MADIAVE: Multi-Agent Debate for Implicit Attribute Value Extraction

Wei-Chieh Huang[1]   **Cornelia Caragea**[1]

[1]University of Illinois Chicago

{whuang80, cornelia}@uic.edu

Abstract

Implicit Attribute Value Extraction (AVE) is essential for accurately representing products in e-commerce, as it infers latent attributes from multimodal data. Despite advances in multimodal large language models (MLLMs), implicit AVE remains challenging due to the complexity of multidimensional data and gaps in vision-text understanding. In this work, we introduce MADIAVE, a multi-agent debate framework that employs multiple MLLM agents to iteratively refine inferences. Through a series of debate rounds, agents verify and update each other's responses, thereby improving inference performance and robustness. Experiments on the ImplicitAVE dataset demonstrate that even a few rounds of debate significantly boost accuracy, especially for attributes with initially low performance. We systematically evaluate various debate configurations, including identical or different MLLM agents, and analyze how debate rounds affect convergence dynamics. Our findings highlight the potential of multi-agent debate strategies to address the limitations of single-agent approaches and offer a scalable solution for implicit AVE in multimodal e-commerce.

1 Introduction

Technological advancements have radically reshaped consumer behavior, establishing e-commerce as an indispensable part of modern life. Recent studies (Mohdhar and Shaalan, 2021) indicate that e-commerce will continue to drive future application development and thrive in an increasingly digital landscape, underscoring its ever-growing importance. However, accurately identifying product attributes remains a significant challenge (Zhang et al., 2024), as incorrect or imprecise labels can lead to customer dissatisfaction and a loss of trust and loyalty (Lin et al., 2021). In response, Attribute Value Extraction (AVE) has emerged as a promising approach to enhance

[Title] RuffleButts Little Girls Teal Ruffled Shirt w/Pink Bow
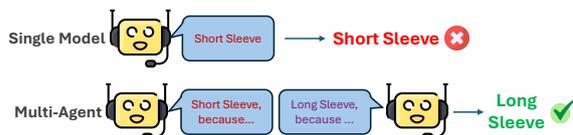
Figure 1: Illustration of the differences between explicit and implicit attributes. A single model may sometimes produce incorrect inferences, whereas a multi-agent setting can potentially provide more accurate reasoning and lead to correct inferences.

product representation and categorization (Zheng et al., 2018; Yan et al., 2021; Yang et al., 2023). Owing to the complexity and enormous volume of data, AVE has attracted significant research interest (Hu et al., 2025; Gong et al., 2025), leading to the development of various methods to more effectively address this challenge.

AVE can generally be classified into two categories: *(1) Explicit* (Sabeh et al., 2024a; Su et al., 2025) and *(2) Implicit* (Gong and Eldardiry, 2024; Jia et al., 2024). Explicit approaches extract attributes that are directly mentioned in the textual description, whereas implicit approaches infer attributes from visual cues and the textual information, as illustrated in Figure 1. Due to the inherent challenges of the task, most research has focused on explicit AVE. For instance, some recent studies (Luo et al., 2022; Loughnane et al., 2024) have proposed model architectures based on transformers and named-entity recognition models to address explicit AVE tasks. However, a key limitation of these approaches is that they rely solely on textual similarity for label extraction rather than inferring

attributes from underlying content or images, potentially overlooking latent attributes. Consequently, researchers have begun exploring implicit AVE methods. Some studies (Zhang et al., 2023; Luo et al., 2023; Zou et al., 2024b) propose advanced methods and architectures aimed at reducing the need for expensive labeled data in implicit AVE. Yet, these methods can still be costly to train, often require labeled data, and achieve only moderate inference performance. Addressing these challenges remains a key area for further research.

To address these challenges, we propose a novel framework, MADIAVE, which integrates textual and visual reasoning across multiple agents to enhance overall performance in a fully *zero-shot setting* with multimodal large language models (MLLMs). Inspired by the success of large language models (LLMs) (Zou et al., 2025a; Wu et al., 2025b) in general inference tasks (Xu et al., 2024; He et al., 2024), and the robust capabilities of agents (Zhang et al., 2025a; Li et al., 2025a; Zou et al., 2025b), we posit that multi-round, multi-agent debate and reasoning (Li et al., 2025b; Zhang et al., 2025b) can refine model outputs, guiding them toward consensus and ultimately achieving correct inferences for the implicit AVE task. To better understand agent behavior during debate, we systematically analyze performance and convergence in debates among both identical and different agents. We expect this framework not only to advance the state of the art in implicit AVE, but also to serve as a representative approach for other multimodal tasks.

Our contributions are summarized as follows:

- **Novel Framework for Implicit AVE:** We present a novel framework that integrates reasoning capabilities of multi-agents through debate to enhance inference performance. To best of our knowledge, this is the first study to investigate multi-agent debate mechanism into multimodal implicit AVE task.

- **Comprehensive Evaluation of Debate Configurations:** We systematically evaluate various debate configurations, including those involving identical, diverse, open-source, and closed-source MLLMs in a zero-shot setting. This analysis provides valuable insights into how different MLLMs contribute to and impact each other's performance.

- **In-depth Analysis of Debate Convergence:**

We present a detailed examination of debate statistics and reasoning consensus to understand how agents adopt, refine, or reject each other's viewpoints during the debate, and how final convergence is ultimately achieved.

2 Related Works

2.1 Explicit and Implicit AVE

Early stages of AVE research primarily focused on explicit extraction methods. For example, early approaches (Zheng et al., 2018; Wang et al., 2020) utilized conventional long short-term memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) to tackle the AVE task. However, the inherent complexity of multimodal data limited the accuracy of these methods. Although recent research has introduced advanced techniques (Yang et al., 2023; Sabeh et al., 2024b), enhanced encoders (Wang et al., 2022a), and innovative architectures (Brinkmann et al., 2024; Loughnane et al., 2024) for explicit AVE, these approaches often overlook potential attributes derivable from visual cues and textual information. Alternatively, the dataset and framework focus on the task of mixed evidence for both explicit and implicit AVE in e-commerce products, making the evaluation of implicit performance unreliable (Wang et al., 2022b; Liu et al., 2023; Wang et al., 2023).

As a result, research has shifted toward implicit AVE. Current studies have employed multimodal generative frameworks (Zhang et al., 2023) and efficient multimodal LLMs (Zou et al., 2024b) to improve inference accuracy in implicit AVE tasks. Despite these advances, implicit AVE remains in its early stages, generally achieving only moderate accuracy and revealing significant potential for improvement. To address this gap, we propose the MADIAVE framework, which enhances inference performance by leveraging multi-agent and multi-round debate and reasoning (Wu et al., 2025c). Our approach integrates insights from multiple agents to better capture latent attributes in multimodal data, thereby advancing the state of the art in implicit AVE.

2.2 MLLM, LLM and Debate

Recent research in LLMs and MLLMs has demonstrated impressive inference capabilities in tasks such as visual question answering (Hu et al., 2024) and cross-modal retrieval (Wang et al., 2025). Despite significant progress in processing multimodal
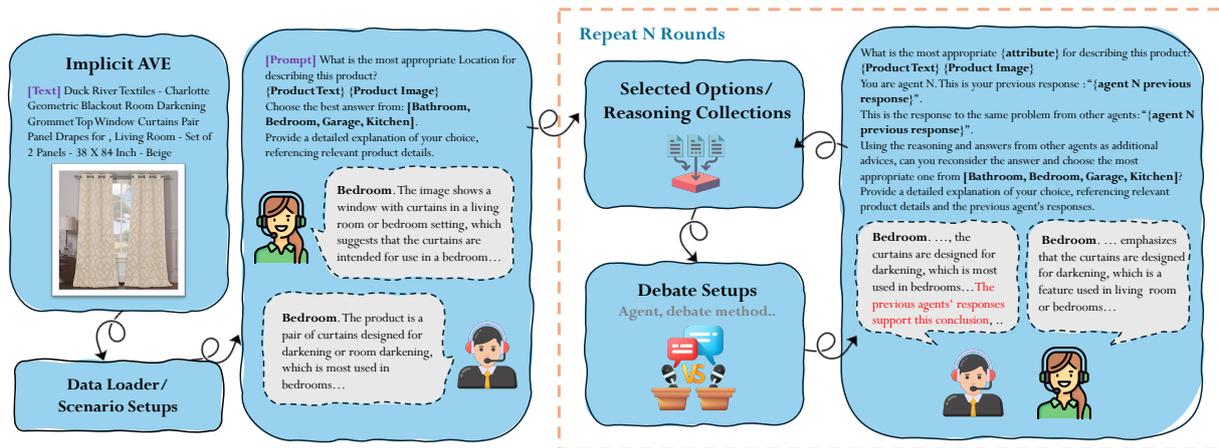
Figure 2: **Overview of the MADIAVE framework**. The ImplicitAVE is ingested by a data loader that applies scenario steps to initialize agent roles and trigger multi-round communication and debate. After each round, the message transcript is recorded and appended to the context, serving as additional input for the next round.

data, achieving high inference accuracy remains challenging due to the complexity of multidimensional data. For example, although previous work on implicit AVE has utilized LLMs and MLLMs for inference,with models like EIVEN (Zou et al., 2024b) reducing reliance on labeled data, they can still produce inaccurate results under certain conditions. Moreover, while LLMs have shown strong performance with minimal in-context examples (Blume et al., 2023), these approaches are often limited by expensive training processes and only moderate overall performance.

