From Discord to Accord: A Consensus-Driven Multi-Agent Framework for Multi-Modal Question Answering

Dongyang Li
dongyangli.ldy@shiep.edu.cn
Shanghai University of Electric Power
Shanghai, China

Luobin Huang
luobinhuang043@gmail.com
East China Normal University
Shanghai, China

Taolin Zhang
tlzhang@hfut.edu.cn
Hefei University of Technology
Hefei, China

Xiaofeng He
hexf@cs.ecnu.edu.cn
East China Normal University
Shanghai, China

Chengyu Wang✉
chengyu.wcy@alibaba-inc.com
Alibaba Group
Hangzhou, China

Abstract

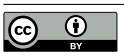
Multi-modal large language models (MLLMs) integrate heterogeneous inputs to solve tasks that require joint reasoning across modalities, such as multi-modal question answering (MQA). Beyond conventional prompt-based inference, multi-agent frameworks can elicit complementary evidence and aggregate diverse reasoning traces to better address challenging queries. Existing multi-agent methods are largely inspired by individual-level cognitive mechanisms (e.g., fast–slow thinking and metacognition), but often underexploit *social* cognitive factors that govern how groups reconcile disagreements. Moreover, while debate-style strategies can improve logical consistency, their final summarization is frequently heuristic, making it prone to overlooking factual details. We propose *CRATER*, a consensus-driven multi-agent framework for MQA that aligns discordant perspectives through structured interactions among heterogeneous MLLM agents. CRATER proceeds in three stages. First, in **Collaborative Multi-Agent Initialization**, a discussion group generates candidate opinions from diverse perspectives and modalities to cover a broad space of viewpoints. Second, **Multi-Dimensional Scaling (MDS) Projection** embeds these opinions into a low-dimensional geometry, facilitating the identification of polarized stances. Third, polarized opinions are reconciled via an iterative **Monotonic Concession Re-discussion** procedure, which progressively narrows disagreements and establishes consensus. Experiments on public MQA benchmarks spanning general capability evaluation, text recognition, and multidisciplinary domains show that CRATER consistently outperforms strong baselines. Additional analyses validate the contributions of each core component.

CCS Concepts

• **Computing methodologies** → **Multi-agent planning**.

Keywords

Multi-Modal Large Language Models; Multi-Agent Collaboration; Consensus Formation; Question Answering



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil.
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835528

ACM Reference Format:
Dongyang Li, Luobin Huang, Taolin Zhang, Xiaofeng He, and Chengyu Wang. 2026. From Discord to Accord: A Consensus-Driven Multi-Agent Framework for Multi-Modal Question Answering. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835528

1 Introduction

Multi-modal question answering (MQA) aims to answer questions grounded in heterogeneous inputs (e.g., images and text), requiring reasoning across modalities [48–50]. Recent progress has been propelled by multi-modal large language models (MLLMs) [5, 20, 47], enabling advances in high-stakes and knowledge-intensive applications such as medical analysis [47, 52], autonomous driving [8, 32], and scientific MQA [15, 40].

Most existing MQA methods rely on a *single* MLLM and achieve strong results by leveraging its large capacity. Broadly, they fall into two families. (1) *Chain-reasoning* approaches [29, 46, 57] use Chain-of-Thought prompting to induce step-by-step inference. (2) *Visual-planning* approaches [6, 39, 58] employ an MLLM to plan and orchestrate actions (e.g., tool use or navigation) and execute a structured workflow. Despite their effectiveness, these methods mainly maximize the capability of a single model, leaving the robustness and complementary reasoning benefits of multi-agent collaboration under-explored.

Motivated by this gap, a natural direction is to integrate multiple MLLMs into a multi-agent framework. However, multi-modal multi-agent research remains less mature than its text-only counterpart [4, 25, 26, 54]. Existing work can be roughly grouped into: (1) *collaboration-based* methods [13, 22, 27, 43] that design agent topologies (e.g., pipelines or star-shaped hubs) for role assignment and task decomposition, and that sometimes draw inspiration from metacognitive mechanisms such as two-stage reasoning and discussion [30, 38]; and (2) *conflict-based* methods [2, 25, 26] that encourage disagreement (supporting vs. opposing views) to elicit critiques and reduce single-agent biases. Nevertheless, many existing systems still depend on a central moderator that consolidates agent outputs via prompting. Such consolidation is often coarse and may fail to faithfully integrate key evidence across viewpoints, resulting in factual omissions [4, 19, 26]. As illustrated in Figure 1, coarse voting followed by shallow consolidation produces the incorrect

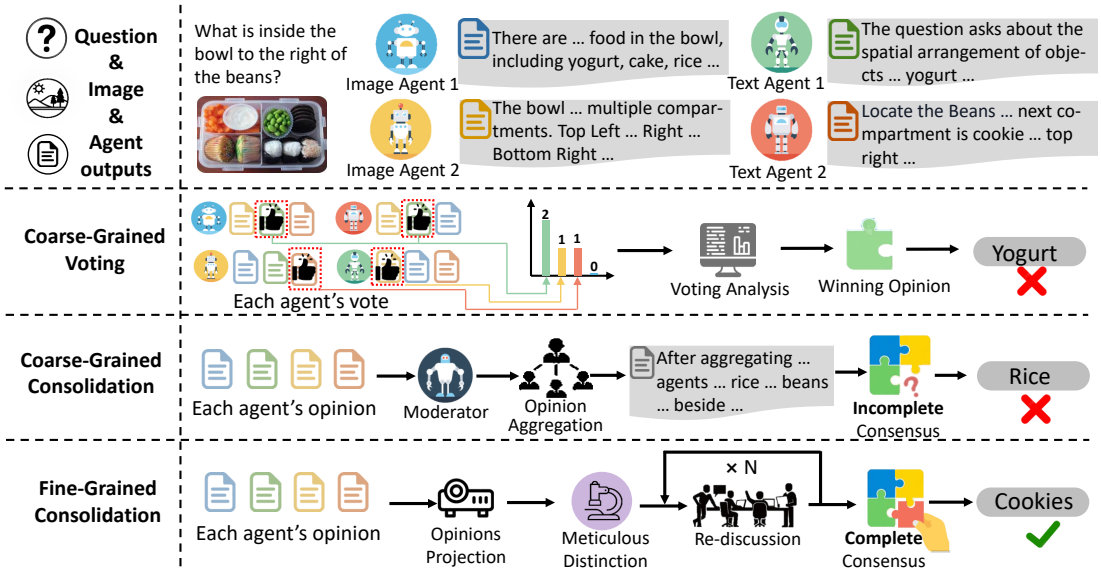


Figure 1: Comparison of consensus construction strategies. Coarse voting and coarse consolidation often yield incomplete consensus and can miss key evidence, leading to incorrect answers. In contrast, fine-grained consolidation via explicit distinction and multiple rounds of re-discussion yields more complete agreement and the correct answer.

answer (*"Yogurt"*). In contrast, a fine-grained process that (i) separates polarized opinions, and (ii) conducts targeted re-discussion can recover the correct answer (*"Cookies"*).

