another, and the scalar it produces carries no record of which modality drove the judgment. This opacity is not incidental; it is structural, and it is what motivates an explicit treatment of disagreement rather than a better-trained aggregator. Ensemble and multi-agent methods [8,9,13,24] take a step toward robustness by combining multiple sources of evidence,

Table 1: VERDICT vs. prior methods across three design axes (Section 3).

Method	Training Free	Disagree. Aware	Plug-in Compat.
Domain Specific Critics [2, 6, 16, 31, 33]	✗	✗	✗
Reward Ensembles [9]	✓	✗	✓
Weaver [24]	✓	✗	✓
Variance Filtering [13]	✓	✗	✓
VERDICT (Ours)	✓	✓	✓

but aggregate through av-
eraging or majority voting, treating all patterns of agreement and disagreement as equivalent; a three-way split and a near-unanimous vote are indistinguishable once reduced to a mean. The richer question- not whether agents disagree, but how and between whom is precisely what averaging discards and what our framework is designed to recover. Multi-agent formulations, prominent in GANs and adversarial training [23], have so far relied on adversarial zero-sum games where agents work against one another by design. Our formulation departs from all of these: we employ a coupled scoring framework in which frozen, modality-specialized judges seek consensus rather than oppose one another, and where the closed-form solution makes the structure of their disagreement explicit rather than averaging it away: a property that, to our knowledge, no existing training-free step-wise verification method provides. Table 1 summarizes how VERDICT relates to prior work along the three axes that motivate its design.

3 Preliminaries

In this section, we formalize the verification setting and define the problem addressed by our approach.

Verification Setting. We consider a setting where a base MLLM generates a reasoning chain step by step, and each step must be evaluated before extending the chain further. Formally, at reasoning step t, the base model is provided with an image I, a question Q, and a partial reasoning trace $r_{1:t-1}$ consisting of all previously accepted steps. From this context, the base model generates n candidate continuations $\{r_t^{(1)}, \ldots, r_t^{(n)}\}$ via sampling. A verifier then evaluates each candidate and returns a binary decision: *accept* the step and continue reasoning, or *reject* it and resample. Crucially, evaluation is *local* to the current step; it does not require access to future reasoning or to the final answer. This locality allows errors to be intercepted at the point where they arise, before they propagate through subsequent steps.

Problem Definition. Given n candidate steps at position t, m specialized judge agents $\{A_1, \ldots, A_m\}$, and context $(I, Q, r_{1:t-1})$, the verification problem is to produce a single verified continuation r_t^*. Each agent A_i independently scores every candidate $r_t^{(j)}$ with a scalar confidence $\hat{s}_i^{(j)} \in [0, 1]$ reflecting its modality-specific assessment. The verifier must aggregate these heterogeneous assessments into a selection decision, accounting for both collective confidence and inter-agent agreement.

We impose three design constraints: (1) *training-free*, requiring no labeled data or parameter updates; (2) *disagreement-aware*, treating cross-modal conflict as a diagnostic signal rather than noise; and (3) *plug-in compatible*, integrating with any base MLLM without architectural modification. Domain-specific critics violate (1); standard aggregation methods violate (2). Constraint (2) deserves emphasis: variance-based rejection treats all agents symmetrically, contributing equally regardless of their modality-specific reliability, and any two score vectors with identical means and per-agent deviations are deemed equivalent even when their disagreement patterns differ qualitatively. What is needed is a mechanism that *couples* agents one in which each agent's adjusted assessment depends on

where every other agent landed. As we show in Section 4 (Proposition 1), this coupling cannot be replicated by any separable function of individual scores, and it is precisely what **VERDICT** provides through a coordination-game formulation.

4 VERDICT: Disagreement-Aware Consensus Verification

We now present **VERDICT**, our verification framework. We first describe the verifier agents and their roles, then formulate their interaction as a coupled scoring problem, derive the closed-form consensus , and finally define the acceptance criterion used to select verified reasoning steps.

Verifier Agents. Verification is carried out by three frozen MLLMs, each prompted to judge a candidate's reasoning step from a distinct modality-specific perspective. We instantiate three agents:

- **Visual Agent (V):** assesses whether objects and spatial relationships mentioned in the reasoning step are visually verifiable in the provided image. It checks whether the step is grounded in what can actually be observed.
- **Logical Agent (L):** evaluates whether the reasoning step follows logically from the preceding steps and makes progress toward answering the question. It checks for coherence, valid inferences, and logical gaps.
- **Contextual Agent (C):** determines whether the step maintains focus on the original question and avoids introducing irrelevant or tangential information. It penalizes unnecessary speculation and off-topic claims.

Each agent receives the image I, question Q, the partial reasoning trace $r_{1:t-1}$ (assumed correct), and the current candidate step $r_t^{(j)}$ to evaluate. It outputs a single scalar score $\hat{s}_i \in [0, 1]$, interpreted as its subjective confidence that the step is valid given its modality-specific evidence. Agents operate in *complete isolation*: they never see other agents' scores, their prompts contain no information about other agents' judgments, and no fine-tuning or calibration is performed.

