Complex & Intelligent Systems (2025) 11:464
https://doi.org/10.1007/s40747-025-02093-3

ORIGINAL ARTICLE

# Enhancing safety of vision-language reasoning through model-to-model deliberation

Sungwoo Kim[1] · Yongjin Lee[1] · Yunsick Sung[1]

Received: 8 July 2025 / Accepted: 1 September 2025 / Published online: 10 October 2025


## Abstract

Traditional vision-language models demonstrate strong performance in tasks such as image captioning and visual question answering, but they remain limited by issues such as hallucination, lack of self-correction, and shallow reasoning. These shortcomings compromise the safety, robustness, and consistency of their reasoning, particularly in ambiguous or high-stakes scenarios. In this paper, we propose three complementary frameworks aimed at enabling more trustworthy visual reasoning through structured deliberation. The first is the self-reflective reasoning single-agent framework, which facilitates iterative self-revision without requiring external supervision. The second is the structured debate agent framework, in which turn-based rebuttals between agents promote contrastive, multi-perspective refinement. The third is the progressive two-stage debate agent framework, which enables efficient yet accurate decision-making through model-to-model deliberation between smaller and larger agents. Experiments on the COCO dataset demonstrate that all three frameworks significantly enhance reasoning performance, achieving up to a 5.4% improvement in Intersection over Union (IoU) and over a 40% reduction in localization error compared to a single-pass baseline. Further evaluation across robustness (IoU), safety (self-revision rate, SRR), and consistency (consistency score, CS) confirms the effectiveness of multi-round, self-corrective, and multi-agent reasoning strategies. These results establish a practical path toward safer, more robust, and more interpretable vision-language models through lightweight, deliberative inference frameworks.

**Keywords** Vision language model (VLM) · Visual question answering (VQA) · Vision reasoning · Object detection · Debate

## Introduction

Traditional multi-model research has positioned vision-language reasoning as a core task, combining image understanding with a natural language framework to enable machines to interpret and answer questions about visual content [1]. Tasks such as visual question answering (VQA) [2], image captioning [3], and image-grounded dialogue [4] have seen significant improvements through the integration of large-scale pre-trained models [5]. These models show strong performance in understanding complex scenes and generating appropriate language, but making their visual reasoning framework safe and trustworthy is still a major [6–8]. In particular, the lack of robustness, safety, and consistency in their decision-making often leads to biased or inconsistent outputs [9]. The safety of vision-language models involves more than just preventing harmful outputs [10, 11], which includes the robustness, safety, and consistency of the visual reasoning framework that leads to a model's final refinement [6]. Traditional vision-language models have been evaluated based on their performance on benchmark tasks. However, their decision-making framework is often unclear and complex to verify [12]. Because it's difficult to understand how these models work, people worry that they may not truly understand the content, but instead rely on patterns in the data. To make these models safer, we need to find and fix problems related to how visual and language features are connected and represented, especially during reasoning. In this section, we explain three main challenges that make it hard for current vision-language models to reason in a safe and reliable way. **Imprecise alignment of vision and language**

✉ Yunsick Sung
sung@dongguk.edu

Sungwoo Kim
sean1255@dgu.ac.kr

Yongjin Lee
lyj.lmmlab@dgu.ac.kr

[1] Department of Computer Science and Artificial Intelligence, Dongguk University-Seoul, Seoul 04620, Republic of Korea

**features** Traditional vision-language models are designed to integrate visual and language inputs into a shared semantic space. However, current approaches often suffer from imprecise alignment between vision and language feature vectors. This misalignment can result in models focusing on irrelevant visual cues or misinterpreting the relationship between image content and linguistic queries, leading to hallucinated or semantically incoherent outputs [13]. **Importance of robust visual preprocessing** Safe and reliable vision-language reasoning begins with the quality of visual input and how effectively relevant features are extracted. Previous work in seismic image analysis [14, 15], omnidirectional vision [16], and biomedical simulation [17] has emphasized the critical role of preprocessing in enabling robust downstream interpretation. These studies address challenges such as noise suppression, spatial consistency, and illumination variance issues that directly affect how visual information is represented and understood by models. Although our focus lies in improving the reasoning layer, these earlier findings highlight that reasoning quality is closely tied to the fidelity of preprocessed visual features. **Limited contextual and situational inference** Traditional vision-language models perform well at detecting objects and describing visible scenes. However, they often struggle with understanding complex situations. These include social context, emotional cues, and how things change over time. Tasks that require deeper inference such as identifying a person's intent, understanding cause-and-effect relationships, or interpreting abstract meanings are still difficult for current models. This highlights a clear gap between machine understanding and human perception [18].

**Single-perspective and shallow reasoning** Traditional vision-language models generate refinement based on a single-pass Reasoning pipeline, lacking the ability to consider alternative viewpoints, self-correction, or iterative reasoning. This limitation results in rigid and narrow responses, particularly when faced with ambiguous or multifaceted inputs, and prevents models from engaging in a deeper cognitive framework akin to human deliberation.

One of the key challenges in current vision-language models is the issue of single-perspective and shallow reasoning. This limitation hinders a model's ability to interact in a deeper reasoning framework, such as considering alternative viewpoints, performing self-correction, or handling ambiguous visual-linguistic inputs. We detail the following core problems caused by this limitation and explain why addressing them is crucial for enhancing the robustness, safety, and consistency of vision-language reasoning.

**Absence of self-reflection and error correction** Traditional vision-language models operate in a static feed-forward manner, generating outputs without the ability to reflect on or revise their reasoning. Once a refinement is made, there is no framework to reassess its validity or explore alternative interpretations. This leads to over-confident errors, especially in complex or ambiguous situations.

**Lack of exposure to contrasting perspectives** Traditional vision-language models make decisions in isolation, without encountering opposing viewpoints or alternative hypotheses. As a result, their reasoning framework becomes unilateral and susceptible to bias. Without a structured, turn-based exchange, models lack the ability to refine conclusions through contrastive analysis or deliberation [19].

**Unreliable and unsafe outputs** Traditional vision-language models often generate outputs that are unreliable or socially unsafe. One prevalent issue is hallucination, where model produces descriptions that are semantically inconsistent with the visual input, such as referencing non-existent objects, attributes, or actions. This often results from over-reliance on learned priors and insufficient grounding in the actual image context [20]. The absence of ethical decision frameworks also makes models susceptible to generating socially harmful content, including biased, offensive, or culturally insensitive statements, particularly in complex social scenarios. These models also tend to demonstrate inconsistent responses when faced with repeated or rephrased queries, revealing a lack of stable reasoning pathways or memory of prior conclusions. Collectively, these issues pose significant challenges to the safe and trustworthy deployment of vision-language systems [21]. They also highlight the urgent need for deeper, multi-round reasoning frameworks [22, 23]. To address the shortcomings of traditional vision-language models such as hallucination, lack of self-correction, and shallow reasoning, we propose frameworks comprising three complementary methods. First, we propose a self-reflective reasoning framework in which a single model iteratively rebuttals and refines its captions without external feedback. Next, we propose a structured debate framework that enables two vision-language agents to interact in turn-based argumentation to mutually improve their reasoning. Finally, we propose a Progressive two-stage caption generation and selection framework, where a smaller model produces diverse candidates and a larger model refines the output through a two-step evaluation. As illustrated in Fig. 1, the overall system takes an original image and an input prompt as inputs, applies one of the three reasoning methods to generate an output prompt, and then extracts coordinate values from the generated prompt to refine and visualize bounding boxes on the original image. This integration of reasoning and spatial grounding enables the model not only to produce semantically coherent and safe outputs, but also to align its predictions more precisely with visual evidence.

First, we propose a self-reflective reasoning single-agent framework as a vision-language model that performs self-reflective reasoning to refine its outputs without external feedback. In this framework, model first generates an opinion for a given image and then internally rebuts the caption

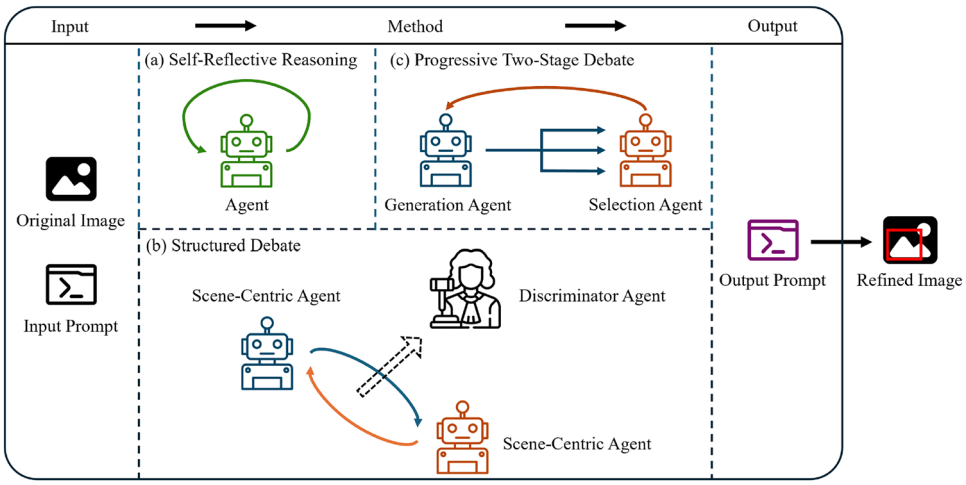



**Fig. 1** Overall architecture of the proposed frameworks for trustworthy vision-language reasoning. The system takes an original image and an input prompt, then applies one of the three proposed methods **a** self-reflective reasoning, **b** structured debate, or **c** progressive two-stage debate to generate an output prompt. From the generated prompt, coordinate values are extracted to refine and visualize bounding boxes on the original image. In self-reflective reasoning, a single model iteratively critiques and revises its captions without external feedback. In structured debate, two agents engage in turn-based argumentation to challenge and improve each other's reasoning. In the progressive two-stage debate, a smaller model generates diverse candidates, and a larger model iteratively selects and refines the most suitable output

based on its fluency, relevance, and factual alignment with the image. It subsequently produces a revised caption informed by this self-assessment. This framework of self-rebuttal and rewriting is repeated for a fixed number of iterations, avoiding the complexity of dynamic termination conditions while ensuring stable convergence. By incorporating a lightweight self-feedback loop, model simulates a self-revisioning framework similar to human reflection, enabling progressive refinement. Second, we propose a structured debate agent framework in which two vision-language agents interact in a turn-based argumentative dialogue. In each turn, a model can challenge, question, or refine the other's response. This turn-based setting enables agents to surface conflicting interpretations, identify potential flaws, and iteratively improve the quality of their reasoning. The overall framework simulates deliberation and introduces a self-corrective dynamic. Third, we propose a progressive two-stage debate agent framework with a consensus-based final refinement module. In the first stage, a smaller vision-language model generates three diverse captions for a given image. A larger model then selects the most semantically appropriate one among them. In the second stage, the smaller model generates two additional captions based on the previously selected caption. The larger model once again selects the final output from these, enabling model-to-model deliberation. This framework enhances the safety and robustness of model's reasoning. In this paper presents the following key contributions to enhance reasoning capabilities in vision-language models:

