DREAM: DYNAMIC RESOLUTION ASSIGNMENT FOR MULTIMODAL MULTI-AGENT DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent debate (MAD) has emerged as an effective paradigm to improve the reasoning capabilities of large language models (LLMs) and is increasingly being extended to multimodal settings. However, existing multimodal MAD frameworks typically expose agents to the same fixed visual input, ignoring substantial variation in the visual scale needed across samples and agents. In addition, these frameworks frequently suffer from *groupthink*—a phenomenon where agents prematurely abandon correct deductions to conform with confident but hallucinated peer responses. To address these bottlenecks, we introduce **DREAM** (**D**ynamic **Re**solution **A**ssignment For **M**ultimodal Multi-Agent Debate), which operates via two core components: (1) Dynamic Resolution Assignment, a zero-shot "probe round" where agents test multiple resolutions, quantify uncertainty using Average Normalized Log-Likelihood (ANLL), and use an adaptive threshold to assign each agent to its empirically optimal resolution; (2) Uncertainty-Guided Rollback Aggregation counters groupthink by tracking each agent's uncertainty over rounds and restoring early low-uncertainty answers overridden by group pressure. On six multimodal datasets, DREAM improves the accuracy–token trade-off over multi-agent debate baselines by 1.5–3.2% accuracy without dataset-specific tuning.

1 INTRODUCTION

Reasoning in large language models (LLMs) has advanced rapidly, with significant progress driven by the orchestration of multiple model instances. Multi-agent debate (MAD) (Du et al., 2024; Liang et al., 2024) has emerged as a widely adopted paradigm where agents iteratively exchange answers and critiques, revising their responses under the assumption that logical errors are more easily detected via peer review. Recently, MAD has been extended to multimodal reasoning, standardizing visual collaboration in benchmarks like M3MAD-Bench (Li et al., 2026). However, while debate effectively refines textual logic, directly porting it to vision-language tasks exposes severe structural bottlenecks, specifically regarding visual scale and consensus integrity.

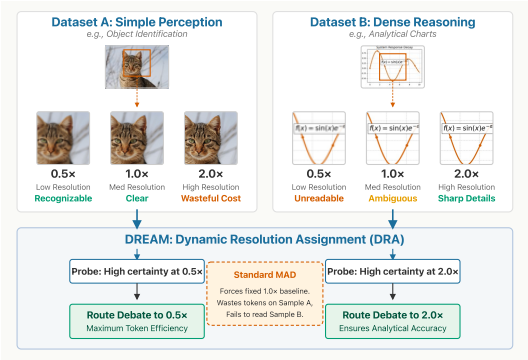


Figure 1: Visual tasks demand different resolutions: simple perception tasks (left) only need $0.5\times$, while $2.0\times$ introducing long-context noise that actively degrades accuracy. Conversely, dense charts (right) require $2.0\times$ to preserve critical fine-grained details, thus $0.5\times$ irreversibly discarding them. Standard MAD uses a fixed resolution, degrading on simple samples and failing on complex ones. DREAM solves this by dynamically assigning each sample to its empirically optimal scale.

The first major bottleneck in existing frameworks is the static handling of visual resolution. Although modern Multimodal LLMs (MLLMs) (OpenAI, 2023; Gemini Team et al., 2024; Bai et al., 2023) natively support dynamic resolution encoding, current multi-agent systems force a rigid, one-size-fits-all visual scale. Simply maximizing resolution is inadequate: excessively high resolutions on simple perception tasks actively degrade performance by injecting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

long-context noise and spurious features (Wang et al., 2024b; Dong et al., 2024), while low resolutions obscure critical text in dense documents (Figure 1). Because the optimal visual scale is highly sample-dependent (Niu et al., 2025; Li et al., 2025), static assignments leave agents fundamentally ungrounded. Crucially, this loss of visual grounding directly deepens a second bottleneck: *sycophantic groupthink*. Standard MAD inherently suffers from herding (Wynn et al., 2025; Sharma et al., 2023; Xiong et al., 2023), and multimodality amplifies this vulnerability. When suboptimal resolution deprives agents of the ability to independently verify visual claims, a single hallucinated assertion can easily trigger a cascade of blind agreement. Under this peer pressure, valid early deductions are displaced, which is a phenomenon measurable as certainty erosion, and the final consensus becomes dominated by sycophancy rather than reasoned deduction.

To jointly resolve these structural issues, this work introduces **DREAM** (**D**ynamic **Re**solution **A**ssignment for **M**ultimodal Multi-Agent Debate), a plug-and-play wrapper featuring two complementary mechanisms: (1) *Dynamic Resolution Assignment (DRA)*: An exploratory "Round 0" probe dynamically determines the optimal visual scale for each sample using Average Normalized Log-Likelihood (ANLL) as a zero-shot uncertainty signal. Based on initial consensus, DRA assigns agents to either a locked single scale to minimize noise, or a mixed-resolution payload for maximal visual coverage. (2) *Uncertainty-Guided Rollback Aggregation (URA)*: A targeted defense against late-stage groupthink that monitors ANLL trajectories throughout the debate. If an agent's uncertainty sharply rises by the final round, URA rescues their most confident prior response. To counteract judge recency bias, rescued answers are strategically prepended to the final prompt, ensuring early, visually grounded deductions are weighed equally against the prevailing groupthink.

Our main contributions are summarized as follows:

- We identify and formally characterize two structural bottlenecks universal to all existing multimodal MAD frameworks: static visual resolution assignment and late-stage sycophantic groupthink, and demonstrate how these two challenges are causally intertwined.
- We propose DREAM, to our knowledge the first multimodal MAD framework to jointly resolve these bottlenecks. DRA introduces a zero-shot uncertainty-driven probing phase that dynamically routes each sample to its empirically optimal resolution strategy, while URA provides a principled rollback mechanism that prevents correct early deductions from being displaced by group pressure.
- We conduct extensive experiments across six multimodal benchmarks and four open-weight MLLMs, demonstrating that DREAM consistently improves the accuracy-token trade-off as a plug-and-play wrapper over existing debate baselines and other multimodal agentic approaches, without requiring any dataset-specific tuning.

2 Related Work

2.1 Multi-Agent Debate for LLMs

Du et al. (2024) introduced multi-agent debate (MAD) to enhance factual accuracy by allowing multiple LLM instances to critique and refine each other's responses. Subsequent research has diversified these structures (Liang et al., 2024; Liu et al., 2025), though some explicitly question whether debate consistently outperforms advanced single-model strategies (Wang et al., 2024c; Smit et al., 2024). Across all configurations, a persistent vulnerability remains groupthink, where agents converge on an answer regardless of correctness due to herding and sycophancy (Sharma et al., 2023; Wynn et al., 2025). While recent approaches attempt to mitigate this using voting mechanisms, verbal exchanges, or confidence-weighted updates (Benedikt Kaesberg et al., 2025; Lin & Hooi, 2025; Tian et al., 2026), they still treat uncertainty strictly as an intra-debate filter. Crucially, none of these methods track how an individual agent's uncertainty evolves across rounds, nor do they leverage it to dynamically assign the optimal visual inputs prior to the debate.

MAD has also been extended to multimodal settings, including deception detection (Fang et al., 2025) and visual misinformation (Lakara et al., 2024), while Huang et al. (2025) separate perception from reasoning via a dedicated perceiver agent. M3MAD-Bench (Li et al., 2026) further standardizes multimodal debate and highlights substantial token inefficiencies. Critically, all existing approaches operate under static initial conditions, assigning identical visual inputs to every agent

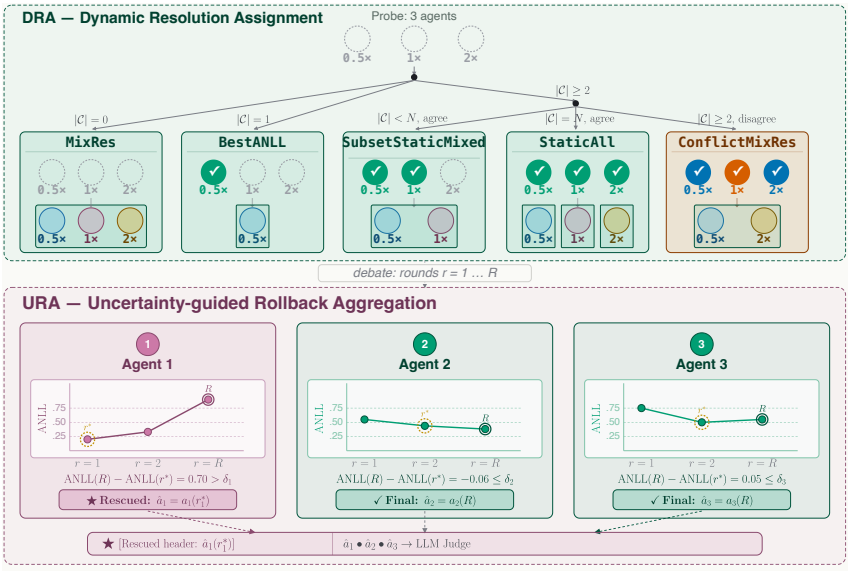


Figure 2: **The DREAM framework pipeline. Top:** DRA's probe-round decision tree, selecting a DRA strategy based first on confident-set size $|\mathcal{C}|$ and then, when $|\mathcal{C}| \geq 2$, on answer agreement within $\mathcal{C}$; the disagreement strategy (CONFLICTMIXRES) is highlighted in orange. **Bottom:** URA's per-agent ANLL trajectory over debate rounds $r = 1, 2, \ldots, R$, comparing each agent's final round to its earlier peak r^* (gold marker) against the rescue threshold δ_i. Agent 1 (pink) exceeds δ_1 and is rescued; Agents 2–3 (green) keep their final answers. Green marks the default path throughout.

and leaving per-sample resolution assignment entirely unaddressed. To our knowledge, DREAM is the first multimodal MAD framework to treat resolution as a per-agent, per-sample decision and to jointly counter groupthink by tracking each agent's uncertainty trajectory rather than relying on fixed aggregation or prompting strategies.