As a consequence, relying on a single LLM may not yield optimal results and can introduce bias. Researchers have explored methods to integrate the intelligence of multiple agents on tasks such as presidential debates (Liu et al., 2024), debate simulations (Taubenfeld et al., 2024), stance and argument analysis (Heinisch et al., 2024), and approaches to improve factuality and reasoning (Du et al., 2023). Recent investigations reveal that a multi-agent debate strategy can further enhance performance by enabling models to converge on more accurate answers in open-ended scenarios (Chan et al., 2023; Khan et al., 2024). However, these studies have focused exclusively on textual input, neglecting the visual component and the integration of visual cues with textual information during inference. In addition, most research has examined debates among identical LLMs under limited input conditions, leaving the potential for collaboration between different MLLMs largely unexplored.

To address these gaps, we propose the MADI-AVE framework as a novel solution for the implicit

AVE task. Our approach leverages multi-round, multi-agent debate among MLLMs to integrate both textual and visual reasoning, thereby improving inference accuracy for implicit AVE tasks.

3 MADIAVE Framework

We present our MADIAVE framework for the implicit AVE task in Section 3.1. This section details the overall inference and debate process, explaining how multiple agents interact to refine their inferences. In Section 3.2, we examine the behavior and convergence of the debate over multiple rounds and how the agents collectively reach consensus. The detailed MADIAVE framework, experimental setup, and workflow are illustrated in Figure 2. The prompts for the initial round and subsequent debate rounds are provided in Table 1.

3.1 Multi-agent Debate

In our context, an "agent" refers to an MLLM that functions much like a human, with the capacity to update its answers dynamically. For instance, when addressing an implicit AVE task such as determining a product's shape from its image and text, an MLLM might infer possible answers like "heart," "oval," or "rectangle." Although this attribute is not explicitly stated in the text, it is derived through both visual and textual cues. While a single MLLM might generate a tentative inference, the reliability of the answer is enhanced when multiple agents interact and converge on the same conclusion. Examining each agent's reasoning and explanations provides insight into diverse perspectives and reveals how inferences are continuously

Table 1: Prompt configurations of the initial and debate rounds for MLLMs .

Condition Setting	Prompt Text
Initial — 1st Round	*What is the most appropriate {attribute} for describing this product?*
	Context: Category: {category}, Title: {title}. Choose the best answer from: {options}.
Debate — 2nd to 5th Round	*What is the most appropriate {attribute} for describing this product?*
	Context: Category: {category}, Title: {title}.
	You are agent {agent_num}. Your previous response: "{previous_response_1}".
	Response from other agents: "{previous_response_2}".
	Using other agents' reasoning as advice, reconsider and choose the best option from: {options}.

refined. Moreover, our MADIAVE framework adopts a zero-shot setting to circumvent extensive training requirements and reliance on a training set, allowing us to evaluate whether MLLMs have sufficient general understanding and reasoning abilities regarding the product. In this multi-agent debate procedure, each MLLM emulates human-like reasoning and revises its answer based on the other agent's insights, ultimately leading to a more robust and reliable solution for the implicit AVE task.

To take advantage of this reasoning process and improve the inference capabilities of MLLMs, the MADIAVE framework first sets up two to five agents to independently analyze the implicit AVE dataset (Zou et al., 2024a) by addressing questions such as "What is the length of this product?". Each agent selects an answer from a predefined option list and provides a justification for its choice, indicating whether its decision is based on visual evidence, textual cues, or both. This corresponds to the first inference block in Figure 2.

After the initial round of inference, the agents proceed to subsequent rounds where they receive not only the original product image and text but also the responses from both agents, including each answer and its accompanying reasoning. For instance, in rounds 2 through 5, agent 1 will receive the image, text, its own previous response, and agent 2's previous response. This iterative process enables each agent to incorporate the reasoning, consensus, or disagreements from the previous round as additional context, ultimately leading to a more reliable and refined answer. This process is shown in the selected option and reasoning collection and the debate block in Figure 2.

3.2 Consensus between Agents in Debate

During debates among agents, several outcomes are possible. When an agent's answer is challenged, its confidence may decrease as it is exposed to alternative perspectives, prompting it to reconsider its evaluation of the product. This interaction may lead an agent to modify its answer, reach consensus with other agents, or even attempt to persuade other agent that its original response is correct. Over multiple rounds of debate, we expect the agents to gradually converge toward a shared consensus; however, achieving consensus does not necessarily guarantee that the final answer is correct for the implicit AVE task. We show empirical findings of the consensus during the debate in Appendix A

4 Experiment

In this section, we describe the experimental setup. Details about the benchmark are provided in Section 4.1. The debate scripts are provided in Section D. The debate constructions and scenarios are described in Section 4.2 and Section 4.3.

4.1 ImplicitAVE Benchmark

The quality of an implicit AVE dataset is essential for achieving accurate inference and reasoning. However, we have identified several issues with current implicit AVE datasets: *(1) implicit AVE research is still in its early stages, so few datasets are available*; *(2) many datasets are not publicly available*; *(3) some datasets conflate explicit and implicit AVE* (Yang et al., 2024; Wang et al., 2022b). To minimize the impact of noise, we rely on a high-quality benchmark ImplicitAVE dataset (Zou et al., 2024a), derived from the MAVE dataset (Yang et al., 2022). Its evaluation set encompasses five domains—Clothing, Footwear, Jewelry&GA, Home, and Food—for a total of 1,610 samples. Each domain includes five to eight attributes and spans 23 to 45 values, providing high-quality images and accompanying text. This evenly distributed evaluation set is ideally suited for the multi-agent debate process and serves as an excellent testbed for the MADIAVE framework. Further details, including the number of attributes and attribute values, can be found in Table 3.

Table 2: Overview of scenario configurations. This table details the agent setups for each scenario.

Scenario	Agent1	Agent2	Scenario Detail
1	GPT-4o	GPT-4o	Same LLM (Closed-Source Model, Zero-Shot) Debate
2	Llama-3.2-11B-Vision-Instruct	Llama-3.2-11B-Vision-Instruct	Same LLM (Open-Source Model, Zero-Shot) Debate
3	Phi-3.5-Vision-Instruct	Phi-3.5-Vision-Instruct	Same LLM (Open-Source Model, Zero-Shot) Debate
4	Llama-3.2-11B-Vision-Instruct	GPT-4o	Different LLM (Closed, Open-Source, Zero-Shot) Debate
5	Llama-3.2-11B-Vision-Instruct	Phi-3.5-Vision-Instruct	Different LLM (Open-Source Model, Zero-Shot) Debate
6	Qwen-2.5-VL-7B	Qwen-2.5-VL-7B	Same LLM (Open-Source Model, Zero-Shot) Debate
7	GPT-o1	GPT-o1	Same LLM (Closed-Source Model, Zero-Shot) Debate
8	Claude-3.5-Sonnet	Claude-3.5-Sonnet	Same LLM (Closed-Source Model, Zero-Shot) Debate

4.2 MLLM agents for Debate

Our MADIAVE framework adopts a *zero-shot* setting throughout the paper. This design choice is deliberate. While few-shot prompting and parameter-efficient fine-tuning methods (e.g., LoRA) can often improve task accuracy, they typically require task- and domain-specific examples that may not reflect the full diversity of real-world e-commerce profiles, queries, and product catalogs. In practice, collecting and curating high-quality labeled demonstrations at scale is expensive, and training on only thousands of examples, compared to the billions of interactions and heterogeneous distributions encountered in real deployments, can introduce undesirable bias and lead to overfitting to a narrow slice of user behaviors. This risk is particularly serious in e-commerce, where long-tail preferences, rapidly changing inventories, and shifting seasonal trends can cause distribution shift and reduce robustness under realistic conditions. To prioritize broad generalizability and minimize additional supervision, we therefore evaluate debate agents purely via zero-shot inference.

Concretely, we instantiate debate agents with six state-of-the-art MLLMs that span both closed-source and open-source models: GPT-4o, Llama-3.2-11B-Vision-Instruct, Phi-3.5-Vision-Instruct, Qwen-2.5-VL-7B, Claude-3.5-Sonnet, and GPT-o1. This selection allows us to assess whether the debate dynamics induced by MADIAVE are consistent across model families and capabilities, rather than being an artifact of a single backbone or a tuned prompt.

Few-shot prompting and instruction-tuning. We emphasize that few-shot prompting and instruction-tuned backbones are *promising* and can further boost absolute performance. However, such gains are conditioned on the availability and representativeness of curated examples: demonstrations that work well for a subset of products, user personas (Wu et al., 2025a), or browsing patterns may not transfer to unseen catalogs or novel preference signals. Because the central goal of this work is to evaluate MADIAVE as a *general* multi-agent debate framework, one that can be applied to new domains and catalogs with minimal additional supervision, we focus on the zero-shot settings as the most conservative and widely applicable evaluation setting.

4.3 Debate Scenarios and Experiments

We thoroughly investigate a variety of debate scenarios related to multimodal data and AVE tasks. By varying the debate agents, we capture a broad range of possibilities. Specifically, we define eight distinct scenarios, as detailed in Table 2. To quantify the benefit of debate, we include a control experiment of majority vote and an ablation that measures performance sensitivity to the number of agents and the number of debate rounds.

5 Debate Results

5.1 Debate Result and Analysis

5.1.1 Domain-Level Result

Table 4 reports domain-level and overall accuracy for three settings: prior baselines, single-model inference, and the MADIAVE self-debate framework. Relative to their single-model counterparts, all listed models improve under MADIAVE. Among single model runs, GPT-o1 has the highest

Table 3: Domain level data statistics for the evaluation set of the ImplicitAVE dataset.