- **Problem identification and motivation:** We analyze the shortcomings of current vision-language models, particularly their dependence on shallow, single-pass reasoning without mechanisms for self-correction or multi-perspective interpretation. This motivates the need for deeper, more transparent reasoning frameworks.
- **Proposal of a unified framework:** We propose a framework that introduces complementary strategies to support deeper reasoning in vision-language tasks. The framework is designed to enhance robustness, reduce hallucination, and improve model's ability to deliberate over complex inputs.
- **First structured debate in vision models:** To our knowledge, this is the first work to implement a structured debate framework directly on top of a vision-language model, enabling explicit visual reasoning through agent-to-agent interaction.
- **Safety-oriented design for reliable outputs:** Through these components, our framework enhances the safety, robustness, and consistency of vision-language reasoning systems, especially in complex or ambiguous scenarios.

In the remainder of this paper, we first review existing efforts to enhance vision-language reasoning, including chain-of-thought prompting and multi-agent debate strategies. We then introduce and detail our three proposed frameworks: self-reflective reasoning, structured debate agent framework, and progressive two-stage debate agent framework. We further detail the design and workflow of each sub-framework with formal definitions and design components. Finally, we present our experimental setup and results, which demonstrate the effectiveness and improved reliability of our approach.

## Related work

Extensive research has been conducted to enhance the inference capabilities of large language models (LLMs) [24]. In this section, we review relevant work in two key areas: improving inference performance and facilitating inter-agent discussion.

### Reasoning

Since the introduction of chain-of-thought (CoT) [25], numerous methods have emerged to advance reasoning in language models.

**Self-consistency improves chain-of-thought reasoning in language models** [26] proposed the self-consistency method, which aggregates multiple sampled reasoning paths to select the most consistent final answer. This approach acknowledges that different inference trajectories can lead to varying conclusions. However, we observe that this framework can occasionally generate incorrect or nonsensical inference paths, indicating a need for further improvements in the model's foundational inference generation.

**Tree of thoughts: deliberate problem solving with large language models** [27] highlights a limitation of traditional CoT approaches: they follow a single reasoning path, risking failure if the initial path is flawed. To address this, the authors extend intermediate inference steps into a tree structure, generating multiple candidate paths and allowing the model to evaluate, explore, and backtrack autonomously. This method improves problem-solving by broadening the search space for solutions.

### Agent debate

To further enhance reasoning performance, recent work has explored multi-agent debate frameworks [21].

**ChatEval: towards better LLM-based evaluators through multi-agent debate** [28] proposes a method to organize a multi-agent group called 'ChatEval' to evaluate the quality of responses generated by different models. Each agent is given a unique persona and autonomously discusses and evaluates answers based on different perspectives or expertise. This can help capture nuances and subtleties that might be missed from a single perspective, but it is important to note that too many discussions do not necessarily improve performance, so it is important to balance the number of discussions appropriately.

**Encouraging divergent thinking in large language models through multi-agent debate** [29] investigates why LLMs such as ChatGPT continue to struggle with complex reasoning tasks, exploring self-reflection as a compensatory strategy [30]. The study identifies a key limitation, the degeneracy of thought, where an incorrect initial position hinders the generation of novel ideas despite reflection. To overcome this, they propose a multi-agent debate framework comprising two debater agents and one judge agent. Their experiments show that GPT−3.5- Turbo can outperform GPT-4 on tasks involving specific word translation and commonsense verification, although more debate rounds are needed for complex queries. The authors also note that substituting the judge with another LLM may compromise fairness.

**Can LLMs beat humans in debating? A dynamic multi-agent framework for competitive debate** [31] addresses challenges LLMs face in sustained confrontational debates, including hallucinations, safety alignment issues, and long-context limitations. To tackle these, they propose an agent debate framework where four specialized agents Searcher, Analyzer, Writer, and Reviewer collaborate dynamically to emulate a human debate team. Their evaluation of 200 real debate competition transcripts suggests this approach achieves competitive human-level argument coherence and persuasiveness.

In this paper, we propose a new reasoning approach using large language and vision assistant (LLaVA) [32], which receives both visual and textual inputs simultaneously, rather than the traditional method of discussing using text to reason about a situation or the correct judgment to make next from visuals. Our goal is to develop a framework that enables rational reasoning through discussion when given image information and input information.

## Safety of vision-language reasoning frameworks

To address the shortcomings of traditional vision-language models, such as hallucination, lack of self-correction, and shallow reasoning. We propose three independent frameworks: self-reflective reasoning single-agent framework, structured debate agent framework, and a progressive two-stage debate agent framework. Each framework is designed to repair one specific shortcoming and, in doing so, to reinforce a matching property that is critical for trustworthy reasoning: mitigating hallucination enhances safety, instituting iterative self-revision boosts consistency, and expanding one-shot reasoning into multi-round deliberation improves robustness. This proposed framework has the goal of enhancing the safety, robustness, and consistency of vision-language reasoning systems.

### Self-reflective reasoning single-agent framework

We propose a self-reflective reasoning single-agent framework using LLaVA as a vision language model $M$ to improve its reasoning ability through self-reflective iteratively. This framework is designed to refine faulty reasoning, such as

when model fails to correctly locate a vehicle in an image. In this framework, we define a initial round where the model receives the original image $i$ and an initial prompt $p$, and generates the first object bounding box prediction $b_0 = (x_{1_0}, y_{1_0}, x_{2_0}, y_{2_0})$ without any prior feedback or visual overlay. This round establishes the starting point for iterative self-refinement in subsequent rounds. From the first round to the last round, the previously predicted coordinates $x, y$ are overlaid on the image as an object bounding box, producing a modified input. Model then receives this updated image, along with the previous coordinates, a follow-up prompt $\dot{p}$, and output refined coordinates $b_n = (x_{1_n}, y_{1_n}, x_{2_n}, y_{2_n})$. The detailed set of prompts for this framework is provided in Appendix 11. Equation (1) shows $b_n$ represents the refined coordinates at round $n$. The initial refined $b_0$ is generated using the original image $i$ and prompt $p$. In subsequent rounds $n \geq 1$, model takes as input the image $i_{b_{n-1}}$, which is the image overlaid with an object bounding box at the previous refined $b_{n-1}$, along with the previous coordinates and the prompt.

$$b_n = M(i_{b_{n-1}}, b_{n-1}, \dot{p})(n \geq 1) \tag{1}$$

This iterative process is repeated for a fixed number of rounds to progressively refine the object bounding box, as shown in Fig. 2

By explicitly visualizing its previous decision and reevaluating the output in each round, model interact in a self-reflective loop that simulates human-like reflection, and this loop is repeated for a fixed number of user-specified iterations to enable progressively more robustness, safety, and consistency object bounding box without requiring external annotations or supervision.

## Structured debate agent framework

We propose a structured debate framework in which three agents, instantiated from the same pre-trained LLaVA model, interact in iterative turn-based argumentation to refine their visual understanding through contrasting multi-perspective. This framework is designed to simulate deliberative reasoning by encouraging the critical examination of alternative perspectives. To operationalize this debate, the three agents are assigned each a distinct interpretive role to promote multi-perspective visual reasoning. Agent $C$ is oriented toward scene-centric prompt $p^c$, Agent $J$ emphasizes object-centric prompt $p^j$, and Agent $D$ serves as the Discriminator agent for the final synthesizer of the debate result in the final round. We fix three role-specific prompts throughout the debate. The scene-centric prompt $p^c$ guides $C$ to reason about global context and spatial layout. The object-centric prompt $p^j$ steers $J$ toward fine-grained object attributes and relations. Then, the consensus prompt $p^d$ instructs $D$ to synthesize a balanced caption from all arguments. These prompts remain constant across rebuttal rounds, ensuring each agent preserves its interpretive stance while iteratively updating the content of its arguments. The detailed set of prompts for this framework is provided in Appendix 12. Here, scene-centric refers to a holistic interpretation focusing on the overall context and layout of the image, while object-centric emphasizes detailed recognition and attributes of individual objects within the scene. In the initial round of the debate, Agent $J$ and Agent $C$ independently receive the same image $i$ and are guided by each centric prompt that reflects their respective roles and reasoning goals. Each centric prompt is tailored to guide the agent's reasoning toward either holistic scene understanding or detailed object-level interpretation, depending on its assigned role. Each agent generates an initial opinion structured textual interpretation of the image from its own perspective. Specifically, Eq. (2) shows, Agent $C$ opinion $o^c$ captures scene-centric prompt $p^c$, while Agent $J$ opinion $o^j$ focuses on object-centric detail and attributes $p^j$. These initial opinions serve as the starting point for the iterative rebuttal phase.

