This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models

Yuhao Dong[1*] Zuyan Liu[2,3*] Hai-Long Sun[3,4] Jingkang Yang[1]
Winston Hu[2] Yongming Rao[2] ✉ Ziwei Liu[1] ✉
[1] S-Lab, Nanyang Technological University [2] Tencent
[3] Tsinghua University [4] Nanjing University
https://github.com/dongyh20/Insight-V

Abstract

Large Language Models (LLMs) demonstrate enhanced capabilities and reliability by reasoning more, evolving from Chain-of-Thought prompting to product-level solutions like OpenAI o1. Despite various efforts to improve LLM reasoning, high-quality long-chain reasoning data and optimized training pipelines still remain inadequately explored in vision-language tasks. In this paper, we present ***Insight-V****, an early effort to* ***1)*** *scalably produce long and robust reasoning data for complex multi-modal tasks, and* ***2)*** *an effective training pipeline to enhance the reasoning capabilities of multi-modal large language models (MLLMs). Specifically, to create long and structured reasoning data without human labor, we design a two-step pipeline with a progressive strategy to generate sufficiently long and diverse reasoning paths and a multi-granularity assessment method to ensure data quality. We observe that directly supervising MLLMs with such long and complex reasoning data will not yield ideal reasoning ability. To tackle this problem, we design a multi-agent system consisting of a reasoning agent dedicated to performing long-chain reasoning and a summary agent trained to judge and summarize reasoning results. We further incorporate an iterative DPO algorithm to enhance the reasoning agent's generation stability and quality. Based on the popular LLaVA-NeXT model and our stronger base MLLM, we demonstrate significant performance gains across challenging multi-modal benchmarks requiring visual reasoning. Benefiting from our multi-agent system, Insight-V can also easily maintain or improve performance on perception-focused multi-modal tasks.*

1. Introduction

The development of artificial general intelligence requires models that can seamlessly understand and respond to multi-modal data. Recent advancements in Large Language Models (LLMs) [9, 36, 37, 41] and Multi-modal LLMs (MLLMs) [6, 21–23, 29, 38, 49] have significantly facilitated this progress across various fields, ranging from common question-answering [6, 17, 22, 27, 38] to autonomous driving [32, 42] and robotics [8, 47]. Despite the substantial progress made in enhancing the performance of MLLMs on a wide range of tasks, enabling MLLMs to perform human-level reasoning remains a key challenge. This area remains underexplored and has yet to fully realize its potential.

Existing efforts [44, 48] to enhance the reasoning capabilities of LLMs through long-chain reasoning have

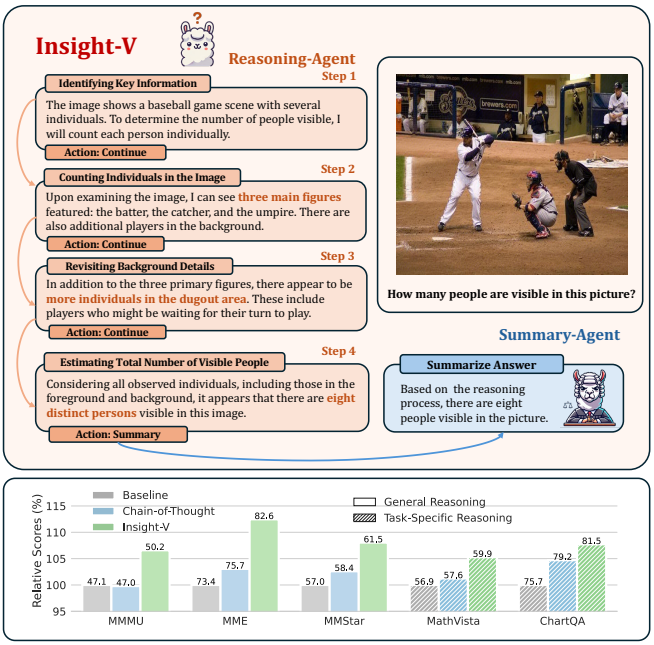

Figure 1. **Illustration and Performance of Insight-V.** Insight-V consists of two agents, one dedicated to reasoning and the other to summarization, driving significant improvements in performance across various visual reasoning benchmarks.

*Authors contributed equally to this research. ✉Corresponding authors.

demonstrated considerable progress, largely benefiting from the availability of structured, high-quality data and well-established training pipelines. In contrast, teaching MLLMs to perform long-chain visual reasoning remains a significant challenge, primarily due to the lack of large-scale, high-quality datasets and efficient and effective training strategies. Compared to text-only data, visual reasoning data is not only more expensive to collect but also requires significant human labor for detailed annotation and validation, due to the absence of an effective data generation pipeline. Moreover, while previous work [58] has demonstrated that directly applying chain-of-thought [44] reasoning can improve the capabilities of MLLMs, other research [56, 57] suggests that current training approaches have limited effectiveness in enhancing CoT reasoning. This highlights the inability of current MLLMs to leverage visual cues for precise step-by-step problem-solving, emphasizing the need for an effective training procedure that enables MLLMs to reason in detail while maintaining clear visual perception.