To address these limitations, we propose *CRATER*, a consensus-driven multi-agent framework for MQA that incorporates principles from social cognition and consensus formation to aggregate answers from heterogeneous MLLM agents.[1] CRATER first builds a collaborative discussion group that elicits diverse interpretations by assigning varied roles and cognitive perspectives, producing a richer pool of candidate opinions. Building on this pool, CRATER introduces a consensus-driven re-discussion strategy that gradually aligns divergent views. At each iteration, we project opinions into a low-dimensional space using multidimensional scaling (MDS) [10] and select the two most polarized participants via a max-margin criterion for targeted re-discussion. To encourage convergence while preserving informativeness, we adopt a monotonic concession protocol [11], an automated negotiation scheme that requires each participant to make a bounded concession relative to its previous proposal. This quantitative consensus construction yields a final consensus that improves answer quality for MQA.

2 Related Work

Our work connects three research threads: MQA, LLM-based multi-agent systems, and computational consensus construction. Prior MQA approaches predominantly optimize a single MLLM, while many multi-agent pipelines (including debate-style variants) still rely on a centralized moderator with coarse aggregation. In contrast, we study multi-agent MQA with iterative, fine-grained consensus formation among heterogeneous MLLM agents.

Multi-Modal Question Answering. Recent MQA progress has been driven by MLLM-based reasoning frameworks. One line improves answer generation through chain-based reasoning. Wu *et al.* [46] combine detection-oriented prompting with CoT for complex inference, while related studies [29, 57] use progressive multi-step prompting to improve correctness. Another line focuses on planning-oriented reasoning for multi-modal tasks [6, 39, 58]. AO-Planner [6], for example, applies foundation models to affordance-driven vision-and-language navigation, while other methods [39, 58] use MLLMs to coordinate tool use and workflow execution. However, these methods generally rely on a single model and cannot systematically leverage complementary perspectives. We instead study MQA in a multi-agent setting, consolidating diverse opinions into more reliable answers.

Multi-Agent Systems. LLM-based multi-agent systems improve reasoning through interaction and specialization. *Collaboration-based* methods typically use a planner or coordinator to decompose tasks, assign roles, and aggregate outputs [13, 14, 27, 43]. MetaGPT [14], for example, organizes agents into SOP workflows, while other systems [27, 43] use centralized coordination for delegation and consolidation. Cognitive-theory-inspired approaches instead employ multi-stage interactions to improve coordination [30, 38]. *Conflict-based* methods generate supporting and opposing views to expose errors and reduce single-agent bias [2, 25, 26]. However, centralized synthesis through summarization or voting may miss crucial fine-grained evidence. Our approach instead builds consensus through iterative re-discussion among polarized agents.

Consensus Construction. Consensus construction is central to multi-agent systems because it aligns participants and reduces unproductive conflict [34, 45]. Existing approaches mainly use *voting-based* or *summarization-based* aggregation. Voting methods [25, 28, 56] select answers by majority vote and can improve

[1] *CRATER* denotes a **C**onsensus-d**R**iven multi-**A**gent framework for mul**T**i-modal qu**E**stion answe**R**ing.

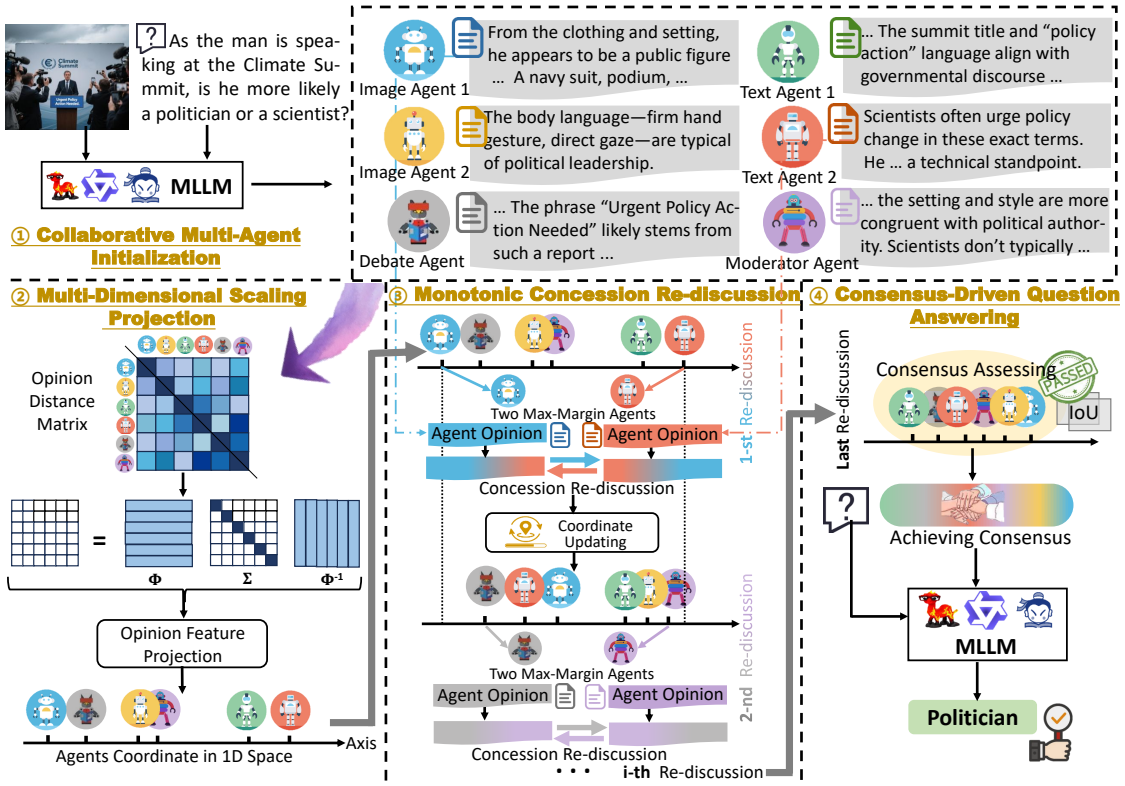


Figure 2: Overview of *CRATER*. We first project agents' opinions into a low-dimensional space via multi-dimensional scaling (MDS) and select a polarized pair using a max-margin criterion. The selected agents then engage in monotonic-concession re-discussion to progressively narrow disagreement and reach consensus. In the example, Image Agent 1 and Text Agent 2 (blue and red) are identified as the most polarized due to their divergent role-conditioned expressions (politician-leaning vs. scientist-leaning), and are thus chosen for re-discussion.

reasoning-intensive tasks [18, 53], but may overlook subtle viewpoint differences and discard correct minority evidence. Summarization methods [4, 26] use a judge or moderator to synthesize discussion traces. For example, ChatEval [4] introduces a "Simultaneous-Talk-with-Summarizer" mechanism that condenses information after each round. However, these methods remain coarse-grained and may only partially integrate diverse evidence. Our framework instead explicitly models disagreement and progressively strengthens consensus through fine-grained interaction.

3 Methodology

An overview of our approach is shown in Figure 2. Our framework comprises four key modules: (1) **Collaborative Multi-Agent Initialization**, which generates diverse outputs by assigning distinct role configurations to each agent; (2) **Multi-Dimensional Scaling (MDS) Projection**, which embeds high-dimensional opinion representations into a unified low-dimensional space to facilitate subsequent agent selection; (3) **Monotonic Concession Re-discussion**, in which the two agents exhibiting the largest opinion margin iteratively incorporate each other's content during re-discussion; and (4) **Consensus-Driven Question Answering**, which uses the resulting consensus to generate more accurate and coherent responses.