$$o^c = C(i, p^c), \quad o^j = J(i, p^j) \tag{2}$$

From the first round, agents begin to exchange rebuttals. In the first round, each agent generates a rebuttal in response to the opponent's initial opinion. Equation (3) shows the opponent's Initial round of the opinion as input to generate their respective rebuttals $r_1^c$ and $r_1^j$. Here, $r_1^c$ means that agent $C$ rebuts agent $J$ initial opinion $o^j$, while $r_1^j$ indicates the rebuttal of agent $J$ against opinion $o^c$. This cross-evaluation encourages both agents to reconsider their reasoning based on the contrasting perspective of the other, laying the foundation for deeper deliberation. Here, $b_{o^j} = (x_{1_{o^j}}, y_{1_{o^j}}, x_{2_{o^j}}, y_{2_{o^j}})$ denotes the corner-based bounding box predicted for the main object referenced in $o^j$ during initial round, and $b_{o^c} = (x_{1_{o^c}}, y_{1_{o^c}}, x_{2_{o^c}}, y_{2_{o^c}})$ is defined analogously for $o^c$. At this stage, the role-specific prompt $p^c$ (or $p^j$) is prepended with "after reviewing the opponent's last statement, provide a rebuttal from your perspective" explicitly directing the agent to output a counter-argument rather than a fresh description. Agent $C$ takes as input the image $i_{b_{o^j}}$, which is the image overlaid with an object bounding box at the previously refined coordinates $b_{o^j}$ together with opinion $o^j$. Similarly, $r_1^j$ continues the dialogic exchange, deepening the reasoning process.

$$r_1^c = C(i_{b_{o^j}}, p^c, o^j), \quad r_1^j = J(i_{b_{o^c}}, p^j, o^c) \tag{3}$$

From the second round onward, each agent continues to generate rebuttals by responding to the previous rebuttal of its opponent. Equation (4) shows how agents respond to each other's first-round interpretations, here each rebuttal $r_n$ is

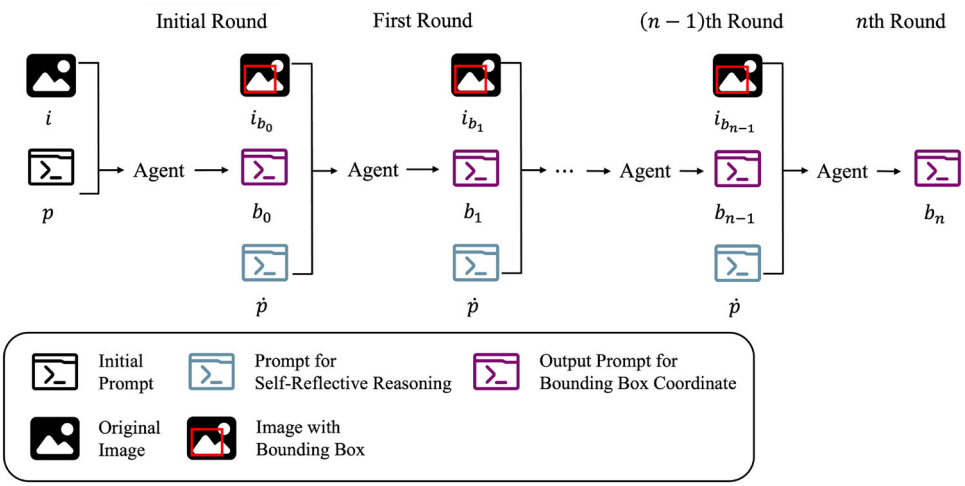



**Fig. 2** Overview of the self-reflective reasoning single-agent framework. Model receives the original image $i$, the initial prompt $p$ in the initial round. In the first round, it receives overlaid coordinates on the image $i_{b_0}$, the object bounding box coordinates $b_0$, and prompt $\dot{p}$. In subsequent rounds, model refines the object bounding box coordinates $b_n$ by overlaying previous coordinates as an object bounding box on the image, iteratively improving its refinement by revisiting the coordinates in each round. This framework is repeated for multiple iterations to improve robustness, safety, and consistency, without requiring external supervision or annotations

generated based on the rebuttal in the previous round. For example, $r_n^c$ indicates that agent $C$ now rebuts agent $J$ previous rebuttal $r_{n-1}^j$ rather than their initial opinion. Similarly, $r_n^j$ continues the dialogic exchange, deepening the reasoning process. This alternative rebuttal mechanism enables both agents to iteratively refine their understanding through multi-perspective contrastive reasoning.

$$r_n^c = C\left(i_{b_{r_{n-1}^j}}, p^c, r_{n-1}^j\right), \quad r_n^j = J\left(i_{b_{r_{n-1}^c}}, p^j, r_{n-1}^c\right) \tag{4}$$

Here, $b_{r_{n-1}^j} = (x1_{r_{n-1}^j}, y1_{r_{n-1}^j}, x2_{r_{n-1}^j}, y2_{r_{n-1}^j})$ and $b_{r_{n-1}^c} = (x1_{r_{n-1}^c}, y1_{r_{n-1}^c}, x2_{r_{n-1}^c}, y2_{r_{n-1}^c})$ denote the corner-based bounding boxes predicted in the previous round's rebuttals $r_{n-1}^j$ and $r_{n-1}^c$, respectively. These boxes are overlaid on the image to form the input image $i_{b_{r_{n-1}^j}}$ and $i_{b_{r_{n-1}^c}}$ for the current round $n$. The updated rebuttal at each round $n$, denoted by $r_n^c$ and $r_n^j$, integrate the scene centric agent input image $i_{b_{r_{n-1}^j}}$ and the object centric agent input image $i_{b_{r_{n-1}^c}}$ respectively with previous rebuttal for each. In the final round, Discriminator agent $D$ receives all previously generated opinions and rebuttals from both agents and synthesizes a final, Eq. (5) shows, consensus-based opinion, denoted as $o^*$. It reflects the most accurate, comprehensive, and balanced understanding of the image. This final opinion $o^*$ serves as the resulting caption for the image. Unlike the Generation agents, $D$ does not perform stochastic decoding or apply a temperature parameter. Instead, it follows a fixed, rule-based decision process defined explicitly in its prompt, ensuring deterministic selection of the final caption. This consensus prompt specifies evaluation criteria such as semantic completeness, factual accuracy, and alignment with both scene-centric and object-centric perspectives. The full text of this prompt is provided in Appendix 12.

$$o^* = D\left(i, p^d, \{o^c, o^j, r_1^c, r_1^j, \ldots, r_n^c, r_n^j\}\right) \tag{5}$$

This iterative rebuttal framework is shown in Fig. 3.

Rather than relying on parallel or isolated generations, this framework structures the interaction as an evolving exchange of rebuttals. By allowing agents to iteratively critique and refine each other's outputs, it promotes multi-perspective reasoning. This process also mitigates issues such as semantic incompleteness, overconfidence, and misinterpretation. Through such layered deliberation, the final output benefits from improved safety, robustness, and consistency.

## Progressive two-stage debate agent framework

We propose a progressive two-stage debate agent framework for opinion generation and selection, using two versions of the same vision-language model (LLaVA-7B and LLaVA-13B). These are designated as Generation agent define by function $G$ (LLaVA-7B), and Selection agent define by function $S$ (LLaVA-13B). The goal of the framework is to enhance three key aspects of semantic alignment, robustness, and accuracy of the reasoning of image opinion through iterative deliberation between agents with different capacities. Our framework operates through multiple rounds, each consisting of two stages: generation stage and selection stage. In each round, Generation agent produces diverse opinions, and Selection agent evaluates and selects the most appro-

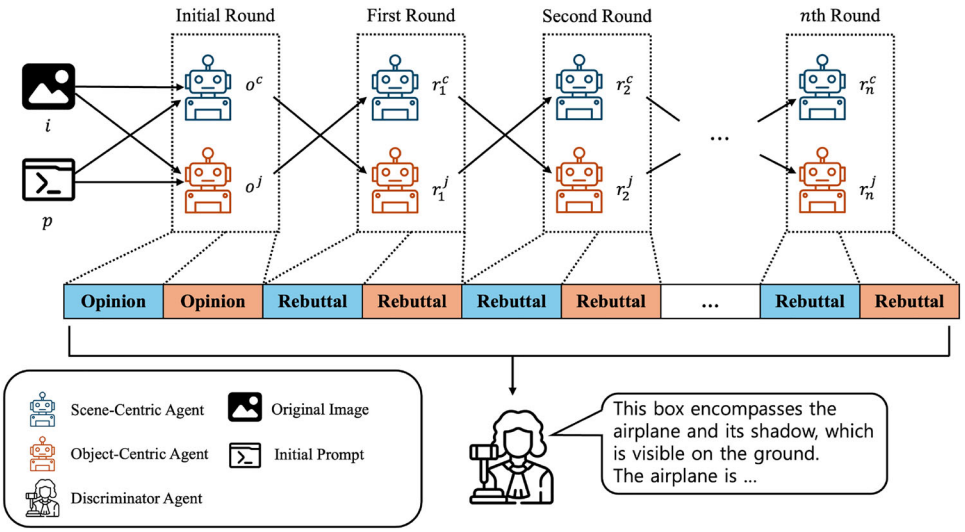



**Fig. 3** Overview of the structured debate agent framework. Two agents generate initial opinions and iteratively rebut each other's opinions over multiple rounds. In each round, one agent challenges the other's interpretation and refines their reasoning based on a multi-perspective approach. The updated image, with the previously refined coordinates overlaid as an object bounding box, is passed on to the next agent for further refinement. The final consensus is synthesized by a discriminator agent, which produces the final, most balanced, and comprehensive understanding of the image

