CorrectFlow: On-the-Spot Correction for Multimodal Reasoning with Multi-Agent Collaboration

Anonymous CVPR submission

Paper ID 2259

Abstract

Multimodal Large Language Models (MLLMs) have shown great potential in addressing complex reasoning tasks. However, their progress is often hindered by misleading or ambiguous internal knowledge resulting from training biases. Although Chain-of-Thought (COT) reasoning and its variants have proven effective in enhancing reasoning task performance, they often fail to correct errors in intermediate steps. Similarly, judge-based methods, while useful for validating reasoning steps, frequently struggle to identify and rectify specific mistakes. To tackle these challenges, we propose ***CorrectFlow*** *framework, a novel approach consisting of two key agents. In CorrectFlow, one agent extracts knowledge from visual and textual modalities to mitigate internal knowledge errors in MLLMs, while the other serves as a multi-level agent responsible for intermediate reasoning and solution generation. The multi-level agent serves as the core of our approach and implements three key strategies: self-suspect mechanism, active-abandonment mechanism, and advanced agent takeover mechanism. Within this framework, when lower-level agents identify potential reasoning errors, they positively abandon the current thought process and delegate it to higher-level agents based on the task's complexity. By incorporating this real-time correction mechanism into a multimodal reasoning framework, CorrectFlow significantly enhances the accuracy and reliability of collaborative agents, particularly in complex reasoning scenarios. Comprehensive qualitative and quantitative experiments on widely used benchmarks demonstrate that CorrectFlow surpasses existing baseline methods, underscoring its effectiveness in improving model performance and addressing both MLLM limitations.*

1. Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) [1, 6, 17, 20, 29–31, 37, 48] have significantly propelled the fields of perception, such as object detection [42], segmentation [15, 32], and video understanding [7, 18, 41], driving the development of specialized MLLMs tailored for these downstream tasks. However, despite their success in perception, MLLMs still face substantial challenges when it comes to complex reasoning tasks, especially those scenarios involving complex and long-horizon problem-solving.

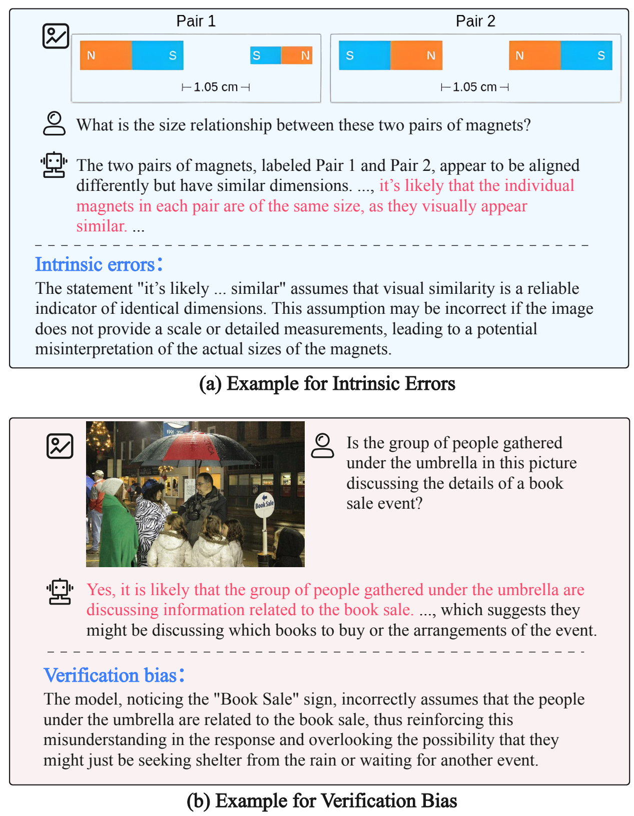

Figure 1. Examples to illustrate two limitations in MLLMs: (a) **intrinsic errors** and (b) **verification bias**.

To tackle such tasks, methods like Chain-of-Thought (CoT) reasoning [36] and its variants [10, 12, 22, 28, 44, 47] have emerged. These methods break down complex reasoning processes into smaller steps and use strategies such as self-correction and self-criticism to evaluate the accuracy of reasoning paths [36]. While these techniques can help mitigate logical errors during inference, two critical issues remain largely unaddressed: (1) **intrinsic errors** that arise from the MLLMs themselves, and (2) **verification bias**, which stems from the limited capability of MLLMs to actively correct errors and instead only verify the reasoning steps.

In Fig. 1, we illustrate both of the aforementioned errors. From Fig. 1 (a), it is evident that intrinsic bias can manifest as a form of hallucination, which is difficult to mitigate in the absence of external knowledge about the problem-solving object. When solving reasoning problems, these intrinsic errors could easily mislead the model into an erroneous reasoning path right from the first step, ultimately leading to failure in subsequent steps. Meanwhile, we also show the verification bias in Fig. 1 (b). This verification bias reveals another characteristic of MLLMs: they can detect or suspect reasoning errors, but they do not actively correct them; instead, they only verify the validity of their reasoning steps.

To overcome these challenges, we introduce **CorrectFlow**, a novel framework that leverages multi-agent collaboration to effectively address both intrinsic and verification biases. CorrectFlow features a two-agent system: the first agent, a *knowledge extractor*, gathers objective knowledge from both image and text data to provide relevant background information about the object being reasoned about. This knowledge is essential, as MLLMs can become confused, especially when dealing with visually similar objects or insufficient background context. By leveraging the knowledge extractor, CorrectFlow mitigates internal errors during the initial reasoning steps.