3.1 Collaborative Multi-Agent Initialization

To capture broad-spectrum perspectives and mitigate conformity bias [44], we initialize the multi-agent system with carefully designed role configurations [16]. For MQA, our framework deploys six specialized agents that share a common MLLM backbone:

Image Agents: Two agents are dedicated to extracting rich visual content. The first emphasizes objects, background, emotional atmosphere, and event occurrence; the second focuses on temporal aspects, location or setting, and spatial relationships.

Text Agents: Two agents mine semantic information from text. One analyzes the question type and defines the answer scope, while the other performs detailed textual reasoning with step-by-step logic based on the question.

Debate Agent: To enhance logical completeness and error detection, this agent generates opposing viewpoints relative to the opinions of the other agents.

Moderator Agent: Acting as a synthesizer, this agent summarizes the perspectives of all other agents into a coherent answer.

Formally, the initial output of the k-th agent is defined as

$$\mathbf{A}_k^0 = \mathrm{MLLM}(\mathbf{R}_k, \mathbf{Q}, \mathbf{I}), \tag{1}$$

where $\mathbf{A}_k^0$ denotes the initial output of the k-th agent, $\mathbf{R}_k$ is the predefined role of the k-th agent, $\mathbf{Q}$ is the input question, and $\mathbf{I}$ denotes the corresponding visual input.

3.2 Multi-Dimensional Scaling Projection

Let $\mathcal{V} = \{a_1, \ldots, a_n\}$ denote the agent set, where $n = 6$ in our framework. Given the initial agent outputs $\{\mathbf{A}_k^0\}_{k=1}^n$, we map each textual opinion to an embedding vector $\mathbf{e}_k^0 \in \mathbb{R}^d$ using a text encoder (e.g., BGE-M3 [7] or Qwen3-VL-Embedding [24]). We then define the initial pairwise opinion-distance matrix $D^0 \in \mathbb{R}^{n\times n}$ by

$$D_{uv}^0 = \|\mathbf{e}_u^0 - \mathbf{e}_v^0\|_2, \qquad 1 \le u, v \le n. \tag{2}$$

To obtain a low-dimensional Euclidean representation that best preserves the pairwise opinion distances, we adopt classical MDS [10]. Let $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ be the centering matrix, where I is the identity matrix and $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector. The associated initial Gram matrix is

$$B^0 = -\frac{1}{2}H(D^0)^{\circ 2}H, \tag{3}$$

where $(D^0)^{\circ 2}$ denotes the element-wise square of D^0. By the classical MDS construction, B^0 is the double-centered inner-product matrix from which the initial embedding is recovered.

Let the eigendecomposition of B^0 be $B^0 = V^0\Lambda^0(V^0)^T$, where the eigenvalues satisfy $\lambda_1^0 \ge \lambda_2^0 \ge \cdots \ge \lambda_n^0$ and the columns of V^0 are the associated orthonormal eigenvectors. Since our goal is to construct a one-dimensional viewpoint axis, we retain the leading eigenpair $(\lambda_1^0, \mathbf{v}_1^0)$ and, assuming $\lambda_1^0 \ge 0$, define the initial coordinate vector

$$P^0 = \sqrt{\lambda_1^0}\mathbf{v}_1^0 \in \mathbb{R}^{n\times 1}. \tag{4}$$

We denote the coordinate of agent a_k at round t by p_k^t, and in particular the initial coordinate of a_k by p_k^0. This one-dimensional embedding can be interpreted as the dominant latent axis of inter-agent opinion variation.

At each round t, the most divergent pair is selected via the max-margin criterion:

$$(i^t, j^t) = \text{argmax}_{u,v\in\{1,\ldots,n\}}|p_u^t - p_v^t|. \tag{5}$$

In later rounds, we preserve this one-dimensional opinion space and update agent coordinates within it via the approximation strategy described in the next subsection, rather than re-running full MDS at every iteration.

3.3 Monotonic Concession Re-discussion

Let $P^t = (p_1^t, \ldots, p_n^t)^T \in \mathbb{R}^{n\times 1}$ denote the one-dimensional coordinate vector of all agents before round t, where p_k^t is the coordinate of agent a_k. We first select the most divergent pair (a_{i^t}, a_{j^t}) using Eq. (5). Without loss of generality, assume $p_{i^t}^t \le p_{j^t}^t$. We then define the current extreme-opinion interval by $S^t = [p_{i^t}^t, p_{j^t}^t]$ and $\Delta^t = \text{diam}(S^t) = p_{j^t}^t - p_{i^t}^t$. The monotonic concession mechanism updates only the two extreme agents, while the remaining agents are updated through a Nyström-style approximation in the same opinion space.

Specifically, let $\alpha_{i^t}^t, \alpha_{j^t}^t \in (0, 1)$ denote the concession rates of the left and right extreme agents, respectively. We define the post-discussion coordinates by the affine operators

$$\bar{p}_{i^t}^t = T_{i^t}^t(p_{i^t}^t, p_{j^t}^t) = (1 - \alpha_{i^t}^t)p_{i^t}^t + \alpha_{i^t}^t p_{j^t}^t, \tag{6}$$

and

$$\bar{p}_{j^t}^t = T_{j^t}^t(p_{i^t}^t, p_{j^t}^t) = \alpha_{j^t}^t p_{i^t}^t + (1 - \alpha_{j^t}^t)p_{j^t}^t. \tag{7}$$

These updates are convex combinations of the pre-discussion endpoints and therefore admit several structural properties. First, both updated coordinates remain in the convex hull of the original pair, i.e., $\bar{p}_{i^t}^t, \bar{p}_{j^t}^t \in [p_{i^t}^t, p_{j^t}^t]$. Second, the endpoint ordering is preserved whenever $\alpha_{i^t}^t + \alpha_{j^t}^t \le 1$, namely $\bar{p}_{i^t}^t \le \bar{p}_{j^t}^t$. Third, the distance between the two updated endpoints contracts according to

$$\bar{p}_{j^t}^t - \bar{p}_{i^t}^t = (1 - \alpha_{i^t}^t - \alpha_{j^t}^t)(p_{j^t}^t - p_{i^t}^t). \tag{8}$$

Hence, if $\alpha_{i^t}^t + \alpha_{j^t}^t \in (0, 1)$, the diameter of the updated selected pair contracts strictly:

$$\bar{\Delta}^t := \bar{p}_{j^t}^t - \bar{p}_{i^t}^t = \rho^t\Delta^t, \qquad \rho^t := 1 - \alpha_{i^t}^t - \alpha_{j^t}^t \in (0, 1). \tag{9}$$

Therefore, the pairwise concession update defines a strict contraction on the current extreme-opinion interval.

At the language level, the two selected agents re-generate their responses under prompts that require partial concession to the counterpart's viewpoint while preserving their own supporting evidence. Their updated positions in the opinion space are directly given by the affine updates in Eqs. (6) and (7), without recomputing the full MDS embedding. For example, after re-discussion, Image Agent 1's original inclination toward *politician* changes to: "Even visually, there is no party branding. A scientist might be the keynote speaker at such a summit, especially in opening remarks or to convey credibility." Similarly, Text Agent 2's original inclination toward *scientist* changes to: "Scientists do not typically face that many cameras alone unless they hold public office." Notably, both revised opinions contain speculative terms such as *might* and *typically*, rather than firm assertions.