priate one, as shown in Fig. 4. This iterative process refines the output across rounds. The reasoning process in both agents is controlled by a temperature parameter that affects the randomness of their outputs. Let $p^g$ denote the generation prompt given to $G$ in round $n$ and $p^s$ the selection prompt given to $S$ in the same round. In initial round, $p^g$ simply requests three diverse opinions on image $i$, whereas $p^s$ instructs $S$ to choose exactly one opinion. From first round onward, $\dot{p}^g$ explicitly asks $G$ to generate new opinions that exclude the previously selected opinion $s_n$, while $p^s$ directs $S$ to select a single opinion from $\{o_{n+1}, o_{n+2}, s_n\}$. The chosen opinion becomes $s_{n+1}$ and is passed to the next round. The detailed set of prompts for this framework is provided in Appendix 13. In the initial round, first stage, Generation agent $G$ receives an input image $i$ together with the generation prompt $p^g$ and produces three diverse opinions. In the initial round, second stage, each opinion $o_i$ reflects a different semantic perspective or interpretation of the input image. Equation (6) shows the Selection agent, guided by the selection prompt $p^s$, then evaluates these candidates in the context of the image and selects the one that best aligns with the semantic content based on its broader understanding and reasoning capacity. The selected opinion is denoted as $s_1$, representing the candidate chosen as the most appropriate opinion:

$$s_1 = S(i, p^s, \{o_1, o_2, o_3\}) \tag{6}$$

In the first round, first stage, Generation agent is prompted again with the same image and the previously selected opinion $s_1$, and $\dot{p}^g$, and generates two new opinions, $o_4$ and $o_5$, which are explicitly instructed to introduce different content or shifted emphasis relative to $s_1$. Equation (7) shows how the Generation agent uses both $i$ and $s_1$ as inputs to generate these diverse outputs:

$$\{o_4, o_5\} = G(i, \dot{p}^g, s_1) \tag{7}$$

In the first round, second stage, Selection agent evaluates the new set of opinions $\{s_1, o_4, o_5\}$ under $p^s$ and selects the most coherent and informative one, denoted as $s_2$. Equation (8) shows this selection process, where the Selection agent takes the image and the candidate opinions as input and outputs the refined final opinion:

$$s_2 = S(i, p^s, \{s_1, o_4, o_5\}) \tag{8}$$

In the final round, second stage, this final select opinion $s_{n+1}$ serves as the resulting caption, as shown in Fig. 4. Through this two-stage deliberation, Selection agent progressively refines the output based on comparative evaluation and broader contextual understanding.

The overall framework supports a user-specified number of debate rounds, allowing flexible control over the depth of deliberation and computational cost. Importantly, our framework operates in a lightweight manner in that it requires no additional training or parameter updates. Instead, it leverages the inherent reasoning capabilities of pre-trained vision-language models through structured intramodel debate. By iteratively generating and selecting opinions, the system achieves improved robustness, safety, and consistency without fine-tuning, demonstrating the potential of inference round deliberation alone to enhance captioning performance.

## Experiments

We conducted experiments to evaluate the effectiveness of each proposed reasoning framework: self-reflective reason-

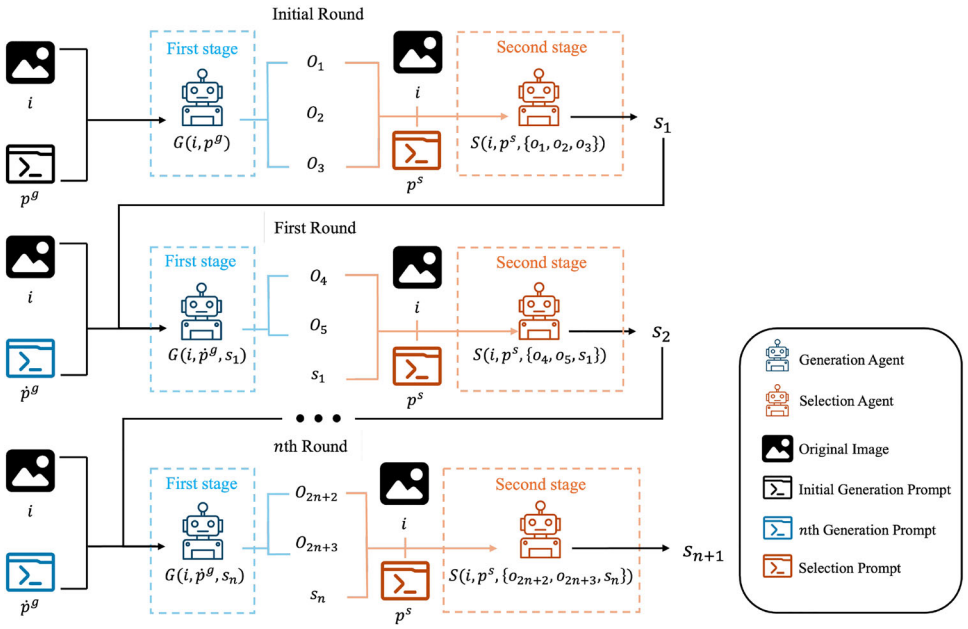



**Fig. 4** Overview of the progressive two-stage debate agent framework. Generation agent (LLaVA-7B) produces three diverse opinions based on the input image and the provided prompt. In the second stage, Selection agent (LLaVA-13B) evaluates the opinions generated by the Generation agent, selects the most appropriate opinion, and refines it further through additional reasoning. This iterative process is repeated, progressively improving the final output by considering alternative perspectives and ensuring semantic coherence in the final caption

ing, structured inter-agent debate, and progressive two-stage debate in improving the precision of object bounding box refinement. Specifically, we focused on refining the bounding box of a target object in an image through multiple rounds of deliberation. The task was formulated as an iterative object bounding box refinement problem. Given an input image, the goal was to accurately localize a specific object by progressively refining the coordinates of its bounding box, defined by the top-left and bottom-right corners. The initial bounding box prediction was generated by the LLaVA-7B vision-language model in response to a prompt such as: "Locate the object 'Airplane' in the image and generate a bounding box that fully contains the object. Return the coordinates in the exact format $(x_1, y_1, x_2, y_2)$. Normalize each coordinate value between 0 and 1, with $x_1 < x_2$ and $y_1 < y_2$." Each reasoning framework then iteratively refined this initial prediction to improve spatial accuracy and semantic alignment through subsequent rounds of reasoning. Experiments were performed on a manually curated subset of the common object in context (COCO) validation set [33], restricted to images containing a single salient object from a fixed set of categories such as car and airplane. Throughout all tables, $R$ denotes the refinement round with $R0$ as the initial prediction, and $N$ indicates the number of repetitions per experiment. The subsequent rounds are $(R1, R2, \ldots)$.

## Evaluation metrics

We evaluated the precision of each proposed visual reasoning framework through a set of quantitative metrics, carefully selected to reflect the most critical aspects of vision-language reasoning: robustness, safety, and consistency. These three dimensions collectively capture not only how well a model performs in terms of spatial prediction accuracy, but also how stable and trustworthy its decision-making process is over time and across repeated executions. Each metric is designed to probe a different property of visual reasoning behavior. Robustness reflects the model's ability to progressively improve its predictions through iterative deliberation. Safety assesses how stable the model's refinements are across visual reasoning rounds, indicating whether it can avoid erratic changes or overcorrections. Consistency measures whether the model can reproduce similar results when given the same input multiple times, which is essential for practical deployment in real-world systems. By evaluating the frameworks across these three complementary perspectives, our goal is to offer a holistic understanding of their effectiveness not just in reaching accurate predictions, but in demonstrating how performance progressively improves for each framework. The following subsections detail how each metric is formally defined and how it is applied in our experiments.

### Intersection over Union (robustness)

In this paper, we used Intersection over Union (IoU) [34] as the primary metric to evaluate the robustness of visual reasoning frameworks. Specifically, IoU measures how much the accuracy of an object bounding box improves through iterative visual reasoning and agent debate. Higher IoU across rounds indicates that model progressively refined its spatial refinement, demonstrating enhanced robustness through deliberative processes. The spatial accuracy of the refined object bounding box was measured relative to the ground truth. Higher IoU indicates better spatial alignment and

**Table 1** IoU performance across frameworks

| Method | R0 | R4 (best) |
|---|---|---|
| LLaVA-7B (baseline) | 0.891 | 0.898 |
| Consultancy | 0.872 | 0.904 |
| Self-reflective framework | 0.879 | 0.936 |
| Structured debate | 0.885 | 0.940 |
| Progressive two-stage debate | 0.898 | 0.961 |

robustness in correcting initial refine errors. To evaluate the robustness of different visual reasoning frameworks, we measured the IoU between the refined object bounding box and ground truth across multiple reasoning rounds. Table 1 summarises the IoU improvement across different frameworks over multiple rounds of visual reasoning. As shown, baseline model (LLaVA-7B) maintains a static IoU value across rounds, indicating a lack of iterative refinement. started at 0.891 and drifted only to 0.898 by the final round, confirming that a single-pass model offers virtually no iterative refinement. Consultancy included as an additional comparison model [35], showed a modest yet consistent rise. It entered at an IoU of 0.872 and climbed to 0.904 at its best round, a relative gain of 3.7%. This steady improvement confirms that even a simple consultancy-style refinement loop can correct initial bounding box errors. However, its final accuracy still fell short of the self-reflective and debate-based frameworks, which suggests that deeper reasoning strategies are more effective for precise spatial alignment. In contrast, self-reflective framework began lower, at 0.879, but increased steadily to 0.936 in the final round an overall gain of 9.7%. This indicated that the lightweight self-rebuttal loop effectively tightened the object bounding box after each iteration. Self-reflective framework shows gradual improvement in IoU over rounds, demonstrating that internal self-rebuttal enables model to progressively refine its spatial refinement. Structured debate framework achieved even higher accuracy, with consistent improvements across rounds as agents exchange rebuttals and collaboratively correct each other's interpretations. Structured debate framework enters at an already strong 0.885, dips slightly in round 2 as agents expose disagreements, and then rebounds to 0.940 once consensus is reached. Overall, every framework surpasses the static LLaVA-7B baseline, with self-reflective reasoning, structured debate, and progressive two-stage achieving the strongest endpoint accuracy. In particular, the progressive two-stage debate increased from an initial IoU of 0.898 to 0.961 after its single refine and select step (6.3%).