2.2 UNCERTAINTY AND CONFIDENCE IN MULTI-AGENT SYSTEMS

Single-model uncertainty estimation spans token-level self-knowledge, semantic uncertainty, and hallucination detection (Kuhn et al., 2023; Xiong et al., 2024). In multi-agent settings, uncertainty metrics are primarily used to filter message propagation between agents (Lin & Hooi, 2025; Tian et al., 2026), although such filtering can inadvertently restrict output diversity (Nguyen et al., 2026). However, a critical connection between uncertainty and groupthink remains unexploited: when an agent is pressured into abandoning a visually grounded answer, its confidence in the coerced response drops, presenting as a measurable rise in uncertainty across debate rounds. In other words, groupthink leaves a detectable uncertainty trace. DREAM leverages this phenomenon in two distinct ways: first, as a zero-shot pre-debate signal to route agents to their optimal visual resolution; and second, as a per-round trajectory tracker that identifies compromised agents, rescuing their most confident earlier deduction when late-stage uncertainty exceeds a calibrated threshold.

3 METHODOLOGY

We introduce DREAM, a multi-agent multimodal debate framework comprising two key innovations: Dynamic Resolution Assignment (DRA) and Uncertainty-Guided Rollback Aggregation (URA). DRA conducts a preliminary round to identify the most suitable resolution for each agent based on the current sample. Meanwhile, URA monitors each agent's responses throughout the debate, allowing it to recover the low-uncertainty answer that might have been overlooked due to groupthink. Figure 2 details the branching logic of the Route step and how its output feeds the unchanged debate phase and the rollback aggregation step. The full four-phase procedure (Probe → Route → Debate → Aggregate & Judge) is given as Algorithm 1.

3.1 OVERVIEW

Given a visual question-answering sample (I, q), a standard MLLM $\mathcal{M}$ produces an answer $a = \mathcal{M}(I_s, q)$, where I_s denotes image I rendered at a resolution scale s drawn from a discrete candidate set $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$. With $\mathcal{M}$, q, and I held constant, both the model's answer and its output uncertainty vary with s, and the direction and magnitude of this change are sample-dependent. Naively assigning the highest resolution to all agents is suboptimal because different tasks and image types benefit from different scales; likewise, feeding *all* resolutions to *every* agent by default inflates the visual token count and can introduce noise-induced hallucinations (we therefore reserve multi-scale inputs for cases where probing indicates that no single scale is reliably confident).

In a multi-agent debate, N agents $\{\mathcal{M}_i\}_{i=1}^{N}$ exchange candidate answers over R rounds. At round r, agent i receives the image at its assigned resolution $s_i^{(r)} \in \mathcal{S}$, its own prior response, and its peers' responses, and produces an updated answer. In standard MAD, $s_i^{(r)}$ is the same fixed value for all i and all r. DREAM relaxes this: DRA selects a per-agent, per-sample resolution s_i^* in a preliminary probe round. More importantly, URA tracks each agent's certainty trajectory $\{\mathrm{ANLL}_i^{(r)}\}_{r=1}^{R}$ across rounds to identify and rescue certainty-diminished answers at aggregation time.

The output certainty of each agent is measured using the ANN. ANLL is preferred because it is straightforward, length-normalized, and inexpensive to compute from the model's token log-probabilities. This makes it easy to track throughout each debate round without requiring additional forward passes or auxiliary models. Although there are many alternative uncertainty metrics available, such as those based on ensembles, self-consistency, or calibrated sampling, we choose ANLL as a lightweight and broadly applicable measure of generation confidence.

$$\mathrm{ANLL}_i^{(r)} = -\frac{1}{|y_i^{(r)}|} \sum_{t=1}^{|y_i^{(r)}|} \log p_\theta\left(y_{i,t}^{(r)} \,\middle|\, y_{i,<t}^{(r)},\, x_i^{(r)}\right) \tag{1}$$

where $x_i^{(r)}$ is the agent's input at round r and $y_i^{(r)}$ is its generated response. A lower ANLL indicates that the model assigned high probability to every token it generated.

3.2 DYNAMIC RESOLUTION ASSIGNMENT

Probe round. Before any debate exchange occurs, DREAM inserts a *Round 0* whose history is discarded and never shared with peers. Each agent i is assigned a distinct resolution $s_i^{(0)} \in \mathcal{S}$ from the candidate set in round-robin order, so the N agents collectively span the resolution range. Every agent answers the question independently at its assigned resolution using greedy decoding (temperature $= 0$), and we record both its extracted answer $\hat{a}_i^{(0)}$ and its $\mathrm{ANLL}_i^{(0)}$.

Adaptive confidence threshold. Given the probe ANLL vector $\{\mathrm{ANLL}_i^{(0)}\}_{i=1}^{N}$, DREAM computes an effective confidence threshold τ_{eff} that adapts to the sample's difficulty. We examine two thresholding schemes. Under *static* thresholding, $\tau_{\text{eff}} = \tau$ is a fixed scalar (the default is set to 0.2). Under *adaptive* thresholding, the threshold is set to one standard deviation below the mean of the probe-round ANLL scores:

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{N} \mathrm{ANLL}_i^{(0)}, \quad \sigma_0 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{ANLL}_i^{(0)} - \mu_0\right)^2}, \quad \tau_{\text{eff}} = \mu_0 - \sigma_0 \tag{2}$$

Agents whose ANLL falls at least one standard deviation below the sample mean are classified as genuinely confident; those above τ_{eff} are not. This formulation is self-calibrating: on easy samples where all ANLLs are low and tightly clustered, τ_{eff} is also low and the bar is appropriately strict; on hard samples with high, spread-out ANLLs, τ_{eff} rises accordingly. When fewer than two valid probe scores are available, DREAM falls back to the static threshold τ. Agent i is classified as *confident* if $\mathrm{ANLL}_i^{(0)} \leq \tau_{\text{eff}}$.

Uncertainty-Guided Resolution Assignment. After the probe round, each agent i holds an answer $\hat{a}_i^{(0)}$ and an uncertainty score $\mathrm{ANLL}_i^{(0)}$. Agents whose score falls below the adaptive threshold τ_{eff}

form the *confident set* $\mathcal{C} = \{i : \mathrm{ANLL}_i^{(0)} \leq \tau_{\text{eff}}\}$, and their probe resolutions are collected as $\mathcal{S}_{\mathcal{C}}^* = \{s_i^{(0)} : i \in \mathcal{C}\}$. The composition and consensus of $\mathcal{C}$ then determine which resolution payload each agent receives for all subsequent debate rounds:

1. **No confident agent ($|\mathcal{C}| = 0$).** No resolution produced a reliable signal. All agents receive the full multi-scale payload $\mathcal{S}$ (MIXRES).
2. **Exactly one confident agent ($|\mathcal{C}| = 1$).** A single resolution has proven reliable. All agents are locked to that agent's probe resolution (BESTANLL).
3. **All agents confident and in agreement ($|\mathcal{C}| = N$ and $\hat{a}_i^{(0)} = \hat{a}_j^{(0)}\ \forall i, j \in \mathcal{C}$).** Every agent both converged and agreed. Each agent retains its own probe resolution, preserving the diversity across scales established during probing (STATICALL).
4. **A strict confident subset, in agreement ($0 < |\mathcal{C}| < N$ and $\hat{a}_i^{(0)} = \hat{a}_j^{(0)}\ \forall i, j \in \mathcal{C}$).** Confident agents have identified reliable resolutions, but unconfident agents have not. Confident agents retain their own probe resolutions, while unconfident agents receive the mixture of $\mathcal{S}_{\mathcal{C}}^*$, exposing them to the scales that produced reliable signals in their peers. When all confident agents were probed at the same scale, this reduces to a single locked resolution (SUBSETSTATICMIXED / BESTANLL).
5. **Confident agents are divided ($|\mathcal{C}| \geq 2$ and $\exists i, j \in \mathcal{C}$ s.t. $\hat{a}_i^{(0)} \neq \hat{a}_j^{(0)}$).** Agreement failed despite low uncertainty, signaling that the query is genuinely ambiguous across scales. A majority vote is taken over the confident agents' answers. If a strict majority exists, the minority resolutions are discarded, and all agents receive the majority-supporting resolutions $\mathcal{S}^\dagger \subseteq \mathcal{S}_{\mathcal{C}}^*$ as a single scale or mixture. If no strict majority exists, all of $\mathcal{S}_{\mathcal{C}}^*$ is used as a mixture to maximize visual coverage (CONFLICTMIXRES).

