Enhancing Multi-Agent Multi-Modal Collaboration with Fine-Grained Reward Modeling

Qian Yang[1,2], **Weixiang Yan**[3], **Aishwarya Agrawal**[1,2,4]
[1] Mila - Québec AI Institute [2] Université de Montréal
[3] University of California, Santa Barbara [4] Canada CIFAR AI Chair
qian.yang@mila.quebec weixiangyan@ucsb.edu aishwarya.agrawal@mila.quebec

Abstract

Multi-Modal Large Language Models (MLLMs) have significantly advanced multi-modal reasoning but still struggle with compositional reasoning tasks. Multi-agent collaboration provides a promising solution by leveraging the distinct capabilities of different agents. Specifically, a decomposer agent to handle task breakdown and an answerer agent to generate responses. While there have been efforts to adaptively decompose tasks based on the answerer agent's capabilities, such as using in-context learning, these methods often prove insufficient for fully effective decomposition. We address this issue by enhancing collaboration through fine-grained reward modeling, where each generated sub-question is assigned a specialized reward without requiring extra annotation or tuning of a reward model. Our proposed method dynamically optimizes the decomposition process, enabling better alignment between agents. Experimental results on four vision-language tasks demonstrate consistent improvements, with a 5.5% absolute increase in mean performance over traditional approaches. These findings highlight the efficacy of fine-grained reward modeling for enhancing multi-agent, multi-modal collaboration.

1 Introduction

The advent of Multi-Modal Large Language Models (MLLMs) has marked a significant milestone in artificial intelligence, enabling sophisticated multi-modal reasoning. However, MLLMs exhibit notable limitations in compositional reasoning compared to their unimodal language model counterparts (Yuksekgonul et al., 2022). Furthermore, high-quality data for multi-modal reasoning is more challenging to acquire than single-modal data, complicating the optimization of MLLMs.

Research has shown that natural language serves as an effective intermediate representation for reasoning, particularly in human cognition (Gentner & Goldin-Meadow, 2003; Forbus et al., 2017). Building on this insight, recent efforts have sought to address complex multi-modal tasks through multi-agent collaboration, where agents use language as a means of communication. This approach allows for the decomposition of complex tasks into smaller, more manageable sub-tasks. A common framework involves using LLMs as decomposition agents that break down intricate problems, while MLLMs sequentially solve these sub-tasks to arrive at a final solution (Lu et al., 2024b; Khan et al., 2024; Yang et al., 2023b; Surís et al., 2023; Gupta & Kembhavi, 2023; Zhang et al., 2023; Zheng et al., 2023). By leveraging the reasoning strengths of LLMs alongside the perceptual capabilities of MLLMs, this multi-agent collaboration framework reduces the need for extensive annotation in multi-modal datasets and harnesses the reasoning abilities of LLMs with the recognition skills of MLLMs. However, despite the promise of this collaboration, a key limitation remains. The pre-decomposition process often fails to incorporate crucial feedback from MLLMs, as current approaches rely solely on predefined strategies set by LLMs. This can result in inappropriate task breakdowns, lacking

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

adaptability to the specific reasoning capabilities of MLLMs, which often leads to generalized rather than tailored decompositions.

Given the diverse reasoning capabilities of different MLLMs and the dynamic nature of visual question answering tasks, a one-size-fits-all decomposition approach is insufficient. Recent studies have proposed interactive strategies in which LLMs generate the next sub-question based on MLLM feedback, dynamically refining the decomposition process (Yang et al., 2023d; You et al., 2023; Yang et al., 2023a). While promising, these interactive methods still face challenges. Our preliminary experiments suggest that pre-decomposition can sometimes outperform interactive decomposition, primarily because LLMs are not explicitly tuned for task decomposition, and a few in-context learning examples are insufficient to derive an optimal decomposition strategy. Moreover, LLMs struggle to adapt to the specific needs of different MLLMs, as in-context learning fails to determine the level of decomposition that best suits each MLLM. To create a more effective collaboration system, adaptability to the distinct capabilities of each MLLM is essential. One potential solution is supervised fine-tuning of LLMs for task decomposition, but this requires extensive annotation across various MLLMs, making it resource-intensive. Another approach involves reinforcement learning, where the correctness of the MLLM's prediction serves as the reward, avoiding explicit label annotation Yang et al. (2023a). However, this reward structure tends to be too sparse, as it focuses only on final accuracy without accounting for the intermediate decomposition steps. Treating all sub-questions with equal importance often fails to incentivize the decomposer to generate meaningful and efficient sub-questions.