However, reasoning tasks often demand more than just knowledge extraction; they require continuous validation and correction. This is where CorrectFlow's *multi-level agent system* comes into play. The system introduces three key strategies to address verification biases: (1) Confidence Check, (2) Path Pruning, and (3) Expert Intervention. These strategies ensure that the MLLM performs self-evaluation, expands reasoning paths, and receives real-time corrections when necessary. In CorrectFlow, the highest-level agent evaluates the intermediate reasoning path from the root to the current step, classifying it as accurate, erroneous, or uncertain. Unlike previous Chain-of-Thought (CoT) methods and their variants, CorrectFlow introduces a novel mechanism: when a lower-level agent experiences self-doubt or detects potential errors in the reasoning path, a higher-level agent takes over, redirecting the reasoning process. This approach ensures more robust and reliable reasoning. In summary, our contributions are as follows:

- We present CorrectFlow, a novel multi-agent collaboration framework designed to overcome the limitations of a single MLLM in mitigating intrinsic errors and verification biases. By separating knowledge extraction from reasoning validation, CorrectFlow enhances robustness and minimizes internal reasoning errors.
- Three core strategies are proposed: (1) Confidence Check which stimulates the MLLM's ability to self-evaluate; (2) Path Pruning to facilitate the expansion of thought paths; and (3) Expert Intervention for providing real-time correction for reasoning paths. These strategies work collectively to ensure reliable and accurate reasoning.
- CorrectFlow pioneers a dynamic escalation mechanism that enables lower-level agents to transfer control to higher-level agents upon identifying potential reasoning errors, thereby surpassing traditional passive validation methods. This active intervention leads to more refined and robust reasoning outcomes.
- We extensively evaluate CorrectFlow on public benchmark datasets, demonstrating its superior performance compared to existing baseline methods, effectively addressing both intrinsic biases and verification limitations in multimodal reasoning tasks.

2. Related Work

Multi-modal Large Language Model. Since the advent of large language models (LLMs), their remarkable success across numerous language-based applications has inspired the development of multimodal large language models (MLLMs). These models aim to bridge the gap between vision and language modalities, enabling richer understanding and reasoning across both domains. In early research, MLLMs are regarded as a special way to extend the capabilities of LLMs to handle diverse tasks and modalities, by connecting specialized vision models. These models mainly include MiniGPT [3, 48], VisualChatGPT [37], HuggingGPT [30], LMDrive [29], and MM-REACT [38], which integrate LLMs with vision models to facilitate complex interactions between visual and textual information. Recently, the focus of MLLMs has shifted towards aligning visual and language representations more effectively. This has been accomplished through extensive training on datasets consisting of image-caption pairs or image-question dialogues. Two main effective approaches have been proposed. The first approach, LLaVA [20], trains an MLP projector to map image tokens to a representation space aligned with pre-trained LLMs, fostering effective modality integration. The second approach, BLIP-2 [17], utilizes a query transformer (Q-Former) to learn image embeddings by employing learnable queries after extracting image features. Besides the model architecture, a two-stage training strategy has been explored and become a popular approach for MLLMs [1, 6, 31, 48]. In the first stage,