Number of agents: The formulation is defined for arbitrary m agents: Eqs. (1)-(4) and Proposition Proposition 1 hold for any $m \geq 2$. We instantiate three because they map onto three largely orthogonal MLLM failure modes: unsupported visual claims [11, 36], logical gaps [27], and contextual drift [25]. The coupled system remains closed-form for larger m, and its cost is dominated by the m scoring calls rather than the $m \times m$ linear solve. The binding constraint for adding agents is semantic, not computational: each must bring a genuinely disparate evaluation perspective.

Coupled Scoring Formulation. Rather than aggregating raw scores directly, we frame verification as a coupled scoring problem among the agents. The motivating intuition is straightforward: if a reasoning step is truly valid, disparate judges evaluating it from different perspectives should be able to reach agreement. If they *cannot* agree even after accounting for each other's views, the step is likely unstable.

We model this interaction as a coupled system in which each agent i selects a reported score $s_i \in [0, 1]$, balancing agreement with the other agents against fidelity to its own modality-specific judgment. The scoring objective to agent i is defined as:

$$u_i(s_i, s_{-i}) = -(s_i - \bar{s}_{-i})^2 - \lambda_i(s_i - \hat{s}_i)^2 \tag{1}$$

where $\bar{s}_{-i} = \frac{1}{m-1}\sum_{j\neq i} s_j$ denotes the mean reported score of the other agents, $\hat{s}_i$ is agent i's raw confidence score, and $\lambda_i > 0$ controls how strongly agent i weights its own evidence relative to group consensus. The first term encourages agreement: it penalizes the agent for deviating from what others report. The second term enforces self-consistency: it penalizes deviation from the agent's original belief, scaled by λ_i. Under the coordination game interpretation, Eq. Eq. (1) is each agent's payoff: it is maximized when the agent finds a score that balances fidelity to its own signal against coordination with others, which is the defining tension of a coordination game. The stubbornness parameter λ_i encodes how resistant each agent should be to consensus pressure. A higher λ_i means agent i trusts its own modality-specific evidence more strongly and will deviate less from its raw score even when others disagree. In our experiments, we set $\lambda_V = 1.5$, $\lambda_L = 1.0$, and $\lambda_C = 0.8$, reflecting the intuition that the visual agent should be most resistant to consensus pressure on perception-heavy tasks, while the contextual agent may be more flexible when visual or logical evidence is strong. These values are fixed across all datasets. Since the objective function in Eq. (1) is strictly concave in each agent's own score (the second derivative $\frac{\partial^2 u_i}{\partial s_i^2} = -2(1 + \lambda_i) < 0$ for all $\lambda_i > 0$), the coupled system admits a unique fixed point (equivalent to the unique Nash equilibrium of the induced coordination game). This follows from Rosen's uniqueness result for concave interaction systems, [22].

Closed-Form Solution. At the consensus solution, each agent's reported score satisfies the following:

$$s_i^* = \frac{\bar{s}_{-i}^* + \lambda_i \hat{s}_i}{1 + \lambda_i} \tag{2}$$

This system of equations can be rewritten as a linear system:

$$(1 + \lambda_i)\, s_i^* - \frac{1}{m-1}\sum_{j\neq i} s_j^* = \lambda_i \hat{s}_i, \quad i = 1, \ldots, m \tag{3}$$

which admits a closed-form solution and can be solved directly from the raw scores $\{\hat{s}_i\}$ without any iterative optimization, learning, or approximation.

A key property of the consensus solution is that it *preserves the mean while dampening disagreement*: the mean consensus score equals the mean raw score, but individual scores converge as disagreement diminishes proportionally to inter-agent conflict. This property is important because it allows us to separate two distinct failure modes. The first is *collective doubt*: steps where all agents assign low confidence, resulting in a low mean score. The second is *conflicting evidence*: steps where the mean confidence is high but agents fundamentally disagree, producing high dispersion. Simple averaging conflates these two

cases; for instance, it treats $(0.6, 0.6, 0.6)$ and $(0.9, 0.3, 0.6)$ identically since both yield the same mean. However, the consensus dispersion reveals that the second case reflects unresolved cross-modal instability, while the first reflects genuine (if modest) consensus.

The consensus solution adjusts individual scores but not their collective level: $\bar{s}^* = \frac{1}{m}\sum_i s_i^* = \frac{1}{m}\sum_i \hat{s}_i$, a direct consequence of the symmetry in Eq. (3) that holds for any $\{\lambda_i\}$. The coupled system cannot inflate confidence. Its effect is confined to redistributing scores around a fixed mean, isolating cross-modal conflict from collective doubt.

Acceptance Criterion and Step Selection. Once consensus scores $\{s_i^*\}$ are computed for each candidate, we derive two summary statistics: the *mean consensus confidence* $\bar{s}^* = \frac{1}{m}\sum_i s_i^*$, measuring collective endorsement of the step, and the *consensus dispersion* $\Delta^* = \frac{1}{m}\sum_i |s_i^* - \bar{s}^*|$, measuring residual disagreement after consensus adjustment.