To address these challenges, we propose Insight-V, which incorporates two innovative designs to enhance reasoning capabilities. First, we introduce a data generation pipeline consisting of two key steps: a progressive strategy to generate structured, long-chain reasoning data with diverse reasoning paths, and a multi-granularity assessment system to evaluate and score these paths at different levels. Through automatic generation, assessment, and ranking strategies, the pipeline effectively operates without the need for human labor and makes the reasoning dataset more scalable for enhancing reasoning capabilities. To further improve MLLM reasoning beyond data scaling, we design a multi-agent system, as illustrated in Figure 1, that decomposes the problem-solving process into two distinct steps: reasoning and summarization. The reasoning agent generates a detailed reasoning process for the input query, while the summarization agent identifies key information within the reasoning process and selectively answers the question. To refine the quality of the reasoning, we employ an iterative DPO approach to enhance reasoning capabilities. The two agents collaborate to further improve the reasoning quality. Our findings demonstrate that this system significantly enhances the performance of various MLLMs across a broad range of visual reasoning benchmarks.

We evaluate Insight-V by integrating our system into the widely used LLaVA-NeXT [22] model. To further highlight the potential of our design in advancing state-of-the-art models, we construct a robust base MLLM and apply our system to demonstrate its ability to enhance SOTA models across various visual reasoning benchmarks. Notably, Insight-V improves LLaVA-NeXT in an average performance of 7.0% across seven challenging visual reasoning benchmarks, while a 2.9% improvement is observed with the strong base MLLM, underscoring the effectiveness and generalizability of Insight-V.

In summary, Insight-V offers **1)** a scalable data generation pipeline for long-chain, high-quality reasoning data, **2)** a multi-agent system that decomposes visual reasoning tasks into reasoning and summarization, and **3)** a two-stage training pipeline to enhance visual reasoning capabilities. Together, these contributions address key challenges in visual reasoning, providing a solid foundation for future research in MLLM reasoning.

2. Related Work

2.1. Vision-Language Reasoning

Recent advancements in MLLMs [1, 17, 20–23, 27, 29, 38] have equipped these models with robust capabilities across diverse domains, including visual understanding [20, 38], mathematics [19], college-level questions [6], and scientific inquiries. In visual understanding, most research [18, 22, 28, 43, 46] emphasizes fine-grained detail analysis and localization, training models to perform visual reasoning with tailored datasets to enhance interpretive capabilities. For mathematics and expert-level reasoning, existing methods [11, 56, 57] predominantly derive from Chain-of-Thought [44] approaches, training MLLMs to generate step-by-step reasoning across various subjects. However, these approaches often focus primarily on improving dataset quality through Chain-of-Thought, overlooking the importance of structured reasoning paths and extended reasoning chains in advancing model reasoning capabilities. Additionally, significant challenges arise when relying on a single model to manage the entire reasoning process for complex tasks, underscoring the need for a multi-agent system to decompose and enhance this process. In this work, we tackle these challenges by introducing a scalable reasoning data generation pipeline and implementing a multi-agent system for reasoning and summarization decomposition, enhancing the overall reasoning capabilities of existing MLLMs.

2.2. Vision-Language Alignment

To align the model more closely with human preferences, several alignment techniques are employed for MLLMs. A widely used approach is Reinforcement Learning from Human Feedback [2] (RLHF), which iteratively refines the model's responses based on human feedback, enhancing both response quality and interpretability. To further improve MLLM capabilities, Direct Preference Optimization [39] (DPO) is introduced to simplify the alignment process. By directly training on human preference data, DPO optimizes the model's outputs to better match human-selected responses. However, traditional DPO is primarily focused on offline scenarios, and as the model evolves, the effectiveness of this approach may significantly diminish. To address this, Iterative DPO [5] has been proposed, which optimizes pref-

erence pairs through DPO at each iteration. It then generates new preference pairs for the next iteration using the updated model and evaluates them with a reward model. In this paper, we use iterative DPO to achieve stronger alignment and enhance the model's reasoning capabilities.

3. Methodology

In this section, we provide a comprehensive description of the proposed Insight-V system, detailing its architecture and key contributions. Section 3.1 presents an overview of Insight-V, highlighting the core concepts of our approach. The design of Insight-V is structured around three primary components: **1)** a carefully constructed pipeline for structured, scalable reasoning data generation, as described in Section 3.2; **2)** a multi-agent MLLM system that supports complex visual reasoning, as detailed in Sections 3.3; and **3)** a streamlined yet effective training pipeline to enhance overall performance, as outlined in Section 3.4. Together, these components form a cohesive system that effectively tackles the challenges of performing detailed, long-chain reasoning while preserving visual perception capabilities.

3.1. Overview

Advancing reasoning capabilities of LLMs has been a focal point of extensive research. Despite these efforts, the reasoning potential within multi-modal LLMs has barely been explored. Most approaches aim to strengthen reasoning at the inference stage, assuming the model has already acquired robust reasoning skills. Other approaches optimize model parameters using chain-of-thought data, enabling models to mimic human reasoning processes. However, these methods present significant challenges for current general-purpose MLLMs, as they require the model to develop reasoning skills while retaining prior capabilities, which often results in only modest performance gains. Additionally, the lack of structured, high-quality training data impedes training models with advanced reasoning capabilities.