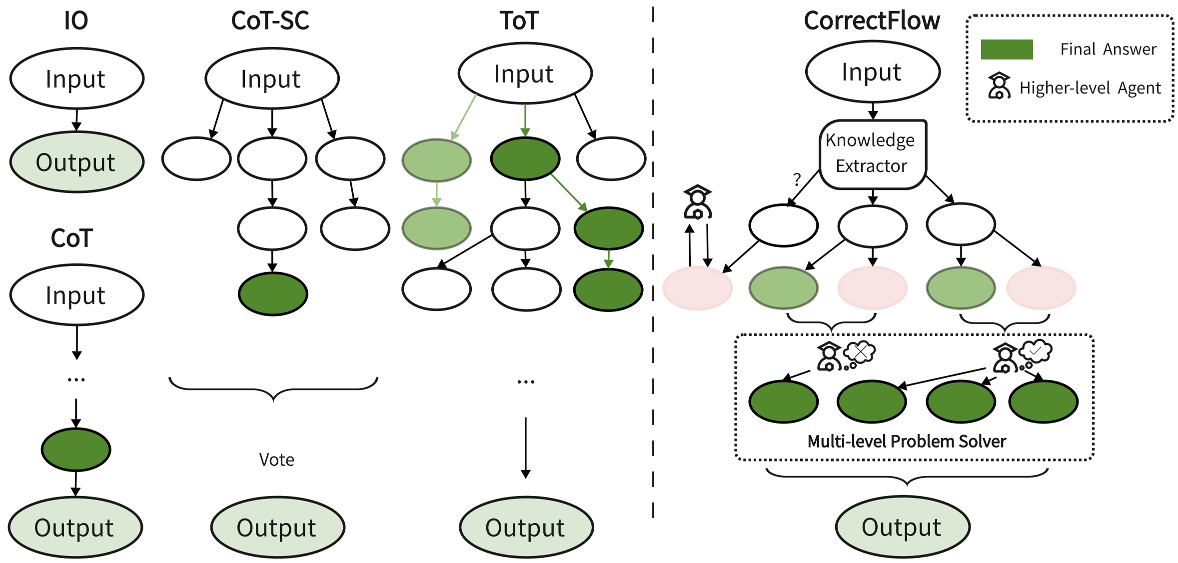

Figure 2. CorrectFlow is a multi-agent collaboration framework that enhances the robustness of MLLMs by separating knowledge extraction from reasoning validation, thereby mitigating intrinsic errors and verification biases. It introduces three core strategies—Confidence Check, Path Pruning, and Expert Intervention—that collectively ensure reliable and accurate reasoning. Compared to CoT and its variants, CorrectFlow features a dynamic escalation mechanism, allowing lower-level agents to transfer control to higher-level agents upon detecting reasoning errors, resulting in more refined and robust outputs.

the models undergo pre-training using large-scale image-caption datasets, laying the foundation for cross-modal understanding. The second stage focuses on refining alignment between modalities using question-answering triplets to ensure nuanced understanding and precise reasoning. With the model structure and training strategies, MLLMs have achieved promising performance in various perception tasks, including fine-grained localization [15, 32], such as object detection [42], video understanding [7, 18, 41], and image generation [13, 27]. Although MLLMs have shown promising results in perception tasks, they still face significant challenges in reasoning tasks, which stem not only from limitations in their perception capabilities but also from biases inherent in the models themselves, leading to misunderstandings.

CoT Reasoning in LLMs and MLLMs. Recent studies have proven using Chain-of-Thought (CoT) reasoning to improve problem-solving skills. CoT prompts encourage LLMs to express intermediate reasoning steps, which considerably enhances their reasoning ability. Studies such as [36] and [14] have demonstrated that simple prompting techniques or a few detailed examples can significantly enhance the reasoning performance of LLMs in both zero-shot and few-shot scenarios. The type methods mainly current research focuses on optimizing these methods through a more refined selection of examples based on factors like similarity, diversity, and complexity [10, 22, 28, 44], while also incorporating structured approaches, including programming [5], problem decomposition [12, 47], and rationale calibration [33].

Similar to LLMs, Chain-of-Thought (CoT) prompting has also shown significant effectiveness in enhancing the performance of multimodal large language models (MLLMs). For example, [45] leveraged visual inputs to generate relevant rationales, thereby improving the model's reasoning capabilities. [46] approached the problem by breaking down questions into sub-questions and utilizing answers from a visual question answering (VQA) model to develop rationales. In addition, [40] directed the model to solve complex questions involving multiple image inputs by assessing similarities and differences across the images. Moreover, [24] was a pioneer in using LLMs to generate scene graphs, subsequently using these models to formulate answers.

3. CorrectFlow

Here we present **CorrectFlow**, an innovative zero-shot prompting approach that leverages a multi-agent framework to enhance the reasoning capabilities of Multimodal Large Language Models (MLLMs) in tackling complex tasks. CorrectFlow enables zero-shot learning solely through prompts, bypassing the need for annotated data for fine-tuning. The core idea is to coordinate multiple MLLM agents to dynamically correct erroneous reasoning paths, thereby broadening and deepening the reasoning process.

Briefly reviewing MLLMs, these models utilize a pre-

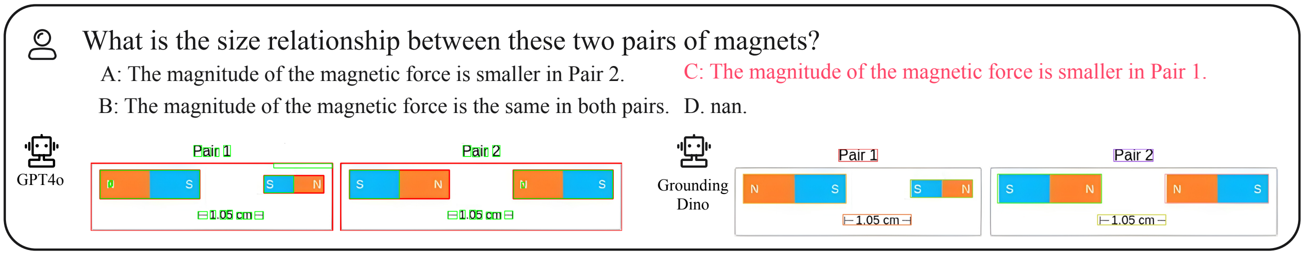

Figure 3. Open-world detectors often struggle to accurately identify relevant objects in images.