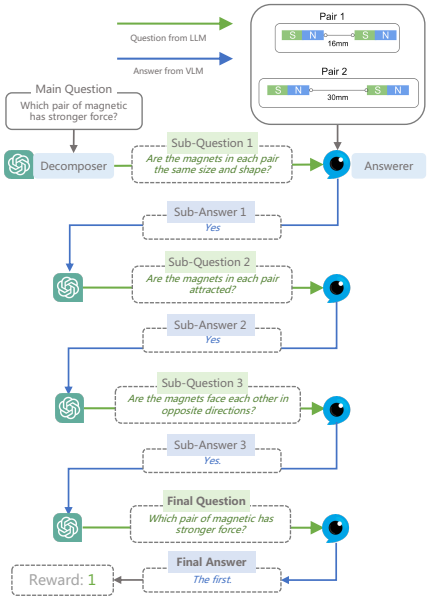

Figure 1: Illustration of interactive decomposition with coarse-grained reward.

In this work, we propose fine-grained reward modeling to address these limitations and enhance multi-agent multi-modal collaboration. Unlike previous approaches, our method assigns a specialized reward to each generated sub-question based on its unique contribution to the overall task, without requiring additional annotations or tuning of a reward model. This allows the LLM to better adapt the decomposition process to match the varying capabilities of different MLLMs. By tailoring the decomposition to the specific strengths and weaknesses of each MLLM, we foster a more balanced and effective collaboration, ultimately leading to improved performance in vision-language tasks. Our experimental results demonstrate the promise of fine-grained reward modeling, with significant improvements in the adaptability and efficiency of the decomposition process compared to traditional methods.

2 Method

To more effectively reward each sub-question, we explore a fine-grained reward design instead of the sparse reward system, providing better guidance for LLM collaboration and task decomposition.

2.1 Interactive Decomposition with Coarse-Grained Reward

As shown in Figure 1, In the interactive decomposition process, the decomposer agent and the answerer agent collaborate iteratively. During each round, the decomposer agent generates a sub-question based on the main question and all priorsub-QA pairs. The answerer agent then responds to this sub-question, utilizing the information from all previous sub-QA pairs along with the image. This iterative process continues until the decomposer agent determines that further decomposition is unnecessary and proceeds to generate the final response to the main question. The answerer agent then attempts to answer the main question based on the accumulated knowledge from all sub-questions. Rewards are assigned based on the correctness of the answerer's response to the final question. To

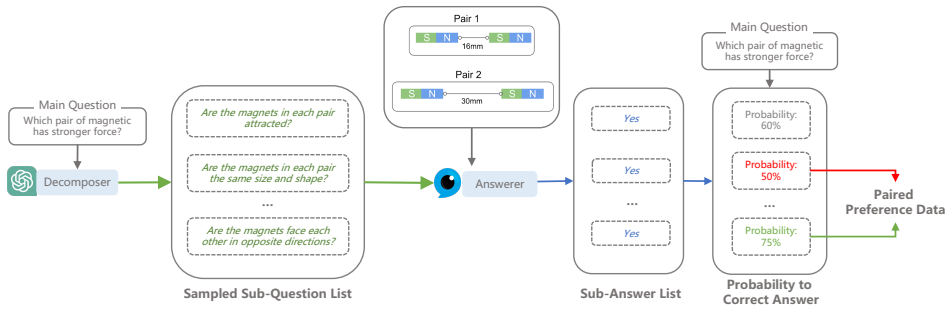

Figure 2: Illustration of DPO with Fine-Grained Reward. In each round, we sample 5 sub-questions and select the one that leads to the highest confidence in the main question's answer as the positive example, and the one that leads to the third highest confidence as the negative example.

implement this reward mechanism, we use Proximal Policy Optimization (Schulman et al., 2017) for fine-tuning the decomposer agent. However, this approach employs a coarse-grained reward system, as each sub-question receives the same reward, regardless of its individual contribution. This means that the reward assignment does not differentiate between helpful and un-helpful sub-questions.

The main limitation of this coarse-grained approach is that it does not account for the varying quality and informativeness of each sub-question. Not all sub-questions contribute equally to the final answer, and some may even be redundant or misleading. By treating all sub-questions with equal importance, this method can fail to incentivize the decomposer agent to generate more meaningful and efficient sub-questions.

2.2 DPO with Fine-Grained Reward

Our aim is to design a fine-grained reward system for each sub-question, acknowledging that each sub-question contributes uniquely to the overall task-solving process. However, determining an appropriate reward for each sub-question poses significant challenges. Different MLLMs exhibit unique strengths and preferences, making it difficult to assess the effectiveness of each sub-question consistently. This variability complicates the development of a reward model capable of assigning absolute rewards to individual sub-questions in a meaningful and reliable manner.