To fully leverage the reasoning capabilities of MLLMs, we propose Insight-V, a novel system comprising two MLLMs dedicated to reasoning and summarization, respectively. The `reasoning model` is tasked with generating a detailed reasoning process to assist in problem-solving, while the `summary model` evaluates this reasoning as supplementary information to assess its relevance and utility for answering the question. We also construct a structured, high-quality dataset to train both agents. We posit that this multi-agent system can enhance the reasoning strengths of MLLMs by decomposing the problem-solving process into distinct reasoning and summarization phases, thereby driving substantial performance improvements.

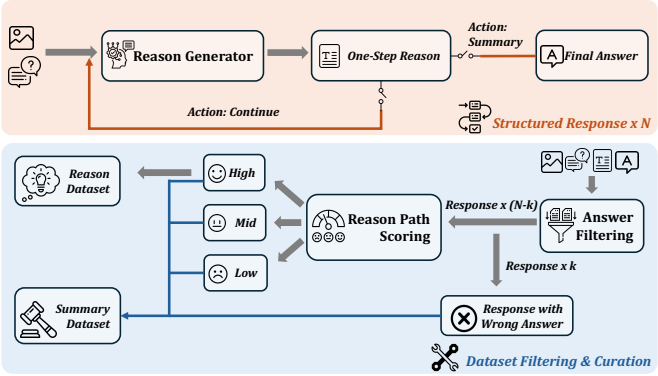

Figure 2. **Data Generation Pipeline of Insight-V.** The reasoning processes are generated progressively through a reasoning generator, and then fed into a multi-granularity assessment system to ensure high-quality reasoning.

3.2. Construction of Structured Reasoning Data

Previous studies [56, 58] have explored the integration of reasoning capabilities into MLLMs. However, training MLLMs to develop robust reasoning skills remains a considerable challenge, particularly due to data limitations. To address this, we introduce our data generation pipeline in this section, designed to produce high-quality, long-chain reasoning data using a progressive generation process and multi-granularity assessment. As shown in Figure 2, this scalable approach enables us to generate high-quality data to enhance the model's reasoning capabilities effectively.

Progressive Long-Chain Reasoning Data Generation. For each input query, we first employ a reasoning generator to produce a structured reasoning process in JSON format to address the problem. At each step, the reasoning generator provides a brief summary of the current step, a detailed reasoning response, and an action for the following step. If the action is $continue$, the model proceeds with an additional reasoning step in the next iteration; if the action is $summary$, the model generates a final summary and answer based on the complete reasoning process in the subsequent iteration. Specifically, For a multi-modal model M, input image I, and question Q, the data generation process for each step is represented as follows:

$$R_t = M(I, Q, [R_1 \cdots R_{t-1}], A),$$
$$R_{ans} = M(I, Q, [R_1 \cdots R_n]),$$

where R_t and R_{ans} denote the response at the t-th step and the final answer, respectively, R_i represents the reasoning generated by the model at the i-th step, n represents the total reasoning steps, and A is the action determined in the previous step. By repeating this process N times, we can iteratively sample N structured responses for each query. The generation parameters are adjusted to encourage the

model to produce outputs with various information and steps, allowing us to identify the most effective reasoning chain for each question.

Multi-Granularity Assessment. After obtaining the structured responses, we utilize an assessment pipeline to ensure data quality. Specifically, we first apply a strong LLM, such as Qwen2 [37], for direct **answer filtering**. As shown in Figure 2, the model is provided with the generated final answer and the ground truth answer and is tasked with determining whether the generated answer is correct, serving as an approximate indicator of the validity of the associated reasoning chain. Once responses with incorrect answers are filtered out, the remaining reasoning processes are passed to a **reasoning path scoring** agent. Here, an advanced multi-modal model, such as Qwen2-VL [38], is supplied with the image, question, reasoning path, and ground truth answer, and is prompted to evaluate the reasoning path. The scoring agent assesses each response based on the step-by-step accuracy of the reasoning path and the level of detail in the reasoning. To ensure consistency in scoring across different data samples, we aggregate all responses for each question and process them in a single pass. The model then generates scores for each response, ranging from 1 to 100.

Through the above two steps, we construct a structured, high-quality dataset that provides detailed reasoning for each question, effectively supporting the training of our models.

3.3. Model Design

After constructing the dataset, we develop a multi-agent framework to enhance overall reasoning capabilities through collaborative agent interaction. Specifically, we first train a reasoning agent to generate a detailed reasoning process for each problem. Then, a summary agent is employed to answer the question, selectively utilizing the reasoning process based on its assessment. Together, these two agents collaborate to improve reasoning performance effectively.

Reasoning Agent. Previous approaches typically combine reasoning and question-answering within a single process, which poses challenges for MLLMs. Generating a long-chain reasoning process can introduce errors, and directly answering questions based on flawed reasoning often leads to poorer results. To address this, we propose a specialized reasoning agent designed to generate a detailed, step-by-step reasoning process in response to an input query. We construct the reasoning dataset by selecting the highest-scoring reasoning path for each question. After training on this dataset, the model transforms into a reasoning agent with enhanced reasoning capabilities, enabling it to generate more detailed, structured reasoning processes.