A candidate step is accepted if and only if it satisfies a dual criterion: the mean confidence must exceed a threshold τ *and* the dispersion must fall below a tolerance ϵ:

$$\text{accept}(r_t^{(j)}) \iff \bar{s}^{*(j)} > \tau \quad \wedge \quad \Delta^{*(j)} < \epsilon \tag{4}$$

This dual criterion enables two complementary verification mechanisms. The dispersion check ($\Delta^* < \epsilon$) filters candidates where cross-modal evidence fundamentally conflicts, while the confidence check ($\bar{s}^* > \tau$) ensures sufficient collective endorsement. Among accepted candidates, the step with the highest $\bar{s}^*$ is selected to extend the reasoning trace, prioritizing the most stable continuation. When no candidate satisfies both criteria, indicating that all options are suboptimal, the system falls back to continuous ranking by $\bar{s}^* - \Delta^*$, selecting the candidate that best balances confidence against remaining disagreement. This fallback ensures that the reasoning process can always continue while still preferring more stable steps.

Numerical Illustration. Consider raw scores $\hat{\mathbf{s}} = (0.9, 0.3, 0.9)$ from the Visual, Logical, and Contextual agents with λ_V=1.5, λ_L=1.0, λ_C=0.8. The linear system in Eq. (3) yields $\mathbf{s}^* \approx (0.80, 0.55, 0.77)$: the outlying Logical agent is pulled from 0.30 to 0.55, the stubborn Visual agent shifts only 0.10, and the flexible Contextual agent shifts 0.13: asymmetry driven directly by the λ_i coupling in Eq. (2). The resulting $\bar{s}^* \approx 0.71$ and $\Delta^* \approx 0.11$: the confidence check passes ($0.71 > \tau$=0.6) but dispersion fails ($0.11 > \epsilon$=0.1), So the step is rejected as cross-modally unstable. Contrast $\hat{\mathbf{s}} = (0.7, 0.7, 0.7)$: consensus returns $\Delta^* = 0$ and the step is accepted: same raw mean, opposite decision, a distinction simple averaging cannot make.

Where the consensus formulation adds value. By construction, $\bar{s}^*$ equals the raw mean: the consensus does not shift collective confidence. Its contribution lies entirely in the individual equilibrium scores $\{s_i^*\}$ and the resulting dispersion Δ^*. One might argue that since Eq. (2) admits a closed-form solution, one could define the same asymmetric dispersion directly, without motivating the coupled system. This is correct computationally, but it misses what the formulation provides. Defining a weighted dispersion requires choosing how much each

agent's deviation should count: a design decision with no principled grounding. The coupled scoring framework *derives* these weights from a single interpretable assumption: each agent trades off agreement against fidelity to its own evidence, governed by stubbornness λ_i. This is precisely the assumption that defines a coordination game. The fixed point uniquely determines the rest.

Crucially, the consensus solution introduces *coupling* that no per-agent measure can replicate. Because agent i's adjusted score s_i^* depends on all other agents' raw scores through the linear system in Eq. (3), its residual $|s_i^* - \bar{s}^*|$ reflects not just its own deviation but how that deviation interacts with where other agents land. We formalize this below.

Proposition 1 (Consensus Dispersion is Not Recoverable from Weighted Averages). *For any fixed weight vector* $\mathbf{w}$*, there exist score vectors* $\hat{\mathbf{s}} \neq \hat{\mathbf{s}}'$ *such that* $\mathbf{w}^\top \hat{\mathbf{s}} = \mathbf{w}^\top \hat{\mathbf{s}}'$ *yet* $\Delta^*(\hat{\mathbf{s}}) \neq \Delta^*(\hat{\mathbf{s}}')$*. As for any agent* i*, the consensus residual* $|s_i^* - \bar{s}^*|$ *depends on the full raw score vector* $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_m)$*, not solely on* $\hat{s}_i$*. Consequently, the consensus dispersion* $\Delta^* = \frac{1}{m}\sum_i |s_i^* - \bar{s}^*|$ *cannot be decomposed as a separable function* $\sum_i f_i(\hat{s}_i)$ *for any choice of per agent functions* $\{f_i\}$*.*

Proof. From Eq. (2), s_i^* depends on $\bar{s}^*_{-i}$, which itself depends on all agents' raw scores through the coupled linear system in Eq. (3). Consequently, $\partial s_i^*/\partial \hat{s}_j \neq 0$ for $i \neq j$, so the residual $|s_i^* - \bar{s}^*|$ is not a function of $\hat{s}_i$ alone and Δ^* is not separable. As a concrete illustration: $\hat{\mathbf{s}} = (0.9, 0.2, 0.9)$ and $\hat{\mathbf{s}}' = (0.7, 0.6, 0.7)$ share the same mean $\frac{2}{3}$, yet the consensus solution yields $\Delta^*(\hat{\mathbf{s}}) \approx 0.13 > \epsilon$ and $\Delta^*(\hat{\mathbf{s}}') \approx 0.04 < \epsilon$, producing opposite acceptance decisions under Eq. (4), a distinction no separable measure can replicate, since the difference arises from the consensus solution coupling agent 2's residual to agents 1 and 3's raw scores.