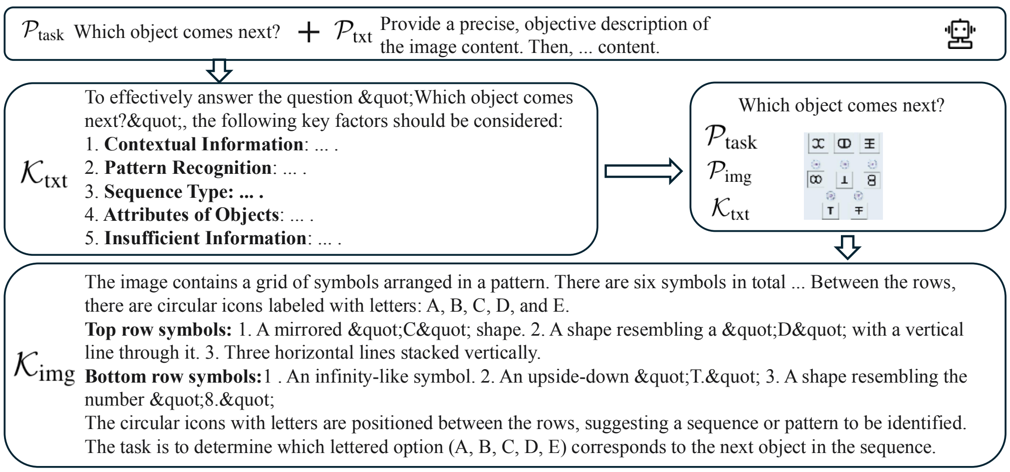

Figure 4. The pipeline of our knowledge extractor.

trained vision encoder $\phi_{\mathbf{w}}(\cdot)$, parameterized by $\mathbf{w}$, to convert an image $\mathbf{I}$ into an embedding, and a language encoder $\psi_{\mathbf{o}}$, parameterized by $\mathbf{o}$, to encode the task prompt $\mathcal{P}_{\text{task}}$ (e.g., a question or caption request). These embeddings are then fed into a pretrained language model $f_{\boldsymbol{\theta}}$, parameterized by $\boldsymbol{\theta}$, to generate a response $\mathcal{R}$:

$$\mathcal{R} = f_{\boldsymbol{\theta}}(\phi_{\mathbf{w}}(\mathbf{I}), \psi_{\mathbf{o}}(\mathcal{P}_{\text{task}})).$$

As the vision encoder $\phi_{\mathbf{w}}(\cdot)$ has been pretrained on paired image-text descriptions, it maps visual input into a shared text embedding space, enabling the LLM $f_{\boldsymbol{\theta}}$ to reason seamlessly within a unified embedding space.

3.1. Knowledge Extractor

As shown in Fig. 1 (upper), MLLMs often display internal errors when reasoning through complex and incomplete tasks. These errors likely stem from training biases that favor certain reasoning paths, leading MLLMs to overlook parts of the user's problem and produce incomplete or incorrect conclusions. Additionally, MLLMs commonly struggle to accurately detect and relate all relevant objects in an image, further complicating reasoning.

To address this, a straightforward solution would be to extract all objects relevant to the user's query to help MLLMs understand object relationships in context. However, in practice, open-world detectors frequently fail to identify these objects accurately as illustrated in Fig. 3. Even when successfully detected, linking these objects to the user's question presents a major challenge, as incorrect initial relationships between objects established during the initial reasoning step can lead to cascading errors in reasoning.

To mitigate these challenges, we introduce a dedicated "knowledge extractor" agent, designed to provide MLLMs with objective contextual knowledge drawn from the image and task prompt. Fig. 4 illustrates this process. This agent gathers supplementary information to guide the reasoning process, enabling a more comprehensive understanding of the user's question. The process begins with extracting key information from both the task prompt $\mathcal{P}_{\text{task}}$ and image $\mathbf{I}$. Given $\mathcal{P}_{\text{task}}$ and a textual key point generation prompt $\mathcal{P}_{\text{txt}}$, we first derive textual key points $\mathcal{K}_{\text{txt}}$ directly from $\mathcal{P}_{\text{task}}$. Next, we use $\mathcal{K}_{\text{txt}}$ and an image key point generation prompt $\mathcal{P}_{\text{img}}$ to identify objective facts $\mathcal{K}_{\text{img}}$ within the image $\mathbf{I}$, thus isolating each modality to avoid cross-modal interference:

$$\begin{aligned}\mathcal{K}_{\text{txt}} &= f_{\boldsymbol{\theta}}(\psi_{\mathbf{o}}(\texttt{cat}(\mathcal{P}_{\text{task}}, \mathcal{P}_{\text{txt}}))), \\ \mathcal{K}_{\text{img}} &= f_{\boldsymbol{\theta}}(\phi_{\mathbf{w}}(\mathbf{I}), \psi_{\mathbf{o}}(\texttt{cat}(\mathcal{P}_{\text{img}}, \mathcal{K}_{\text{txt}}, \mathcal{P}_{\text{task}}))),\end{aligned} \tag{1}$$

where the operation cat denotes concatenating the inputs.

This knowledge extractor agent systematically identifies the query's objects and extracts objective attributes related to them from both the image and text. This enriched context allows the MLLM to holistically interpret the problem, focusing on the full image rather than isolated objects or partial conclusions. By integrating this knowledge, the MLLM can better align its reasoning steps with the intended solution, significantly reducing the chance of errors in initial reasoning steps.

3.2. Multi-level Problem Solver

At the core of our CorrectFlow framework is a novel multi-level problem-solving approach that enhances the robustness and reliability of reasoning in complex tasks. Inspired by automated curriculum learning, we introduce a hierarchical structure consisting of multiple low-level problem-solving agents and a high-level problem-solving agent. In this framework, the low-level agents act as "students," while the high-level agent functions as a "teacher," guiding the reasoning process in real time. This relationship allows students to leverage the teacher's higher expertise, ensuring more accurate and efficient reasoning. The motivation behind this setup is intuitive: when students encounter uncertainty, they can request help from the teacher to resolve potential errors. This teacher-student dynamic mirrors real-world learning, where a teacher helps correct misunderstandings, fostering more accurate decision-making and deeper understanding.