Summary Agent. Summarization plays a critical role in enabling models to accurately answer questions. After generating multi-step reasoning, summarization provides a cohesive understanding of the reasoning process, ultimately guiding the model to the final answer. However, since the response generated by the reasoning agent may contain errors, we develop a summarization model robust to inaccuracies in the reasoning path, selectively incorporating or disregarding elements as needed. This approach maximizes the reasoning model's effectiveness while minimizing the risk of introducing misleading information.

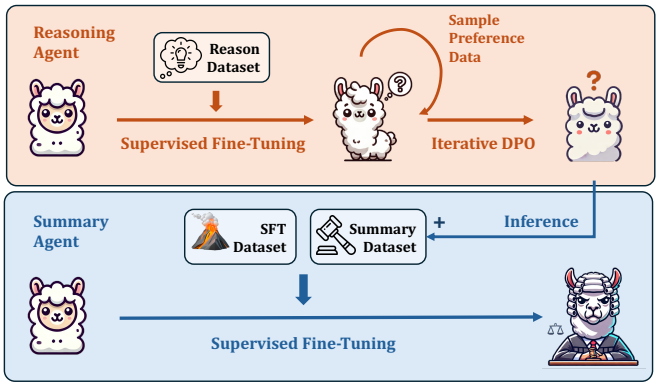

Figure 3. **Overview of Insight-V Model Design.** We derive a multi-agent system from a single model. By decomposing the task into reasoning and summarization, the two agents collaborate to enhance the overall reasoning capability.

To enhance the robustness of the summary agent, we carefully curate its training dataset. We utilize the collected dataset, which is comprised of two types: data with optimal reasoning processes and data with flawed reasoning processes for the summarization task. This method prevents the model from simply copying reasoning outcomes and encourages critical evaluation of reasoning quality. To further promote critical analysis by the summary agent, we select flawed reasoning samples based on their performance scores. Specifically, we draw flawed responses from varying score ranges to create a dataset with different levels of error, prompting the model to assess reasoning processes at various granularities. To better align the summary model with the reasoning agent, we also incorporate question-reasoning pairs generated by the reasoning agent to enhance collaboration between the two agents. Additionally, to preserve the original multi-modal capabilities, we supplement the dataset with standard question-answering data to sustain the summary agent's performance in direct question-answering.

3.4. Training Pipeline

The training pipeline for Insight-V is designed to be straightforward and efficient, utilizing a two-stage strategy. For both the reasoning agent and the summary agent, we begin with a well-trained MLLM. In the first stage, we apply supervised fine-tuning to train the agents to fulfill their designated roles using corresponding datasets. In the second stage, we

implement direct preference optimization, following prior research [45, 57]. This optimization is applied to the reasoning model, aligning it with human reasoning processes in a simple yet effective manner. These two stages enable the development of a robust system with enhanced visual reasoning capabilities.

3.4.1. Supervised Fine-tuning for Multi-agent System

To perform supervised fine-tuning and obtain the two agents, we first train a base multi-modal model, following established methodologies. This model can address general visual question-answering tasks and gain foundational vision-language skills. We compile a high-quality image-text dataset focused on knowledge learning to train the base model. This data is sourced from various open-source academic datasets, including LLaVA-NeXT [22], Cauldron [15], and Cambrian-1 [43]. Once the base model is trained, we further fine-tune the two agents, initializing them from the base model. For the reasoning agent, we utilize the curated reasoning dataset to develop step-by-step reasoning capabilities. For the summary agent, we formulate a dataset as outlined in Section 3.3 and sample about one million general image-text pairs from the dataset used for the base model, preserving its original visual perception abilities.

3.4.2. Enhanced Reasoning with RL

Preference learning has gained increasing focus in the field of large language models. The primary aim is to fine-tune model outputs to align better with human (or expert) preferences, creating outputs more suited to real-world applications. Let's assume a preference dataset defined as $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1,\ldots,|\mathcal{D}|}$, where each $x^{(i)}$ is prompt, and $y_w^{(i)}$ and $y_l^{(i)}$ represent the preferred and less preferred responses, respectively. We denote $y_w \succ y_l \mid x$ to signify that y_w is preferred over y_l for prompt x.

Since the true distribution of human preferences cannot be directly observed, we approximate it with a latent reward model $r^*(x, y)$, assuming that higher rewards correspond to stronger preferences. Following the approach by Rafailov et al. [39], we can model the human preference distribution p^* using the Bradley-Terry (BT) model [3]:

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} = \sigma(r^*(x, y_1) - r^*(x, y_2)),$$

where σ denotes the logistic function.

To estimate the parameters of the reward model, we can apply maximum likelihood estimation by minimizing the negative log-likelihood:

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))],$$

where r_ϕ is a parameterized reward model. This approach allows us to approximate the preference distribution and fine-tune the model to capture human-like preferences effectively.

The traditional DPO algorithm operates in an offline setting. During DPO training, as model parameters continuously change, the preference dataset generated offline can gradually diverge from the model's current distribution, which weakens the effectiveness of the DPO algorithm. To address this issue, we employ an iterative DPO algorithm. By conducting multiple rounds of DPO training and sampling, this approach enables the model to better approximate an online setting during training, thus further enhancing its performance. Specifically, our approach involves training a sequence of models $M_1, \ldots, M_T$, where each subsequent model M_{t+1} utilizes preference data $\mathcal{D}_t$ generated by the t-th model. We apply this complete training process to the fine-tuned reasoning agent, enabling the model to better align with human preferences and produce structured, detailed reasoning steps for complex questions, which supports the summary agent more effectively.