Two properties follow. First, disagreement from a high-λ agent weighs more heavily in Δ^* than equivalent disagreement from a flexible one, an asymmetry that raw variance treats identically. Second, this coupling is what lets the dual criterion in Eq. (4) distinguish genuine cross-modal conflict from ordinary confidence fluctuations, even when two score vectors share identical means and per-agent deviations.

Game-theoretic interpretation: The coupled scoring formulation admits a natural reading as a coordination game in which each agent is a player, its reported score is a strategy, and . Eq. (1) defines the payoff. The fixed point in Eq. (2) is then the unique Nash equilibrium, guaranteed by Rosen's theorem for concave n-person games [22]. This connection is more than cosmetic: it supplies a principled rationale for why the stubbornness parameters λ_i govern the influence structure, as each agent trades off fidelity to its own evidence against pressure to coordinate exactly the tension a coordination game formalizes. The analogy has limits: our agents score in isolation and never observe one another; the equilibrium is computed in closed form from one-shot scores rather than reached through iterated play, and should be understood as interpretive scaffolding that motivates the formulation rather than a claim about strategic interaction.

Table 2: Accuracy (%) across multimodal reasoning benchmarks under different verification strategies. **Avg.** is the unweighted mean across all six benchmarks. Values in blue indicate improvement of our method relative to the Base Model. Our detailed evaluation setup is given in the supplementary.

Dataset	3DSRBench	CV-Bench-3D	CV-Bench-2D	BLINK	MMStar	AI2D	Avg.
Base	56.12	76.39	74.27	48.31	61.25	81.52	66.31
Domain-Specific Critics							
LLaVA Critic	52.71 ±0.83	81.58 ±0.91	67.52 ±0.82	49.22 ±0.77	64.07 ±0.61	81.81 ±0.44	66.15
Critic-V$_{\text{CVPR, 25}}$	53.25 ±0.73	77.66 ±0.81	75.38 ±0.79	46.17 ±0.86	55.83 ±0.59	80.17 ±0.75	64.74
Sherlock$_{\text{NeurIPS, 25}}$	48.11 ±0.93	58.13 ±1.10	68.78 ±1.98	49.07 ±0.89	57.26 ±0.96	82.77 ±0.52	60.69
VisionSR1$_{\text{ICLR, 26}}$	53.05 ±1.82	54.18 ±1.05	73.10 ±1.91	30.09 ±1.45	57.20 ±2.73	80.25 ±0.97	57.98
DreamPRM$_{\text{NeurIPS, 25}}$	53.08 ±1.88	63.39 ±0.98	75.58 ±1.61	49.89 ±0.82	61.23 ±0.59	81.21 ±0.49	64.06
Domain-Agnostic Baselines							
Variance	57.21 ±0.57	78.16 ±0.68	76.43 ±0.53	49.61 ±0.49	63.09 ±0.47	81.41 ±0.34	67.65
Mean	58.34 ±0.53	79.77 ±0.61	77.57 ±0.51	50.17 ±0.58	64.14 ±0.44	82.18 ±0.36	68.70
Min	56.11 ±0.58	77.21 ±0.63	75.17 ±0.59	49.39 ±0.55	62.41 ±0.46	81.09 ±0.42	66.90
Majority	57.58 ±0.62	77.37 ±0.73	75.53 ±0.68	48.43 ±0.66	62.10 ±0.52	81.78 ±0.48	67.13
Max	57.37 ±0.55	78.16 ±0.56	76.46 ±0.63	49.17 ±0.43	63.42 ±0.58	82.12 ±0.54	67.78
VERDICT (Ours)	**59.02** ±0.45	**82.34** ±0.51	**79.22** ±0.53	**51.32** ±0.48	**65.88** ±0.37	**83.14** ±0.32	**70.15**
Gains	+2.90	+5.95	+4.95	+3.01	+4.63	+1.62	+3.84

5 Experiments and Results

Experimental Setup. We evaluate our verification framework on six benchmarks that collectively assess different aspects of visual and multimodal reasoning. **3DSRBench** [20], **CV-Bench** [28] split into CV-Bench-2D (spatial relationships and object counting) and CV-Bench-3D (depth ordering and relative distance). **BLINK** [10], **MMStar** [3], **AI2D** [12].

Our approach employs three verification agents: Visual (V), Logical (L), and Contextual (C), all built on Qwen2.5-VL-7B-Instruct [1], the same model used as the base reasoner. At each reasoning step, the base model generates n=3 candidate continuations via temperature sampling (T=0.8, top-p=0.6). Each agent disparately scores every candidate, and consensus scores are computed using agent-specific stubbornness parameters (λ_V=1.5, λ_L=1.0, λ_C=0.8). A candidate is accepted if its consensus dispersion satisfies $\Delta^* < \epsilon$=0.1 and mean confidence satisfies $\bar{s}^* > \tau$=0.6. Among accepted steps, we select the one with the highest $\bar{s}^*$. These hyperparameters are fixed across all six benchmarks with no task-specific tuning. The final answer is extracted from the complete reasoning trace using (Gemma-3:12B via Ollama). The use of Ollama was only for answer extraction, and is not for any part of the verification pipeline. All agent prompts are provided in the supplementary.