To update the remaining agents without rerunning full MDS at every round, we employ a Nyström-style approximation [12]. Let $\mathcal{L}^t = \{i^t, j^t\}$ and $\mathcal{N}^t = \{1, \ldots, n\} \setminus \mathcal{L}^t$ denote the landmark and non-landmark index sets, respectively. We treat the two updated extreme agents as landmark points and define the Gaussian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. The resulting landmark kernel matrix is

$$W^t = \begin{bmatrix} k(\bar{p}_{i^t}^t, \bar{p}_{i^t}^t) & k(\bar{p}_{i^t}^t, \bar{p}_{j^t}^t) \\ k(\bar{p}_{j^t}^t, \bar{p}_{i^t}^t) & k(\bar{p}_{j^t}^t, \bar{p}_{j^t}^t) \end{bmatrix} \in \mathbb{R}^{2\times 2}. \tag{10}$$

Let $\mathcal{N}^t = \{m_1, \ldots, m_{n-2}\}$. We then compute the cross-kernel matrix between the non-landmark agents and the updated landmarks:

$$Z^t = \begin{bmatrix} k(p_{m_1}^t, \bar{p}_{i^t}^t) & k(p_{m_1}^t, \bar{p}_{j^t}^t) \\ k(p_{m_2}^t, \bar{p}_{i^t}^t) & k(p_{m_2}^t, \bar{p}_{j^t}^t) \\ \vdots & \vdots \\ k(p_{m_{n-2}}^t, \bar{p}_{i^t}^t) & k(p_{m_{n-2}}^t, \bar{p}_{j^t}^t) \end{bmatrix} \in \mathbb{R}^{(n-2)\times 2}. \tag{11}$$

Let the eigendecomposition of W^t be $W^t = U^t\Lambda^t(U^t)^T$, where $\Lambda^t = \text{diag}(\lambda_1^t, \lambda_2^t)$ with $\lambda_1^t \ge \lambda_2^t \ge 0$. Since the opinion space is one-dimensional, we retain only the leading eigenpair $(\lambda_1^t, \mathbf{u}_1^t)$ and compute the approximate one-dimensional coordinates of the non-landmark agents as

$$\Gamma^t = \frac{1}{\sqrt{\lambda_1^t}}Z^t\mathbf{u}_1^t \in \mathbb{R}^{(n-2)\times 1}. \tag{12}$$

This step provides an efficient out-of-sample approximation of the updated non-landmark coordinates conditioned on the revised landmark pair. We then assemble the full coordinate vector after

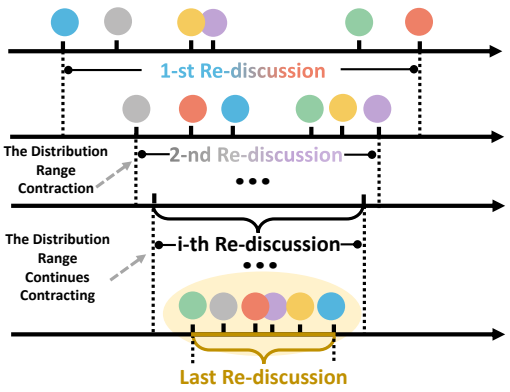


Figure 3: Dynamic process of social opinion formation. The gradually narrowing distribution of viewpoints indicates convergence toward consensus.

round t by combining $\bar{p}_{i^t}^t$, $\bar{p}_{j^t}^t$, and the entries of Γ^t according to the original agent indices. The same max-margin rule in Eq. (5) is then applied to select the pair for the next round.

We next formalize the contraction behavior of the selected extreme pair.

Proposition 1. *Suppose the two selected extreme agents at round t are updated according to Eqs.* (6) *and* (7)*, with concession rates satisfying $\alpha_{i^t}^t, \alpha_{j^t}^t \in (0,1)$ and $\alpha_{i^t}^t + \alpha_{j^t}^t \in (0,1)$. Then the diameter of the updated selected pair contracts strictly:*

$$\bar{\Delta}^t = (1 - \alpha_{i^t}^t - \alpha_{j^t}^t)\Delta^t < \Delta^t. \tag{13}$$

Proof. By substituting Eqs. (6) and (7), we obtain

$$\bar{p}_{j^t}^t - \bar{p}_{i^t}^t = \left(\alpha_{j^t}^t p_{i^t}^t + (1-\alpha_{j^t}^t)p_{j^t}^t\right) - \left((1-\alpha_{i^t}^t)p_{i^t}^t + \alpha_{i^t}^t p_{j^t}^t\right) \tag{14}$$

$$= (1 - \alpha_{i^t}^t - \alpha_{j^t}^t)(p_{j^t}^t - p_{i^t}^t). \tag{15}$$

Since $\alpha_{i^t}^t + \alpha_{j^t}^t \in (0,1)$, it follows that

$$0 < 1 - \alpha_{i^t}^t - \alpha_{j^t}^t < 1. \tag{16}$$

Therefore,

$$0 \le \bar{\Delta}^t = (1 - \alpha_{i^t}^t - \alpha_{j^t}^t)\Delta^t < \Delta^t, \tag{17}$$

which proves strict contraction. □

Corollary 1. *If there exists a constant $\underline{\alpha} > 0$ such that $\alpha_{i^t}^t + \alpha_{j^t}^t \ge \underline{\alpha}$ for all t, then the diameter of the updated selected pair satisfies*

$$\bar{\Delta}^t \le (1 - \underline{\alpha})\Delta^t. \tag{18}$$

Proposition 2. *Under the update rule in Eqs.* (6) *and* (7)*, the interval of the selected pair is forward invariant, i.e.,*

$$[\bar{p}_{i^t}^t, \bar{p}_{j^t}^t] \subseteq [p_{i^t}^t, p_{j^t}^t]. \tag{19}$$

Proof. Since $\bar{p}_{i^t}^t$ and $\bar{p}_{j^t}^t$ are convex combinations of $p_{i^t}^t$ and $p_{j^t}^t$, both must lie in the closed interval $[p_{i^t}^t, p_{j^t}^t]$. Hence, the updated interval is contained in the original interval. □

Remark. The above results establish strict contraction and forward invariance for the selected max-margin pair. In the complete multi-agent system, the global evolution of all opinions additionally depends on repeated max-margin reselection and the Nyström-style approximation used to update non-landmark agents. We therefore treat full-system consensus as an empirical property of the overall framework, rather than claiming a stronger theorem that does not directly follow from the pairwise update rule alone.

3.4 Consensus-Driven Question Answering

To evaluate the consensus status among agent opinions, we conduct a dual verification from both text-grounded consistency and image-grounded consistency. For text-grounded consistency, we adopt the DBSCAN clustering algorithm [35] to examine whether the opinions are sufficiently close in the semantic space. Let the vector after round t be $P^t = (p_1^t, \ldots, p_n^t)^T \in \mathbb{R}^{n\times 1}$. DBSCAN induces a partition $\mathcal{G}^t = \{G_1^t, \ldots, G_{M_t}^t\}$ over the non-noise points, where M_t denotes the number of clusters at round t. We declare that text-grounded consistency is achieved if DBSCAN returns a single non-noise cluster satisfying $M_t = 1$. To avoid indefinite re-discussion, we additionally impose an upper bound $T_{\max}$ on the number of interaction rounds; hence, the procedure terminates and the moderator agent summarizes the latest available agent opinions. For image-grounded consistency, we require the image agents to provide all primary entities they identify, along with the corresponding bounding boxes for each entity. Image-grounded consistency is considered achieved when the same entities are identified across image agents and the average Intersection over Union (IoU) of the bounding boxes for all matched entities is greater than η. Consensus is considered achieved only when both text-grounded consistency and image-grounded consistency are satisfied.