In all cases, the probe history is discarded before the debate begins, ensuring agents enter without anchoring bias from Round 0 responses.

3.3 UNCERTAINTY-GUIDED ROLLBACK AGGREGATION

Per-agent ANLL trajectory. During every debate round $r \in \{1, \ldots, R\}$, DREAM records each agent's ANLL score alongside its response, forming a trajectory $\{\mathrm{ANLL}_i^{(r)}\}_{r=1}^R$. This trajectory captures how certainty evolves under peer pressure: an agent that begins confident but is persuaded by an incorrect argument will typically exhibit a rising ANLL in later rounds, which is a signal of *certainty erosion*.

Debate-calibrated rescue threshold. For each agent i, we compute the standard deviation of its ANLL trajectory across the R debate rounds:

$$\sigma_i^{\text{debate}} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\Big(\mathrm{ANLL}_i^{(r)} - \bar{\mu}_i\Big)^2}, \quad \bar{\mu}_i = \frac{1}{R}\sum_{r=1}^{R}\mathrm{ANLL}_i^{(r)} \tag{3}$$

The per-agent rescue threshold is then $\delta_i = \alpha \cdot \sigma_i^{\text{debate}}$, where α is a scalar hyperparameter (default 0.5). A larger σ_i^{debate} means agent i's confidence fluctuated widely during debate, so δ_i is raised to avoid triggering rescue on natural variation. A small σ_i^{debate} means the agent was consistently certain throughout, so δ_i is tight, and any late-round erosion is readily detected.

Answer rescue. At aggregation time, let $r_i^* = \arg\min_{r<R} \mathrm{ANLL}_i^{(r)}$ be agent i's most confident round strictly before the final. Rescue triggers for agent i if:

$$\mathrm{ANLL}_i^{(R)} - \mathrm{ANLL}_i^{(r_i^*)} > \delta_i \tag{4}$$

When rescue triggers, agent i's contribution is replaced by its answer from round r_i^* rather than round R. By construction, rescue can only point to an earlier round.

Anti-recency bias. The judge receives exactly one answer per agent: the rescued answer $a_i^{(r_i^*)}$ if rescue was triggered, or the final-round answer $a_i^{(R)}$ otherwise. Since LLM judges tend to overweight answers that appear later in the prompt due to position bias, DREAM prepends a structured header for each rescued answer at the top of the judge's prompt:

`[Agent` i `Rescued Answer (Round` r_i^*`): ``answer'']`

The judge is then instructed to weigh rescued answers equally with all others, ensuring that an earlier, lower-uncertainty answer is not discounted simply because it originates from a prior round.

4 Experimental Setup

Datasets. The DREAM framework is evaluated on the multimodal evaluation subset of M3MAD-Bench (Li et al., 2026): six benchmarks spanning knowledge (MME), mathematics (MathVista, MathVision), medicine (PathVQA), and reasoning (MME-Reasoning, VisualPuzzles). Appendix A provides full dataset descriptions and the experimental setup.

Models. To assess the framework across diverse architectures and model sizes, the experiments employ four primary multimodal large language models (MLLMs): **Gemma-3-4B**, **Qwen2.5-VL-7B**, **InternVL3-8B**, and **GPT-4o-mini**. These models form the core analysis presented in Table 1. More evaluated models are in Appendix Table 5.

Baselines. Comparisons are made against standard inference methods and existing multi-agent frameworks. Single-agent baselines include **IO** (direct answer), **CoT** (Wei et al., 2022), and **SC** (Wang et al., 2024d) at native resolution. Multi-agent debate baselines include **LLM Debate** (Du et al., 2024), **Div-MAD** (Liang et al., 2024), **DMAD** (Liu et al., 2025), and **Debate with Images (DwI)** (Fang et al., 2025). Finally, to isolate dynamic resolution assignment from the underlying debate logic, the textual debate baselines are evaluated under three distinct **Resolution Wrappers**: bare (fixed resolution), **Perceiver** (Huang et al., 2025), and the proposed **DREAM** framework.

Implementation Details. All experiments utilize greedy decoding (temperature $= 0$) for the initial probe round to ensure deterministic, reliable uncertainty measurement, and temperature $= 0.5$ for subsequent debate rounds to encourage conversational diversity. The rescue scaling hyperparameter $\alpha = 0.5$ is fixed across all datasets; the per-agent rescue threshold is defined as $\delta_i = \alpha \cdot \sigma_i^{\text{debate}}$, where σ_i^{debate} is the standard deviation of agent i's ANLL trajectory across the debate rounds (Eq. 3). ANLL is computed over the full response sequence, excluding special tokens. To rigorously account for generation variance during the debate rounds, all reported metrics represent the mean accuracy computed across three independent random seeds. Full implementation details are reported in Appendix B.

5 Results and Analysis

5.1 DREAM Outperforms Baselines on M3MAD-Bench

Table 1 shows the results from three small-scale models (with up to 8 billion parameters), one frontier model, and six benchmarks. Instead of just evaluating isolated performance, our analysis explores the structural limitations of current multi-agent frameworks, emphasizing how DREAM effectively addresses these challenges. GPT-4o-mini with Perceiver was excluded because it could not be used.

The Structural Failures of Div-MAD, Perceiver, and DwI. Among the evaluated baselines, LLM Debate and DMAD hit a strict performance plateau, while Div-MAD (Liang et al., 2024), Perceiver (Huang et al., 2025), and DwI (Fang et al., 2025) exhibit poor performance. Div-MAD consistently ranks as the weakest baseline in our scope (e.g., dropping to a mere 24.5% on Gemma-3-4B). We attribute this to its early termination mechanism; in complex multimodal tasks, premature halting interrupts discussions before agents can resolve visual ambiguities and reach consensus. Perceiver-based approaches fail due to the inherent spatial compression bottlenecks (Huang et al., 2025), where mapping visual inputs to a fixed-capacity latent space aggressively destroys fine-grained details. Finally, while DwI attempts to improve visual grounding through spatial zooming, it continues to force the model to process these crops at a fixed, static resolution. This rigid reliance on zooming without dynamic resolution scaling leaves agents unable to effectively balance dense local details with global image context.

DREAM Shows Consistent, Architecture-Robust Generalization. Dynamic Resolution Assignment (DRA) and Uncertainty-guided Rollback Aggregation (URA) directly address these bottle-

Table 1: Performance on M3MAD-Bench across four distinct model families (results for an additional model are provided in Appendix Table 5). Best results are **bolded**, second-best underlined. Δ denotes DREAM's average improvement over the LLM Debate baseline. *MVt*: MathVista, *MVs*: MathVision, *PVQA*: PathVQA, *MME-R*: MME-Reasoning, *VP*: VisualPuzzles.

Model	Method	Know.	Mathematics		Med.	Reasoning		Avg.	Δ
		MME	MVt	MVs	PVQA	MME-R	VP		
Gemma-3-4B	IO	79.3	37.4	15.0	24.9	21.8	22.9	33.5	
	CoT [33]	80.2	37.4	15.7	24.9	21.8	22.9	33.8	
	SC [32]	80.1	37.1	15.1	24.9	20.8	22.8	33.5	
	DwI [7]	58.9	23.7	8.7	26.1	23.0	18.6	26.5	
	LLM Debate [6]	78.4	57.0	24.3	40.6	26.5	36.3	43.9	
	Div-MAD [17]	32.7	32.7	17.8	17.6	19.7	26.3	24.5	
	DMAD [19]	79.8	51.6	21.6	39.5	25.4	36.1	42.3	
	Perceiver [11]	**82.5**	52.4	23.1	35.0	26.0	32.7	41.9	
	DREAM	80.7	**57.3**	**27.3**	**41.6**	**28.2**	**37.0**	**45.4**	↑1.5
Qwen2.5-VL-7B	IO	73.4	64.8	25.2	45.4	27.8	34.4	45.2	
	CoT [33]	67.8	66.6	23.0	26.0	32.0	34.8	41.7	
	SC [32]	76.8	67.0	27.8	27.0	30.0	36.8	44.2	
	DwI [7]	73.2	67.3	20.8	26.9	26.8	33.6	41.4	
	LLM Debate [6]	72.3	68.3	26.3	42.2	33.2	39.5	47.0	
	Div-MAD [17]	57.8	62.2	24.2	17.1	26.4	33.2	36.8	
	DMAD [19]	74.2	68.6	27.1	**46.2**	32.3	35.3	47.3	
	Perceiver [11]	54.1	65.4	24.4	39.1	30.4	38.9	42.1	
	DREAM	**81.9**	**73.3**	**28.5**	41.6	**34.5**	**41.6**	**50.2**	↑3.2
InternVL3-8B	IO	85.0	65.8	27.0	**52.0**	30.4	38.2	49.7	
	CoT [33]	85.0	66.6	27.6	44.4	29.0	34.8	47.9	
	SC [32]	86.0	65.0	29.0	32.0	28.8	37.8	46.4	
	DwI [7]	81.6	66.3	23.6	38.5	30.9	37.3	46.4	
	LLM Debate [6]	88.6	70.6	28.7	37.4	30.8	37.5	48.9	
	Div-MAD [17]	42.4	44.7	29.0	32.6	26.3	31.3	34.4	
	DMAD [19]	85.2	67.2	**29.4**	42.2	29.8	37.0	48.5	
	Perceiver [11]	84.8	70.0	25.5	39.1	30.7	34.7	47.5	
	DREAM	**89.6**	**73.3**	**29.4**	42.9	**31.6**	**41.7**	**51.4**	↑2.5
GPT-4o-mini	IO	72.6	57.0	32.0	39.6	26.0	38.8	44.3	
	CoT [33]	65.2	57.6	26.2	14.4	30.0	37.4	38.5	
	SC [32]	63.2	59.4	30.8	18.0	27.4	40.6	39.9	
	LLM Debate [6]	71.2	59.6	28.8	29.4	29.8	**43.4**	43.7	
	Div-MAD [17]	61.8	51.4	28.0	21.0	27.6	37.8	37.9	
	DMAD [19]	**76.0**	56.2	27.4	37.5	28.9	40.4	44.4	
	DREAM	73.0	**61.2**	**32.9**	**41.3**	**31.4**	41.6	**46.9**	↑3.2