Domain	# Data	# Attribute	# Value	Attributes
Clothing	226	4	23	Neckline, Sleeve Style, Length, Shoulder Style
Footwear	317	5	29	Athletic Shoe Style, Boot Style, Shaft Height, Heel Height, Toe Style
Jewelry&GA	220	3	20	Pattern, Material, Shape
Home	457	8	45	Material, Special Occasion, Location, Animal Theme, Season, Shape, Size, Attachment Method
Food	390	5	41	Flavor, Form, Candy Variety, Occasion, Container

overall score at 86.27, closely followed by GPT-4o. In terms of by domain, GPT-4o leads *Clothing* and *Jewelry&GA*, while GPT-o1 leads *Footwear* and *Food* and achieves the best *Overall* accuracy. The only exception is *Home*, where the earlier GPT-4V baseline remains the top score at 89.93. Overall, the table shows that self-debate reliably boosts performance across models, with GPT-4o and GPT-o1 providing the strongest domain and aggregate results under MADIAVE. To test the stability of the performance, we run the experiment three times and the result can be found in the Appendix E.

We present domain-level debate results in Figure 3, highlighting performance differences among the MLLMs in various scenarios. In the same MLLM debate in Figures 3(a), 3(b), 3(c), and 3(f), we observe that the F1 score increases during the first one or two rounds. In particular, for GPT-4o, the debate results show a general improvement in F1 score from 0.2% to 3% across all domains. However, after round two, the accuracy decreases or fluctuates within a limited range, indicating that although the debate converges, additional rounds do not yield further improvements. However, the inference performance after two to five rounds remains superior to that of a single agent.

Similarly, we observe an initial increase in inference performance after one round of debate for Llama-3.2-11B-Vision-Instruct, with improvements ranging from 0.12% to 2.55%. However, this gain is accompanied by some noise and uncertainty, as evidenced by a 0.75% decrease in the Home domain. Overall, Llama-3.2-11B-Vision-Instruct's improvement is more moderate and stable compared to GPT-4o, and as the debate continues, its performance fluctuates within a limited range, mirroring GPT-4o's behavior. On the other hand, for Phi-3.5-Vision-Instruct and Qwen-2.5-VL-7B, F1 improves after the first debate round, yet further rounds bring no additional gains and instead cause performance to fluctuate within a narrow and fixed range.

These findings consistently indicate that the MADIAVE framework improves inference performance relative to single model approaches, with one or two rounds of debate yielding notable improvement, whereas additional rounds do not provide further benefits. In many instances, agents converge on incorrect inferences, possibly because MLLMs tend to prioritize an opposing agent's advice over their own initial judgment,
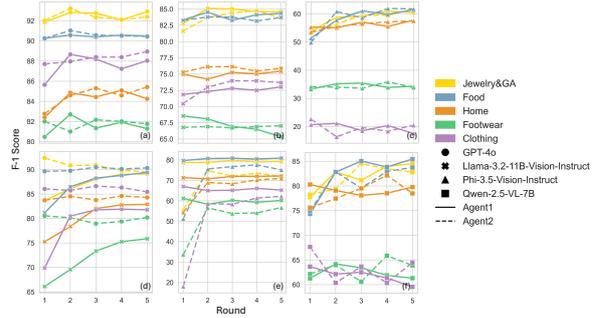

Figure 3: Domain-level F1 trajectories during debate, reported per agent across rounds. (a) Scenario 1: GPT-4o debating with GPT-4o; (b) Scenario 2: Llama-3.2-11B-Vision-Instruct debating with Llama-3.2-11B-Vision-Instruct; (c) Scenario 3: Phi-3.5-Vision-Instruct debating with Phi-3.5-Vision-Instruct; (d) Scenario 4: Llama-3.2-11B-Vision-Instruct debating with GPT-4o; (e) Scenario 5: Llama-3.2-11B-Vision-Instruct debating with Phi-3.5-Vision-Instruct(f) Scenario 6: Qwen-2.5-VL-7B debating with Qwen-2.5-VL-7B.

thus creating confusion. Moreover, the results show that GPT-4o outperforms the other models with F1-scores ranging from 80% to 93%, peaking at 93.23% in the Jewelry & GA domain at round two. Llama-3.2-11B-Vision-Instruct achieves moderate performance, ranging from 65% to 85%, whereas Phi-3.5-Vision-Instruct exhibits the lowest performance. GPT-4o demonstrates a more stable and pronounced improvement compared to Llama-3.2-11B-Vision-Instruct and Qwen-2.5-VL-7B, mirroring real-life scenarios in which knowledgeable participants lead the discussion toward better consensus while less-informed individuals may inadvertently lead to suboptimal convergence or wrong directions.

Additionally, a domain-specific analysis reveals that the Jewelry & GA domain generally exhibits superior inference performance compared to other domains. In contrast, the Clothing and Footwear domains consistently demonstrate lower performance. This observation suggests that either the MLLM agents have a training bias favoring certain domains, or that the Clothing and Footwear domain is inherently more challenging for all MLLM agents—potentially even for human reviewers. We analyze attribute-level inference performance and present the result in Appendix F.

5.2 Convergence Statistic

Detailed convergence information is provided in Figures 4. In the same MLLM debates involv-

3040

Method	Language Model	Clothing	Footwear	Jewelry&GA	Food	Home	Overall
Previous Baseline (Zou et al., 2024a)							
BLIP-2	FlanT5XXL-11B	55.31	55.21	82.72	71.02	71.33	67.39
InstructBLIP	FlanT5XXL-11B	62.83	63.41	83.18	73.58	73.96	71.49
LLaVA-1.5	Vicuna-13B	49.12	63.72	81.81	76.15	80.31	71.86
Qwen-VL	Qwen-7B	68.14	62.14	84.09	76.92	73.96	70.86
GPT	GPT-4V	77.43	81.39	90.45	90.77	**89.93**	86.77
Single Model							
GPT	GPT-4o	85.66	80.48	91.87	90.26	82.41	85.68
Llama-3.2	Llama-3.2-11B-Vision-Instruct	70.44	66.8	81.62	83.28	74.98	75.64
Phi-3.5	Phi-3.5-Vision-Instruct	20.67	33.27	53.24	49.86	53.6	44.02
Qwen2.5	Qwen-2.5-VL-7B	64.16	61.83	77.72	74.87	75.71	71.43
Claude	Claude-3.5-Sonnet	71.68	69.40	79.55	82.05	76.57	76.21
GPT	GPT-o1	84.96	81.08	91.36	90.26	83.81	86.27
MADIAVE Framework (Multi-Agent Debate)							
GPT	GPT-4o	**88.95**	82.71	**93.23**	91.04	85.78	87.91
Llama-3.2	Llama-3.2-11B-Vision-Instruct	73.96	68.09	85.1	84.52	76.14	77.50
Phi-3.5	Phi-3.5-Vision-Instruct	22.67	35.87	60.99	61.93	57.79	49.98
Qwen2.5	Qwen-2.5-VL-7B	61.95	65.93	85.00	85.13	81.84	77.14
Claude	Claude-3.5-Sonnet	76.10	66.88	80.45	84.61	78.77	77.70
GPT	GPT-o1	88.50	**83.28**	92.27	**91.80**	86.88	**88.32**

Table 4: The MADIAVE framework is compared to previous baseline methods and single model inference on the ImplicitAVE dataset. The aggregated system performance is evaluated across five individual domains as well as overall accuracy. The **bold green** entries indicate the overall best performance among the methods.

ing GPT-4o, Llama-3.2-11B-Vision-Instruct, Phi-3.5-Vision-Instruct, and the Qwen faimily Qwen-2.5-VL-7B, as shown in Figures 4(a), 4(b), 4(c), 4(f), respectively, we observe that the agents exhibit the most changes during round two, the first round of debate, with the frequency of changes decreasing in subsequent rounds. For example, in Figure 4(a) for GPT-4o, the total number of changes drops from 70 to 20, indicating that the debate steadily converges until both agents reach consensus on the inference. Additionally, the number of cases where an agent changes its answer yet still arrives at an incorrect inference decreases over time, suggesting that agents increasingly tend to follow other agent's selection rather than its own options. However, although there is an initial improvement in correct inferences during round two, the number of worsened inferences begins to exceed the number of improved ones by round three, implying that the convergence process can sometimes deviate in the wrong direction. In such cases, agents appear to become confused by flawed reasoning, failing to properly assess whether to adopt correct or incorrect reasoning. A similar pattern is observed in Figures 4(b) and 4(c). It is also noteworthy that weaker agents, due to their limited ability to evaluate reasoning, tend to change their answers more frequently. For instance, when Phi-3.5-Vision-Instruct debates with itself, both agents change their answers in 600 out of a total of 1,610 instances. This number remains

around 300 even after four rounds of debate.

For different MLLM debates, we observe that the convergence patterns differ from those seen in the same agent debates. In Figures 4(d) and 4(e), the stronger agent, acting as a "teacher", often ends up with worsened answers, likely due to the influence of inaccurate reasoning from the weaker "student" agent. In contrast, the student agent consistently corrects more inferences in each round, indicating that it benefits significantly from the teacher's input during the debate. In Figure 4(e), although the weaker learner, Phi-3.5-Vision-Instruct, does not exhibit strong reasoning or a clear understanding of the product image and text, it changes its answers approximately 800 times in round 2 and corrects about 500 data points, by the guidance and support of Llama-3.2-11B-Vision-Instruct model. Conversely, the teacher agent shows more deteriorated answers across rounds, suggesting it benefits less from the exchange and can be pulled off course by the student's confident but weakly grounded arguments. This indicates a reverse-influence effect: prompts that encourage incorporating peer reasoning may cause the teacher to over-weight unreliable feedback and drift from correct choices. Overall, while the teacher's guidance is invaluable for improving the weaker agent, the teacher agent may pay an additional performance cost by partially adopting or accepting the student agent's inaccurate or imprecise reasoning, especially in cases or

Table 5: The performance comparison on GPT-4o: single inference, majority vote (10 independent runs), and two-agent debate (2 agents, 5 rounds).