Once the stopping criterion is met, the moderator agent synthesizes the current viewpoints into a cohesive consensus representation. Unlike prior one-shot moderator-based methods, our moderator is invoked only after iterative consensus construction has substantially reduced inter-agent disagreement. The resulting consensus content is then combined with the question and visual input and fed into the MLLM to generate the final answer:

$$\mathbf{A}_{\text{final}} = \text{MLLM}(\mathbf{C}, \mathbf{Q}, \mathbf{I}), \tag{20}$$

where $\mathbf{C}$ denotes the consensus content, $\mathbf{Q}$ is the input question, and $\mathbf{I}$ denotes the associated visual input.

4 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our method and compare it against strong baselines.

4.1 Datasets

We evaluate on six public MQA benchmarks that cover comprehensive capability evaluation, text recognition, and multi-disciplinary reasoning.

Comprehensive Evaluation Datasets. GQA [17] is a large-scale benchmark for real-world visual reasoning and compositional question answering. It emphasizes logical and compositional understanding and is widely used to evaluate vision–language reasoning models. **SEED** [23] is a large-scale benchmark for evaluating the generative comprehension abilities of MLLMs, featuring human-annotated multiple-choice questions spanning both image and video understanding.

Text Recognition Datasets. EST-VQA [41] is a bilingual scene-text VQA dataset designed to promote generalization by reducing reliance on superficial data correlations. Following common practice, we use only the English subset of this dataset. **OCR-VQA** [31]

Table 1: General results of *CRATER* compared with baselines in terms of accuracy. Baseline names prefixed with "M" indicate that we adapt the original text-modal mechanisms into multi-modal counterparts within the MLLM reasoning process. The best results are highlighted by bold.

Method↓ Dataset→	Comprehensive Evaluation		Text Recognition		Multi-disciplinary	
	GQA	SEED	EST-VQA	OCR-VQA	ScienceQA	MMMU
Single-Agent Reasoning						
Qwen2.5-VL-7B [1]	59.30	62.34	70.62	77.83	67.09	58.56
M-Self-Reflection [33]	62.23	65.04	72.78	78.63	69.45	60.07
M-CoT [42]	61.37	65.91	71.15	79.42	70.03	59.68
Wiki-LLaVA [3]	63.53	68.90	77.04	80.37	73.56	62.33
FAST [38]	61.42	62.14	71.69	78.25	68.91	59.45
Voting-based Multi-Agent Reasoning						
M-Should-MAD [37]	65.14	71.57	78.43	84.78	75.16	63.50
M-GEDI [56]	67.31	71.23	77.86	85.67	77.42	63.05
M-DebUnc [53]	65.74	70.45	78.82	85.28	76.90	64.73
M-Sparse-MAD [25]	62.81	72.37	76.21	83.49	77.24	63.85
Summarization-based Multi-Agent Reasoning						
M-MAD [26]	65.62	72.91	79.48	84.87	80.43	65.14
M-VorC [18]	66.82	78.28	80.16	82.37	81.30	64.82
M-ChatEval [4]	65.34	78.84	80.71	85.86	82.45	66.38
M-BELLE [55]	71.91	82.56	83.89	86.11	84.74	67.23
CRATER (Ours)	**75.28**	**85.24**	**90.13**	**89.84**	**88.01**	**69.32**

focuses on questions that require understanding and reasoning over text in images. By coupling OCR with VQA, it introduces unique challenges for both text recognition and visual grounding.

4.2 Baselines

We compare *CRATER* with a strong single-agent MLLM baseline, as well as a broad set of multi-agent approaches covering self-reflection, debate, voting, and summarization paradigms. **Qwen2.5-VL-7B** [1] is a representative MLLM from the Qwen series, with strong capabilities in image recognition and object localization. **M-Self-Reflection** [33][2] prompts an agent to inspect and revise its own reasoning process, improving subsequent problem solving and reducing repeated mistakes. **M-CoT** [42] encourages divergent thinking and multi-step reasoning via multi-round discussions in which agents propose and rebut candidate ideas. **Wiki-LLaVA** [3] enhances question answering by integrating external knowledge from multi-modal sources via a hierarchical retrieval pipeline. **FAST** [38] is a visual agent framework that incorporates Fast–Slow Thinking mechanisms, dynamically adapting problem-solving strategies based on task complexity. **M-Should-MAD** [37] proposes a debate prompting strategy that dynamically modulates the level of agreement between agents during discussions. **M-GEDI** [56] is an electoral collective decision-making strategy that outputs ordinal rankings, providing richer signals about collective preferences than simple majority voting. **M-DebUnc** [53] incorporates uncertainty-aware confidence assessment into debate-based aggregation to reduce misleading outputs. **M-Sparse-MAD** [25] studies sparse communication topologies in multi-agent debate; in our experiments, we adopt its sparse neighbor-connected communication scheme. **M-MAD** [26] is a debate framework that uses "tit-for-tat" arguments to enhance reasoning and prevent thought stagnation. **M-VorC** [18] employs a consensus mode that prompts agents to converge on a shared solution. **M-ChatEval** [4] enables LLMs to collaboratively evaluate generated responses, and includes the Simultaneous-Talk-with-Summarizer mechanism for consolidation. **M-BELLE** [55] is a bi-level multi-agent reasoning framework that follows a fast–slow thinking paradigm, using debates to plan and execute adaptive strategies.

[2]Since few methods directly address multi-agent MQA, we also include several text-based multi-agent approaches as baselines. Baseline names prefixed with "M-" indicate that we adapt the original text-only mechanisms into multi-modal counterparts within the MLLM reasoning process.

4.3 Implementation Details

Our framework consists of six agents, each assigned a distinct role, and all role prompts are kept consistent across multi-agent baselines to ensure a fair comparison. The low-dimensional embedding space is set to one dimension, projecting diverse opinions onto a single viewpoint axis. The Gaussian kernel hyperparameter σ is set to 1. The concession rates are sampled randomly during discussion, and the maximum number of interaction rounds is set to $T_{\max} = 5$. For DBSCAN, we set `eps`=0.15 and `MinPts`=2. The threshold η for the average IoU of all matched entities is set to 0.7.

All baseline methods are implemented using Qwen2.5-VL-7B as the backbone, with the same decoding and model hyperparameter

Table 2: Results of the ablation study in terms of accuracy. CMAI: Collaborative Multi-Agent Initialization; MCR: Monotonic Concession Re-discussion; CDQA: Consensus-Driven Question Answering.