### Self-revision rate (safety)

To complement the safety evaluation, we proposed self-revision rate (SRR) as a supporting metric for the safety analysis. SRR uses IoU scores computed between consecutive object bounding-box. A higher SRR indicates higher decision safety, which we interpret as an indicator of safer and more consistent visual reasoning behavior. Equation (9) shows, Let $b_n$ denote the refined box in round $n$. We define SRR over $T$ total rounds. Here, $\delta$ is a fixed threshold set to 0.7 in our experiments, below which two bounding boxes are considered significantly different. The indicator function $\left[\text{IoU}(b_n, b_{n+1}) \geq \delta\right]$ returns 1 if the condition is satisfied and 0 otherwise.

$$\text{SRR} = \frac{1}{T-1} \sum_{n=1}^{T-1} \begin{cases} 1, & \text{if IoU}(b_n, b_{n+1}) \geq \delta, \\ 0, & \text{if IoU}(b_n, b_{n+1}) < \delta. \end{cases} \tag{9}$$

In particular, a higher SRR indicates that model's refine are more stable across rounds, reflecting safer and more reliable decision-making behavior. In particular, a low revision rate paired with a high IoU suggests that model not only stabilizes its refinement but also accurately aligns with the ground truth. This combination highlights model's ability to avoid overconfident hallucinations and output variability. It also demonstrates model's capability to maintain stable reasoning paths across multiple rounds, thereby enhancing its overall safety in visual reasoning tasks. In the SRR computation, the IoU between bounding boxes $b_n$ and $b_{n+1}$ quantifies the degree of spatial overlap between consecutive reasoning rounds. where $|b_n \cap b_{n+1}|$ denotes the area of intersection between the two bounding boxes and $|b_n \cup b_{n+1}|$ denotes the area of their union. A lower IoU indicates less overlap, suggesting a larger change between consecutive refine. Therefore, if IoU falls below $\delta$, the transition is considered a significant revision. We evaluated SRR on a set of 20 validation images. For each image, Eq. (9) was applied between all successive rounds, and the values were averaged to obtain per-round SRR scores. Table 2 presents these scores across all frameworks. As shown in Table 2, the two baseline methods exhibit the lowest safety scores. The LLaVA-7B baseline records the poorest SRR, averaging 0.434 across all rounds, indicating highly unstable and unsafe bounding-box updates. The Consultancy model shows modest improvement with an average SRR of 0.620, thanks to its feedback mechanism. However, it still falls short compared to our proposed frameworks. In contrast, our three frameworks achieve clearly higher SRR values, reflecting stronger stability. Both the self-reflective framework and progressive two-stage debate average 0.676, while the structured debate framework attains the highest overall average of 0.682. Notably, the progressive two-stage debate achieves the best single-round score of 0.81 in the final round, validating the effectiveness of the select-and-revise procedure can maximize safety once agent consensus is reached. These results suggest that deeper self-assessment and multi-agent debate mechanisms lead to safer

**Table 2** Self-revision rate (SRR) improvement across reasoning rounds (higher is better)

| Method | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| LLaVA-7B (baseline) | 0.49 | 0.39 | 0.43 | 0.41 | 0.45 |
| Consultancy | 0.50 | 0.67 | 0.66 | 0.62 | 0.65 |
| Self-reflective framework | 0.51 | 0.68 | 0.71 | 0.76 | 0.72 |
| Structured debate | 0.49 | 0.69 | 0.73 | 0.72 | 0.78 |
| Progressive two-stage debate | 0.53 | 0.63 | 0.67 | 0.74 | 0.81 |

**Table 3** Consistency score (CS) across independent runs (higher is better)

| Method | N2 | N3 | N4 | N5 |
|---|---|---|---|---|
| LLaVA-7B (baseline) | 0.78 | 0.75 | 0.72 | 0.76 |
| Consultancy | 0.80 | 0.78 | 0.81 | 0.81 |
| Self-reflective framework | 0.84 | 0.83 | 0.86 | 0.91 |
| Structured debate | 0.81 | 0.82 | 0.84 | 0.86 |
| Progressive two-stage | 0.85 | 0.81 | 0.87 | 0.94 |

and more consistent visual reasoning than the two existing baselines. Iterative self-reflection, inter-agent discussion, and progressive refinement all contribute to improved stability and reliability in decision-making.

In contrast, self-reflective reasoning, structured debate, and progressive two-stage debate significantly increase the SRR, suggesting that iterative self-assessment, debate agent, or Selection agent interaction leads to safer, more stable refinement. Overall, the results confirm that deeper or collaborative reasoning strategies enhance the safety of vision-language models.

### Consistency score (consistency)

To assess reasoning stability, we introduced the consistency score (CS), defined as one minus the standard deviation of final IoU values obtained from repeated executions under identical conditions. We compute the IoU using the final object bounding box generated at the end of the 5 rounds of the reasoning process. This procedure was repeated $K = 10$ times, yielding ten final IoU values. The standard deviation of these IoU values reflects the variability of the model's final outputs across repeated runs. A higher CS indicates that model makes stable refine across repeated inferences, reflecting strong internal reasoning prior and robustness to sampling variance or latent randomness within model. Equation (10) defines the standard deviation of the final IoU values obtained from $K$ repeated runs. Here, $\text{IoU}^j$ represents the IoU score of the j-th run, and $\bar{\mu}$ denotes the mean IoU score across all $K$ runs. This Eq. (10) quantifies the variability of the model's final IoU predictions value under identical input conditions and serves as the basis for computing the CS. To express this as a consistency score where higher values indicate more stable behavior, we define CS as one minus the standard deviation:

$$\text{CS} = 1 - \sqrt{\frac{1}{K}\sum_{j=1}^{K}\left(\text{IoU}^j - \bar{\mu}\right)^2} \tag{10}$$

As shown in Table 3, the two pre-existing baselines record the lowest consistency scores. LLaVA-7B (Baseline) averages 0.75 across the four independent runs (highest 0.78), indicating that its final refinements fluctuate considerably under identical inputs. Consultancy raises the mean CS to 0.80, confirming that a simple feedback loop can dampen variance, yet its stability still lags behind every framework we propose. Our three reasoning-centric frameworks achieve markedly higher repeatability. Self-reflective framework reaches a mean CS of 0.86, showing that an internal self-revision loop already provides strong run-to-run stability. Structured debate records 0.83 on average, surpassing consultancy and approaching the other proposed methods by having agents expose and resolve divergent views. Progressive two-stage debate delivers the best overall consistency, averaging 0.87 and peaking at 0.94 in Run 5, which confirms that a select then refine procedure can drive near-deterministic outcomes once consensus is reached. These findings align with the IoU and SRR analyses: frameworks equipped with deeper self-assessment or multi-agent discussion not only improve spatial accuracy and safety but also yield more consistent final refinements than the two baselines. Consequently, in deployment scenarios where repeatability is critical, such as autonomous perception, our three proposed frameworks offer a more trustworthy visual-reasoning process.

In conclusion, these results demonstrate that models with structured internal or collaborative reasoning processes not only achieve better spatial accuracy (IoU) and decision safety (SRR) but also show higher inference consistency (CS), as reflected in consistently reproducible refinement. In the practical value of such frameworks in deployment-critical settings where repeatability is crucial, such as auto-perception, higher consistency enhances the trustworthiness of the reasoning process.

## Results

This section presents the results of evaluating each of the reasoning frameworks introduced in "Safety of vision-language reasoning frameworks", applied over multiple rounds of refinement. We focused on the "Airplane" class from the COCO dataset, and assessed performance using the Intersection-over-Union (IoU) metric, which quantifies the overlap between the refined bounding box and the ground-

truth annotation. To make the refinement process visually interpretable, each refinement round was overlaid on the output image from the previous round, enabling a step-by-step visualization of bounding box adjustments. As a baseline, we conducted ten independent single-pass inferences using the LLaVA-7B model and recorded the resulting bounding box coordinates from scratch without iterative refinement. The average IoU scores from these ten runs are reported as the baseline rows in Tables 5, 7, and 9, yielding an average IoU of 0.877. For each set of experimental results, we describe the average improvement in reasoning for object detection. The performance achieved by our visual reasoning frameworks indicates how much each framework's loss is reduced compared to the baseline. To better quantify the performance improvement over this baseline, we measure the reduction in loss, where loss is defined as 1-IoU. The specific loss reduction rate is calculated as shown in Eq. (11).

$$\text{Loss reduction} = \frac{\text{Baseline loss} - \text{Proposed framework's loss}}{\text{Baseline loss}} \tag{11}$$

### Self-reflective reasoning single-agent framework

In the self-reflective reasoning single-agent experiment, the iterative refinement process is visualized in Fig. 5. The bounding boxes generated during each self-revision round are shown in blue, while the final refined box is shown in red. As shown in Table 4, a single forward pass of the LLaVA-7B model yields an initial IoU of 0.876. Beginning from this same initial bounding box, the self-reflection loop improves the IoU to 0.910 by the third round and further to 0.929 in the fifth round. Beyond five rounds, performance begins to fluctuate. Therefore, we fixed the maximum number of refinement rounds at five and conducted ten independent runs of the experiment. The aggregated outcomes, presented in Table 5, show that the self-reflective reasoning single-agent framework achieves an average IoU of 0.927, which corresponds to an approximate 5.0% improvement over the 0.877 baseline. According to Eq. (11), this translates to a 40.6% loss reduction.