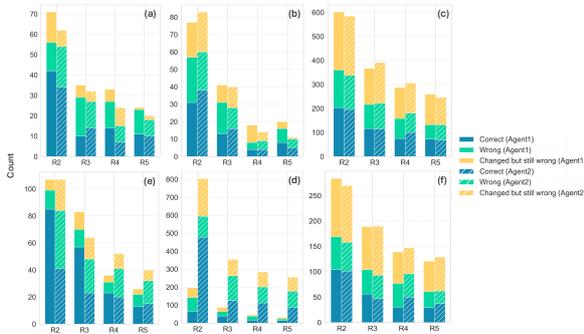| Method (GPT-4o)	Clothing	Footwear	Jewelry & GA	Food	Home	Overall
Single Inference	85.66	80.48	91.87	90.26	82.41	85.68
Majority Vote	87.17	**82.88**	92.72	90.26	83.15	86.69
Debate (2 agents, 5 rounds)	**88.95**	82.71	**93.23**	**91.04**	**85.78**	**87.91**

Figure 4: Debate statistics for (a) Scenario 1: GPT-4o debating with GPT-4o; (b) Scenario 2: Llama-3.2-11B-Vision-Instruct debating with Llama-3.2-11B-Vision-Instruct; (c) Scenario 3: Phi-3.5-Vision-Instruct debating with Phi-3.5-Vision-Instruct; (d) Scenario 4: Llama-3.2-11B-Vision-Instruct debating with GPT-4o; (e) Scenario 5: Llama-3.2-11B-Vision-Instruct debating with Phi-3.5-Vision-Instruct; (f) Scenario 6: Qwen-2.5-VL-7B debating with Qwen-2.5-VL-7B.

environment that are not well-defined.

5.3 The Performance Comparison Between Single Inference, Majority Vote, and Dsebate

To differentiate the advantages of debate from additional sampling, we aligned the inference budget by executing a *single* GPT-4o agent ten times (equivalent to two agents across five rounds, $2 \times 5 = 10$) and compared *Single Inference*, *Majority Vote*, and *Debate*. Table 5 demonstrates that debate achieves the highest *Overall* accuracy at 87.91%, surpassing majority vote at 86.69% and single inference at 85.68%. Domain-specific improvements are uniform, with the most significant increase observed in *Home* (85.78% compared to 83.15% majority and 82.41% single). Debate enables agents to examine heuristic assumptions and integrate cross-modal evidence (e.g., correlating packing dimensions with visual indicators), while majority vote aggregation of independent executions lacks a framework and procedure for reasoning exchange, resulting in the continuation of initial errors.

5.4 Ablations on Agents and Rounds

To study how MADIAVE trades-off *interaction depth* (rounds) versus *agent breadth* (number of agents), we follow a standard inference-time budget control used in multi-agent LLM systems: we hold the *number of model calls* fixed and vary only how those calls are orchestrated. Concretely, we fix the budget to 10 total calls and compare settings such as two agents over five rounds versus five agents over two rounds. This call-level control isolates the effect of debate structure, whereas token-level matching is difficult to enforce in practice and can be misleading due to variable context growth and heterogeneous reasoning behavior across models. Table 6 shows that with two agents, **2–3 rounds** yield the strongest performance (Overall 87.42–87.39%), while extending to 4–5 rounds provides no additional benefit (86.90–86.83%). Under the same 10-call budget with two rounds, increasing breadth from 2 to 3–5 agents *reduces* accuracy (86.77%, 85.22%, 86.15%). Overall, these results suggest that a small team with brief exchanges best leverages debate: early rounds help reconcile cross-modal evidence, but additional rounds or larger committees introduce redundancy and noise, leading to diminishing returns and occasional drift.

Combined with the call-matched comparison in Section 5.3, these findings indicate that MADIAVE's gains persist when controlling for the number of communications. MADIAVE outperforms both single inference and majority vote under the same call budget, particularly on harder implicit cues (e.g., *Home*). We note that debate typically incurs additional context tokens as transcripts accumulate; this is an inherent trade-off of richer orchestration rather than a confound we can reliably equalize. Practitioners should therefore view MADIAVE as *call-comparable* to majority vote and decide whether the extra context cost is justified under their latency and token constraints. As a practical default, we recommend **two agents with two rounds** (optionally three), and scaling up only for difficult slices after cost–latency considerations.

Table 6: Agents × Rounds ablation under a fixed compute budget of 10 model calls (interaction depth vs. agent breadth)

Rounds	Agents	Clothing	Footwear	Jewelry & GA	Food	Home	Overall
5	2	88.95	81.39	92.27	90.26	85.34	86.83
4	2	87.61	81.70	91.82	90.26	84.92	86.90
3	2	88.05	82.02	92.27	90.52	**85.78**	87.39
2	2	88.50	**82.71**	**93.23**	**91.04**	84.25	**87.42**
2	3	**89.82**	81.39	90.91	89.74	84.46	86.77
2	4	87.17	80.44	90.91	87.69	82.71	85.22
2	5	86.28	82.02	91.82	88.98	83.81	86.15

Table 7: One round debates (R2) and single round (R1) baseline across six agents. Accuracy gains come with higher latency but modest token growth. Efficiency (Gain/Δt) highlights improvement per added second.

Model	R1 Lat. (s)	R1 Acc. (%)	R2 Lat. (s)	R2 Acc. (%)	ΔAcc. (%)	Lat. ×	ΔAcc/Δt (% s^{-1})	Init Tok	R2 Tok	Tok ×
GPT-4o	3.72	85.68	7.95	87.42	+1.74	2.14	0.41	80.31	99.60	1.24
Llama-3.2-11B-Vision-Instruct	4.32	75.64	8.96	77.02	+1.38	2.07	0.30	79.32	97.22	1.23
Phi-3.5-Vision-Instruct	3.29	44.02	6.84	48.63	+4.61	2.08	1.30	85.17	95.33	1.12
Qwen-2.5-VL-7B	2.32	71.43	6.08	75.90	+4.47	2.62	1.19	76.55	94.32	1.23
Claude-3.5-sonnet	7.16	76.21	12.77	77.20	+0.99	1.78	0.18	133.25	155.96	1.17
GPT-o1	8.73	86.27	19.21	87.52	+1.25	2.20	0.12	157.41	189.25	1.20

5.5 Latency Analysis and Accuracy-Cost Trade-offs of Debate

In a comparison of six agents, the most informative metric is efficiency, defined as gain per additional second $\Delta\text{Acc}/\Delta t$. Two rounds of debate reveal that weak models has a higher gain, with Phi-3.5-Vision-Instruct at $1.30\%\,\text{s}^{-1}$ and Qwen-2.5-VL-7B at $1.19\%\,\text{s}^{-1}$; more robust models exhibit just moderate returns, such as GPT-4o at $0.41\%\,\text{s}^{-1}$ and GPT-o1 at $0.12\%\,\text{s}^{-1}$. Absolute accuracy improvements range from $+0.99$ to $+4.61\%$, although latency increases by a factor of 1.78 to 2.62 compared to a single round. Therefore, the overall quality–time trade-off favors more on weak models under time constraints. Improvements at the category level focus on the significance of implicit cues and cross-modal anchoring, such as interpreting packaging measures in text and validating them with a visual scale. The efficiency profile indicates that one should select debate based on $\Delta\text{Acc}/\Delta t$ within a specified latency budget, rather than solely on absolute gain, as minor absolute improvements in robust models can be critical, while more substantial and time-efficient gains are characteristic of the weak models.

We recommend a selective implementation approach with precise budget management for deployment. Use a single-round debate by default and initiate a second-round debate only if necessary. Implement an adaptive stopping criterion that ceases when the last iteration enhances accuracy by less than a user-defined threshold in relation to the latency budget. In summary, Table 7 endorses a pragmatic operating point: a one-round debate achieves an optimal equilibrium between quality and cost for weak models, while simultaneously addressing complex cases for more robust models. Additionally, meticulous triggering and early cessation retain the majority of the advantages with minimal time investment.

6 Conclusion

In this study, we introduced MADIAVE, a multi-agent debate framework that enhances implicit AVE performance in multimodal e-commerce by integrating textual and visual reasoning across multiple MLLMs. Through a series of debate rounds, we discovered that agents collectively refine their responses in just one or two rounds of discussion, significantly improving inference compared to single MLLM approaches and previous baselines. Specifically, we observed that low-accuracy attributes benefit the most; however, excessive rounds can lead to confusion, underscoring the need to balance the number of debate iterations. Additionally, our experiments show that weaker agents substantially improve through interaction with stronger agents, often reaching performance comparable to that of the stronger agents. These findings highlight the potential of multi-agent debate strategies to overcome single MLLM limitations and provide scalable solutions for implicit AVE and related multimodal tasks. In the future, we plan to extend this approach to other e-commerce domains.