Type↓ Dataset→	GQA	EST-VQA	MMMU
CRATER (Ours)	**75.28**	**90.13**	**69.32**
w/o CMAI	67.33	85.57	61.48
w/o MDS Projection	72.15	87.34	66.90
w/o MCR	71.76	85.19	63.61
w/o CDQA	73.55	88.62	67.43

settings as those used in our *CRATER* experiments. For M-Self-Reflection and M-CoT, we follow the original inference procedures described in their papers. For M-GEDI, M-DebUnc, M-MAD, and M-BELLE, we strictly follow the original discussion rounds and multi-agent interaction protocols. For M-Sparse-MAD, we adopt the "3/5" sparse connection scheme described in the original paper while keeping all other settings consistent with the original implementation. M-Should-MAD is implemented using the collaborative mechanism of Solo Performance Prompting. For M-VorC, the consensus mechanism is implemented by prompting agents to converge to a shared solution. M-ChatEval adopts the Simultaneous-Talk-with-Summarizer mechanism. We use Qwen3-VL-Embedding-2B to encode agent opinions into embedding vectors.

We report accuracy as the evaluation metric. Statistical significance is assessed via t-tests; improvements of *CRATER* over the compared methods are significant with $p < 0.05$. All experiments are conducted on 8× NVIDIA A100 GPUs.

4.4 General Results

Table 1 reports the performance of *CRATER* across six datasets. We summarize several key observations. (1) *CRATER* consistently achieves the best overall performance among both voting-based and summarization-based multi-agent methods, including representative systems such as M-Should-MAD [37] and M-ChatEval [4]. We attribute these gains to our explicit and quantitative consensus construction mechanism, which models inter-agent disagreement, selects polarized pairs for re-discussion, and leverages the resulting consensus for final reasoning. (2) *CRATER* also outperforms FAST [38], highlighting the benefit of incorporating social-cognition-inspired consensus formation into multi-agent reasoning, beyond individual-level cognitive heuristics. (3) FAST [38] outperforms the single-agent baseline Qwen2.5-VL-7B [1], suggesting that introducing structured interaction mechanisms can improve reasoning robustness under multi-modal evaluation. (4) Overall, multi-agent methods outperform the single-agent baseline, likely because interactions among multiple agents provide complementary evidence and reduce idiosyncratic errors. Although *CRATER* achieves the strongest results, failures still occur in some cases; we conjecture that these failures are primarily caused by hallucinated evidence or other factors that disrupt effective consensus formation during discussion.

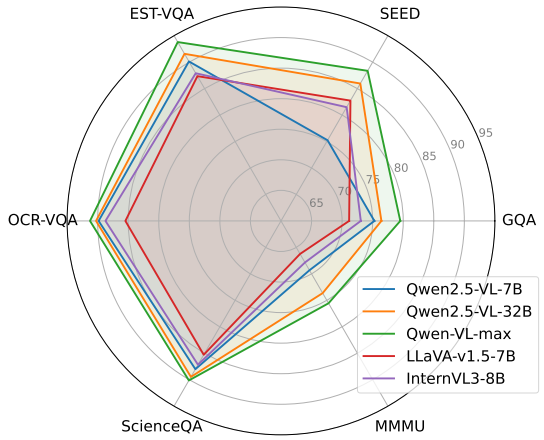


Figure 4: Performance comparison of different backbones.

4.5 Ablation Study

To evaluate the contribution of each component in *CRATER*, we conduct ablation studies on three datasets, with results reported in Table 2. We summarize the main observations as follows. (1) Removing **Collaborative Multi-Agent Initialization** keeps the same number of agents but discards role-specific perspectives and specialized analysis patterns. The resulting performance drop suggests that prompting agents to analyze the image and text from diverse viewpoints is important for capturing complementary evidence and achieving more comprehensive understanding. (2) Replacing the **MDS Projection** step with PCA [36] yields consistently worse results, indicating that MDS is better suited for our setting as it preserves pairwise (dis)similarities and the relative geometric structure of opinions in the low-dimensional space, which is critical for polarization-aware agent selection. (3) Omitting the **Monotonic Concession Re-discussion** module prevents the polarized agents from explicitly incorporating counterpart evidence and making structured concessions during re-discussion, leading to suboptimal performance. This validates the importance of interaction that encourages viewpoint adjustment, rather than rigidly maintaining one-sided conclusions. (4) Removing the **Consensus-Driven Question Answering** stopping mechanism by fixing the number of re-discussion rounds instead of using the clustering-based termination strategy also degrades performance. This highlights the advantage of DBSCAN in adaptively identifying convergence (or lack thereof) under varying question difficulty and disagreement patterns.

4.6 Detailed Analysis

Influence of Different Backbone Selection. We further vary the backbone MLLM and report the corresponding results in Figure 4. We draw two observations. (1) Backbones with larger parameter counts generally yield better performance across all six datasets; however, the improvements are relatively modest and come with substantially higher computational cost. (2) Among backbones with similar parameter scales, Qwen2.5-VL-7B achieves the best overall performance, which supports our backbone choice in the main experiments.

Table 3: Token consumption and speedup of different methods across datasets.

Method	GQA		EST-VQA		MMMU	
	# Tokens	Speedup	# Tokens	Speedup	# Tokens	Speedup
CRATER	**10382**	**1.52×**	**12360**	**1.78×**	**15650**	**1.44×**
M-BELLE	13491	1.37×	17417	1.41×	21594	1.35×
M-ChatEval	14159	1.23×	18058	1.33×	22626	1.12×
M-MAD	16864	1.00×	19110	1.00×	22681	1.00×

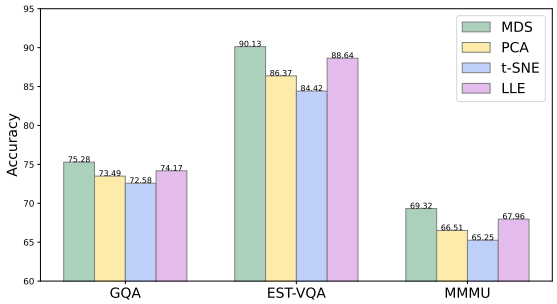


Figure 5: Performance comparison of different opinion projection methods.

Token and Time Consumption Analysis. We analyze token usage and wall-clock latency of different baselines on three datasets. As shown in Table 3, we observe that: (1) M-BELLE incurs higher token usage and time consumption than *CRATER* because its fast–slow discussion paradigm repeatedly incorporates extensive historical context, increasing both prompt length and decoding overhead; (2) M-ChatEval consumes fewer resources than M-MAD but still more than our method, largely due to its fully connected discussion process in which propagated content is repeatedly summarized by a moderator agent; and (3) M-MAD involves all agents in multiple rounds of discussion and requires historical records to be shared among all participants. Since the number of discussion rounds is determined dynamically by the MLLM, it exhibits the highest resource consumption among the compared methods.

Influence of Different Opinion Projection Methods. We further evaluate alternative projection methods for agent opinion embeddings. As shown in Figure 5, we find that: (1) methods that explicitly exploit distance/neighbor geometry, such as MDS and Locally Linear Embedding (LLE), achieve relatively strong performance, suggesting that pairwise opinion structure is an important signal for identifying polarized viewpoints in our framework; and (2) the slightly lower performance of LLE compared with MDS further supports the suitability of our chosen projection method for preserving global disagreement structure in a low-dimensional space.