4. Experiments

We conduct extensive experiments across multiple vision-language benchmarks to validate the effectiveness of our method. In this section, we first introduce the implementation details of Insight-V in Section 4.1. Then we present a comparison with state-of-the-art MLLMs, outlining the primary results of our method on visual reasoning tasks as well as additional results on general image understanding in Section 4.2. Moreover, we offer further analytical experiments and essential ablation studies on design choices in Section 4.3, along with qualitative results shown in Section 4.4 for more insights.

4.1. Implementation Details

We integrate the Insight-V system with various MLLMs to demonstrate the broad applicability of our approach. Our initial implementation with Insight-V on LLaVA-NeXT-LLaMA3 [22] illustrates the method's effectiveness. To further validate its generalizability and establish a solid baseline against state-of-the-art MLLMs, we additionally train a base multi-modal model using the Qwen-2.5-7B [41] LLM. During pretraining, we utilize the 558K captioning dataset from LLaVA-1.5 [21], unfreezing the connector parameters. This is followed by supervised fine-tuning with a curated dataset of approximately 4 million images, using a learning rate of 2e-5 as guided by prior researches [21, 22]. This two-stage training process equips the baseline model with essential visual perception abilities, achieving competitive results on vision-language benchmarks.

We then initialize two agents from the baseline model, performing targeted fine-tuning to obtain the final agents. For the reasoning agent, we compile a dataset of 200K images and train the model over 2 epochs with a learning rate of 5e-6. For the summary agent, we use a dataset of 1.2 million images, applying a learning rate of 1e-5 and train-

Table 1. **Results on Visual Reasoning Tasks.** We conduct evaluation experiments across 7 benchmarks, covering both general reasoning and task-specific reasoning assessments. Insight-V exhibits notable effectiveness and generalizability when applied to LLaVA-NeXT and our baseline model, surpassing other state-of-the-art MLLMs by a large margin.

Model	Size	MMMU	MMMU-Pro	MMBench	MME	ChartQA	MMStar	MathVista	Average
DeepSeek-VL [29]	7B	35.4	-	73.5	-/-	59.1	37.1	36.1	-
VILA-1.5 [20]	8B	38.6	-	75.3	1634.9/-	-	39.7	-	-
Cambrian-1 [43]	8B	42.7	-	75.9	1547.1/-	73.3	-	49.0	-
InternLM-XComposer2 [7]	7B	41.1	-	77.6	2220.4	71.8	56.2	59.5	-
POINTS [26]	7B	51.4	-	78.0	2184.1	-	60.9	63.0	-
IXC-2.5 [55]	7B	42.9	-	79.4	2233.1	82.2	59.9	**63.7**	-
Bunny-LLaMA3 [12]	8B	43.4	-	77.2	1588.9/321.1	-	-	34.4	-
MM-1.5 [54]	7B	41.8	-	-	1514.9/346.4	78.6	-	47.6	-
MiniCPM-LLaMA3-V 2.5 [49]	8B	45.8	19.6	77.2	2024.6	-	51.8	54.3	-
MiniCPM-V-2.6 [50]	7B	49.8	**27.2**	78.0	2268.7	-	57.5	60.6	-
Qwen2-VL [38]	7B	**53.7**	-	81.0	-	**83.0**	60.7	61.4	-
Idefics3-LLaMA3 [14]	8B	46.6	22.9	77.5	1937.4	74.8	55.9	58.4	48.1
Ovis1.5-LLaMA3 [31]	8B	48.3	23.6	76.6	1948.5	76.4	57.3	63.0	49.4
LLaVA-NeXT-LLaMA3 [22]	8B	36.9	13.2	72.3	1611.1/346.0	69.4	43.1	45.9	40.2
+ Multi-Agent	8B	40.8	17.8	77.6	1603.7/469.3	74.6	52.6	47.4	44.5
+ Iterative DPO (**Insight-V-LLaVA**)	8B	42.0	21.0	81.7	1583.9/485.4	77.4	57.4	49.8	**47.2** (+7.0)
Our Base Model	7B	47.1	22.6	81.3	1573.7/482.5	75.7	57.0	56.9	48.7
+ Multi-Agent	7B	49.7	23.8	82.2	1662.2/**629.3**	81.2	58.6	58.7	50.7
+ Iterative DPO (**Insight-V**)	7B	50.2	24.9	**82.3**	**1685.1**/627.0	81.5	**61.5**	59.9	**51.6** (+2.9)

Table 2. **Results on other multimodal benchmarks.** Insight-V enhances reasoning capabilities without compromising general visual perception and even achieves improvements on benchmarks requiring perception ability more.