In our multi-level problem-solving framework, we adopt an O1-inspired approach, combining self-correction and cross-validation. At each reasoning step, multiple independent paths are generated, allowing the system to evaluate the state from different perspectives or knowledge sources. This multi-perspective validation enables a thorough reassessment, comparing outcomes across paths. When significant discrepancies are detected, the system analyzes and adjusts the reasoning to correct biases or errors. While theoretically effective, MLLMs often fail to identify the root causes of mistakes, allowing errors to propagate along the reasoning path.

To overcome this limitation, we introduce three pivotal strategies at each step of reasoning: (1) Confidence Check, (2) Path Pruning, and (3) Expert Intervention. These strategies work together to ensure robustness, accuracy, and adaptability in the reasoning process.

Confidence Check. It enables the reasoning agent to self-assess the validity of each reasoning step. When an agent detects potential flaws or inconsistencies, it generates a "self-suspect" signal. This signal prompts further investigation or assistance from a higher-level agent. This process is inspired by iterative questioning, where doubts lead to deeper scrutiny, ensuring more reliable conclusions.

Path Pruning. It discards the unreliable path before errors can propagate when an agent is uncertain about the correctness of a reasoning path. This ensures the system only progresses along valid reasoning paths, preventing the system from getting stuck or moving forward with flawed reasoning.

Expert Intervention. When a self-suspect signal or error is detected, control is escalated to the high-level agent, the "teacher," which performs a more thorough analysis of the reasoning process. The high-level agent evaluates the reasoning path and decides on the next action:

- **Correct**: If the reasoning path is validated, no further action is needed.
- **Wrong**: If a logical error is identified, teacher agent corrects the reasoning path based on prior steps.
- **Uncertain**: If teacher agent is unsure, "Path Pruning" is triggered to discard the uncertain path.

In Appendix, we provide a detailed figure to summarize the reasoning steps in our multi-level problem-solving agents.

These mechanisms create a robust feedback loop that continuously refines the reasoning process, minimizing error propagation. The dynamic interplay between low-level agents (students) and the high-level agent (teacher) creates an adaptable, self-correcting system that is both efficient and reliable. This makes it ideal for complex problem-solving tasks where precision is critical. Our multi-level problem-solving approach bridges the gap between theoretical advancements and real-world applications, empowering agents to tackle challenging tasks with confidence and accuracy.

4. Experiment

4.1. Implementation Details

GPT-4O. The architectural and pretraining details of GPT-4O [26] are not publicly available. Nevertheless, we use GPT-4O as the MLLM backbone due to its state-of-the-art language reasoning capabilities. This allows us to evaluate the performance of our proposed method on an LMM with advanced reasoning skills, providing insights into its effectiveness in solving complex multi-step problems. In addition to using GPT-4O, we also conducted experiments with other (MLLMs) as our base models. Detailed results of these additional experiments are provided in the supplementary materials.

4.2. Multimodal Reasoning Benchmarks

The implementation of CorrectFlow has undergone rigorous evaluation using several benchmark datasets, including MME, MathVista [23], BLINK [9], MMStar [4], CCBench [21], and RealWorldQA [43]. These benchmarks are specifically designed to assess the multimodal perception and reasoning capabilities of large multimodal language models (LMMs). Both MME and MathVista feature different splits that evaluate general visual perception and reasoning. For instance, MME includes perception tasks

Table 1. Performance Comparison on Mathvista dataset.

Method	Overall	SCI	TQA	NUM	ARI	VQA	GEO	ALG	GPS	MWP	LOG	FQA	STA
LLaVA-OneVision-72B (SI)	66.9	64.8	63.3	51.4	61.2	54.2	75.3	70.8	77.4	77.4	21.6	62.1	71.4
InternVL2-Llama3-76B	65.6	63.1	66.5	41.7	62.0	49.2	66.1	65.8	67.8	75.8	32.4	67.3	76.7
Ovis1.5-Gemma2-9B	65.6	64.8	60.1	50.7	66.3	54.7	62.8	58.7	63.5	87.1	13.5	62.8	74.1
InternVL2-40B	64.0	60.7	63.3	41.7	64.9	58.7	56.5	56.9	57.2	71.0	21.6	68.4	76.7
NVLM-D-72B	63.9	66.4	68.4	40.3	50.7	41.9	76.2	73.0	78.4	65.1	16.2	63.9	71.8
InternLM-XComposer2.5	63.7	55.7	55.7	43.8	64.0	53.1	62.8	56.9	62.0	83.3	16.2	63.2	73.4
Ovis1.5-Llama3-8B	63.0	63.1	60.8	48.6	65.7	57.5	61.1	57.3	61.1	79.6	21.6	58.0	67.1
POINTS-Qwen2.5-7B	63.0	61.5	61.4	49.3	61.5	58.1	70.7	66.5	72.6	71.0	13.5	54.3	63.5
POINTS-Yi-1.5-9B	63.0	61.5	59.5	46.5	61.5	55.9	70.3	66.2	72.1	73.1	10.8	55.8	65.8
LLaVA-OneVision-7B	62.3	65.6	60.8	45.1	57.5	47.5	68.6	64.1	70.2	76.9	16.2	56.9	66.1
Claude3.5-Sonnet	61.6	75.4	74.1	31.2	53.5	45.8	58.6	61.6	57.7	59.1	35.1	69.5	77.7
RBDash-v1.2-72B	61.6	59.8	65.8	40.3	53.3	41.9	68.2	66.9	69.7	74.2	24.3	57.2	69.1
Qwen2-VL-7B	61.4	66.4	63.3	41.0	58.9	57.0	51.0	51.6	51.0	66.1	27.0	68.0	73.8
GPT-4o (0806, high)	62.7	71.3	75.3	42.4	56.9	48.0	65.7	68.3	65.9	68.3	32.4	58.7	69.1
CorrectFlow (GPT-4o$_{(0806,\ high)}$)	67.0	72.1	73.4	49.3	63.2	55.9	69.5	70.1	70.2	72.0	37.8	64.7	74.4