Influence of Modality-Specific Agent Counts. Balancing performance and resource consumption, our main experiments adopt two image agents and two text agents (along with a debate agent and a moderator agent). To examine how modality-specific allocation affects performance under limited computational budgets, we vary the number of agents per modality between two and three. As shown in Table 4, we make three observations: (1) the "3–3" configuration achieves the best overall results, but also incurs substantially higher token usage and computational cost; (2) uneven modality allocation leads to a performance drop, likely because imbalanced modality-specific evidence biases the discussion and shifts the final aggregation toward over-relying on one modality; and (3) allocating more image agents for MMMU yields a slight improvement, plausibly because MMMU contains visually complex, fine-grained tasks where additional visual specialists extract more detailed cues. However, the cross-modality imbalance also constrains the magnitude of this gain.

Table 4: Comparison of assigning different numbers of agents to different modalities in terms of accuracy.

# Image Agents	# Text Agents	GQA	EST-VQA	MMMU
2	2	75.28	90.13	69.32
3	2	73.65	89.02	69.81
2	3	71.67	85.39	68.57
3	3	75.91	90.84	71.26

Table 5: Performance comparison of different coordinate updating methods.

Method↓ Dataset→	GQA	EST-VQA	MMMU	Avg.
CRATER (Ours)	75.28	90.13	69.32	78.24
Anchor Graph Hashing	72.65	87.36	65.87	75.29
Geometric Harmonics	72.28	86.64	64.51	74.48
Random Fourier Features	71.46	84.52	61.75	72.58

Influence of Different Coordinate Updating Methods. To investigate the impact of alternative coordinate updating strategies within the Monotonic Concession Re-discussion module, we replace the Nyström extension with several representative techniques, including Anchor Graph Hashing [21], Geometric Harmonics [9], and Random Fourier Features [51]. As shown in Table 5, the Nyström extension achieves the best overall performance, further supporting the effectiveness of our design choices in *CRATER*. Random Fourier Features yield lower scores, likely because they typically require a larger number of landmarks to avoid underfitting in low-sample regimes (here, only a small number of agents are available).

5 Conclusion

We propose *CRATER*, a consensus-driven multi-agent framework for multi-modal question answering, grounded in principles from social cognitive science. CRATER initializes participating agents with broad-spectrum role configurations to elicit diverse perspectives and complementary evidence. To facilitate consensus formation, we identify the two agents with the greatest opinion divergence by projecting agent outputs with Multi-Dimensional Scaling (MDS), and guide them through an iterative re-discussion process under a monotonic concession protocol. Once inter-agent consensus is achieved, the consensus content, together with the original question and image, is fed into an MLLM to generate a more accurate and coherent final response. Experiments on six benchmarks demonstrate the effectiveness of our approach. In future work, we plan to explore more efficient interaction protocols and automated role assignment strategies to further improve flexibility and scalability.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *CoRR* abs/2502.13923 (2025). arXiv:2502.13923 doi:10.48550/ARXIV.2502.13923

[2] Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective Multi-Agent Collaboration based on Large Language Models. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2024/hash/fa54b0edce5eef0bb07654e8ee800cb4-Abstract-Conference.html

[3] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *CVPR Workshops*. 1818–1826. doi:10.1109/CVPRW63382.2024.00188

[4] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *ICLR*. https://openreview.net/forum?id=FQepisCUWu

[5] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024. LION : Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge. In *CVPR*. 26530–26540. doi:10.1109/CVPR52733.2024.02506

[6] Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K. Wong. 2025. Affordances-Oriented Planning Using Foundation Models for Continuous Vision-Language Navigation. In *AAAI*. 23568–23576. doi:10.1609/AAAI.V39I22.34526

[7] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *CoRR* abs/2402.03216 (2024). arXiv:2402.03216 https://doi.org/10.48550/arXiv.2402.03216

[8] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. In *ICRA*. 14093–14100. doi:10.1109/ICRA57147.2024.10611018

[9] Ronald R Coifman and Stéphane Lafon. 2006. Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. *Applied and Computational Harmonic Analysis* 21, 1 (2006), 31–52.

[10] Michael AA Cox and Trevor F Cox. 2008. Multidimensional scaling. In *Handbook of data visualization*. Springer, 315–347.

[11] Ulle Endriss. 2006. Monotonic concession protocols for multilateral negotiation. In *AAMAS*. 392–399. doi:10.1145/1160633.1160702

[12] Charless C. Fowlkes, Serge J. Belongie, Fan R. K. Chung, and Jitendra Malik. 2004. Spectral Grouping Using the Nyström Method. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 2 (2004), 214–225. doi:10.1109/TPAMI.2004.1262185

[13] Liu He, Yizhi Song, Hejun Huang, Daniel G. Aliaga, and Xin Zhou. 2024. Kubrick: Multimodal Agent Collaborations for Synthetic Video Generation. *CoRR* abs/2408.10453 (2024). arXiv:2408.10453 doi:10.48550/ARXIV.2408.10453

[14] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *ICLR*. https://openreview.net/forum?id=VtmBAGCN7o

[15] Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024. mPLUG-PaperOwl: Scientific Diagram Analysis with the Multimodal Large Language Model. In *ACM MM*. 6929–6938. doi:10.1145/3664647.3681294

[16] Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-Write: A Persona-Driven Multi-Agent Framework for Diverse Argument Generation. In *COLING*. 4689–4703. https://aclanthology.org/2025.coling-main.314/

[17] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*. 6700–6709. doi:10.1109/CVPR.2019.00686

[18] Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or Consensus? Decision-Making in Multi-Agent Debate. *CoRR* abs/2502.19130 (2025). arXiv:2502.19130 doi:10.48550/ARXIV.2502.19130

[19] Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. 2025. Tree-of-Debate: Multi-Persona Debate Trees Elicit Critical Thinking for Scientific Comparative Analysis. *CoRR* abs/2502.14767 (2025). arXiv:2502.14767 doi:10.48550/ARXIV.2502.14767

[20] Jihyung Kil, Farideh Tavazoee, Dongyeop Kang, and Joo-Kyung Kim. 2024. II-MMR: Identifying and Improving Multi-modal Multi-hop Reasoning in Visual Question Answering. In *Findings of ACL*. 10698–10709. doi:10.18653/V1/2024.FINDINGS-ACL.636

[21] Saehoon Kim and Seungjin Choi. 2013. Multi-view anchor graph hashing. In *ICASSP*. 3123–3127. doi:10.1109/ICASSP.2013.6638233

[22] Devang Kulshreshtha, Wanyu Du, Raghav Jain, Srikanth Doss, Hang Su, Sandesh Swamy, and Yanjun Qi. 2025. The Subtle Art of Defection: Understanding Uncooperative Behaviors in LLM based Multi-Agent Systems. *arXiv preprint arXiv:2511.15862* (2025).