Acknowledgments

7 Limitations

Our multi-agent debate framework, MADIAVE, demonstrates a strong capability for improving performance on the implicit AVE task compared to previous baseline methods and single MLLM approaches. However, our work has limitations. We compared our approach with only a limited number of baselines, as implicit AVE is a relatively new task and current research is still in its early stages. Additionally, many related works are not open-sourced or are unavailable. Therefore, we hope that MADIAVE can serve as a milestone for multimodal implicit AVE, providing a clear direction for future improvements.

8 Ethics Statement

We use only publicly released benchmarks and do not collect or release any user data or personally identifiable information. Since wrong attributes may mislead users and affect search and recommendation, debate outputs should be treated as probabilistic predictions and validated (e.g., confidence thresholds or human review) in high-impact settings. Multi-agent debate also increases compute usage; practitioners should consider early stopping and selective triggering to reduce unnecessary cost.

References

Ansel Blume, Nasser Zalmout, Heng Ji, and Xian Li. 2023. Generative models for product attribute extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 575–585, Singapore. Association for Computational Linguistics.

Alexander Brinkmann, Roee Shraga, and Christian Bizer. 2024. Extractgpt: Exploring the potential of large language models for product attribute value extraction. *Preprint*, arXiv:2310.12537.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multi-agent debate. *arXiv preprint arXiv:2305.14325*.

Jiaying Gong, Ming Cheng, Hongda Shen, Pierre-Yves Vandenbussche, Janet Jenq, and Hoda Eldardiry. 2025. Visual zero-shot e-commerce product attribute value extraction. *arXiv preprint arXiv:2502.15979*.

Jiaying Gong and Hoda Eldardiry. 2024. Multi-label zero-shot product attribute-value extraction. In *Proceedings of the ACM Web Conference 2024*, pages 2259–2270.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. 2024. LLM factoscope: Uncovering LLMs' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10218–10230, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Heinisch, Lorik Dumani, Philipp Cimiano, and Ralf Schenkel. 2024. "tell me who you are and I tell you how you argue": Predicting stances and arguments for stakeholder groups. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1968–1982, Mexico City, Mexico. Association for Computational Linguistics.

Jiazhen Hu, Jiaying Gong, Hongda Shen, and Hoda Eldardiry. 2025. Hypergraph-based zero-shot multimodal product attribute value extraction. In *Proceedings of the ACM on Web Conference 2025*, pages 4853–4862.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264.

Wei-Chieh Huang, Henry Peng Zou, Yaozu Wu, Dongyuan Li, Yankai Chen, Weizhi Zhang, Yangning Li, Angelo Zangari, Jizhou Guo, Chunyu Miao, et al. 2025. Deepresearchguard: Deep research with open-domain evaluation and multi-stage guardrails for safety. *arXiv preprint arXiv:2510.10994*.

Meihuizi Jia, Lei Shen, Luu Anh Tuan, Meng Chen, Jing Xu, Lejian Liao, Shaozu Yuan, and Xiaodong He. 2024. Mujo-sf: Multimodal joint slot filling for attribute value prediction of e-commerce commodities. *IEEE Transactions on Multimedia*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktaschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *Preprint*, arXiv:2402.06782.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. 2025a. Towards agentic RAG with deep reasoning: A survey of rag-reasoning systems in llms. *CoRR*, abs/2507.09477.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, et al. 2025b. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms. *arXiv preprint arXiv:2507.09477*.

Rongmei Lin, Xiang He, Jie Feng, Nasser Zalmout, Yan Liang, Li Xiong, and Xin Luna Dong. 2021. Pam: Understanding product images in cross product category attribute extraction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 3262–3270, New York, NY, USA. Association for Computing Machinery.

Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. 2023. Multimodal pre-training with self-distillation for product understanding in e-commerce. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1039–1047.

Zhengliang Liu, Yiwei Li, Oleksandra Zolotarevych, Rongwei Yang, and Tianming Liu. 2024. Llm-potus score: A framework of analyzing presidential debates with large language models. *Preprint*, arXiv:2409.08147.

Robyn Loughnane, Jiaxin Liu, Zhilin Chen, Zhiqi Wang, Joseph Giroux, Tianchuan Du, Benjamin Schroeder, and Weiyi Sun. 2024. Explicit attribute extraction in e-commerce search. In *Proceedings of the Seventh Workshop on e-Commerce and NLP @ LREC-COLING 2024*, pages 125–135, Torino, Italia. ELRA and ICCL.

Chen Luo, Rahul Goutam, Haiyang Zhang, Chao Zhang, Yangqiu Song, and Bing Yin. 2023. Implicit query parsing at amazon product search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3380–3384, New York, NY, USA. Association for Computing Machinery.

Chen Luo, William Headden, Neela Avudaiappan, Haoming Jiang, Tianyu Cao, Qingyu Yin, Yifan Gao, Zheng Li, Rahul Goutam, Haiyang Zhang, and Bing Yin. 2022. Query attribute recommendation at amazon search. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 506–508, New York, NY, USA. Association for Computing Machinery.

Aisha Mohdhar and Khaled Shaalan. 2021. The future of e-commerce systems: 2030 and beyond. *Recent Advances in Technology Acceptance Models and Theories*, pages 311–330.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kassem Sabeh, Mouna Kacimi, Johann Gamper, Robert Litschko, and Barbara Plank. 2024a. Exploring large language models for product attribute value identification. *arXiv preprint arXiv:2409.12695*.

Kassem Sabeh, Robert Litschko, Mouna Kacimi, Barbara Plank, and Johann Gamper. 2024b. An empirical comparison of generative approaches for product attribute-value identification. *Preprint*, arXiv:2407.01137.

Yindu Su, Huike Zou, Lin Sun, Ting Zhang, Haiyang Yang, Liyu Chen, David Lo, Qingheng Zhang, Shuguang Han, and Jufeng Chen. 2025. Taclr: A scalable and efficient retrieval-based method for industrial product attribute value identification. *arXiv preprint arXiv:2501.03835*.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 251–267. Association for Computational Linguistics.

Kai Wang, Jianzhi Shao, Tao Zhang, Qijin Chen, and Chengfu Huo. 2023. Mpkgac: Multimodal product attribute completion in e-commerce. In *Companion Proceedings of the ACM Web Conference 2023*, pages 336–340.

Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D. Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 47–55, New York, NY, USA. Association for Computing Machinery.

Qifan Wang, Li Yang, Jingang Wang, Jitin Krishnan, Bo Dai, Sinong Wang, Zenglin Xu, Madian Khabsa, and Hao Ma. 2022a. SMARTAVE: Structured multimodal transformer for product attribute value extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 263–276, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qifan Wang, Li Yang, Jingang Wang, Jitin Krishnan, Bo Dai, Sinong Wang, Zenglin Xu, Madian Khabsa, and Hao Ma. 2022b. Smartave: Structured multimodal transformer for product attribute value extraction. In *Findings of the association for computational linguistics: EMNLP 2022*, pages 263–276.

3045

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2025. Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE*, pages 1–39.

Yaozu Wu, Jizhou Guo, Dongyuan Li, Henry Peng Zou, Wei-Chieh Huang, Yankai Chen, Zhen Wang, Weizhi Zhang, Yangning Li, Meng Zhang, et al. 2025a. Psgagent: Personality-aware safety guardrail for llm-based agents. *arXiv preprint arXiv:2509.23614*.

Yaozu Wu, Dongyuan Li, Yankai Chen, Renhe Jiang, Henry Peng Zou, Liancheng Fang, Zhen Wang, and Philip S. Yu. 2025b. Multi-agent autonomous driving systems with large language models: A survey of recent advances. *CoRR*, abs/2502.16804.

Yaozu Wu, Dongyuan Li, Yankai Chen, Renhe Jiang, Henry Peng Zou, Wei-Chieh Huang, Yangning Li, Liancheng Fang, Zhen Wang, and Philip S Yu. 2025c. Multi-agent autonomous driving systems with large language models: A survey of recent advances, resources, and future directions. *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Liyan Xu, Zhenlin Su, Mo Yu, Jin Xu, Jinho D. Choi, Jie Zhou, and Fei Liu. 2024. Identifying factual inconsistencies in summaries: Grounding LLM inference via task taxonomy. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14626–14641, Miami, Florida, USA. Association for Computational Linguistics.

Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. 2021. Adatag: Multi-attribute value extraction from product profiles with adaptive decoding. *Preprint*, arXiv:2106.02318.

Li Yang, Qifan Wang, Jianfeng Chi, Jiahao Liu, Jingang Wang, Fuli Feng, Zenglin Xu, Yi Fang, Lifu Huang, and Dongfang Liu. 2024. Eave: Efficient product attribute value extraction via lightweight sparse-layer interaction. *arXiv preprint arXiv:2406.06839*.

Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabsa, Sinong Wang, Zenglin Xu, and Dongfang Liu. 2023. MixPAVE: Mix-prompt tuning for few-shot product attribute value extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9978–9991, Toronto, Canada. Association for Computational Linguistics.

Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. Mave: A product dataset for multi-source attribute value extraction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pages 1256–1265, New York, NY, USA. Association for Computing Machinery.

Tao Zhang, Chenwei Zhang, Xian Li, Jingbo Shang, Hoang Nguyen, and Philip S Yu. 2024. Stronger, lighter, better: Towards life-long attribute value extraction for e-commerce products. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8631–8643.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Yankai Chen, Chunkit Chan, Peilin Zhou, Xinyang Zhang, Chenwei Zhang, Jingbo Shang, Ming Zhang, Yangqiu Song, Irwin King, and Philip S. Yu. 2025a. From web search towards agentic deep research: Incentivizing search with reasoning agents. *CoRR*, abs/2506.18959.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, et al. 2025b. From web search towards agentic deep research: Incentivizing search with reasoning agents. *arXiv preprint arXiv:2506.18959*.

Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. 2023. Pay attention to implicit attribute values: A multi-modal generative framework for AVE task. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13139–13151, Toronto, Canada. Association for Computational Linguistics.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1049–1058, New York, NY, USA. Association for Computing Machinery.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, et al. 2025a. Llm-based human-agent collaboration and interaction systems: A survey. *arXiv preprint arXiv:2505.00753*.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Chunyu Miao, Dongyuan Li, Aiwei Liu, Yue Zhou, Yankai Chen, Weizhi Zhang, Yangning Li, Liancheng Fang, Renhe Jiang, and Philip S. Yu. 2025b. A call for collaborative intelligence: Why human-agent systems should precede AI autonomy. *CoRR*, abs/2506.09420.

Henry Peng Zou, Vinay Samuel, Yue Zhou, Weizhi Zhang, Liancheng Fang, Zihe Song, Philip S. Yu, and Cornelia Caragea. 2024a. Implicitave: An open-source dataset and multimodal llms benchmark for implicit attribute value extraction. *Preprint*, arXiv:2404.15592.

Henry Peng Zou, Gavin Heqing Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia, and Cornelia Caragea. 2024b. Eiven: Efficient implicit attribute

value extraction using multimodal llm. *Preprint*, arXiv:2404.08886.

<div align="center">

Appendix Contents

</div>

A Converge Examples

Our empirical findings indicate that multi-agent MLLM debates frequently reach consensus after just one or two rounds. We identified three distinct convergence outcomes, including correct convergence, incorrect convergence, and no convergence, as illustrated in Figure C.1. Correct convergence occurs when both agents consistently maintain correct predictions or when they correct an initially incorrect prediction through debate. Incorrect convergence arises when the agents persist in erroneous predictions or transition from a correct to an incorrect inference. No convergence occurs if one or both agents repeatedly change their answers without reaching a stable agreement. Although our goal is to achieve correct convergence, the process by which consensus is reached remains uncertain. In fact, convergence behavior varies depending on the debate configuration (e.g., whether the debate involves identical agents, different agents, or varied input conditions).

Our observations further reveal that agents tend to incorporate and accept feedback from their counterparts across different scenarios. We believe this behavior is influenced by the underlying instruction tuning and reinforcement learning mechanisms governing the agents' reasoning processes (Ouyang et al., 2022). In many cases, agents prioritize the prompt input over their own initial stance.

B Model Selections

Our experiments were completed before the public release of newer Qwen3-family models (e.g.,

Qwen3-VL and Qwen3-Omni). In addition, Qwen2.5-Omni and Qwen3-Omni are primarily optimized for real-time multimodal streaming (text, images, audio, video) and are generally comparable to same-sized Qwen2.5 or Qwen3 text or VL counterparts on standard benchmarks. As a sanity check, we additionally evaluated Qwen3-VL-7B on a subset of our tasks. Although the absolute performance improves slightly, the relative behavior across models and the qualitative trends discussed in the paper remain consistent. Therefore, replacing our backbone (Qwen-2.5-VL-7B) with Qwen2.5-Omni or Qwen3-VL or Omni would not materially change our main findings.

C Analysis and Potential Improvement

According to the MADIAVE study, the multi-agent debate framework generally leads the models to agree on an answer within one or two rounds. In many of those cases, the agreed answer is indeed correct. Empirically, this is reflected in the rapid rise in F1 scores for most domains, as both agents adopt each other's reasoning to correct initial mistakes. However, the study also documents instances of incorrect convergence, where both agents agree on the wrong attribute, because one agent's flawed reasoning can unduly influence its partner. By the third round, the number of cases that worsened begins to exceed the number that improved, indicating that further rounds can sometimes push the consensus in the wrong direction. Finally, a small fraction of debates never reach a stable answer (no convergence), with agents repeatedly changing their responses without reaching an agreement.

To improve both the accuracy and stability of convergence, we suggest several refinements. First, introduce a stricter validation step during debates so that an agent only changes its answer when it can explicitly verify its peer's justification, rather than simply deferring to it. Second, mitigate the teacher-student imbalance observed when a stronger model adopts a weaker model's error by weighting updates according to model reliability, allowing the stronger model to critique without being forced to revise unless there is clear, corroborated evidence. Third, apply an early stopping criterion once consensus is reached or the F1 score gains plateau, since most benefits occur in the first two rounds. Extra rounds can introduce confusion. Finally, further instruction that tunes the agents toward a more skeptical debate style, encouraging them to main-

tain their initial answers unless the opposing argument is highly convincing. Together, these measures would help ensure that when agents agree, they are far more likely to converge on the correct inference and less likely to oscillate or jointly adopt wrong answers.

C.1 Compatibility with Stronger Settings

Importantly, MADIAVE is *compatible* with stronger settings. One can directly plug in few-shot prompts, instruction-tuned backbones, or lightly fine-tuned LoRA variants as debate agents, and we expect the *absolute* accuracy to increase. Meanwhile, our main qualitative findings about multi-agent debate. For example, that most gains occur within the first 1–2 rounds and that excessive debating can yield diminishing returns or drift, are expected to remain stable, since these effects primarily arise from interaction structure and information exchange rather than from a specific parameterization of the backbone. A full, controlled comparison across zero-shot, few-shot, and instruction-tuned variants would require additional large-scale training and careful curation of domain-labeled data (including coverage of long-tail profiles and continuously updated catalogs), which is beyond the scope of this paper.

C.2 Weighted Aggregation of Debate Agents

In the main paper, we aggregate agents' outputs with *equal weighting*. This is a conscious design choice: it keeps MADIAVE simple, transparent, and broadly applicable across domains, without introducing extra assumptions about which agent should be trusted a priori. Nevertheless, *weighted* aggregation could potentially further improve performance by exploiting heterogeneity among agents and arguments.

A natural extension is to assign higher influence to "expert" agents, such as (i) stronger backbones, (ii) agents specialized in different evidence modalities (e.g., visual vs. textual grounding), or (iii) agents that consistently provide higher-quality arguments on specific task subsets. Another complementary direction is *confidence- or evidence-aware* weighting, where arguments supported by stronger evidence (e.g., explicit citations to relevant product attributes, clearer visual cues, or internally consistent reasoning) or higher calibrated confidence receive more weight in the final decision.

Systematically studying weighting strategies requires an additional learning and calibration layer—

including how to estimate confidence, how to validate "expertise" across tasks, and how to prevent over-reliance on spurious signals. This introduces new design choices and hyperparameters, and it requires careful validation to ensure robustness under distribution shifts (e.g., new product categories or evolving catalogs). Because our goal is to evaluate MADIAVE as a general debate framework with minimal supervision, we leave this exploration beyond the scope of the current paper.

We view at least three directions as particularly promising future work: (i) **expertise-aware weighting**, where per-agent weights are determined by backbone strength or specialization; (ii) **confidence- or evidence-aware weighting**, where weights depend on calibrated uncertainty or measurable evidence quality; and (iii) **learning debate weights**, where the weighting policy (or even the debate graph or structure) is treated as a reinforcement learning problem that optimizes downstream task performance subject to constraints such as cost or debate length. We will clarify these points in the revised version and explicitly state that equal weighting is a deliberate baseline for generality, while weighted aggregation offers a principled avenue for further improvements.

D Debate Scripts

In this appendix, we present concise multimodal debate traces and scripts between `Llama-3.2-11B-Vision-Instruct` (Agent 1) and `Phi-3.5-Vision-Instruct` (Agent 2). Each case begins with a brief summary, then shows an *Inference Round* followed by a *Debate Round*. The scripts illustrate how agents integrate textual cues (e.g., "Room Darkening," measurements) with visual evidence (scene context, shape/fit).