Table 2. Performance Comparison of reasoning task on MME dataset.

Method	Overall	Code Reasoning	Numerical Calculation	Text Translation	Commonsense Reasoning
Qwen-VL-Max-0809	723.9	177.5	170.0	200.0	176.4
InternVL2-Llama3-76B	658.6	152.5	185.0	162.5	158.6
NVLM-D-72B	655.7	160.0	162.5	162.5	170.7
GPT-4o (0513, low)	719.3	182.5	170.0	192.5	174.3
LLaVA-OneVision-72B	583.9	145.0	177.5	100.0	161.4
GPT-4o (0513, high)	696.1	177.5	147.5	192.5	178.6
Qwen-VL-Plus-0809	633.9	157.5	125.0	200.0	151.4
InternVL2-40B	572.1	137.5	117.5	170.0	147.1
JT-VL-Chat	608.2	145.0	170.0	132.5	160.7
Qwen-VL-Max	576.1	132.5	107.5	192.5	143.6
CongRong	521.8	97.5	100.0	185.0	139.3
MiniCPM-V-2.6	597.9	155.0	117.5	177.5	147.9
GPT-4o (0806, high)	696.4	185.0	147.5	192.5	171.4
CorrectFlow (GPT-4o$_{(0806,\ high)}$)	766.1	185.0	200.0	188.57	192.5

Table 3. Performance Comparison of perception task on MME dataset.

Method	Overall	OCR	Artwork	Color	Count	Existence	Landmark	Position	Posters	Scene
Qwen-VL-Max-0809	1585.5	177.5	156.2	190.0	170.0	200.0	183.5	155.0	189.1	164.2
InternVL2-Llama3-76B	1572.4	147.5	173.2	180.0	180.0	195.0	179.8	173.3	188.4	164.2
NVLM-D-72B	1586.2	185.0	141.8	190.0	170.0	200.0	179.5	168.3	187.1	164.5
GPT-4o (0513, low)	1562.7	192.5	144.0	180.0	190.0	195.0	175.5	145.0	192.2	148.5
LLaVA-OneVision-72B	1570.5	162.5	153.2	185.0	170.0	200.0	178.8	178.3	183.7	159.0
GPT-4o (0513, high)	1546.2	192.5	145.2	185.0	185.0	185.0	182.0	133.3	191.2	147.0
Qwen-VL-Plus-0809	1513.3	155.0	150.0	180.0	158.3	180.0	185.0	160.0	182.0	163.0
InternVL2-40B	1565.0	162.5	170.0	188.3	180.0	190.0	180.2	153.3	189.5	151.2
JT-VL-Chat	1535.0	117.5	161.5	185.0	170.0	195.0	185.0	173.3	184.7	163.0
Qwen-VL-Max	1528.6	177.5	150.2	168.3	160.0	190.0	191.0	140.0	187.8	163.8
CongRong	1576.8	177.5	151.0	176.7	175.0	195.0	187.2	168.3	171.1	175.0
MiniCPM-V-2.6	1519.4	192.5	149.0	168.3	160.0	195.0	177.5	146.7	177.9	152.5
GPT-4o (0806, high)	1550.3	200.0	139.5	178.3	190.0	195.0	189.2	113.3	193.5	151.5
CorrectFlow (GPT-4o$_{(0806,\ high)}$)	1540.7	192.5	148.3	190	180	195.0	148.3	153.3	190.4	143.0

that assess an LMM's ability to identify instances and understand instance attributes, as well as higher-order reasoning tasks such as scene understanding and instance interaction. MathVista, on the other hand, contains complex mathematical problems, often requiring extensive inference steps. We evaluate our method on MME, excluding the instance identification task, and on the entirety of MathVista. Additionally, we use the reasoning sets of BLINK, MMStar, CCBench, and RealWorldQA to further evaluate our approach, focusing on the LMMs' ability to provide detailed, long-form answers to visual questions.

Table 4. Performance Comparison on MMbenc benchmark.