necks. Across four diverse model families and all three debate frameworks (yielding 12 distinct model×family combinations), equipping the baseline with DREAM improves the average accuracy in **all 12 evaluated cases**. These results suggest that DREAM is not an overfitted heuristic, but a robust, plug-and-play wrapper that consistently improves visual scaling and uncertainty-aware aggregation across a range of MLLMs.

5.2 Generalization Across Debate Frameworks

To demonstrate that our approach addresses fundamental flaws in multimodal reasoning rather than overfitting to a specific configuration, we apply DREAM as a modular upgrade to three distinct general debate backbones: standard LLM Debate (Du et al., 2024), Div-MAD (Liang et al., 2024), and DMAD (Liu et al., 2025). Because DREAM operates entirely as a plug-and-play wrapper, it integrates seamlessly without altering their core messaging logic. As shown in Table 2, equipping these frameworks with DREAM yields universal accuracy improvements across all evaluated MLLMs. In contrast, the Perceiver (Huang et al., 2025) wrapper, which isolates perception by compressing the

Table 2: Performance of the evaluated models under different wrapper methods (*Perceiver* vs. *DREAM*). Best results are **bolded**, second-best underlined. Δ denotes the average accuracy improvement of DREAM over the corresponding bare debate baseline. *MVt*: MathVista, *MVs*: MathVision, *PVQA*: PathVQA, *MME-R*: MME-Reasoning, *VP*: VisualPuzzles.

Model	Method	Know.	Mathematics		Med.	Reasoning		Avg.	Δ
		MME	MVt	MVs	PVQA	MME-R	VP		
Gemma-3-4B	Perceiver [11]								
	w/ LLM Debate [6]	**82.5**	52.4	23.1	35.0	26.0	32.7	41.9	
	w/ Div-MAD [17]	43.0	45.5	20.6	18.3	20.5	30.3	29.7	
	w/ DMAD [19]	81.7	56.1	22.1	38.9	25.8	30.0	42.4	
	DREAM								
	w/ LLM Debate [6]	80.7	**57.3**	**27.3**	41.6	**28.2**	**37.0**	**45.4**	**↑1.5**
	w/ Div-MAD [17]	33.4	39.5	18.4	18.0	20.6	26.1	26.0	**↑1.5**
	w/ DMAD [19]	80.1	52.3	22.9	**43.0**	25.7	36.9	43.5	**↑1.2**
Qwen2.5-VL-7B	Perceiver [11]								
	w/ LLM Debate [6]	54.1	65.4	24.4	39.1	30.4	38.9	42.1	
	w/ Div-MAD [17]	52.3	63.1	21.9	15.5	25.7	32.5	35.2	
	w/ DMAD [19]	39.3	53.5	27.8	28.0	28.5	36.1	35.5	
	DREAM								
	w/ LLM Debate [6]	**81.9**	**73.3**	28.5	41.6	**34.5**	**41.6**	**50.2**	**↑3.2**
	w/ Div-MAD [17]	59.1	63.5	25.8	18.6	27.6	33.9	38.1	**↑1.3**
	w/ DMAD [19]	76.7	71.6	**29.9**	**47.9**	33.4	37.3	49.5	**↑2.2**
InternVL3-8B	Perceiver [11]								
	w/ LLM Debate [6]	84.8	70.0	25.5	39.1	30.7	34.7	47.5	
	w/ Div-MAD [17]	58.9	46.1	21.5	34.2	26.3	31.7	36.5	
	w/ DMAD [19]	80.0	64.4	23.1	41.3	26.7	35.2	45.1	
	DREAM								
	w/ LLM Debate [6]	**89.6**	**73.3**	29.4	42.9	**31.6**	**41.7**	**51.4**	**↑2.5**
	w/ Div-MAD [17]	43.4	45.1	30.7	33.9	28.0	34.0	35.9	**↑1.5**
	w/ DMAD [19]	85.9	70.3	**31.9**	**44.9**	31.4	38.1	50.4	**↑1.9**
GPT-4o mini	**DREAM**								
	w/ LLM Debate [6]	73.0	**61.2**	**32.9**	**41.3**	**31.4**	41.6	**46.9**	**↑3.2**
	w/ Div-MAD [17]	62.3	51.9	29.0	20.9	29.3	39.5	38.8	**↑0.9**
	w/ DMAD [19]	**77.8**	58.2	28.8	38.6	30.9	**42.1**	46.1	**↑1.7**

image into a fixed number of latent tokens, consistently degrades performance. This confirms that aggressive spatial compression destroys necessary fine-grained details, whereas DREAM preserves better grounding through dynamic resolution scaling.

Crucially, analyzing these combinations reveals how different backbones interact with visual assignment. The standard LLM Debate backbone pairs best with DREAM, achieving the highest absolute averages (e.g., 51.4% on InternVL3-8B) because its unrestricted, multi-round exchanges allow DREAM's URA to capture the full trajectory of certainty erosion and roll back groupthink effectively. Conversely, Div-MAD remains the weakest backbone; its early termination mechanism prematurely cuts debates short, preventing agents from resolving visual ambiguities. While DREAM still improves Div-MAD over its bare baseline (e.g., ↑1.3% on Qwen), it cannot fully reverse the structural damage of halting reasoning too early. Finally, DMAD—which uses adversarial roles to break textual groupthink—receives strong, stackable boosts from DREAM (e.g., reaching 49.5% on Qwen), proving that our uncertainty-guided visual rescue provides orthogonal benefits even when the underlying backbone already attempts to mitigate consensus errors.

6 Ablation Study

6.1 URA Rescues Certainty-Erosion Cases

Table 3: Ablation of DREAM components and their contribution to average accuracy (%).

Method	Gemma-3-4B	Qwen2.5-VL-7B	InternVL3-8B
Baseline (LLM Debate)	43.9	47.0	48.9
+ DRA (Resolution Probing)	44.7 (↑**0.8**)	48.3 (↑**1.3**)	49.6 (↑**0.7**)
+ URA (Adaptive Threshold)	45.1 (↑**1.2**)	48.9 (↑**1.9**)	50.5 (↑**1.6**)
+ URA (Rollback)	45.4 (↑**1.5**)	50.2 (↑**3.2**)	51.4 (↑**2.5**)

Table 3 ablates the contribution of each DREAM module to the final debate consensus. The addition of DRA consistently improves upon the baseline (e.g., ↑ 1.3% on Qwen2.5-VL). Implementing URA further compounds these gains; applying the Adaptive Threshold provides a solid intermediate boost, while the complete Confidence-Guided Rescue mechanism achieves peak accuracy across all evaluated models. Crucially, these continuous improvements are observed across fundamentally different model architectures. This demonstrates that DREAM generalizes seamlessly across diverse MLLM families, proving that dynamically rescuing early, high-confidence answers is a universally effective strategy for mitigating late-stage groupthink and hallucination.

To ensure the robustness of these modules, the sensitivity of DREAM's core hyperparameters was systematically evaluated (detailed configurations and full sweep results are provided in Appendix Table 8). For URA, the rescue sensitivity multiplier α strictly governs the threshold for rollback. Empirical evaluations indicate that $\alpha = 0.5$ yields the optimal balance across all tested models. Lower values trigger rollbacks too aggressively in response to natural token probability variations, thereby disrupting valid late-round consensus. Conversely, higher values (e.g., $\alpha \geq 1.0$) render the rollback mechanism overly conservative, allowing subtle but destructive instances of certainty erosion to slip through uncorrected. Similarly, for DRA, the adaptive thresholding strategy ($\mu_0 - \sigma_0$) systematically outperforms fixed static scalars by dynamically calibrating the confidence boundary to the inherent visual difficulty of each individual query. Appendix G.2 breaks this behavior down further by individual DRA branch, showing that no single branch dominates uniformly across benchmarks and that each trades off accuracy against token cost differently, which motivates assigning per-sample rather than committing to one fixed strategy.