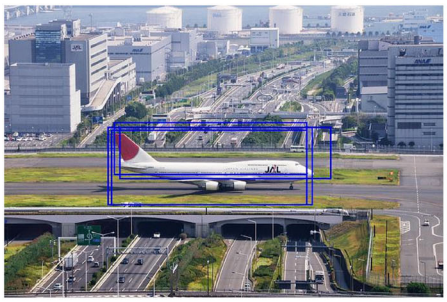
### Structured debate agent framework

In the structured debate agent framework experiment, Fig. 6 illustrates the object bounding box results at various stages: the initial coordinates independently detected by the Scene-centric and Object-centric agents; the intermediate coordinates proposed by both agents during the debate rounds (shown in blue); and the final coordinates selected by the discriminator (shown in red). At the outset, each agent focuses on its respective cue scene context for the Scene-centric agent and object-specific details for the Object-centric agent to identify candidate objects. The agents then share their findings and engage in a turn-based debate to iteratively refine their proposals. The per-round refinements from a single debate session are summarized in Table 6, where the discriminator value for each round corresponds to the IoU score of the bounding box selected by the discriminator based on the debate and refutation process in that round. Notably, since the IoU of the refined boxes began to decline after the fourth rebuttal round, only results up to Round 3 were submitted to the discriminator. This discriminator then aggregated the debate history to determine the final bounding box coordinates. We conducted ten independent runs of this experiment, with the aggregated results presented in Table 7. while also contributing to a reduction in the overall error rate by roughly 40.9%. The structured debate agent framework achieves an average IoU of 0.928, which is an absolute gain of 0.051 over the 10-run baseline average of 0.877. The final performance, as shown in Eq. (11), improved by 40.9%.

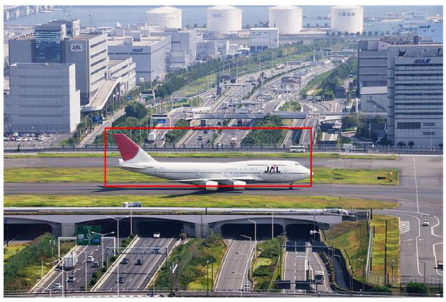
### Progressive two-stage debate agent framework

In the progressive two-stage debate agent framework experiment, Fig. 7 visualizes the candidate bounding boxes generated during the reasoning process. The blue boxes indicate the coordinates proposed during the first stage, while the red boxes show the coordinates selected in the second stage. As shown in Fig. 7a, the agent generates three initial candidate bounding boxes for the target object during the first stage. In the following iterations depicted in Fig. 7b, c two new candidate boxes are proposed in each round, while retaining the box previously selected in the second stage. Table 8 lists the candidate bounding box coordinates explored during each round of the first stage. The final IoU score results from ten independent experiments are summarized in Table 9. The progressive two-stage debate agent framework achieves an average IoU of 0.929. reflecting a 5.2 percentage point improvement over the 0.877 baseline. This corresponds to a loss reduction of 42.3%, as calculated using Eq. (11).

In the Generation stage, the temperature value is set to 1.2 to encourage a broader range of possible interpretations and diverse opinions. In contrast, in the Selection stage, the temperature value is set to 1.0 to ensure that the Selection agent makes more confident, deterministic choices based on the generated opinions. As shown in Table 10, the combination of a higher Generate temperature (1.2) and a moderate Select temperature (1.0) achieved the highest IoU score of 0.948. These values were determined through an ablation study exploring different combinations of Generate and Select temperatures, which showed that (Generate temperature 1.2, Select temperature 1.0) achieved the highest IoU among the tested settings. This configuration promotes diversity in the Generation stage while maintaining sufficient determinism in the Selection stage, leading to more accurate final



**(a)** Self-Reflective Reasoning



**(b)** Repeat the self-reflective reasoning process 5 rounds

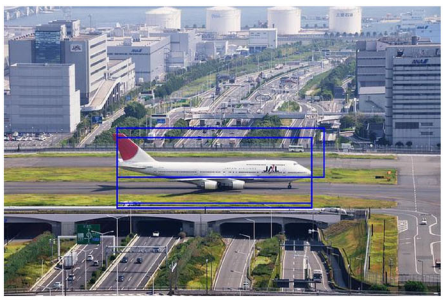
**Fig. 5** **a** Self-reflective refinement of the initial bounding box (blue) over multiple rounds. **b** Final bounding box after the last refinement step (red)

**Table 4** Self-reflective reasoning—IoU per round

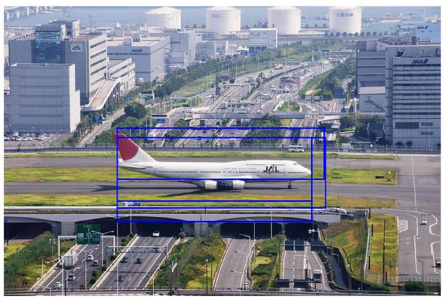
| Method | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Self-reflective | 0.876 | 0.864 | 0.908 | 0.910 | 0.929 | 0.925 | 0.910 | 0.908 | 0.781 | 0.914 |

**Table 5** Self-reflective reasoning—final IoU across 10 runs

| Method | N1 | N2 | N3 | N4 | N5 | N6 | N7 | N8 | N9 | N10 |
|---|---|---|---|---|---|---|---|---|---|---|
| LLaVA (baseline) | 0.876 | 0.892 | 0.884 | 0.861 | 0.894 | 0.862 | 0.837 | 0.886 | 0.895 | 0.879 |
| Self-reflective | 0.929 | 0.956 | 0.925 | 0.929 | 0.931 | 0.926 | 0.910 | 0.956 | 0.917 | 0.934 |



**(a)** Scene-Centric agent`s third rebuttal



**(b)** Object-Centric agent`s third rebuttal

**(c)** Discriminator Agent`s final decision

**Fig. 6** Bounding boxes proposed during debate rounds (blue). **b** Final bounding box selected by the discriminator (red)

**Table 6** Structured debate—IoU per round

| Method | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Scene-centric agent | 0.925 | 0.935 | 0.923 | 0.877 | 0.868 | 0.868 | 0.868 | 0.875 | 0.868 | 0.875 |
| Object-centric agent | 0.930 | 0.940 | 0.909 | 0.868 | 0.868 | 0.861 | 0.875 | 0.868 | 0.868 | 0.868 |
| Discriminator | 0.925 | 0.923 | 0.935 | 0.926 | 0.868 | 0.875 | 0.868 | 0.861 | 0.853 | 0.909 |

bounding boxes. Lower Generate agent temperatures tended to reduce diversity, while higher Select temperatures introduced inconsistency in final choices, both of which degraded performance.

**Table 7** Structured debate—final IoU across 10 runs

| Method | N1 | N2 | N3 | N4 | N5 | N6 | N7 | N8 | N9 | N10 |
|---|---|---|---|---|---|---|---|---|---|---|
| LLaVA (baseline) | 0.876 | 0.892 | 0.884 | 0.861 | 0.894 | 0.872 | 0.862 | 0.881 | 0.873 | 0.878 |
| Discriminator | 0.930 | 0.934 | 0.923 | 0.917 | 0.930 | 0.925 | 0.935 | 0.914 | 0.926 | 0.941 |

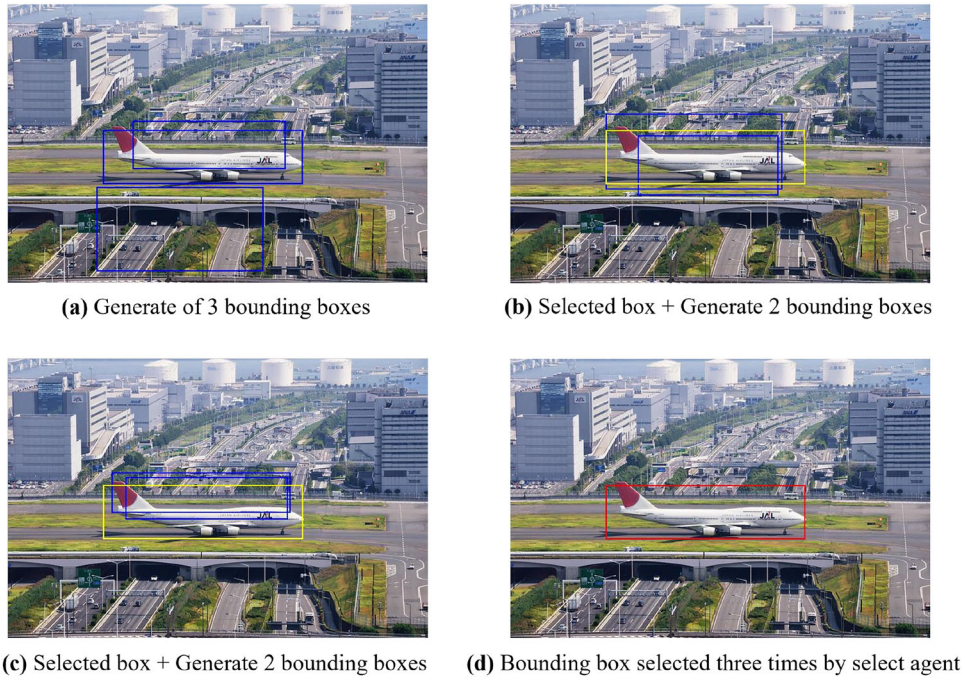


**Fig. 7** As in **a**, three candidate bounding boxes (in blue) are generated to find the corresponding object in the image. After **a**, the Selection agent selects the most appropriate bounding box among the generated coordinates and generates two more coordinates in addition to the bounding box selected in **b** (in yellow). After that, each stage is repeated as in **b** and **c**. Finally, if the Selection agent determines that the same coordinate is correct three times, the final selection is made as in **d**