To address these challenges, we adopt Direct Preference Optimization (DPO) (Rafailov et al., 2024), a method that directly optimizes a language model to align with human-like preferences without requiring explicit reward modeling or reinforcement learning. DPO works by increasing the relative log probability of preferred responses over dispreferred ones:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]. \quad (1)$$

DPO requires only the construction of paired preference data, thereby eliminating the complexities of tuning a dedicated reward model. Instead of determining absolute values for each sub-question's reward, we focus on generating preference pairs that compare the relative effectiveness of different sub-questions. This allows us to automatically construct preference pairs for each round based on the answerer's confidence in the correct answer, typically indicated by the probability assigned to that answer. As shown in Figure 2, during each decomposition round, we sample several sub-questions generated by the decomposer agent. The MLLMs then answers all of these sampled sub-questions, and subsequently attempts to answer the main question using each sub-QA pair as additional context. By comparing the main question's answer confidence after considering different sub-questions, we can identify which sub-questions have had a more positive impact. This information enables us to construct paired preference data, where relatively more effective sub-questions are paired against less effective ones. This preference-based learning strategy provides a robust mechanism for guiding the LLM towards generating more informative sub-questions. Unlike traditional reward modeling, which often struggles with sparse or poorly differentiated feedback, DPO provides a more continuous and context-sensitive signal that enhances the adaptability of the decomposition agent. DPO with fine-grained reward not only simplifies the reward structure by avoiding explicit value assignments but

	Model	SNLI-VE†	VCR	Winoground	MathVista	Mean
1	Base MLLM	39.3	62.3	50.5	48.0	50.0
2	Base MLLM + Sample	39.5	62.5	49.3	48.2	49.9
3	Base MLLM + Chain-of-Thought	43.6	63.0	49.3	47.2	50.8
4	Base MLLM + Chain-of-Thought + Sample	44.3	62.1	49.0	48.1	50.9
5	Pre-Decomposition	53.0	64.0	53.5	49.0	54.9
	Interactive Decomposition					
6	Interactive Decomposition	54.1	61.1	55.8	48.4	54.9
7	SFT$_{VCR_{7K}+SNLI_{13K}}$	54.1	61.9	55.3	48.4	54.9
8	SFT + PPO$_{SNLI_{3K}}$ with Coarse-Grained Reward	53.7	65.2	55.3	47.8	55.5
9	DPO$_{SNLI_{50k}}$ with Fine-Grained Reward	56.3	61.5	55.8	48.5	55.5

Table 1: Accuracy across various datasets utilizing Idefics2-8B as the MLLM for answering. † indicates that samples from these datasets are used for few-shot prompting in the Decomposition.

also enhances the interaction between the LLM and MLLMs, leading to more effective sub-question generation and ultimately improving the quality of the final answer.

3 Experiments

3.1 Datasets and Baselines

We conduct experiments on six vision-language tasks, including SNLI-VE (Xie et al., 2019), VCR (Zellers et al., 2019), A-OKVQA (Schwenk et al., 2022), Winoground (Thrush et al., 2022), and MathVista (Lu et al., 2024a). For our experiments, we use OpenHermes-2.5-Mistral-7B[1] as the Decomposer Agent, and Idefics2-8B (Laurençon et al., 2024) serves as the Answerer Agent.

We evaluate several baseline approaches to benchmark the effectiveness of different decomposition and answering strategies. These baselines are summarized as follows: **Base MLLM (Line 1)**: We first evaluate the performance of the base MLLM without any enhancement strategies. **MLLM with Sampling (Line 2)**: In this setting, we generate five potential answers from the base MLLM model and select the final answer through majority voting. **MLLM with Chain-of-Thought (Line 3)**: We prompt the MLLM to "Think step by step and then answer the question." **MLLM with Chain-of-Thought and Sampling (Line 4)**: This strategy combines Chain-of-Thought prompting with the sampling approach. **Pre-Decomposition (Line 5)**: For this strategy, we follow the settings proposed by Yang et al. (2023c). The LLM is prompted to pre-generate four sub-questions conditioned on the given main question, without any iterative interaction. Few-shot prompting is used for the LLM, selecting the same four examples as used in interactive decomposition, ensuring a fair comparison. The LLMs decompose the problem in a static manner before any interaction with the MLLM. **Interactive Decomposition without Tuning (Line 6)**: We directly prompt the decomposer agent to engage interactively with the answerer agent without any further tuning. **Interactive Decomposition with Supervised Fine-Tuning (Line 7)**: In this strategy, we fine-tune the LLM using a supervised dataset that combines samples from VCR and SNLI-VE. Initially, we apply interactive decomposition across the training set and retain only those samples that successfully aid the answerer agent in providing the correct answer to the main question. This curated set is then used to fine-tune the LLM for better decomposition. **Interactive Decomposition with PPO Fine-Tuning and Coarse-Grained Reward (Line 8)**: After initial supervised fine-tuning, we further optimize the model using Proximal Policy Optimization (PPO) (Schulman et al., 2017) with Coarse-Grained Reward introduced in Section 2.1. We do not use few-shot prompting with the MLLM itself, as it lacks the ability to handle multiple images simultaneously. For decomposition, we utilize few-shot prompting with LLMs, randomly selecting four samples from SNLI-VE as examples to guide the decomposer. The remaining datasets are approached using a zero-shot decomposition strategy by the LLMs, while all MLLM operate in a zero-shot setting across all datasets.