6.2 Accuracy vs. Computation

The primary objective of DREAM is not strict token reduction, but rather the optimization of the accuracy-versus-tokens trade-off. Figure 3 illustrates this trade-off by plotting the bare debate baselines against their *w/ DREAM*-wrapped counterparts for Qwen2.5-VL-7B. While accuracy consistently improves across all three debate families (yielding gains of ↑1.3% to ↑3.2%), this performance is accompanied by an expected increase in token consumption due to multi-scale visual processing. Consequently, the wrapped configurations incur higher computational costs than their bare counterparts. Ultimately, the integration of DREAM systematically shifts the trade-off upward, demonstrating that the strategic investment of additional visual tokens translates directly into superior multimodal reasoning capabilities that standard debate structures cannot achieve.

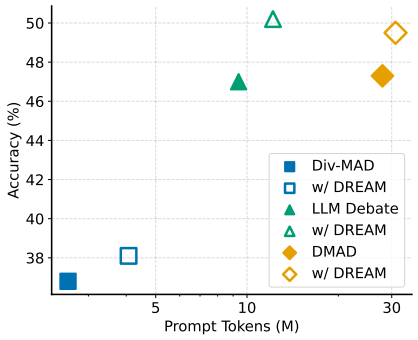


Figure 3: Accuracy vs. prompt token consumption for Qwen2.5-VL-7B between baselines and DREAM.

7 Conclusion

We have presented DREAM, a framework designed to overcome critical efficiency and consensus bottlenecks in multi-agent multimodal debate. DREAM introduces two core mechanisms: Dynamic Resolution Assignment (DRA), which transforms image resolution scaling into a data-driven, per-sample decision using zero-shot ANLL uncertainty signals; and Uncertainty-guided Rollback Aggregation (URA), which mitigates debate groupthink by tracking per-agent uncertainty trajectories and rescuing high-confidence answers before they are overridden by group pressure. Across six multimodal reasoning benchmarks, wrapping existing debate baselines with DREAM consistently improves their accuracy-vs-computation trade-off without requiring dataset-specific tuning.

REPRODUCIBILITY STATEMENT

The approach is described in Section 3 and provided as pseudocode in Appendix F. Details of the implementation and hyperparameter choices are reported in Appendix B. All experiments use frozen models and publicly available checkpoints, with inference run using three different random seeds. The code and checkpoints will be made available.

ETHICS STATEMENT

Our work focuses on improving the reasoning accuracy of multimodal AI systems through multi-agent debate. We do not introduce new datasets or collect human data. The benchmarks used in our evaluation are publicly available and widely used in the research community. We acknowledge that improvements in multimodal reasoning capabilities could have both beneficial applications (e.g., medical image analysis, educational tools) and potential risks (e.g., more convincing misinformation generation). We encourage responsible deployment of such systems with appropriate safeguards.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing. We have not used generative AI tools to formulate the core research ideas, conduct the empirical experiments, or generate the evaluation data. Specifically, all AI-refined text was manually verified by the authors to ensure it perfectly aligns with our scientific claims and empirical findings, and all AI-generated formatting code was tested for compilation correctness. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. *arXiv e-prints*, pp. arXiv–2502, 2025.

Ding Chen, Qingchen Yu, Pengyuan Wang, Wentao Zhang, Bo Tang, Feiyu Xiong, Xinchi Li, Minchuan Yang, and Zhiyu Li. xverify: Efficient answer verifier for reasoning model evaluations. *arXiv preprint arXiv:2504.10481*, 2025.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=YfZ4ZPt8zd.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024.

Sitong Fang, Shiyi Hou, Kaile Wang, Boyuan Chen, Donghai Hong, Jiayi Zhou, Juntao Dai, Yaodong Yang, and Jiaming Ji. DEBATE WITH IMAGES: DETECTING DECEPTIVE BEHAVIORS IN MULTIMODAL LARGE LANGUAGE MODELS, 2025. URL https://mm-deception.github.io.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models, 2024.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

James Y Huang, Sheng Zhang, Qianchu Liu, Guanghui Qin, Tinghui Zhu, Tristan Naumann, Muhao Chen, and Hoifung Poon. Be My Eyes: Extending Large Language Models to New Modalities Through Multi-Agent Collaboration, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023.

Kumud Lakara, Georgia Channing, Christian Rupprecht, Juil Sock, Philip Torr, John Collomosse, and Christian Schroeder de Witt. Mad-sherlock: Multi-agent debate for visual misinformation detection. *arXiv preprint arXiv:2410.20140*, 2024.

LangChain. Langchain. `https://www.langchain.com/langchain`, 2025. Accessed: 2025-05-09.

Ao Li, Jinghui Zhang, et al. M3mad-bench: Are multi-agent debates really effective across domains and modalities?, 2026.

Chenxu Li, Zhicai Wang, Yuan Sheng, Xingyu Zhu, Yanbin Hao, and Xiang Wang. Res-bench: Benchmarking the robustness of multimodal large language models to dynamic resolution input, 2025.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL `https://aclanthology.org/2024.emnlp-main.992/`.

Zijie Lin and Bryan Hooi. Enhancing multi-agent debate system performance via confidence expression. *arXiv preprint arXiv:2509.14034*, 2025.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=t6QHYUOQL7`.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.

Tien Anh Nguyen, Khanh-Binh Nguyen, Van Dai Do, Svetha Venkatesh, and Hung Le. Creative generation via multi-agent debate: Does debate suppress diversity? *arXiv preprint arXiv:2609.00683*, 2026.

Junbo Niu, Yuanhong Zheng, Ziyang Miao, Hejun Dong, Chunjiang Ge, Hao Liang, Ma Lu, Bohan Zeng, Qiahao Zheng, Conghui He, and Wentao Zhang. Native visual understanding: Resolving resolution dilemmas in vision-language models, 2025.

OpenAI. Gpt-4v(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023. Accessed: 2024-06-01.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. In *International Conference on Machine Learning*, pp. 45883–45905. PMLR, 2024.

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

Hongduan Tian, Xiao Feng, Ziyuan Zhao, Xiangyu Zhu, Rolan Yan, and Bo Han. Multi-agent debate with memory masking. *arXiv preprint arXiv:2603.20215*, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pp. 6106–6131. Association for Computational Linguistics (ACL), 2024c.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2024d.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7572–7590, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, et al. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms. *arXiv preprint arXiv:2505.21327*, 2025.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3bq3jsvcQ1.

A Dataset Details

Full descriptions of the six benchmarks used in evaluation (multimodal evaluation subset of M3MAD-Bench (Li et al., 2026)):

- **MME** (Fu et al., 2025): 14 subtasks spanning perception (OCR, color recognition) and cognition (commonsense reasoning).
- **MathVista** (Lu et al., 2024): 6,141 visual mathematical reasoning tasks from 31 source datasets.
- **MathVision** (Wang et al., 2024a): 3,040 visual math problems over geometric figures, plots, and diagrams.
- **PathVQA** (He et al., 2020): pathology-image VQA requiring microscopic feature identification.
- **MME-Reas.** (Yuan et al., 2025): abductive, deductive, and analogical reasoning, extending MME.
- **VisualPuzzles** (Song et al., 2025): spatial and abstract visual puzzle reasoning.

B Implementation Details.

To ensure cost efficiency, we randomly sampled 500 examples from each dataset for evaluation, following the approach outlined in M3MAD (Li et al., 2026). We set the temperature to 0.5 and the maximum token length to 4096. To achieve more accurate answer assessments, we employed a model specifically designed for evaluation, called x-Verify (Chen et al., 2025). This model utilizes an LLM-as-a-judge paradigm to provide scalable and consistent evaluations across various tasks and modalities.

For the methods evaluated, we maintained consistent hyperparameter settings specific to each method (e.g., the number of agents and debate rounds) in accordance with their official implementations. Specifically, the Self-Consistency (SC) method uses $N = 5$ independent reasoning paths with majority voting. LLM Debate employs $N = 3$ agents over $N = 3$ debate rounds. The Div-MAD approach involves two role-based agents ("Affirmative" and "Negative") with $N = 3$ rounds. Similarly, DMAD also utilizes $N = 3$ agents across $N = 3$ rounds. In alignment with the original paper, DMAD assigns different reasoning strategies to its agents to promote diversity.

For unimodal tasks, we employed the following strategies: Chain-of-Thought (CoT) (LangChain, 2025), Step-Back Prompting (SBP) (Zheng et al., 2024), and Program-of-Thought (PoT) (Chen et al., 2023). For multimodal tasks, we utilized IO inference, Compositional Chain-of-Thought (CCoT) (Mitra et al., 2024), and Duty-Distinct Chain-of-Thought (DDCoT) Zheng et al. (2023). The experiments were conducted on eight NVIDIA A100 GPUs with 40GB of memory, while access to GPT-4o-mini was facilitated through OpenAI API calls.

C Assignment Strategy Definitions

The Dynamic Resolution Assignment (DRA) step selects an image payload for each agent to use across all debate rounds. We define the five named strategies below, reusing the main text's notation: $\mathcal{S} = \{s_1, \ldots, s_K\}$ is the full resolution candidate set, $\mathcal{C}$ the confident agent set after thresholding, $\mathcal{S}^*_{\mathcal{C}} = \{s_i^{(0)} : i \in \mathcal{C}\}$ the resolutions at which confident agents were probed, and $\mathcal{S}^\dagger \subseteq \mathcal{S}^*_{\mathcal{C}}$ the majority-agreeing subset (used only in Case 5). A *multi-scale payload* of a set $\mathcal{X}$ means the image is encoded at every resolution in $\mathcal{X}$ and all resulting visual tokens are concatenated into a single input.