Model	TextVQA	DocVQA	OCRBench	AI2D
LLaVA-NeXT-LLaMA3	65.2	78.2	553	71.5
+ Multi-Agent	68.9	81.8	631	75.7
+ Iterative DPO (**Insight-V-LLaVA**)	**70.5**	**82.9**	**663**	**77.3**
Our Base Model	75.4	90.2	713	79.7
+ Multi-Agent	**77.0**	91.4	**738**	**80.1**
+ Iterative DPO (**Insight-V**)	76.8	**91.5**	735	79.8

ing for 1 epoch. Additionally, we apply Direct Preference Optimization (DPO) to the reasoning agent, using approximately 15K preference data and training for 1 epoch at a learning rate of 5e-7. This DPO process is iteratively conducted across 3 rounds by using the model from the previous stage to generate preference data, thus enhancing the agent's reasoning capabilities. The lmms-eval [16] is utilized for fast evaluation. Further training details are provided in the appendix.

4.2. Main Results on Visual Reasoning

Setup. We present comprehensive experimental results on various visual reasoning benchmarks, which require the model to handle complex and challenging questions, demonstrating strong reasoning capabilities. We select several well-known and representative benchmarks, covering comprehensive evaluation, chart understanding, and mathematical problem-solving. MMMU [52] and MMMU-Pro [53] evaluate the model's expert-level perception and reasoning abilities across various topics. MMBench [24] is a widely used benchmark for comprehensive MLLM evaluation on images. MME [10] includes 14 challenging subtasks for assessing

Table 3. **Ablations on the Insight-V Design Choice.** The multi-agent design outperforms other configurations, highlighting the critical role of reasoning and summarization decomposition.

Model	MMMU	ChartQA	MathVista	MMStar	Avg
Baseline	47.1	75.7	56.9	57.0	59.2
Vanilla - Direct SFT	47.0	79.2	57.6	58.4	60.6
Multi-Turn Supervised	48.1	79.6	57.9	58.2	61.0
Summary Agent Only	47.5	76.3	57.3	57.9	59.8
Multi-Agent	49.7	81.2	58.7	58.6	62.1

visual perception and cognition. ChartQA [33] focuses on logical reasoning with charts, while MathVista [30] assesses problem-solving skills in math, using GPT-4-0613 [35] as the evaluator, following standard practice. MMStar [4] tests a range of tasks with varying difficulty levels.

Main Results. The experimental results in Table 1 confirm the effectiveness and generalizability of Insight-V on visual reasoning tasks. Applied to LLaVA-NeXT and our baseline model, the Insight-V system substantially boosts performance across challenging visual reasoning benchmarks, achieving competitive outcomes against state-of-the-art methods. We report an average improvement across all benchmarks: Insight-V yields an average increase of 4.9% and 3.2% when applied to LLaVA-NeXT and the base model, underscoring the system's effectiveness. Following Direct Preference Optimization, an additional improvement of 2.6% and 1.0% was observed, illustrating the value of reinforcement learning algorithms in enhancing reasoning capabilities. Specifically, on general reasoning benchmarks, MMMU, MMMU-Pro, and MMBench, we noted gains of 3.1%, 2.3%, and 1.0%, respectively. The highest improvement was on MME, with a 9.1% increase, highlighting Insight-V's strength in tasks requiring both perception and

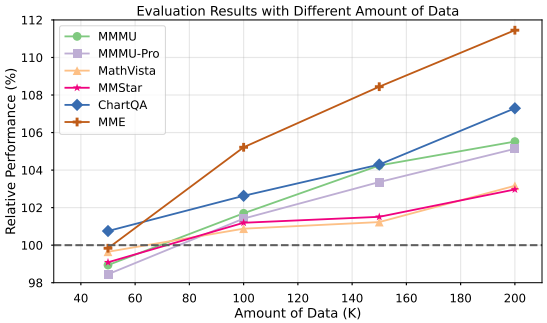

Figure 4. **Ablations on the amount of training data.** The reasoning agent benefits from data scaling, providing more valuable insights for the summary agent.

reasoning. On ChartQA, which entails complex chart-based reasoning, we achieved a 5.8% increase. Performance gains on MMStar and MathVista, with improvements of 4.5% and 3.0%, further validate our approach's versatility in tackling math problems that demand sophisticated reasoning and numerical calculation skills. This strong performance across all benchmarks underscores the efficacy of Insight-V and its potential to advance visual reasoning capabilities.

Results on Other Multi-modal Benchmarks. To verify that Insight-V enhances reasoning capabilities without compromising basic visual perception skills, we conduct experiments on benchmarks requiring only fundamental image understanding. We selected TextVQ [40], DocVQA [34], OCRBench [25], and AI2D [13] for this purpose, as these benchmarks focus on basic visual interpretation rather than complex reasoning.

As shown in Table 2, integrating Insight-V enables models to maintain strong performance across general visual perception benchmarks. Additionally, applying Insight-V yields performance improvements on both LLaVA-NeXT and our base model, as it enhances general visual understanding by guiding the model's attention to relevant regions, thereby improving visual perception accuracy. Following DPO, the model demonstrates advanced reasoning capabilities while achieving competitive results on standard image understanding benchmarks compared to baseline models, underscoring Insight-V's robustness. Consistent results of both models further confirm Insight-V's generalizability.

4.3. Further Analysis

In this section, we present comprehensive experiments to validate the design choices of Insight-V, emphasizing our approach's key contributions. Additionally, we include a case study to further demonstrate the qualitative effectiveness of Insight-V.