**(a)** Generate of 3 bounding boxes

**(b)** Selected box + Generate 2 bounding boxes

**(c)** Selected box + Generate 2 bounding boxes

**(d)** Bounding box selected three times by select agent

**Table 8** Progressive two-stage debate agent—IoU per round

| Method | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 |
|---|---|---|---|---|---|---|---|---|---|---|
| First candidate | 0.926 | 0.932 | 0.932 | 0.945 | 0.951 | 0.907 | 0.907 | 0.901 | 0.901 | 0.937 |
| Second candidate | 0.930 | 0.914 | 0.914 | 0.951 | 0.914 | 0.881 | 0.888 | 0.888 | 0.946 | 0.946 |
| Third candidate | 0.932 | 0.923 | 0.945 | 0.885 | 0.906 | 0.908 | 0.908 | 0.908 | 0.933 | 0.933 |
| Selected | 0.932 | 0.932 | 0.945 | 0.951 | 0.906 | 0.906 | 0.908 | 0.901 | 0.946 | 0.937 |

**Table 9** Progressive two-stage debate agent—final IoU Scores (10 rounds)

| Method | N1 | N2 | N3 | N4 | N5 | N6 | N7 | N8 | N9 | N10 |
|---|---|---|---|---|---|---|---|---|---|---|
| LLaVA (baseline) | 0.876 | 0.892 | 0.884 | 0.861 | 0.894 | 0.862 | 0.837 | 0.886 | 0.895 | 0.879 |
| Select | 0.936 | 0.967 | 0.903 | 0.912 | 0.917 | 0.915 | 0.911 | 0.937 | 0.953 | 0.938 |

**Table 10** Generate vs select temperature—IoU scores

| Temperature | Select 0.8 | Select 1.0 | Select 1.2 | Select 1.4 |
|---|---|---|---|---|
| Generate 0.8 | 0.933 | 0.844 | 0.923 | 0.889 |
| Generate 1.0 | 0.898 | 0.903 | 0.888 | 0.827 |
| Generate 1.2 | 0.939 | 0.948 | 0.896 | 0.845 |
| Generate 1.4 | 0.863 | 0.706 | 0.863 | 0.884 |

Across all three frameworks, the results consistently demonstrate that deliberate, multi-perspective reasoning strategies significantly outperform the static single-pass baseline. These methods yield notable improvements in robustness, safety, and consistency of object localization.

## Discussion

The experimental results confirm that the proposed self-reflective reasoning single-agent framework, structured debate agent framework, and progressive two-stage debate agent framework significantly enhance the robustness, safety, and consistency of vision-language reasoning. Each framework addresses key challenges observed in traditional mod-

els, including hallucinations, inconsistent outputs, and shallow reasoning, through distinct mechanisms. The self-reflective reasoning single-agent framework exhibits clear advantages by enabling the model to autonomously refine its predictions. Incorporating a self-feedback loop allows the model to simulate internal cognitive revision, leading to more accurate object bounding box alignment as evidenced by improved IoU scores and SRR values [36]. These findings suggest that even without external agents, vision-language models can enhance their reasoning quality through internal deliberation, making this approach well-suited for resource-constrained or low-supervision environments. The structured debate agent framework outperforms the self-reflective model across all evaluation metrics. This supports the hypothesis that turn-based interaction between agents introduces constructive friction between differing perspectives, promoting deeper reasoning and reducing semantic bias. Furthermore, the consistently high CS indicates that multi-agent debate not only enhances reasoning depth but also results in more stable and reproducible refinements across repeated inferences. The progressive two-stage debate agent framework achieves the highest overall performance. By separating the generation and selection phases between models of different sizes (7B and 13B), it enables an efficient division of cognitive labor, reducing computational cost while preserving high reasoning quality. These results highlight that even lightweight, strategically structured deliberation can yield substantial improvements in visual reasoning. Across all three experimental frameworks, we observe consistent and substantial quantitative improvements. In terms of accuracy, all approaches significantly outperform the single-pass baseline (mean IoU = 0.877). The self-reflective, structured debate, and progressive two-stage frameworks achieve peak IoU scores of 0.927 (+5.0 pp), 0.940 (+6.3 pp), and 0.961 (+8.4 pp), respectively, reflecting a 38 42% reduction in average localization error. In terms of safety, the same trend holds in the SRR, the baseline's instability (SRR = 0.44) is nearly doubled by the self-reflective (0.72), structured debate (0.76), and progressive two-stage (0.78) frameworks. This indicates that over 70% of the refinements remain consistent under a strict overlap threshold ($\delta = 0.7$), requiring no corrective adjustments. Regarding consistency, the consistency score (CS) improves from 0.76 for the baseline to 0.91, 0.94, and 0.95 for the self-reflective, structured debate, and progressive two-stage frameworks, respectively, confirming that their predictions remain stable and reproducible when the same prompt–image pair is processed multiple times. Taken together, these results validate that deliberate, multi-perspective reasoning, whether through self-reflection, structured agent debate, or progressive staged reasoning, not only improves prediction quality but also enhances the safety, trustworthiness, and reliability of vision-language decision-making.

## Conclusion

In this paper, we addressed key limitations of traditional vision-language models, including hallucinations, lack of self-correction, and shallow reasoning, by introducing three novel frameworks designed to support deeper, safer, and more reliable visual reasoning: the self-reflective reasoning single-agent framework, the structured debate agent framework, and the progressive two-stage debate agent framework. Each proposed framework is designed to enhance a specific dimension of safe reasoning. The self-reflective framework emulates cognitive self-assessment through iterative self-revision; the structured debate framework enables multi-agent, rebuttal-based refinement by leveraging contrastive viewpoints; and the progressive two-stage framework distributes reasoning tasks between smaller and larger models to enable efficient yet accurate decision-making. Extensive experiments conducted on the COCO dataset demonstrate that the three frameworks significantly outperform the single-pass baseline in terms of robustness (measured by IoU), safety (measured by SRR), and consistency (measured by CS). Among them, the progressive two-stage framework achieved the highest overall performance, confirming that structured deliberation, even with limited interaction rounds, can produce outputs that are not only more accurate but also more stable and trustworthy. These findings suggest that safe and interpretable visual reasoning can be substantially improved by incorporating multi-round, self-corrective, and multi-agent deliberation strategies at inference time, without requiring additional training or supervision. This opens a promising direction for building more accountable and trustworthy vision-language systems, especially in high-stakes domains where interpretability and safety are essential. In future work, we plan to extend these frameworks to more complex, multi-object visual scenes and explore the integration of ethical and social constraints directly into the reasoning loop to enable socially aware visual intelligence.

## Appendix A

In this section, we present some example prompts used by our three frameworks. Tables 11, 12, and 13 cover input prompts, role-specific rebuttal prompts, and the generation-selection prompts with output format requirements, respectively.

**Table 11** Prompts used in the self-reflective framework

| Agent | Prompt |
|---|---|
| Initial round | Locate the object 'Airplane' in the image and generate a bounding box that fully contains the object<br>Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 < x2 and y1 < y2 |
| *n*th Round | The previous bounding box for 'Airplane' has top-left (x1:.3f, y1:.3f) and bottom-right (x2:.3f, y2:.3f). TASK: Directly output ONE improved bounding box that more accurately encloses 'Airplane'. Rules: - State ONE flaw of the old box (≤10 words). - Normalized to [0,1]; x1 < x2; y1 <y2; up to 4 decimals. - Completeness: the airplane must be fully inside (no truncation). - Tightness: reduce background margins compared to the previous box while preserving completeness. - Adjustment: modify at least two coordinates by ≥0.02 unless such a change would cause truncation. - Deterministic: do not sample; do not propose multiple candidates. Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 <x2 and y1 <y2<br>OUTPUT: Return exactly one line in the format: [x1, y1, x2, y2] |

**Table 12** Prompts used in the structured debate agent framework

| Agent | Prompt |
|---|---|
| Scene-centric agent at initial round | STRICT RULES:<br>1. You are a Scene-centric analyst<br>2. Reason about global context and spatial layout to localize the target robustly<br>3. Avoid overfitting to tiny parts unless necessary to ground the scene<br>4. You are Scene-centric agent. Locate exactly one tight bounding box that encloses all of 'Airplane' in the image<br>Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 < x2 and y1 < y2. Use up to 4 decimals |
| Object-centric agent at initial round | STRICT RULES:<br>1. You are an Object-centric analyst<br>2. Focus on fine-grained attributes, parts, and local relations to localize the target precisely<br>3. vague scene-level talk not helpful to localization<br>4. You are an Object-centric agent. Locate exactly one tight bounding box that encloses all of 'Airplane' in the image<br>Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 < x2 and y1 < y2. Use up to 4 decimals |