Baselines. We compare **VERDICT** against two categories of baselines. The first category consists of *domain-specific critics*(so called because each requires labeled data from the target domain to train): LLaVA-Critic [33], Sherlock [6], VisionSR1 [15], DreamPRM [2], Critic-V [35] These represent the current state of the art in step-level verification, but each requires task-specific training data or learned reward signals. All domain-specific critics are evaluated under comparable conditions. The second category consists of *domain-agnostic baselines*(which require no domain-specific training data, instead relying on modality-specialized agents with task-general rubrics) that use the same three frozen agents as our method but differ in how they combine the raw scores. These include: **Mean** (select the candidate with the highest average raw score), **Max** (select by highest

maximum score across agents), **Min** (select by highest minimum score), **Majority** (accept candidates where a majority of agents score above 0.5, then rank by mean), and **Variance** (reject candidates with high score variance, then rank by mean among the rest). These baselines isolate the contribution of our consensus formulation by controlling for the agent architecture and scoring procedure.

Tab. 2 presents accuracy across six benchmarks under all verification strategies.We organize the discussion around ***three findings.***

VERDICT consistently improves over the base model, without degradation on any benchmark. Our method improves over the unverified base model on every benchmark, with gains ranging from +2.90 on 3DSRBench to +5.95 on CV-Bench-3D. This consistency holds across qualitatively different tasks: spatial reasoning (3DSRBench, CV-Bench-3D), visual grounding (CV-Bench-2D, AI2D), and multimodal abstraction (BLINK, MMStar).

Domain-specific critics show cross-task fragility. Every domain-specific critic we evaluate degrades below baseline on at least two benchmarks. Sherlock [6] drops 18.26 points on CV-Bench-3D and 8.01 points on 3DSRBench. VisionSR1 [15] collapses by 22.21 points on CV-Bench-3D and 18.22 points on BLINK. Even LLaVA-Critic [33], which achieves the highest single-benchmark score among domain-specific methods on CV-Bench-3D (81.58), simultaneously degrades by 6.75 points on CV-Bench-2D, two splits of the *same* benchmark family. This pattern is consistent with capability tensions documented in prior work [11, 25, 36]: a verifier trained to reward one reasoning style can actively penalize another. Our method sidesteps this entirely by aggregating frozen, modality-specialized judgments through consensus computation rather than learning a single score function that must generalize across heterogeneous evaluation criteria.

VERDICT adds value beyond domain-agnostic aggregation. To isolate the contribution of our consensus formulation, we compare against five aggregation baselines that use identical agents and scoring procedures. The only difference is how raw scores are combined into a selection decision. Among these, Mean scoring emerges as the strongest simple baseline, followed by Max and Variance. However, **VERDICT** consistently outperforms Mean by +0.68 to +2.57 across all benchmarks. This margin confirms that the disagreement structure captured by consensus computation, specifically, the coupling between agents introduced through the coupled scoring framework, provides a verification signal that score averaging discards. It is worth noting that the Variance baseline, which explicitly attempts to incorporate disagreement by rejecting high-variance candidates, still underperforms our method. This indicates that treating all variance equally, without distinguishing genuine cross-modal conflict from ordinary confidence fluctuations, is insufficient (a finding consistent with Proposition 1).

6 Discussion and Analysis

We conduct a series of ablation studies to understand the contribution of each component in our framework.

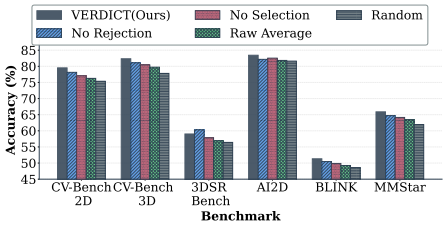

Fig. 3: VERDICT achieves the highest accuracy on five of six benchmarks, with each component (rejection and selection) contributing disparately.

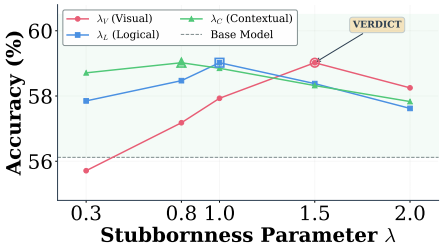

Fig. 4: Stubbornness parameter sensitivity on 3DSRBench. Each curve varies one parameter while holding the others fixed at our configuration. Open markers indicate our chosen values.

6.1 Rejection vs. Selection

To understand whether improvements come primarily from filtering unreliable candidates or from better ranking among viable options, we decompose our method into its constituent mechanisms. We evaluate five strategies: **VERDICT** (our complete method: filter by acceptance criterion, then rank by $\bar{s}^*$), **No Rejection** (skip filtering, rank all candidates by $\bar{s}^* - \Delta^*$), **No Selection** (keep filtering, randomly select among accepted), **Raw Average** (replace consensus scores with simple means and dispersions, keeping the same dual-criterion structure), and **Random** (no filtering, no ranking). Fig. 3 presents the results. **Three findings emerge:** First, rejection and selection contribute differently: removing either degrades accuracy, but removing intelligent ranking (No Selection, -1.17 to -2.17) costs more than removing filtering (No Rejection, -0.26 to -1.08), indicating that consensus-adjusted ranking is the primary driver. Second, Raw Average consistently underperforms **VERDICT** by 2.59-2.95 points despite using the same dual-criterion architecture, confirming that the coupled scoring framework produces better-calibrated scores than naive aggregation independent of the filtering and ranking structure built on top. Third, Random selection performs only marginally above the base model, establishing that verification-aware scoring, not candidate diversity alone drives the gains.