Effectiveness of Multi-agent System. To evaluate the effectiveness of the multi-agent system, we comprehensively compare Insight-V with alternative design choices. We begin by reporting the performance of the summary agent without a reasoning process to underscore the importance of the reasoning agent. For a comprehensive comparison, we also restructure the collected data using a Chain-of-Thought (CoT) template and train a model to perform sequential reasoning and subsequently answer questions. Additionally, we enable a model that can handle multi-turn conversations, facilitating reasoning and summarization in a multi-turn format. We use MMMU, ChartQA, MathVista, and MMStar as representative benchmarks to assess the performance across different methods.

Table 4. **Ablations on the DPO training strategy.** Iterative DPO progressively enhances the model's reasoning capabilities, leading to improved performance.

Model	MMMU	ChartQA	MathVista	MMStar	Avg
Insight-V (Multi-Agent)	49.7	81.2	58.7	58.6	62.1
+ RLAIF	49.5	81.4	59.1	59.2	62.3
+ DPO	50.8	80.8	59.3	59.9	62.7
+ Iterative DPO	50.2	81.5	59.9	61.5	63.3

As shown in Table 3, the results indicate that the multi-agent system plays a critical role in enhancing the system's visual understanding capabilities. Using only the summary agent without a reasoning process results in limited improvements in reasoning tasks, as the model lacks the necessary reasoning framework for optimal performance. Training the model to perform chain-of-thoughts reasoning, which is denoted as Insight-V (Vanilla-Direct SFT), yields modest gains, as this approach does not sufficiently emphasize critical judgments within the reasoning process, merging reasoning and summarization into a single task. Furthermore, models trained with multi-turn conversations still produce sub-optimal results, underscoring the importance of a multi-agent system that separates reasoning and summarization.

Data Scaling Law of Reasoning Agent. To assess the effectiveness of the reasoning agent, we perform ablation experiments on the data volume for training. As shown in Figure 4, we compare reasoning agents trained on varying data sizes from 50K to 200K. For a fair and comprehensive comparison, we employ the same summary agent across evaluations and report results on six benchmarks. The findings clearly indicate that the reasoning agent benefits from increased data. With limited data, the reasoning agent struggles to generalize and fails to provide useful input to the summary model, resulting in performance even worse than baseline models. In contrast, training in larger datasets improves the capabilities of the reasoning agent, allowing it to perform step-by-step reasoning and provide valuable insights that help the summary agent solve tasks effectively.

Effects of RL Algorithms. We assess DPO algorithms to identify optimal strategies for enhancing the reasoning process. To examine the effects of dataset composition on DPO

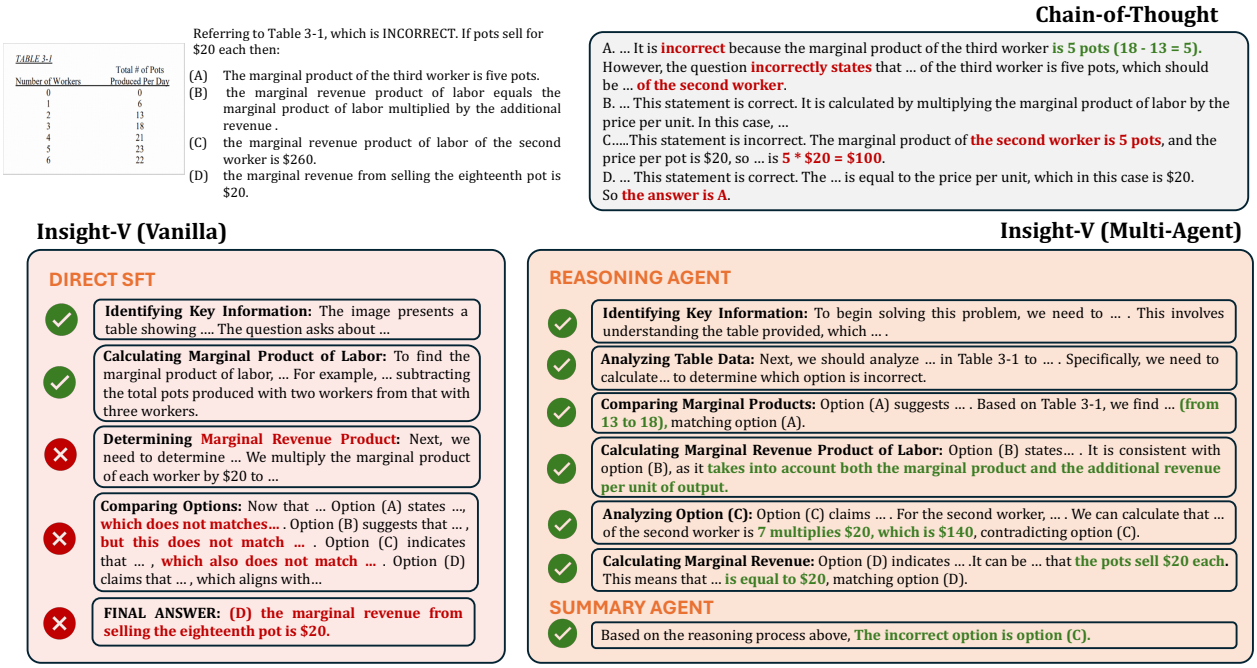

Figure 5. **Qualitative Results of Insight-V.** We present qualitative comparisons of Insight-V with Chain-of-Thought and Insight-V with direct SFT (Vanilla). For the Insight-V system, the reasoning agent delivers a more coherent and structured reasoning process, guiding the summary agent to the correct answer, whereas other methods struggle with complex reasoning and fail to solve such challenging problems.