Method	Code Reasoning	Numerical Calculation	Text Translation	Commonsense Reasoning
CoT	185.0	192.5	177.5	176.4
CoT-SC	192.5	192.5	185.0	180.7
SoT	177.5	132.5	177.5	172.1
ToT	192.5	200	177.5	172.5
CorrectFlow	185	200	188.57	192.5

4.3. Baseline

In our experiments, we compared our proposed CorrectFlow methodology with two prompting baselines. The first baseline aimed to evaluate the added value of our method to pretrained LMMs [2, 8, 11, 16, 19] by applying the model to the benchmark without any prompt engineering as shown in Table 1, Table 2, and Table 3. The second baseline utilized a zero-shot (ZS) Chain-of-Thought (CoT) prompting method to assess the benefits of CorrectFlow compared to a state-of-the-art (SOTA) CoT [35] approach. The ZS-CoT method involves two main steps: (i) given the input question and text, the reasoning prompt ”Let’s think step-by-step.” is appended after the question to guide the model in generating reasoning for an answer, and (ii) since the answer is implicitly embedded in the generated reasoning, the second step involves passing the image, question, generated reasoning, and an answer extraction phrase to produce the response in the desired format. We also compared CorrectFlow to recent SOTA multimodal CoT prompting methods, including COT-SC [34], SoT [25], and ToT [39], on the reasoning split of the MME benchmark, as summarized in Table 4.

4.4. Result

Results are presented in Table 1, Table 2, and Table 3. One notable advantage of our method is its significant improvement in performance on several multimodal reasoning tasks, including the complex MathVista benchmark. We demonstrate that applying CorrectFlow to GPT-4O outperforms the base models across various benchmarks, highlighting the effectiveness of our approach. Figure 5 provides specific examples where CorrectFlow enhances performance over the baselines, as well as instances where it still encounters challenges. Additional results can be found in the supplementary materials.

Multimodal Reasoning Tasks. CorrectFlow outperforms the baselines in the reasoning test category across RealWorldQA, BLINK, MMStar, and CCBench in Table 5, Table 6, Table 7 and Table 8. From these datasets, we observe that CorrectFlow significantly improves performance on complex tasks involving mathematical reasoning and logical understanding, particularly those requiring extended reasoning steps. Notably, our method also proves effective on Chinese datasets. These results provide strong evidence that our approach enhances LMMs’ long inference capabilities in general multimodal reasoning tasks.

Table 5. Performance Comparison on RealWorldQA benchmark.

Method	Overall
Qwen2-VL-72B	76.7
GPT-4o (0513, high)	75.4
Qwen-VL-Max-0809	74.2
LLaVA-OneVision-72B	73.9
LLaVA-OneVision-72B (SI)	73.7
Molmo-72B	73.7
InternVL2-Llama3-76B	72.7
Ovis1.6-Gemma2-9B	70.7
Qwen-VL-Plus-0809	70.1
InternVL2-40B	70.1
LLaVA-OneVision-7B	69.9
NVLM-D-72B	69.9
OmChat-v2.0-13B	69.8
Step-1.5V	69.7
LLaVA-OneVision-7B (SI)	69.5
GPT-4o (0806, high)	76.5
CorrectFlow (GPT-4o$_{(0806,\ high)}$)	77.3

Table 6. Performance Comparison on Blink benchmark.

Method	Multi-view Reasoning	Spatial Relation
Qwen-VL-Max-0809	40.6	88.1
Gemini-1.5-Pro	53.4	79.7
Phi-3.5-Vision	48.1	69.2
Gemini-1.5-Flash	57.1	77.6
InternVL2-26B	42.9	84.6
Yi-Vision	48.1	82.5
MiniCPM-V-2.6	55.6	81.1
LLaVA-OneVision-7B	54.1	80.4
LLaVA-Next-Interleave-7B	44.4	71.3
GPT-4o (0806, high)	45.1	82.5
CorrectFlow (GPT-4o$_{(0806,\ high)}$)	47.4	83.2

Table 7. Performance Comparison on MMstar benchmark.

Method	Logical Reasoning	Math
Qwen-VL-Max-0809	72.4	76.0
Qwen2-VL-72B	72.4	72.8
InternVL2-Llama3-76B	72.4	75.2
LLaVA-OneVision-72B	68.8	74.4
LLaVA-OneVision-72B (SI)	67.2	72.0
Step-1.5V	68.4	64.4
InternVL2-40B	69.2	70.0
JT-VL-Chat-V3.0	69.6	76.8
GPT-4o (0513, high)	72.0	66.4
NVLM-D-72B	68.8	70.8
Molmo-72B	65.2	60.8
GPT-4o (0806, high)	72.0	67.6
CorrectFlow (GPT-4o$_{(0806,\ high)}$)	73.6	72.4

We further conducted a comparative evaluation against Chain-of-Thought (CoT) and its variants, including CoT-SC, SoT, and ToT, within the reasoning category of the