MixRes. All N agents receive the full multi-scale payload $\mathcal{S}$. Applied when $|\mathcal{C}| = 0$ (Case 1): no agent was confident at any resolution, so all scales are provided as a fallback rather than discarding any signal.

BESTANLL. All N agents are locked to the single resolution in $\mathcal{S}_{\mathcal{C}}^*$ (or, in the Case 5 majority-filter branch below, in $\mathcal{S}^\dagger$). Applied in two situations: (i) exactly one confident agent (Case 2), where $\mathcal{S}_{\mathcal{C}}^*$ is already a singleton; and (ii) the confident-subset-agrees branch of Case 4 or the majority-filter branch of Case 5, where $|\mathcal{S}_{\mathcal{C}}^*| = 1$ or $|\mathcal{S}^\dagger| = 1$ collapses the mixture to a single scale.

STATICALL. Each agent i retains its own probe-round resolution $s_i^{(0)}$; no agent changes scale. Applied when all N agents were confident and in agreement (Case 3), preserving the full resolution diversity that the probe was designed to explore.

SUBSETSTATICMIXED. Confident agents ($i \in \mathcal{C}$) keep their own probe resolutions $s_i^{(0)}$. Unconfident agents ($i \notin \mathcal{C}$) receive the multi-scale mixture of $\mathcal{S}_{\mathcal{C}}^*$ (the set of resolutions endorsed by the confident, agreeing subset). Applied when a strict confident subset agrees and spans more than one resolution (Case 4, $|\mathcal{S}_{\mathcal{C}}^*| > 1$). When $|\mathcal{S}_{\mathcal{C}}^*| = 1$, SUBSETSTATICMIXED reduces to BESTANLL.

CONFLICTMIXRES. All N agents receive a multi-scale mixture of a resolution subset. Applied in Case 5 (confident agents disagree):

- *Strict majority exists* (at least two confident agents share the same answer): dissenting agents are excluded; all agents receive the mixture of $\mathcal{S}^\dagger$ (the majority-agreeing resolutions only). When $|\mathcal{S}^\dagger| = 1$ this again reduces to BESTANLL.
- *No strict majority*: all confident resolutions $\mathcal{S}_{\mathcal{C}}^*$ are used without filtering, since no subset can be identified as more reliable.

D TOKEN EFFICIENCY DETAIL

As detailed in Section 6.2, DREAM fundamentally improves the accuracy-efficiency trade-off for multimodal multi-agent systems. Rather than optimizing for a naive, uniform reduction in token usage, which typically starves models of critical visual context, DREAM strategically reallocates computational bandwidth to where it is needed most. Table 4 presents the comprehensive token economics for Qwen2.5-VL-7B (with accuracy figures corresponding to the averages in Table 1). This breakdown definitively shows that by accepting a marginal token overhead during the initial probing phase, DREAM successfully prevents the massive, compounding token waste caused by hallucination loops in later rounds. Ultimately, DREAM establishes a commanding accuracy-efficiency trade-off, achieving state-of-the-art accuracy through highly intelligent, cost-efficient token assignment.

Table 4: Prompt token consumption and accuracy for Qwen2.5-VL-7B.

Method	Prompt Tokens (M)	Accuracy (%)
Div-MAD	2.57	36.8
w/ DREAM	4.07	38.1
LLM Debate	9.38	47.0
w/ DREAM	12.18	50.2
DMAD	27.96	47.3
w/ DREAM	30.88	49.5

The LLM Debate row's bare-token figure uses the 3-round debate configuration, matching DREAM's own $R = 3$ debate budget. Under *w/ DREAM*, all three families improve in accuracy over their own bare row (↑1.3 to ↑3.2), matching Table 1's Δ column, but token consumption also increases for all three, consistent with the added cost of the probing round and, in some cases, the routed resolution set; see Section 6.2 for the resulting accuracy-vs-tokens trade-off.

Table 5: Performance on M3MAD-Bench multimodal benchmarks for LLaVA-Next-7B. Best result per column among rows with reported values **bolded**, second-best underlined. Δ denotes *w/ DREAM*'s average accuracy change over its bare debate baseline. *MVt*: MathVista, *MVs*: MathVision, *PVQA*: PathVQA, *MME-R*: MME-Reasoning, *VP*: VisualPuzzles.

Model	Method	Know.	Mathematics		Med.	Reasoning		Avg.	Δ
		MME	MVt	MVs	PVQA	MME-R	VP		
LLaVA-Next-7B	IO	40.2	24.0	8.0	11.2	12.8	18.0	19.0	
	CoT [33]	30.2	24.0	7.2	4.6	14.8	16.0	16.1	
	SC [32]	29.2	24.8	9.8	3.4	13.4	20.0	16.8	
	DwI [7]	58.9	27.2	9.3	29.5	16.0	20.8	26.9	
	LLM Debate [6]	72.0	**34.5**	9.9	23.7	17.3	22.4	30.0	
	w/ Perceiver [11]	68.0	29.5	9.8	20.9	16.3	18.9	27.2	
	w/ DREAM	**72.1**	34.7	10.7	25.1	17.3	**25.1**	**30.1**	**↑0.1**
	Div-MAD [17]	51.5	26.3	9.4	22.6	16.7	19.5	24.3	
	w/ Perceiver [11]	32.7	15.4	6.5	5.6	7.3	8.1	12.6	
	w/ DREAM	50.1	27.0	9.3	23.1	16.6	18.6	23.6	**↓0.7**
	DMAD [19]	59.3	29.5	11.4	27.1	16.3	21.9	27.6	
	w/ Perceiver [11]	57.4	26.6	8.3	21.4	16.0	16.7	24.4	
	w/ DREAM	62.3	30.7	**11.7**	**28.8**	**17.8**	23.1	29.1	**↑1.5**

E Extended Results: Additional Model

Table 5 extends our evaluation to LLaVA-Next-7B. Consistent with the near-uniform gains observed on the main-table models, *w/ DREAM* improves two of its three debate families: LLM Debate rises by ↑0.1% and DMAD by ↑1.5%. Div-MAD is the one exception, regressing by ↓0.7%.

F Algorithm Pseudocode

Algorithm 1 details the full DREAM procedure, including the DRA and URA phases described in the main methodology.

G Extended Ablations

This appendix collects two secondary ablations referenced from the main text: the resolution-scale ablation and the uncertainty-threshold ablation.

Table 6: Resolution scale ablation (Accuracy %). †4× Scale is not stable.

Resolution Strategy	InternVL3-14B	InternVL3-8B	Qwen2.5-VL-7B	Average
0.5× Scale	49.84	49.93	**49.45**	49.74
2× Scale	49.78	49.48	49.31	49.52
4× Scale	**50.24**	**50.35**†	49.23	**49.94**†

Resolution-scale ablation. Table 6 compares fixed-resolution policies applied to the same DMAD + URA backbone: committing every agent to a single resolution for the whole debate, rather than dynamically assigning resolution per agent. We exclude the $4\times$-Scale row from the headline comparison as it frequently encounters maximum token overflow, resulting in truncated context and failed generations, meaning its nominal average is not a reliable comparison point. Excluding it, $0.5\times$ is the stronger fixed policy on all three models. This ablation isolates the cost of committing to one resolution in advance; it does not include DREAM's full dynamic-assignment configuration, which is evaluated against the bare debate baselines directly in Table 1.

Algorithm 1 DREAM: Per-agent Resolution-Informed Scalable Multi-agent debate

Input: Image I, question q, resolution set $\mathcal{S}$, agents $\{1, \ldots, N\}$, debate rounds R, rescue threshold α

Output: Final answer a^*

// Phase 1: Dynamic Resolution Assignment (DRA)
for each agent $i \in \{1, \ldots, N\}$ **do**
  Assign resolution $s_i^{(0)} \in \mathcal{S}$ (round-robin)
  $\hat{a}_i^{(0)}, \mathrm{ANLL}_i^{(0)} \leftarrow \mathcal{M}(I_{s_i^{(0)}}, q)$
end for
$\tau_{\text{eff}} \leftarrow \mu_0 - \sigma_0$ *// Adaptive threshold, Eq. equation 2*
$\mathcal{C} \leftarrow \{i : \mathrm{ANLL}_i^{(0)} \leq \tau_{\text{eff}}\}$ *// Confident set*
$\mathcal{S}^* \leftarrow \mathrm{AssignResolutions}(\mathcal{C})$ *// DRA branch, Appendix C*
// Phase 2: Multi-Agent Debate
for round $r = 1$ to R **do**
  for each agent $i \in \{1, \ldots, N\}$ **do**
    $s_i^{(r)} \leftarrow \mathcal{S}^*[(i-1) \bmod |\mathcal{S}^*|]$ *// round-robin over routed set*
    $a_i^{(r)}, \mathrm{ANLL}_i^{(r)} \leftarrow \mathcal{M}(I_{s_i^{(r)}}, q, \text{debate_context}^{(r-1)})$
  end for
end for
// Phase 3: Uncertainty-Guided Rollback Aggregation (URA)
for each agent i **do**
  $\sigma_i^{\text{debate}} \leftarrow \mathrm{std}(\{\mathrm{ANLL}_i^{(r)}\}_{r=1}^{R})$ *// Eq. equation 3*
  $\delta_i \leftarrow \alpha \cdot \sigma_i^{\text{debate}}$
  $r_i^* \leftarrow \arg\min_{r<R} \mathrm{ANLL}_i^{(r)}$ *// Most confident round before the final*
  if $\mathrm{ANLL}_i^{(R)} - \mathrm{ANLL}_i^{(r_i^*)} > \delta_i$ **then**
    $\hat{a}_i \leftarrow a_i^{(r_i^*)}$ *// Rescue answer, Eq. equation 4*
  else
    $\hat{a}_i \leftarrow a_i^{(R)}$ *// Keep final answer*
  end if
end for
$a^* \leftarrow \mathrm{JudgeWithAntiRecency}(\{\hat{a}_i\}, \text{transcript})$
return a^*

Table 7: Ablation on uncertainty thresholds across models (Accuracy %). Δ is each row's average minus the best static-threshold average in hindsight (49.74, achieved at 0.20).