**Table 12** continued

| Agent | Prompt |
|---|---|
| Scene-centric agent at *n*th round | After reviewing the opponent's last statement (proposal: ans2_1), provide a rebuttal from your Scene-centric analyst perspective<br>Explicitly address ONE flaw in the opponent's proposal in your sentence<br>On a new last line, output your improved box as [x1, y1, x2, y2]<br>Keep normalization/order, Scene-centric<br>Target: 'Airplane'. Re-evaluate the image and provide your corrected box<br>Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 < x2 and y1 < y2. Use up to 4 decimals |
| Object-centric agent at *n*th round | After reviewing the opponent's last statement (proposal: ans1_1), provide a rebuttal from your Object-centric analyst. perspective<br>Explicitly address ONE flaw in the opponent's proposal in your sentence<br>On a new last line, output your improved box as [x1, y1, x2, y2]<br>Keep normalization/order, Object-centric<br>Target: 'Airplane'. Re-evaluate the image and provide your corrected box<br>Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 <x2 and y1 <y2. Use up to 4 decimals |
| Discriminator agent | STRICT RULES:<br>1. Candidates are the boxes proposed in the last round: cand1, cand2, cand3,cand4, cand5, cand6, cand7, cand8. 2. You are the Discriminator/Synthesizer<br>3. Select the single best box based on visual evidence<br>4. No sampling, do not invent or refine, select exactly one from candidates<br>5. prefer claims supported by both sides or strong cues; avoid hallucination<br>6. Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 <x2 and y1 < y2. Use up to 4 decimals<br>SELECTION CRITERIA (apply in order of priority): (1) Completeness: the box must fully enclose the airplane without truncating salient parts (e.g., wings, tail). (2) Tightness: minimize background margins while preserving completeness. (3) Edge alignment: box borders should align closely with the visible object edges/contours; aspect ratio must be reasonable. (4) Consensus preference: if both agents propose similar boxes, prefer the common overlap. (5) Tie-breaking (if scores are equal): prefer the box supported by both sides > smaller background area > Candidate A |

**Table 12** continued

| Agent | Prompt |
|---|---|
| | `OUTPUT FORMAT: Return ONLY the chosen bounding box on a single line in the exact format: [x1, y1, x2, y2]` |
| | `Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 < x2 and y1 < y2. Use up to 4 decimals` |

**Table 13** Prompts used in the progressive two-stage framework

| Agent | Prompt |
|---|---|
| Generation agent at initial round | `STRICT RULES:` |
| | `1. Output EXACTLY three bounding boxes` |
| | `2. Each line MUST be in the form: [x1, y1, x2, y2]` |
| | `3. Each of the three boxes must differ from the others by ≥0.02 on ≥2 axes. Recommend exactly 3 different bounding boxes that could locate 'Airplane' in the image` |
| | `3. Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 < x2 and y1 < y2. Use up to 4 decimals` |
| Generation agent at *n*th round | `The previously selected bounding box is {seed_str}` |
| | `STRICT RULES:` |
| | `1. For finding 'Airplane', generate 2 brand-new boxes DIFFERENT from it,` |
| | `2. Then output EXACTLY 3 boxes TOTAL: the new ones FIRST,` |
| | `3. Each new box must differ from the seed by ≥0.02 on ≥2 axes; the two new boxes must also differ from each other by ≥0.02 on ≥2 axes` |
| | `Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 < x2 and y1 < y2. Use up to 4 decimals` |
| Selection agent at all rounds | `You are a deliberative agent that must pick ONE bounding box from the given list` |
| | `Here are the candidate boxes (normalized): 1. {cands[0]} 2. {cands[1]} 3. {cands[1]}` |
| | `Choose the single box that most likely encloses an 'Airplane'` |
| | `You must deterministically select exactly ONE box from this list. No sampling; no new boxes` |
| | `GOAL: Choose the box that would maximize IoU w.r.t. the true object (proxy by visual evidence of an 'Airplane')` |
| | `SCORING RUBRIC:` |
| | `1. no sampling, do not invent or refine, select exactly one from candidates` |
| | `2. Completeness: the airplane is fully inside; no truncation of salient parts` |
| | `3. Tightness: minimal background margins while preserving completeness` |

**Table 13** continued

| Agent | Prompt |
|---|---|
| | `4. Edge alignment: borders align with visible object edges/contours; reasonable aspect/size`<br>`5. Stability preference: if the seed box (previous selection) is present among candidates,`<br>`prefer it unless a new box strictly improves (1) and (2) by ≥0.02 on at least two axes`<br>`width or height ≤0.02; obvious truncation or excessive background`<br>`TIE-BREAKING (deterministic): higher Completeness > higher Tightness > seed (if present) > lower index`<br>`OUTPUT: ONLY the chosen box in the exact format [x1, y1, x2, y2]. No explanations`<br>`Return the coordinates in the exact format [x1, y1, x2, y2], normalized between 0 and 1, with x1 < x2 and y1 < y2. Use up to 4 decimals` |

**Acknowledgements** Icons in this paper were generated by Freepik.

**Author Contributions** Sungwoo Kim: Methodology, Yongjin Lee: Investigation. Yunsick Sung: Reviewing, Supervision.

**Funding** This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00254592) grant funded by the Korea government(MSIT). Korea Institute of Police Technology (KIPoT) funded by Korean Government [Korean National Police Agency (KNPA)] (AI Driving Ability Test Standardization and Evaluation Process Development) (Grant number: 092021D75000000).

**Data availability** The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Declarations

**Conflict of interest** The authors declare no conflict of interest.



## References

1. Everitt T, Lea G, Hutter M (2018) AGI safety literature review. arXiv preprint arXiv:1805.01109
2. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) VQA: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433
3. Herdade S, Kappeler A, Boakye K, Soares J (2019) Image captioning: transforming objects into words. In: Advances in neural information processing systems, vol 32
4. Mostafazadeh N, Brockett C, Dolan B, Galley M, Gao J, Spithourakis GP, Vanderwende L (2017) Image-grounded conversations: multimodal context for natural question and response generation. arXiv preprint arXiv:1701.08251
5. Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, Aslanides J, Henderson S, Ring R, Young S, et al (2021) Scaling language models: methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446
6. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv preprint arXiv:1606.06565
7. Dong H, Xiong W, Goyal D, Zhang Y, Chow W, Pan R, Diao S, Zhang J, Shum K, Zhang T (2023) Raft: reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767
8. Ouyang L, Jeffrey W, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A et al (2022) Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 35:27730–27744
9. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I (2023) Improving factuality and reasoning in language models through multiagent debate. In: Forty-first international conference on machine learning
10. Anwar U, Saparov A, Rando J, Paleka D, Turpin M, Hase P, Lubana ES, Jenner E, Casper S, Sourbut O, et al (2024) Foundational challenges in assuring alignment and safety of large language models. arXiv preprint arXiv:2404.09932
11. Gabriel I (2020) Artificial intelligence, values, and alignment. Mind Mach 30(3):411–437
12. Zhou K, Yang J, Loy CC, Liu Z (2022) Learning to prompt for vision-language models. Int J Comput Vis 130(9):2337–2348

13. Bonwoo G, Sung Y (2023) UX framework including imbalanced UX dataset reduction method for analyzing interaction trends of agent systems. Sensors 23(3):1651
14. Kumar A, Sato A, Oishi T, Ono S, Ikeuchi K (2014) Improving GPS position accuracy by identification of reflected GPS signals using range data for modeling of urban structures. Seisan Kenkyu 66(2):101–107
15. Kumar A, Banno A, Ono S, Oishi T, Ikeuchi K (2013) Global coordinate adjustment of the 3D survey models under unstable GPS condition. Seisan Kenkyu 65(2):91–95
16. Aggarwal AK, Chauhan APS (2025) Robust feature extraction from omnidirectional outdoor images for computer vision applications. Int J Instrum Meas 10:8–13
17. Kumar A (2009) Light propagation through biological tissue: comparison between Monte Carlo simulation and deterministic models. Int J Biomed Eng Technol 2(4):344–351
18. Li S, Sung Y (2023) MRBERT: pre-training of melody and rhythm for automatic music generation. Mathematics 11(4):798
19. Zhang Y, Sung Y (2023) Hybrid traffic accident classification models. Mathematics 11(4):1050
20. Rahman S, Issaka S, Suvarna A, Liu G, Shiffer J, Lee J, Parvez MR, Palangi H, Feng S, Peng N, et al (2025) AI debate aids assessment of controversial claims. arXiv preprint arXiv:2506.02175
21. Brown-Cohen J, Irving G, Piliouras G (2023) Scalable AI safety via doubly-efficient debate. arXiv preprint arXiv:2311.14125
22. Deepak P (2024) AI safety: necessary, but insufficient and possibly problematic. AI Soc 40(2):1143–1145
23. Irving G, Christiano P, Amodei D (2018) AI safety via debate. arXiv preprint arXiv:1805.00899
24. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang J, Jin S, Zhou E et al (2025) The rise and potential of large language model based agents: a survey. Sci China Inf Sci 68(2):121101
25. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D et al (2022) Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 35:24824–24837
26. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A, Zhou D (2022) Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171
27. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K (2023) Tree of thoughts: deliberate problem solving with large language models, 2023. https://arxiv.org/abs/2305.10601, 3
28. Chan C-M, Chen W, Su Y, Yu J, Xue W, Zhang S, Fu J, Liu Z (2023) ChatEval: towards better LLM-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201
29. Liang T, He Z, Jiao W, Wang X, Wang Y, Wang R, Yang Y, Shi S, Tu Z (2023) Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118
30. Wang B, Yue X, Sun H (2023) Can ChatGPT defend its belief in truth? Evaluating LLM reasoning via debate. arXiv preprint arXiv:2305.13160
31. Zhang Y, Yang X, Feng S, Wang D, Zhang Y, Song K (2024) Can LLMs beat humans in debating? a dynamic multi-agent framework for competitive debate. arXiv preprint arXiv:2408.04472
32. Liu H, Li C, Wu Q, Lee YJ (2023) Visual instruction tuning. Adv Neural Inf Process Syst 36:34892–34916
33. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context. In: Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6–12, 2014, proceedings, part V 13. Springer, pp 740–755
34. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 658–666
35. Khan A, Hughes J, Valentine D, Ruis L, Sachan K, Radhakrishnan A, Grefenstette E, Bowman SR, Rocktäschel T, Perez E (2024) Debating with more persuasive LLMs leads to more truthful answers. arXiv preprint arXiv:2402.06782
36. Ren R, Basart S, Khoja A, Gatti A, Phan L, Yin X, Mazeika M, Pan A, Mukobi G, Kim R et al (2024) Safetywashing: do AI safety benchmarks actually measure safety progress? Adv Neural Inf Process Syst 37:68559–68594