3.2 Results and Discussion

In our baseline models, we observe that employing a decomposition strategy significantly improves performance across various vision-language tasks compared to using only an MLLM. Interestingly, the pre-decomposition strategy performs comparably to interactive decomposition without any tuning,

[1] `https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B`

achieving the same average performance across four datasets. This suggests that, while interactive decomposition introduces adaptability, its potential is not fully realized without proper tuning. SFT yields a slight performance improvement on the VCR dataset. However, this gain comes at the cost of reduced generalization capability for the decomposer agent, leading to decreased performance on more diverse tasks like Winoground. This trade-off highlights a limitation of SFT: while it can enhance targeted performance, it risks overfitting, thereby impairing broader applicability. Combining SFT with coarse-grained reward optimization using PPO provides an increase on only VCR. This suggests that the coarse-grained reward approach lacks the granularity needed to consistently enhance decomposition across varied tasks, possibly due to its uniform treatment of all sub-questions irrespective of their specific contributions. Line 9 demonstrates the performance using DPO with a fine-grained reward model. Notably, DPO-tuned models achieve performance improvements on all datasets except VCR compared to using only an MLLM. Importantly, this approach yields the highest mean performance across all tasks, showing an absolute increase of 5.5% w.r.t. to the base MLLM. This improvement is comparable to Line 8 in the case of SFT + PPO, which required both supervised SFT and PPO. In contrast, DPO achieves similar results with a simpler, more targeted preference-based optimization approach. These findings suggest that DPO with fine-grained rewards provides a more effective way to guide the decomposition agent compared to traditional SFT or PPO methods. By focusing on the relative effectiveness of each sub-question, DPO ensures a more adaptive and context-sensitive optimization, which in turn leads to more balanced performance gains across diverse datasets. This adaptability is particularly valuable in multi-agent, multi-modal collaboration settings, where task complexity and the effectiveness of individual contributions can vary significantly.

4 Conclusion

In this work, we explored enhancing the collaboration between MLLMs using a novel fine-grained reward modeling approach. We identified key challenges in current multi-agent frameworks, particularly the inefficiencies in task decomposition and the lack of adaptive interaction between decomposer and answerer agents. To address these issues, we proposed the use of DPO to implement a fine-grained reward system, allowing the decomposer agent to iteratively refine its sub-question generation based on the effectiveness of each sub-question in aiding the answerer agent's performance. Our experiments, conducted across four vision-language tasks, demonstrated that fine-grained reward modeling significantly enhances the efficiency and adaptability of the decomposition process. Future work will explore further refinement of reward mechanisms, aiming for more generalised adaptability across diverse tasks.

Acknowledgement

We are grateful to Mila's IDT team for their technical support with the computational infrastructure. During this project, Aishwarya Agrawal was supported by the Canada CIFAR AI Chair award.

References

Kenneth D Forbus, Chen Liang, and Irina Rabkina. Representation and computation in cognitive models. *Topics in cognitive science*, 9(3):694–718, 2017.

Dedre Gentner and Susan Goldin-Meadow. Language in mind: Advances in the study of language and thought. 2003.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Zaid Khan, Vijay Kumar BG, Samuel Schulter, Manmohan Chandraker, and Yun Fu. Exploring question decomposition for zero-shot vqa. *Advances in Neural Information Processing Systems*, 36, 2024.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024a.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Diji Yang, Kezhen Chen, Jinmeng Rao, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. Tackling vision language tasks through learning inner monologues. *arXiv preprint arXiv:2308.09970*, 2023a.

Kaiwen Yang, Tao Shen, Xinmei Tian, Xiubo Geng, Chongyang Tao, Dacheng Tao, and Tianyi Zhou. Good questions help zero-shot image reasoning. *arXiv preprint arXiv:2312.01598*, 2023b.

Kaiwen Yang, Tao Shen, Xinmei Tian, Xiubo Geng, Chongyang Tao, Dacheng Tao, and Tianyi Zhou. Good questions help zero-shot image reasoning, 2023c.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023d.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. IdealGPT: Iteratively decomposing vision and language reasoning via large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

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