Threshold	InternVL3-14B	InternVL3-8B	Qwen2.5-VL-7B	Avg.	Δ
Static 0.10	50.25	49.66	49.29	49.73	-0.01
Static 0.20	50.12	50.16	48.93	49.74	0.00
Static 0.30	50.51	49.57	48.90	49.66	↓**0.08**
Static 0.50	50.36	49.60	49.05	49.67	↓**0.07**
Adaptive ($\mu-\sigma$)	**50.70**	**50.40**	**49.50**	**50.20**	↑**0.46**

Uncertainty-threshold ablation. Table 7 compares static thresholds against our adaptive mechanism. Notably, the adaptive rule does not merely act as a convenient fallback; it strictly outperforms fixed thresholds. No single static threshold is optimal across all three models (e.g., 0.10 wins on the 7B, 0.20 wins on the 8B, and 0.30 wins on the 14B). However, the adaptive rule dynamically adjusts to the per-sample uncertainty distribution, yielding higher accuracy on *every individual model* than even a "tuned oracle" that picks the best static threshold in hindsight. This demonstrates that our data-driven thresholding is a strict performance upgrade over static hyperparameter tuning.

G.1 Hyperparameters Study

Table 8: URA rescue-multiplier (α) sensitivity sweep (Accuracy %, Qwen2.5-VL-7B).

Rescue Multiplier (α)	**Qwen2.5-VL-7B**
0.10	42.2
0.25	43.3
0.50	**50.2**
0.75	47.2
1.00	45.3

Rescue-multiplier (α) sweep. Table 8 sweeps the rescue multiplier α, which sets the per-agent rollback threshold $\delta_i = \alpha \cdot \sigma_i^{\text{debate}}$ in Eq. (5). Accuracy rises sharply from $\alpha = 0.10$ to $\alpha = 0.50$ and falls off past that point: low α narrows the threshold so much that ordinary round-to-round fluctuation in token probabilities triggers a rollback, discarding valid late-round consensus, while $\alpha \geq 0.75$ widens the threshold enough that genuine certainty erosion no longer crosses it and goes uncorrected. $\alpha = 0.5$ is the peak on this model, consistent with its use as the fixed value across all reported results.

G.2 Branch-Level Analysis of Dynamic Resolution Assignment

To characterize how the five DRA branches defined in Appendix C (MixRes, BestANLL, StaticAll, SubsetStaticMixed, ConflictMixRes) actually behave in practice, we break down their firing frequency, accuracy, and token cost on Qwen2.5-VL-7B. The pooled DRA accuracy reported below (50.2%) reproduces the $\alpha = 0.5$ row of Table 8, confirming both figures draw from the same run.

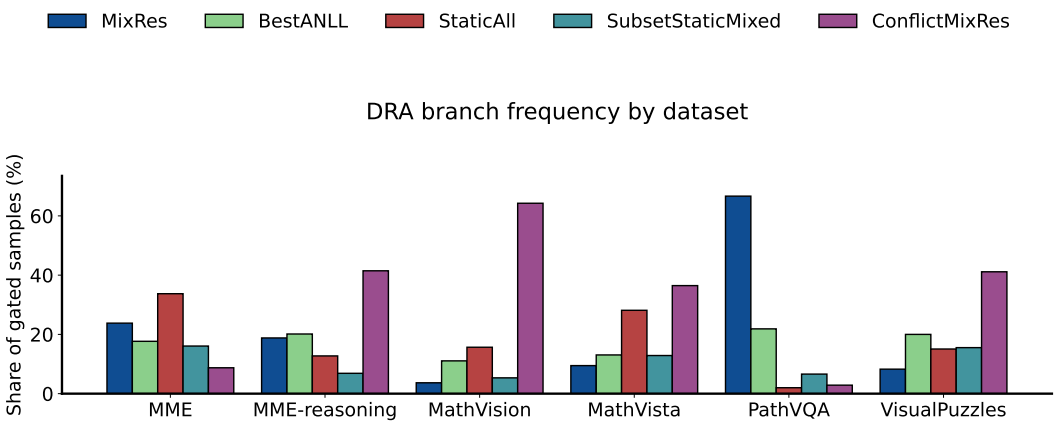


Figure 4: Share of gated samples routed to each DRA branch, by benchmark. ConflictMixRes dominates on MathVision (64%), MME-Reasoning (41%), and VisualPuzzles (41%), while MixRes dominates on PathVQA (67%); StaticAll is comparatively more common on MME (34%) and MathVista (28%). No single branch dominates uniformly across benchmarks, consistent with the probe resolving a genuinely per-sample, per-dataset routing decision rather than converging to one default behavior.

Together, these four views indicate that the DRA branches are not interchangeable: they fire under different diagnostic conditions (Appendix C), at different rates per benchmark, and trade off accuracy against token cost differently. The pattern that incorrect answers cost more tokens than correct ones within most branches, including the no-DRA baseline, suggests token overrun is itself a symptom of the model struggling on a sample rather than a cause distinct to any one branch.

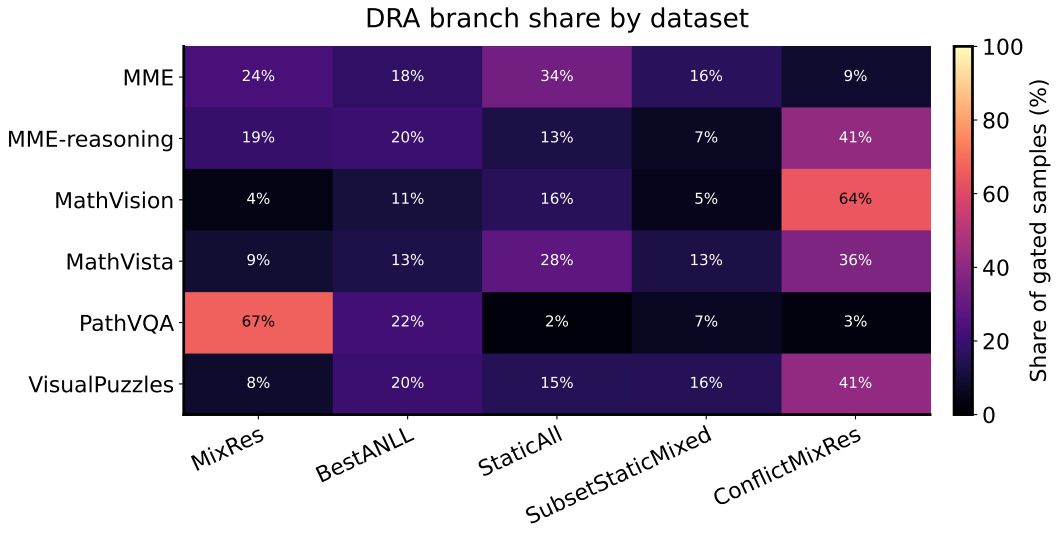


Figure 5: Same branch-share data as Figure 4, shown as a benchmark-by-branch heatmap for direct cross-dataset comparison.

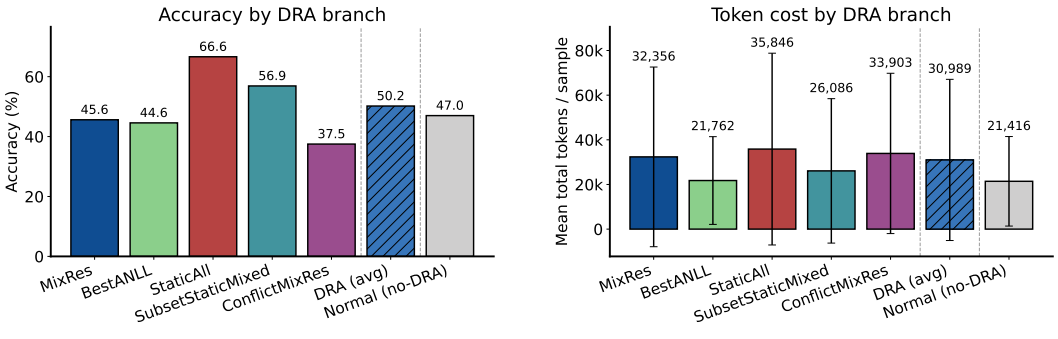


Figure 6: Accuracy (left) and mean token cost (right) per DRA branch, pooled across all six benchmarks, against the DRA-average and the no-DRA baseline. STATICALL is the most accurate individual branch (66.6%) but also among the most expensive (35,846 tokens/sample); CONFLICTMIXRES is the least accurate (37.5%) despite a high token cost (33,903), reflecting that it fires precisely when confident agents disagree and no majority resolution can be trusted. The DRA-average (50.2%, 30,989 tokens) exceeds the no-DRA baseline (47.0%, 21,416 tokens) in accuracy at roughly $1.45\times$ the token cost.