D.1 Case A — Curtain Placement

In Case A, the agents weigh product text (the "Room Darkening" cue, which typically signals bedroom use for blocking light) against scene context (a sofa and coffee table implying a living room). The purpose is to frame the cross-modal tension so the reader knows the lens for evaluation; after integrating both modalities, the debate converges on **Bedroom**.
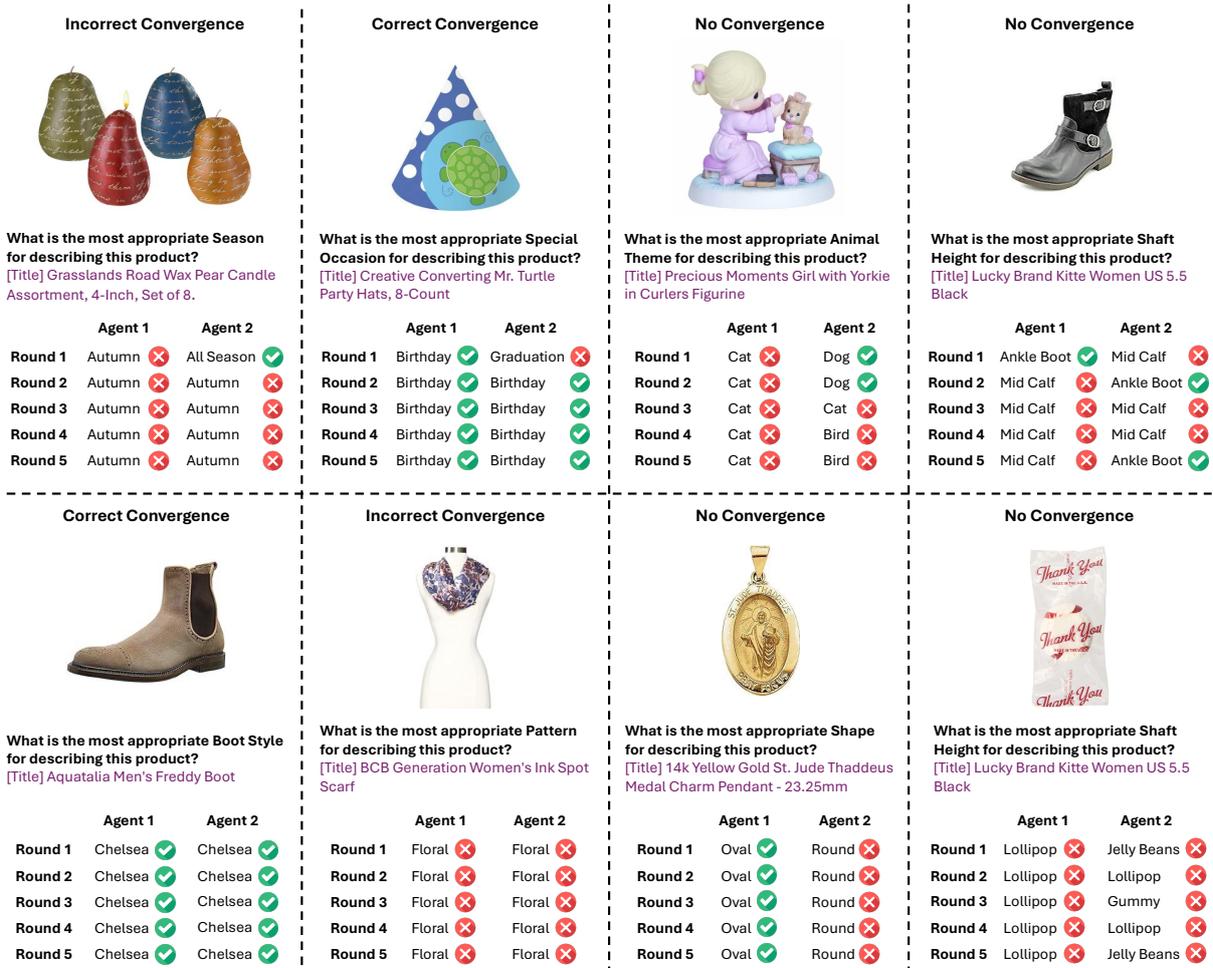
Incorrect Convergence

What is the most appropriate Season for describing this product?
[Title] Grasslands Road Wax Pear Candle Assortment, 4-Inch, Set of 8.

	Agent 1	Agent 2
Round 1	Autumn ✗	All Season ✓
Round 2	Autumn ✗	Autumn ✗
Round 3	Autumn ✗	Autumn ✗
Round 4	Autumn ✗	Autumn ✗
Round 5	Autumn ✗	Autumn ✗

Correct Convergence

What is the most appropriate Special Occasion for describing this product?
[Title] Creative Converting Mr. Turtle Party Hats, 8-Count

	Agent 1	Agent 2
Round 1	Birthday ✓	Graduation ✗
Round 2	Birthday ✓	Birthday ✓
Round 3	Birthday ✓	Birthday ✓
Round 4	Birthday ✓	Birthday ✓
Round 5	Birthday ✓	Birthday ✓

No Convergence

What is the most appropriate Animal Theme for describing this product?
[Title] Precious Moments Girl with Yorkie in Curlers Figurine

	Agent 1	Agent 2
Round 1	Cat ✗	Dog ✓
Round 2	Cat ✗	Dog ✓
Round 3	Cat ✗	Cat ✗
Round 4	Cat ✗	Bird ✗
Round 5	Cat ✗	Bird ✗

No Convergence

What is the most appropriate Shaft Height for describing this product?
[Title] Lucky Brand Kitte Women US 5.5 Black

	Agent 1	Agent 2
Round 1	Ankle Boot ✓	Mid Calf ✗
Round 2	Mid Calf ✗	Ankle Boot ✓
Round 3	Mid Calf ✗	Mid Calf ✗
Round 4	Mid Calf ✗	Mid Calf ✗
Round 5	Mid Calf ✗	Ankle Boot ✓

Correct Convergence

What is the most appropriate Boot Style for describing this product?
[Title] Aquatalia Men's Freddy Boot

	Agent 1	Agent 2
Round 1	Chelsea ✓	Chelsea ✓
Round 2	Chelsea ✓	Chelsea ✓
Round 3	Chelsea ✓	Chelsea ✓
Round 4	Chelsea ✓	Chelsea ✓
Round 5	Chelsea ✓	Chelsea ✓

Incorrect Convergence

What is the most appropriate Pattern for describing this product?
[Title] BCB Generation Women's Ink Spot Scarf

	Agent 1	Agent 2
Round 1	Floral ✗	Floral ✗
Round 2	Floral ✗	Floral ✗
Round 3	Floral ✗	Floral ✗
Round 4	Floral ✗	Floral ✗
Round 5	Floral ✗	Floral ✗

No Convergence

What is the most appropriate Shape for describing this product?
[Title] 14k Yellow Gold St. Jude Thaddeus Medal Charm Pendant - 23.25mm

	Agent 1	Agent 2
Round 1	Oval ✓	Round ✗
Round 2	Oval ✓	Round ✗
Round 3	Oval ✓	Round ✗
Round 4	Oval ✓	Round ✗
Round 5	Oval ✓	Round ✗

No Convergence

What is the most appropriate Shaft Height for describing this product?
[Title] Lucky Brand Kitte Women US 5.5 Black

	Agent 1	Agent 2
Round 1	Lollipop ✗	Jelly Beans ✗
Round 2	Lollipop ✗	Lollipop ✗
Round 3	Lollipop ✗	Gummy ✗
Round 4	Lollipop ✗	Lollipop ✗
Round 5	Lollipop ✗	Jelly Beans ✗

Figure C.1: Illustration of correct, incorrect, and no convergence outcomes observed during agent debate.

Inference Round

Llama-3.2-11B-Vision-Instruct (*Agent 1*)
ANSWER: *Bedroom*
EXPLANATION: The title explicitly mentions "Room Darkening," typically for sleeping spaces. The image shows curtains near a headboard and nightstand, which are common bedroom elements.

Phi-3.5-Vision-Instruct (*Agent 2*)
ANSWER: *Living Room*
EXPLANATION: A sofa and coffee table appear in the background, suggesting a living-room placement.

Debate Round

Llama-3.2-11B-Vision-Instruct (*Agent 1*)
ANSWER: *Bedroom*
EXPLANATION: "Room Darkening" is uncommon in living rooms where natural light is preferred. Listing metadata emphasizes "master bedroom" usage.

Phi-3.5-Vision-Instruct (*Agent 2*)
ANSWER: *Bedroom*
EXPLANATION: Re-checking text and image supports bedroom as the intended location.

D.2   Case B — Pendant Shape

In Case B, visual inspection suggests a circle, but the summary signals that careful multimodal grounding, combining the title's *23.25 mm* numeric detail with visible height-to-width proportion, drives the outcome. This primes the reader to look for numeric–visual alignment during the debate, which ultimately converges on **Oval**.

3050

Inference Round

Llama-3.2-11B-Vision-Instruct (*Agent 1*)
ANSWER: *Oval*
EXPLANATION: Height exceeds width, yielding an elongated silhouette. The 23.25 mm figure in the title denotes the longer axis.

Phi-3.5-Vision-Instruct (*Agent 2*)
ANSWER: *Round*
EXPLANATION: At a glance, the pendant appears nearly circular without precise measurement.

Debate Round

Llama-3.2-11B-Vision-Instruct (*Agent 1*)
ANSWER: *Oval*
EXPLANATION: Compared with the model's finger, the medal is taller than wide. A single numeric value indicates length, not diameter.

Phi-3.5-Vision-Instruct (*Agent 2*)
ANSWER: *Oval*
EXPLANATION: Combining the visual ratio with the title's measurement confirms an oval shape.

D.3 Case C — Footwear Category: Overtrust Example

This section highlights the *failure mode* the case is designed to illustrate: a stronger agent can be impacted by a confident but modality-biased claim. The initial evidence supports **Ankle Boot** (shaft at ankle level), yet during debate the stronger model flips to *Knee High* after over-weighting a visual cue. The purpose is to warn that, in multimodal debate, assertive but unbalanced reasoning can degrade accuracy—here, the correct label remains **Ankle Boot**.

Inference Round

Llama-3.2-11B-Vision-Instruct (*Agent 1*)
ANSWER: *Ankle Boot*
EXPLANATION: The shaft stops just above the ankle bone, a hallmark of ankle boots.

Phi-3.5-Vision-Instruct (*Agent 2*)
ANSWER: *Knee High*
EXPLANATION: Styling suggests coverage up to mid-calf, interpreted as knee-high.

Debate Round

Llama-3.2-11B-Vision-Instruct (*Agent 1*)
ANSWER: *Knee High*
EXPLANATION: On re-check, the shaft appears to extend well above the ankle, suggesting knee-high.

Phi-3.5-Vision-Instruct (*Agent 2*)
ANSWER: *Ankle Boot*
EXPLANATION: Revisiting cues shows the shaft sits at ankle level; "Ankle Boot" is correct.