training, we compare our curated dataset against the widely-used RLAIF-V [51] dataset, which contains 80K DPO data pairs for alignment. We also investigate the potential benefits of iterative DPO in advancing the model's reasoning capabilities. For an unbiased comparison, we subsample the RLAIF-V dataset to approximately 15K data, aligning it with the size of our dataset. The results, summarized in Table 4, show an average improvement of approximately 0.2% on the evaluated benchmarks when trained with the RLAIF-V dataset, whereas our curated dataset achieves greater gains of 0.6%. This suggests that a DPO dataset based on model-generated rationales, rather than externally sourced data, more effectively boosts reasoning accuracy. Moreover, conducting two additional rounds of DPO training using the same methodology yields further performance gains of 0.6%, underscoring that iterative DPO training progressively refines the model's ability to generate precise and high-quality reasoning processes compared to a single training pass.

4.4. Qualitative Results

We present qualitative comparisons in Figure 5 to illustrate the improvements introduced by integrating the Insight-V system. Specifically, we provide an example that highlights advancements in the reasoning process. In this example, the model is challenged with a multiple-choice question based on a table. We compare our approach with a direct Chain-of-Thought application, and the model undergoes supervised fine-tuning without incorporating a multi-agent system, which is denoted as Insight-V (Vanilla). It is evident that applying Chain-of-Thought directly results in suboptimal reasoning and leads to incorrect answers. Although fine-tuning a single model yields somewhat better reasoning, the model begins to falter as the reasoning chain lengthens, ultimately arriving at incorrect answers. This outcome arises because the model is required to generate both reasoning steps and the final answer simultaneously, limiting its judgment capabilities and weakening its robustness in handling flawed reasoning chains. In contrast, employing Insight-V enables more logical step decomposition and a structured reasoning chain. The model can analyze each option step-by-step and perform detailed calculations, which the other two methods cannot achieve. The fine-tuned summary model is able to evaluate the reasoning process and determine whether to derive the final answer based on it, significantly enhancing system robustness and ensuring correct answers.

5. Conclusion

In this paper, we introduce Insight-V, a novel system that combines a scalable data generation pipeline for long-chain reasoning data and an effective multi-agent training pipeline to enhance the reasoning capabilities of MLLMs. By developing this system, we provide a scalable approach to model training aimed at improving reasoning performance. Our extensive evaluation across various benchmarks demonstrates the effectiveness of our approach, paving the way for equipping MLLMs with enhanced reasoning capabilities.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 2

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 2

[3] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 5

[4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 6

[5] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024. 2

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 1, 2

[7] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 6

[8] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 6

[11] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 2

[12] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024. 6

[13] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251. Springer, 2016. 7

[14] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 6

[15] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 5

[16] Bo Li. Xinrun du, yuhao dong, haotian liu, yuanhan zhang, ge zhang, chunyuan li, and ziwei liu. *Lmms-eval: Accelerating the development of large multimoal models*, 2024. 6

[17] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

[18] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2

[19] Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. Unimath: A foundational and multimodal mathematical reasoner. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7133, 2023. 2

[20] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 2, 6

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 1, 5

[22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 2, 5, 6

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 1, 2

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 6

[25] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 7

[26] Yuan Liu, Zhongyin Zhao, Ziyuan Zhuang, Le Tian, Xiao Zhou, and Jie Zhou. Points: Improving your vision-language model with affordable strategies. *arXiv preprint arXiv:2409.04828*, 2024. 6

[27] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 1, 2

[28] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024. 2

[29] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 1, 2, 6

[30] 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. *arXiv preprint arXiv:2310.02255*, 2023. 6

[31] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024. 6

[32] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. *arXiv preprint arXiv:2312.00438*, 2023. 1

[33] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 6

[34] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 7

[35] OpenAI. Gpt-4 technical report. *ArXiv:abs/2303.08774*, 2023. 6

[36] OpenAI. Hello gpt-4o — openai. *OpenAI Blog*, 2024. 1

[37] QwenTeam. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1, 4

[38] QwenTeam. Qwen2-vl: To see the world more clearly. *Wwen Blog*, 2024. 1, 2, 4, 6

[39] 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. 2, 5

[40] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 7

[41] Qwen Team. Qwen2.5: A party of foundation models, 2024. 1, 5

[42] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 1

[43] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 5, 6

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1, 2

[45] Tianyi Xiong, Bo Li, Dong Guo, Huizhuo Yuan, Quanquan Gu, and Chunyuan Li. Llava-onevision-chat: Improving chat with preference learning, 2024. 5

[46] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 2

[47] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*, 2023. 1

[48] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[49] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint 2408.01800*, 2024. 1, 6

[50] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6

[51] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 8

[52] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 6

[53] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 6

[54] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 6

[55] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang,

Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 6

[56] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024. 2, 3

[57] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024. 2, 5

[58] 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. 2, 3