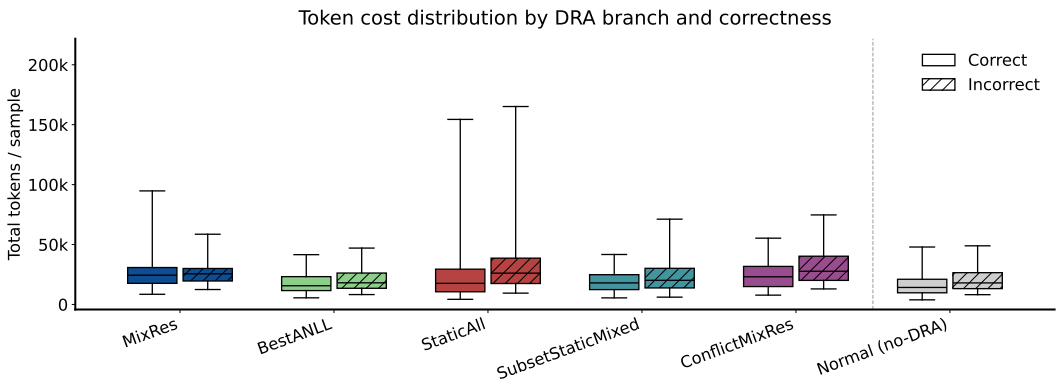


Figure 7: Distribution of total tokens per sample for each DRA branch, split by whether the final answer was correct. Within most branches, incorrect samples show a higher median and a longer upper tail than correct samples, most visibly for STATICALL, whose incorrect-sample whisker extends past 165k tokens; the no-DRA baseline shows the same correct/incorrect gap at a lower absolute cost.

H CASE STUDIES

H.1 QUALITATIVE ANALYSIS

Mitigating Visual Hallucinations. In Figure 8, we observe a catastrophic failure mode of standard multi-agent debate on the MME benchmark. Bound to a static resolution, the baseline agents lose grounding and enter a compounding hallucination loop, generating verbose descriptions detailing non-existent "cosmos flowers" and a "bumblebee" rather than simply identifying the Ringkirche. Conversely, DREAM's Dynamic Resolution Assignment (DRA) probes the image and identifies $2.0\times$ as the optimal, lowest-uncertainty scale (ANLL: 0.106). By aggressively assigning the debate to this single optimal resolution (CONSENSUSSCALING), DREAM completely suppresses the hallucination loop and returns the correct answer with high precision.

Dismantling Sycophancy and Groupthink. Figure 9 highlights DREAM's effectiveness in dismantling sycophancy-a notorious vulnerability in standard multi-agent debate configurations. On a dense spatial reasoning task from MathVision, an early baseline agent hallucinates a flawed logical trajectory, which subsequent agents blindly validate without independent visual verification. DREAM intercepts this failure mode during its Round 0 probing phase. Upon detecting conflicting answers across the independent resolution probes (D, B, and C), the DRA module flags the query's ambiguity and triggers a CONFLICTMIXRES payload. By forcing the agents to reconcile distinct, multi-scale perspectives ($0.5\times$, $1.0\times$, and $2.0\times$), DREAM effectively breaks the groupthink, enabling the agents to systematically verify the spatial layout and converge on the correct answer.

Robust Multi-Scale Clinical Synthesis. The advantages of mixed-resolution reasoning extend beyond breaking groupthink, proving highly beneficial for dense domain-specific tasks. As shown in Figure 10 (PathVQA), processing a severe burn image at a single fixed resolution limits the baseline's ability to synthesize macro-level tissue damage with micro-level textural cues, resulting in a generic and suboptimal diagnosis. DREAM's DRA probing phase processes the image across multiple scales simultaneously, allowing the $0.5\times$ probe to successfully identify the specific etiology ("thermal injury"). Recognizing the complexity of the medical query via its uncertainty scores, DREAM routes the debate to a MIXRES strategy. This allows the debate to seamlessly integrate distinct clinical cues extracted across all three scales, culminating in a highly precise and comprehensive medical evaluation.

Case Study: Resolving Visual Hallucinations (MME) - *Query:* Is this a photo of Ringkirche? (GT: **Yes**)



Standard MAD Baseline

Fixed Resolution → Hallucination & Herding

Debate (Fixed 1.0x)

"Okay, here's a detailed description of the image: The image is a close-up shot of a vibrant garden scene, focusing on a cluster of pink cosmos flowers and a busy bumblebee..."
✗ Severe Visual Hallucination

Final Consensus: ✗ Incorrect (Failed to answer)

DREAM (Ours)

Dynamic Resolution → Accurate Consensus

Round 0: DRA Probing (`consensus_scaling`)

Probe 0.5x → ANLL: 0.236 — Probe 1.0x → ANLL: 0.258
Probe 2.0x → **ANLL: 0.106 (Lowest Uncertainty)**
*Decision: Route debate to optimal **2.0x** resolution.*

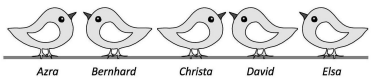
Debate at 2.0x

Agents: "Yes, based on the architectural style and design of the building in the image, it appears to be Ringkirche..."
✓ Accurate Visual Verification

Final Consensus: ✓ Correct (Yes)

Figure 8: **Mitigating Visual Hallucinations.** Qualitative comparison on the MME dataset. While the standard baseline enters a compounding hallucination loop (detailing non-existent flowers and bumblebees), DREAM's DRA module detects high confidence at the $2.0\times$ resolution and routes the debate optimally, yielding the correct answer with high precision.

Case Study: Breaking Sycophancy (MathVision) - *Query:* Which sparrow has turned around? (GT: **B: Bernhard**)



Standard MAD Baseline

Fixed Resolution → Sycophancy & Incorrect Logic

Debate (Fixed 1.0x)

Agent 1: "...This happens when the sparrow was originally at the end of the line (Elsa)."
Agent 2: "The argument suggests... Based on this logic, the sparrow who turned around should be the one at the end, which is Elsa."
✗ Blind Sycophancy (Groupthink)

Final Consensus: ✗ Incorrect (Elsa)

DREAM (Ours)

Dissent-Filtered assignment → Accurate Logic

Round 0: DRA Probing (`conflict_mix_res`)

Probe 0.5x → **D** — Probe 1.0x → **B** — Probe 2.0x → **C**
Decision: Disagreement Detected. Triggering mixed-resolution payload $(2.0\times, 1.0\times, 0.5\times)$

Debate (Mixed Res Payload)

Agents: "Let's analyze the situation step by step... Based on the detailed analysis provided, the sparrow that turned around to make the total whistles increase correctly points to Bernhard."
✓ Multi-Scale Verification Succeeds

Final Consensus: ✓ Correct ((B) Bernhard)

Figure 9: **Dismantling Groupthink.** Qualitative comparison on MathVision. The standard baseline succumbs to agreement bias, where subsequent agents blindly agree with an initial flawed assumption. DREAM detects early disagreement during probing and deploys a mixed-resolution payload, forcing the agents to successfully ground their spatial reasoning rather than relying on unverified assumptions.

Case Study: Multi-Scale Clinical Synthesis (PathVQA) - *Query:* what burn? (GT: **thermal**)



Standard MAD Baseline

Fixed Resolution → Suboptimal Synthesis

Debate (Fixed 1.0x)

"Based on the descriptions provided, the burn in the image appears to be a severe one... The skin is blackened and charred... indicate that this is a full-thickness burn, most likely a third-degree burn."

✗ Static Context Processing

Final Consensus: ✗ Generic Diagnosis

DREAM (Ours)

Multi-Scale Probing → Robust Clinical Synthesis

Round 0: DRA Probing (`mix_res`)

Probe 0.5x → ANLL: 0.529 (Identifies "**thermal injury**")
Probe 1.0x → ANLL: 0.547 — Probe 1.5x → ANLL: 0.599
Decision: Complex clinical query. Trigger mixed-resolution assignment.

Debate (Mixed Res Payload)

Agents: "The skin is discolored, primarily in shades of black and dark brown, which are indicative of severe **thermal injury**... highly likely that these are third-degree burns."

✓ Comprehensive Clinical Context

Final Consensus: ✓ Precise Etiology Identified

Figure 10: **Robust Clinical Synthesis.** Qualitative comparison on PathVQA. While the fixed-resolution baseline struggles to extract specific etiology (providing a generic "third-degree" diagnosis), DREAM's multi-scale probing seamlessly extracts complementary clinical cues across distinct visual resolutions. The resulting mixed-resolution debate yields a far more precise and comprehensive medical evaluation.