These case-specific summaries orient readers to the evidence types and intended lesson *before* the transcripts: (A) text vs. scene context with convergence on bedroom use; (B) numeric measurement plus visual proportion resolving shape ambiguity; (C) overtrust of a single modality leading to an incorrect flip. Used this way, summaries make the ensuing debate easier to scan and evaluate.

E   Debate Stability

Table E.1 reports overall accuracy as mean ± SD for a single round and a one round debate. A capacity effect is evident: weaker models show larger variability and larger average gains, while stronger models remain comparatively stable and still improve. To make variability scale invariant, we examine the coefficient of variation (SD divided by mean). For `Llama-3.2-11B-Vision-Instruct`, the relative SD drops from about 0.70% to 0.60 %; for `GPT-4o`, it drops from about 0.55% to 0.45%; for `GPT-o1`, it drops from about 0.58% to 0.50%. In contrast, `Phi-3.5-Vision-Instruct` rises from about 5.70% to 7.31%, `Qwen-2.5-VL-7B` rises from about 1.09% to 1.74%, and `Claude-3.5-Sonnet` rises from about 0.56% to 1.12%. Thus, debate tends to reduce relative variability for the stronger trio and increase it for the weaker model. There is also an association between mean lift and change in SD: the largest gains, such as `Phi-3.5-Vision-Instruct` at +7.24% and `Qwen-2.5-VL-7B` at +5.16 %, coincide with higher SDs, which suggests that debate introduces productive exploration for weaker systems at the cost of stability.

Debate consistently raises the performance for all models, and its effect on variability follows model capacity. For strong systems, debate both improves accuracy and slightly stabilizes outcomes,

Table E.1: Overall accuracy stability: mean ± SD (%) for initial single round and one round debate. ΔMean is debate minus initial; ΔSD is debate SD minus initial SD.

Model	Initial Round (mean ± SD)	Debate Round (mean ± SD)	ΔMean (%)	ΔSD
Llama-3.2-11B-Vision-Instruct	76.17 ± 0.53	77.97 ± 0.47	+1.80	−0.06
Phi-3.5-Vision-Instruct	46.68 ± 2.66	53.92 ± 3.94	+7.24	+1.28
Qwen-2.5-VL-7B	70.66 ± 0.77	75.82 ± 1.32	+5.16	+0.55
GPT-4o	85.21 ± 0.47	87.52 ± 0.39	+2.31	−0.08
Claude-3.5-Sonnet	76.64 ± 0.43	78.58 ± 0.88	+1.94	+0.45
GPT-o1	84.47 ± 0.49	87.77 ± 0.44	+3.30	−0.05

which is attractive for production settings that value predictable behavior. For weaker systems, debate yields larger average gains but can increase run to run spread, so it is best used with safeguards (Huang et al., 2025). In practice we recommend routing by expected efficiency and variance: enable a round of debate when the predicted gain per added second exceeds a latency budget and when the historical variance on the same slice is acceptable. Use fixed prompts and decoding settings, log seeds, and consider a temperature schedule that cools later rounds to limit variance growth.

F Attribute-Level Result

We present the attribute-level results in Figures F.1. In the same MLLM debate in Figures F.1(a), F.1(b), F.1(c), and F.1(f), the attributes *Size*, *Season*, and *Shaft Height* have great improvement across all attributes. In several scenarios, these underperforming attributes show marked improvements through the debate process. For instance, when GPT-4o debates with itself, the *Size* attribute improves from 26.55% to 54% after four rounds of debate. However, during four rounds of debate, Llama-3.2-11B-Vision-Instruct sees the *Size* attribute decrease from 40% to 38.48%, and Phi-3.5-Vision-Instruct experiences a decline in the F1-score for *Size* from 37.8% to 13.87%. A similar pattern, though with more modest changes, is observed for the *Season* and *Shaft Height* attributes. These findings suggest that debate-induced improvements are concentrated on attributes with lower initial accuracy, which likely require diverse viewpoints and enhanced reasoning to achieve better inference. However, these improvements are only realized when the debating agent possesses sufficient knowledge. If an agent misinterprets the product image or text, its flawed reasoning may negatively influence its own performance and that of its debate partner in subsequent rounds, ultimately further degrad-

ing performance on low-accuracy attributes, as observed with Phi-3.5-Vision-Instruct and Qwen-2.5-VL-7B.

A similar pattern is also observed in the cross MLLM debates shown. In Figure F.1(d), the left-hand side represents Llama-3.2-11B-Vision-Instruct while the right-hand side represents GPT-4o. In this scenario, Llama-3.2-11B-Vision-Instruct exhibits significant improvements in several attributes; for example, *Attachment Method*, *Form*, *Heel Height*, and *Neckline* improve by 30.39%, 19.77%, 25.6%, and 15.32%, respectively. These gains indicate that the weaker agent benefits considerably from the debate, particularly on attributes with lower initial accuracy. In contrast, GPT-4o experiences a slight decline in certain attributes: *Shape*, *Length*, and *Shaft Height* decrease by 0.76%, 1.17%, and 3.32%, respectively. Although GPT-4o's negative changes are less pronounced than the positive improvements observed in Llama-3.2-11B-Vision-Instruct, the attributes that show the largest gains for the weaker agent do not necessarily correspond to those with the most significant declines for the stronger agent, indicating that these effects are not directly correlated.

Similarly, in the same debate settings between Llama-3.2-11B-Vision-Instruct and Phi-3.5-Vision-Instruct, as shown in Figure F.1(e). In this configuration, Phi-3.5-Vision-Instruct's inference improves markedly on several attributes—*Neckline*, *Shape*, *Candy Variety*, and *Sleeve Style* increase by 53.01%, 35.99%, 35.92%, and 42.78%, respectively. In contrast, some attributes for Llama-3.2-11B-Vision-Instruct decline, with *Size*, *Attachment Method*, and *Shoulder Style* decreasing by 11.12%, 20.49%, and 13.52%. These findings suggest that low-accuracy attributes are particularly sensitive to the quality of
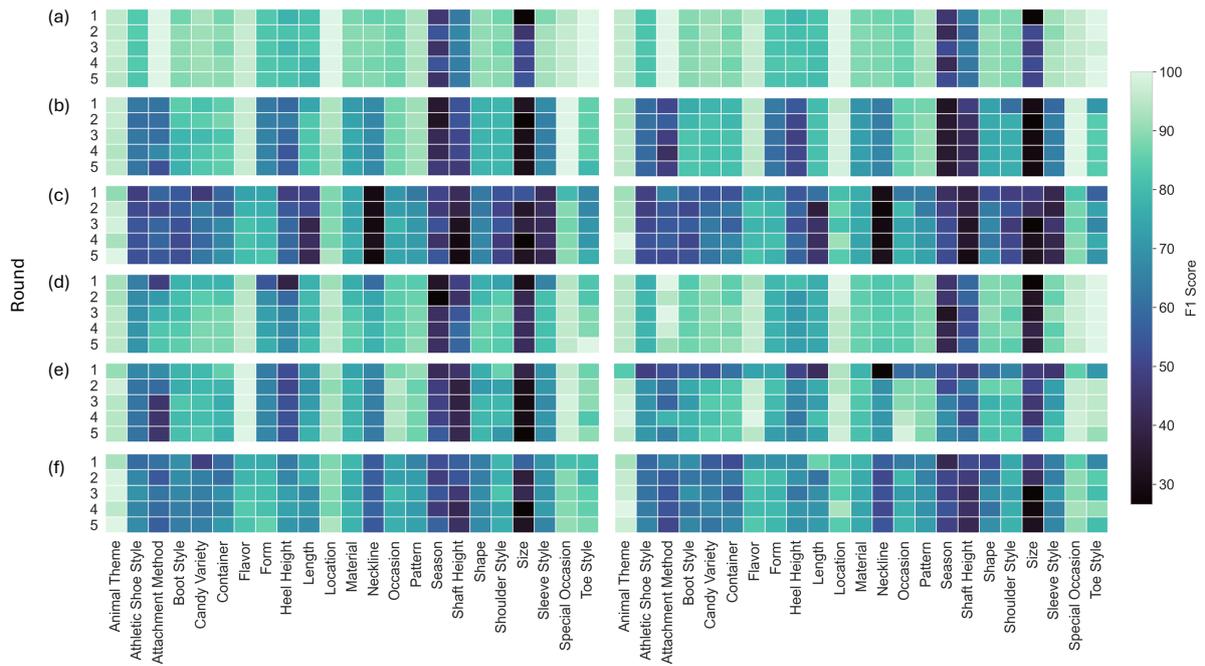
Figure F.1: Attribute-level heatmap of the F1 score changes during the debate process for (a) Scenario 1: GPT-4o debating with GPT-4o; (b) Scenario 2: Llama-3.2-11B-Vision-Instruct debating with Llama-3.2-11B-Vision-Instruct; (c) Scenario 3: Phi-3.5-Vision-Instruct debating with Phi-3.5-Vision-Instruct; (d) Scenario 4: Llama-3.2-11B-Vision-Instruct debating with GPT-4o; (e) Scenario 5: Llama-3.2-11B-Vision-Instruct debating with Phi-3.5-Vision-Instruct; (f) Scenario 6: Qwen-2.5-VL-7B debating with Qwen-2.5-VL-7B.

reasoning: when accurate reasoning is applied, these attributes benefit significantly, but flawed reasoning, even from a stronger agent, can lead to notable performance decrements.

G Use of LLMs

We used LLMs only to refine wording (grammar, clarity, and style) of text. It was not used to generate ideas, write substantive content, design experiments, analyze data, or produce results. No non-public data were provided to the model, and all content was verified and finalized by the authors.