6.2 Stubbornness Sensitivity

We ablate each stubbornness parameter separately on 3DSRBench, varying it over $\{0.3, 0.8, 1.0, 1.5, 2.0\}$ while holding the other two fixed. Fig. 4 reveals a clear sensitivity hierarchy: λ_V has the widest performance range (2.90 percentage points), λ_L an intermediate range (1.40), and λ_C the narrowest (1.19). This ordering directly mirrors our asymmetric design where the parameter with the most impact on spatial reasoning is set highest, while the least impactful one is given the most flexibility. Critically, no setting degrades below the base model, and the three curves converge at their respective optimal values to the same peak accuracy (59.02%), confirming that the chosen configuration sits at a jointly optimal point rather than a compromise. The range of Fig. 4 $\lambda \in [0.3, 2.0]$ spans

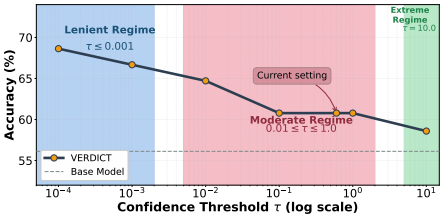

Fig. 5: Confidence threshold sensitivity on 3DSRBench with $\epsilon = 0.1$ fixed. Performance follows a descending gradient.

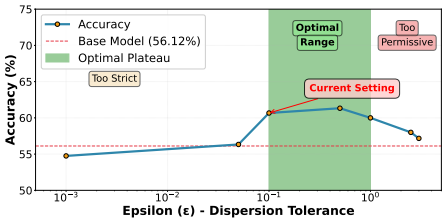

Fig. 6: Dispersion Tolerance Sensitivity on 3DSRBench with $\tau = 0.6$ fixed.

the full coupling transition: at $\lambda = 0.3$ an agent places 77% weight on the group mean (Eq. (2)), making it nearly consensus-driven; at $\lambda = 2.0$ it places 67% on its own evidence, yet the remaining 33% consensus pull still preserves meaningful inter-agent coupling which is the defining property of the framework.

Stubbornness Assignment. The per-parameter sensitivity above establishes *which* parameter matters most; it does not test whether the *assignment* of values to agents matters. We permute the stubbornness triple (λ_V=1.5, λ_L=1.0, λ_C=0.8) via pairwise swaps on 3DSRBench. Swapping L↔C (λ_V unchanged) costs only −0.41 points (58.61%), while swapping V↔L costs −1.19 (57.83%) and V↔C costs −1.76 (57.26%). The degradation is proportional to the reduction in the visual agent's stubbornness. As long as λ_V retains the highest value, **VERDICT** still exceeds the Mean baseline (58.34%); reducing it below either other agent drops performance below Mean. The assignment is not arbitrary: on perception-heavy tasks, visual evidence is the least negotiable signal, and the consensus formulation amplifies this only when the stubbornness structure reflects it.

6.3 Threshold Sensitivity

Confidence threshold τ. We ablate the confidence threshold τ while holding the dispersion tolerance fixed at $\epsilon = 0.1$ on 3DSRBench, and Fig. 5 present results across $\tau \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.6, 1.0, 10.0\}$, extended analysis across operating regimes is provided in the supplementary.For τ in Fig. 5, performance follows a descending gradient across three regimes: permissive ($\tau \leq 0.01$, accuracy 65–69%), intermediate ($\tau = 0.1$–0.6, around 60–61%), and restrictive ($\tau \geq 1.0$, ∼60.6% via fallback ranking). The permissive regime is the most revealing: with the confidence filter effectively disabled, the consensus-adjusted ranking alone substantially exceeds the Mean baseline (58.34%), confirming that the consensus formulation itself, not the filtering architecture, drives the improvement.

Cross-benchmark threshold analysis. We extend the τ sensitivity analysis across all six benchmarks (full results in supplementary). The permissive regime ($\tau \leq 0.01$) achieves higher accuracy on 3DSRBench (68.71% vs. 59.02%), but $\tau = 0.6$ is individually optimal on the remaining five benchmarks.