[23] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *CoRR* abs/2307.16125 (2023). arXiv:2307.16125 doi:10.48550/ARXIV.2307.16125

[24] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *CoRR* abs/2601.04720 (2026). arXiv:2601.04720 https://doi.org/10.48550/arXiv.2601.04720

[25] Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving Multi-Agent Debate with Sparse Communication Topology. In *Findings of the EMNLP*. 7281–7294. doi:10.18653/V1/2024.FINDINGS-EMNLP.427

[26] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *EMNLP*. 17889–17904. doi:10.18653/V1/2024.EMNLP-MAIN.992

[27] Xinyao Liao, Xianfang Zeng, Liao Wang, Gang Yu, Guosheng Lin, and Chi Zhang. 2025. MotionAgent: Fine-grained Controllable Video Generation via Motion Field Agent. *CoRR* abs/2502.03207 (2025). arXiv:2502.03207 doi:10.48550/ARXIV.2502.03207

[28] Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024. GroupDebate: Enhancing the Efficiency of Multi-Agent Debate Using Group Discussion. *CoRR* abs/2409.14051 (2024). arXiv:2409.14051 doi:10.48550/ARXIV.2409.14051

[29] Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. 2024. TextCoT: Zoom In for Enhanced Multimodal Text-Rich Image Understanding. *CoRR* abs/2404.09797 (2024). arXiv:2404.09797 doi:10.48550/ARXIV.2404.09797

[30] David Henry Mguni, Haojun Chen, Taher Jafferjee, Jianhong Wang, Longfei Yue, Xidong Feng, Stephen Marcus McAleer, Feifei Tong, Jun Wang, and Yaodong Yang. 2023. MANSA: Learning Fast and Slow in Multi-Agent Systems. In *ICML*, Vol. 202. 24631–24658. https://proceedings.mlr.press/v202/mguni23a.html

[31] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual Question Answering by Reading Text in Images. In *ICDAR*. 947–952. doi:10.1109/ICDAR.2019.00156

[32] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. 2024. Reason2Drive: Towards Interpretable and Chain-Based Reasoning for Autonomous Driving. In *ECCV (Lecture Notes in Computer Science, Vol. 15084)*. 292–308. doi:10.1007/978-3-031-73347-5_17

[33] Matthew Renze and Erhan Guven. 2024. Self-Reflection in LLM Agents: Effects on Problem-Solving Performance. *CoRR* abs/2405.06682 (2024). arXiv:2405.06682 doi:10.48550/ARXIV.2405.06682

[34] Cecilia L Ridgeway and Shelley J Correll. 2006. Consensus and the creation of status beliefs. *Social forces* 85, 1 (2006), 431–453.

[35] Erich Schubert, Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.* 42, 3 (2017), 19:1–19:21. doi:10.1145/3068335

[36] Jonathon Shlens. 2014. A Tutorial on Principal Component Analysis. *CoRR* abs/1404.1100 (2014). arXiv:1404.1100 http://arxiv.org/abs/1404.1100

[37] Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs. In *ICML*. https://openreview.net/forum?id=CrUmgUaAQp

[38] Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. 2025. Visual Agents as Fast and Slow Thinkers. In *ICLR*. https://openreview.net/forum?id=ncCuiD3KJQ

[39] Georgios Tziafas and Hamidreza Kasaei. 2024. Towards Open-World Grasping with Large Vision-Language Models. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, Vol. 270. PMLR, 3304–3332. https://proceedings.mlr.press/v270/tziafas25a.html

[40] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024. T-SciQ: Teaching Multimodal Chain-of-Thought Reasoning via Large Language Model Signals for Science Question Answering. In *AAAI*. 19162–19170. doi:10.1609/AAAI.V38I17.29884

[41] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. 2020. On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering. In *CVPR*. 10123–10132. doi:10.1109/CVPR42600.2020.01014

[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[43] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. 2024. Editable Scene Simulation for Autonomous Driving via Collaborative LLM-Agents. In *CVPR*. 15077–15087. doi:10.1109/CVPR52733.

2024.01428

[44] Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as We Do, Not as You Think: the Conformity of Large Language Models. In *ICLR*. https://openreview.net/forum?id=st77ShxP1K

[45] Ruth Wodak. 2003. From conflict to consensus? The co-construction of a policy paper. In *European Union discourses on un/employment: An interdisciplinary approach to employment policy-making and organizational change*. John Benjamins Publishing Company, 73–114.

[46] Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. 2024. DetToolChain: A New Prompting Paradigm to Unleash Detection Ability of MLLM. In *ECCV (Lecture Notes in Computer Science, Vol. 15090)*. Springer, 164–182. doi:10.1007/978-3-031-73411-3_10

[47] Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and Yu Huang. 2024. MLeVLM: Improve Multi-level Progressive Capabilities based on Multimodal Large Language Model for Medical Visual Question Answering. In *Findings of ACL*. 4977–4997. doi:10.18653/V1/2024.FINDINGS-ACL.296

[48] Quan Yan, Junwen Duan, and Jianxin Wang. 2024. Multi-modal Concept Alignment Pre-training for Generative Medical Visual Question Answering. In *Findings of ACL*. 5378–5389. doi:10.18653/V1/2024.FINDINGS-ACL.319

[49] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do Large Language Models Latently Perform Multi-Hop Reasoning?. In *ACL*. 10210–10229. doi:10.18653/V1/2024.ACL-LONG.550

[50] Shuwen Yang, Luwei Xiao, Xingjiao Wu, Junjie Xu, Linlin Wang, and Liang He. 2024. Simple contrastive learning in a self-supervised manner for robust visual question answering. *Comput. Vis. Image Underst.* 241 (2024), 103976. doi:10.1016/J.CVIU.2024.103976

[51] Tianbao Yang, Yufeng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. 2012. Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison. In *Neurips*. 485–493. https://proceedings.neurips.cc/paper/2012/hash/621bf66ddb7c962aa0d22ac97d69b793-Abstract.html

[52] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho de Castro, Shruthi Bannur, Stephanie L. Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew P. Lungren, Javier Alvarez-Valle, Aditya V. Nori, and Anja Thieme. 2024. Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. In *CHI*. 444:1–444:22. doi:10.1145/3613904.3642013

[53] Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. 2024. DebUnc: Mitigating Hallucinations in Large Language Model Agent Communication with Uncertainty Estimations. *CoRR* abs/2407.06426 (2024). arXiv:2407.06426 doi:10.48550/ARXIV.2407.06426

[54] Taolin Zhang, Dongyang Li, Chen Chen, Qizhou Chen, Jiuheng Wan, Xiaofeng He, Chengyu Wang, and Richang Hong. 2026. AMATA: Adaptive Multi-Agent Trajectory Alignment for Knowledge-Intensive Question Answering. In *ACL*. 7879–7900. https://doi.org/10.18653/v1/2026.acl-long.359

[55] Taolin Zhang, Dongyang Li, Qizhou Chen, Chengyu Wang, and Xiaofeng He. 2025. BELLE: A Bi-Level Multi-Agent Reasoning Framework for Multi-Hop Question Answering. *CoRR* abs/2505.11811 (2025). arXiv:2505.11811 doi:10.48550/ARXIV.2505.11811

[56] Xiutian Zhao, Ke Wang, and Wei Peng. 2024. An Electoral Approach to Diversify LLM-based Multi-Agent Collective Decision-Making. In *EMNLP*. 2712–2727.

[57] Enshen Zhou, Yiran Qin, Zhenfei Yin, Yuzhou Huang, Ruimao Zhang, Lu Sheng, Yu Qiao, and Jing Shao. 2024. MineDreamer: Learning to Follow Instructions via Chain-of-Imagination for Simulated-World Control. *CoRR* abs/2403.12037 (2024). arXiv:2403.12037 doi:10.48550/ARXIV.2403.12037

[58] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models. *CoRR* abs/2405.13872 (2024). arXiv:2405.13872 doi:10.48550/ARXIV.2405.13872