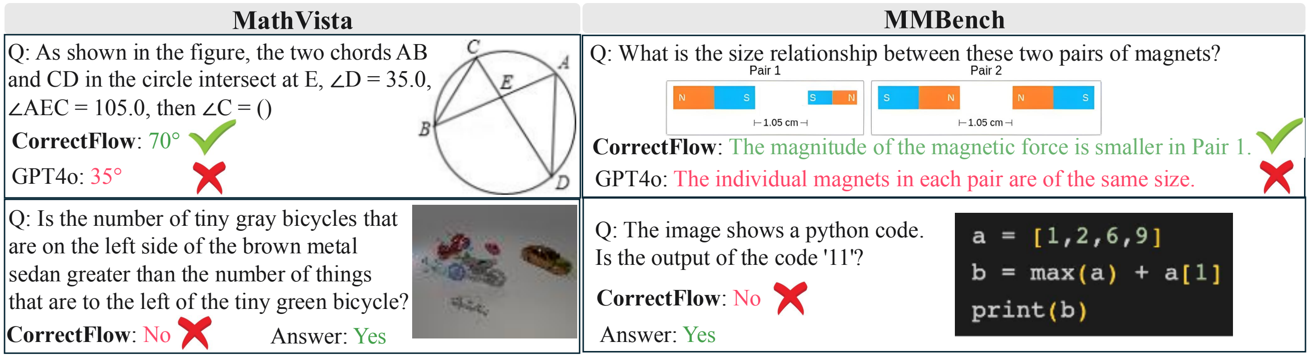

Figure 5. Comparative visualization of our method's performance on the MathVista and MMBench datasets.

Table 8. Performance Comparison on CCBench benchmark.

Method	Sketch Reasoning
InternVL2-Llama3-76B	91.1
InternVL2-8B	91.1
InternVL2-1B	86.7
BlueLM-V-3B	91.1
Step-1.5V	91.1
Qwen-VL-Max-0809	88.9
MMAlaya2	91.1
Qwen2-VL-72B	86.7
CongRong	91.1
GPT-4o (0806, high)	88.9
CorrectFlow (GPT-4o$_{(0806,\ \text{high})}$)	92.2

Table 9. Effects of Knowledge Extractor (KE) and Multi-level Problem Solver (MPS) on the MME benchmark.

Method	Code Reasoning	Numerical Calculation	Text Translation	Commonsense Reasoning
w/o-KE	185.0	200.0	188.5	176.4
w/o-MPS	185.0	192.5	185.0	180.7
Our	185.0	200.0	188.57	192.5

MMbenc benchmark. Table 4 provides a detailed performance comparison across various reasoning tasks, such as code reasoning, numerical calculation, text translation, and commonsense reasoning. This analysis demonstrates the advantages of CorrectFlow in handling complex reasoning tasks, which can be attributed to its real-time correction mechanism during agent collaboration.

Multimodal Perception Tasks. Table 3 presents the experimental results for perception tasks. From the table, we observe that CorrectFlow has minimal impact on perception task performance.

4.5. Ablation Study

We conducted a comprehensive ablation study on reasoning tasks in MME-Bench using our GPT-4o-CorrectFlow model. The study highlights the effectiveness of our knowledge extractor (KE) and multi-level problem solver (MPS), as presented in Table 9. Without KE, the model's performance dropped significantly in the commonsense category due to internal errors. Similarly, without MPS, relying solely on methods like COT-SC, the performance in several categories deteriorated, attributed to the lack of a thorough consideration of the reasoning path. For more ablation results, please refer to the supplementary materials.

4.6. Visualization Analysis

Figure 5 presents sample outputs from our method. On the left, we highlight successful cases of CorrectFlow, demonstrating its effectiveness in accurately handling complex reasoning tasks through agent collaboration. On the right, we display failure cases, offering insights into the current limitations and potential areas for improvement of our approach. For additional qualitative visualizations and a detailed analysis, please refer to the supplementary materials.

5. Conclusion

Our CorrectFlow offers a robust solution for addressing the intrinsic limitations of single MLLMs in multimodal reasoning tasks. By introducing a two-agent framework that separates knowledge extraction from reasoning validation, CorrectFlow significantly enhances accuracy and reliability. The implementation of core strategies such as the self-suspect mechanism, active abandonment, and advanced agent takeover enables dynamic intervention and escalation, effectively reducing reasoning errors and overcoming verification biases. Our extensive evaluations on several public benchmark datasets show that CorrectFlow outperforms existing methods, paving a road in the pursuit of dependable multimodal reasoning systems.

Limitation. CorrectFlow has extra computational overhead due to error correction, may impacting efficiency in rapid-response system. Future work will optimize these costly processes.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023. 7

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[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. 5

[5] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022. 3

[6] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[7] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238, 2023. 7

[9] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: multimodal large language models can see but not perceive. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIII*, pages 148–166. Springer, 2024. 5

[10] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[11] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395, 2024. 7

[12] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[13] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[14] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 3

[15] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 1, 3

[16] 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. *CoRR*, abs/2408.03326, 2024. 7

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[18] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 1, 3

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 7

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[21] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, pages 216–233. Springer, 2024. 5

[22] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[23] 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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 5

[24] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 3

[25] Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. Skeleton-of-thought: Prompting llms for efficient parallel generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 7

[26] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 5

[27] Shengju Qian, Huiwen Chang, Yuanzhen Li, Zizhao Zhang, Jiaya Jia, and Han Zhang. Strait: Non-autoregressive generation with stratified image transformer. *arXiv preprint arXiv:2303.00750*, 2023. 3

[28] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022. 1, 3

[29] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 1, 2

[30] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[31] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1, 2

[32] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[33] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[34] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 7

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 7

[36] 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, 3

[37] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 1, 2

[38] 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*, 2023. 2

[39] 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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 7

[40] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*, 2024. 3

[41] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1, 3

[42] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 1, 3

[43] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? *CoRR*, abs/2408.13257, 2024. 5

[44] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[45] 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. 3

[46] 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. 3

[47] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2