Dispersion tolerance ϵ. We ablate the dispersion tolerance while holding $\tau = 0.6$ fixed. Fig. 6 presents accuracy across $\epsilon \in \{0.001, 0.05, 0.1, 0.5, 1.0, 2.5, 3.0\}$. For ϵ in Fig. 6 , performance peaks in an optimal plateau at $\epsilon \in [0.1, 1.0]$,

with accuracy around 60–61%. Too strict ($\epsilon < 0.1$) rejects legitimate steps with natural cross-modal tension; too permissive ($\epsilon > 1.0$) admits candidates where judges have not genuinely converged. The decline is gradual in both directions because the fallback mechanism ($\bar{s}^* - \Delta^*$ ranking) and the confidence threshold ($\bar{s}^* > \tau$) respectively provide safety nets, preventing catastrophic failure at any operating point. [3]

6.4 Disentangling Judge Quality from Algorithmic Gain

We replace the 7B verification agents with 4B and 2B variants while holding the base reasoner fixed, isolating scoring quality from candidate quality. Fig. 7 reports results on 3DSRBench. Absolute accuracy degrades monotonically with judge capability for both VERDICT and Mean, as expected. However, the margin between VERDICT and Mean widens from +0.68 at 7B to +0.89 at 4B to +1.21 at 2B, consistent with Proposition 1: the coupled scoring system surfaces noise-induced disagreement through the dispersion filter rather than passing it to the selection decision. At the 2B scale, Mean drops 1.65 points below the unverified baseline while VERDICT with the same judges limits this degradation to 0.44 points, a 1.21 point recovery attributable entirely to the consensus formulation, confirming that the reported gains reflect algorithmic structure rather than judge quality.Additional results can be found in the supplementary materials.

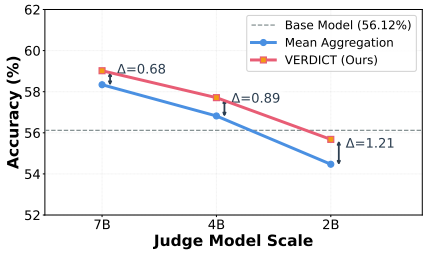

Fig. 7: Disentangling judge quality from algorithmic gain on 3DSRBench.Both methods degrade with weaker judges, but the margin between VERDICT and Mean widens.

7 Conclusion

Multimodal reasoning in MLLMs fails gradually as locally plausible steps drift toward incorrect answers. Our results across six benchmarks show that this drift leaves a detectable trace in the *structure* of cross-modal disagreement. **VERDICT** formalizes this signal through disagreement-aware consensus computation, achieving consistent improvements without task-specific tuning. Our **key takeaway** is that when disparate judges disagree about a reasoning step, the pattern of that disagreement is itself a verification signal that principled aggregation can exploit. The framework has certain limitations. When all agents are confidently wrong, consensus converges on an incorrect assessment; and no filtering recovers from a base model that produces no viable candidates. Characterizing how often this failure mode occurs in practice is deferred to future

[3] Across all six benchmarks at the operating point(τ=0.6, ϵ=0.1), the fallback path is triggered on approximately 15% of reasoning steps, confirming that the dual acceptance criterion admits at least one viable candidate on the large majority of steps

work. On the practical side, **VERDICT** is fully parallelizable and dominated by generation cost, scoring requires 1–5 tokens per step versus 50–200 per reasoning step, yielding 3.80× wall-clock overhead under sequential execution. This is non-trivial but fundamentally cheaper than regenerating entire trajectories, and reducible via adaptive verification at high-uncertainty steps, a direction we leave to future work. Per-agent scores further provide a modality-level diagnostic trace pinpointing failures in visual grounding, logical coherence, or contextual relevance. Unlike domain-specific critics, this overhead is the *entire* cost. Because the coordination-game formulation requires only that heterogeneous evaluators trade off private evidence against consensus, **VERDICT** generalizes beyond multimodal reasoning to reward model ensembles [9, 24], multi-agent evaluation, and compositional code verification.

References

1. 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 (2025), https://arxiv.org/abs/2511.21631
2. Cao, Q., Wang, R., Zhang, R., Somayajula, S.A., Xie, P.: DreamPRM: Domain-reweighted process reward model for multimodal reasoning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=ZyiBk1ZinG
3. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? (2024), https://arxiv.org/abs/2403.20330
4. Chen, Q., Qin, L., Zhang, J., Chen, Z., Xu, X., Che, W.: M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought (2024), https://arxiv.org/abs/2405.16473
5. Coste, T., Anwar, U., Kirk, R., Krueger, D.: Reward model ensembles help mitigate overoptimization (2024), https://arxiv.org/abs/2310.02743
6. Ding, Y., Zhang, R.: Sherlock: Self-correcting reasoning in vision-language models (2025), https://arxiv.org/abs/2505.22651
7. Du, L., Meng, F., Liu, Z., Zhou, Z., Luo, P., Zhang, Q., Shao, W.: Mm-prm: Enhancing multimodal mathematical reasoning with scalable step-level supervision (2025), https://arxiv.org/abs/2505.13427
8. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate (2023), https://arxiv.org/abs/2305.14325
9. Eisenstein, J., Nagpal, C., Agarwal, A., Beirami, A., D'Amour, A., Dvijotham, D., Fisch, A., Heller, K., Pfohl, S., Ramachandran, D., Shaw, P., Berant, J.: Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking (2024), https://arxiv.org/abs/2312.09244

10. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive (2024), https://arxiv.org/abs/2404.12390
11. Kang, W., Kuen, J., Ren, M., Wei, Z., Yan, Y., Liu, K.: Vgent: Visual grounding via modular design for disentangling reasoning and prediction (2025), https://arxiv.org/abs/2512.11099
12. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images (2016), https://arxiv.org/abs/1603.07396
13. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf
14. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
15. Li, Z., Yu, W., Huang, C., Liu, R., Liang, Z., Liu, F., Che, J., Yu, D., Boyd-Graber, J., Mi, H., Yu, D.: Self-rewarding vision-language model via reasoning decomposition (2025), https://arxiv.org/abs/2508.19652
16. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: NeurIPS (2023)
17. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Hallucination in large vision-language models: A survey. arXiv preprint arXiv:2402.00253 (2024)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf
19. Luo, L., Liu, Y., Liu, R., Phatale, S., Guo, M., Lara, H., Li, Y., Shu, L., Zhu, Y., Meng, L., Sun, J., Rastogi, A.: Improve mathematical reasoning in language models by automated process supervision (2024), https://arxiv.org/abs/2406.06592
20. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C.M., Yuille, A.: 3dsrbench: A comprehensive 3d spatial reasoning benchmark (2025), https://arxiv.org/abs/2412.07825
21. Nash, J.F.: Equilibrium points in n-person games. Proceedings of the National Academy of Sciences **36**(1), 48–49 (1950)
22. Rosen, J.B.: Existence and uniqueness of equilibrium points for concave n-person games. Econometrica: Journal of the Econometric Society pp. 520–534 (1965)
23. Roughgarden, T.: Twenty Lectures on Algorithmic Game Theory. Cambridge University Press (2016)
24. Saad-Falcon, J., Buchanan, E.K., Chen, M.F., Huang, T.H., McLaughlin, B., Bhathal, T., Zhu, S., Athiwaratkun, B., Sala, F., Linderman, S., Mirhoseini, A., Re, C.: Weaver: Shrinking the generation-verification gap by scaling compute for verification. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=dRjt4vlYVQ
25. Shen, S., Hou, L., Zhou, T., Yang, S., Wang, Q., Kweon, I.S., Tombari, F., Shen, Y.: Multi-modal preference alignment remedies degradation of visual instruction tuning on language model. arXiv preprint arXiv:2402.10884 (2024)

26. Sun, L., Liang, H., Wei, J., Yu, B., Li, T., Yang, F., Zhou, Z., Zhang, W.: Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. In: arXiv preprint arXiv:2502.13383 (2025)
27. Tian, X., Zou, S., Yang, Z., He, M., Waschkowski, F., Wesemann, L., Tu, P., Zhang, J.: More thought, less accuracy? on the dual nature of reasoning in vision-language models (2025), https://arxiv.org/abs/2509.25848
28. Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., Wang, Z., Fergus, R., LeCun, Y., Xie, S.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms (2024), https://arxiv.org/abs/2406.16860
29. Wang, P., Li, L., Shao, Z., Xu, R.X., Dai, D., Li, Y., Chen, D., Wu, Y., Sui, Z.: Math-shepherd: Verifying and reinforcing mathematical reasoning. In: ACL (2023)
30. Wang, P., Li, L., Shao, Z., Xu, R., Dai, D., Li, Y., Chen, D., Wu, Y., Sui, Z.: Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9426–9439. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.510, https://aclanthology.org/2024.acl-long.510/
31. Wang, W., Gao, Z., Chen, L., Chen, Z., Zhu, J., Zhao, X., Liu, Y., Cao, Y., Ye, S., Zhu, X., Lu, L., Duan, H., Qiao, Y., Dai, J., Wang, W.: Visualprm: An effective process reward model for multimodal reasoning (2025), https://arxiv.org/abs/2503.10291
32. Wang, W., Gao, Z., Chen, L., Chen, Z., Zhu, J., Zhao, X., Liu, Y., Cao, Y., Ye, S., Zhu, X., Lu, L., Duan, H., Qiao, Y., Dai, J., Wang, W.: Visualprm: An effective process reward model for multimodal reasoning (2025), https://arxiv.org/abs/2503.10291
33. Wang, X., Li, C., Yang, J., Zhang, K., Liu, B., Xiong, T., Huang, F.: Llava-critic-r1: Your critic model is secretly a strong policy model (2025), https://arxiv.org/abs/2509.00676
34. Zhang, D., Li, J., Lei, J., Wang, X., Liu, Y., Yang, Z., Li, J., Wang, W., Yang, S., Wu, J., Ye, P., Ouyang, W., Zhou, D.: Critic-v: Vlm critics help catch vlm errors in multimodal reasoning (2025), https://arxiv.org/abs/2411.18203
35. Zhang, D., Li, J., Lei, J., Wang, X., Liu, Y., Yang, Z., Li, J., Wang, W., Yang, S., Wu, J., Ye, P., Ouyang, W., Zhou, D.: Critic-v: Vlm critics help catch vlm errors in multimodal reasoning (2025), https://arxiv.org/abs/2411.18203
36. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Zhang, L., Li, C., Yang, J.: Llava-grounding: Grounded visual chat with large multimodal models. In: European Conference on Computer Vision (ECCV) (2024)
37. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models (2024), https://arxiv.org/abs/2302.00923