

CoMT: A Novel Benchmark for Chain of Multi-modal Thought on Large Vision-Language Models

Zihui Cheng^{1,2*}, Qiguang Chen^{3*}, Jin Zhang³, Hao Fei⁴,
Xiaocheng Feng³, Wanxiang Che³, Min Li¹, Libo Qin^{1,2†},

¹School of Computer Science and Engineering, Central South University, China

²Key Laboratory of Data Intelligence and Advanced Computing in Provincial Universities, Soochow University, China

³Research Center for SCIR, Harbin Institute of Technology, Harbin, China

⁴National University of Singapore, Singapore
czh_up@csu.edu.cn, qgchen@ir.hit.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have recently demonstrated amazing success in multi-modal tasks, including advancements in Multi-modal Chain-of-Thought (MCoT) reasoning. Despite these successes, current benchmarks still follow a traditional paradigm with multi-modal input and text-modal output, which leads to significant drawbacks such as *missing visual operations* and *vague expressions*. Motivated by this, we introduce a novel Chain of Multi-modal Thought (CoMT) benchmark to address these limitations. Different from the traditional MCoT benchmark, CoMT requires both multi-modal input and multi-modal reasoning output, aiming to mimic human-like reasoning that inherently integrates visual operations. Specifically, CoMT consists of four categories: (1) Visual Creation, (2) Visual Deletion, (3) Visual Update, and (4) Visual Selection to comprehensively explore complex visual operations and concise expression in real scenarios. We evaluate various LVLMs and strategies on CoMT, revealing some key insights into the capabilities and limitations of the current approaches. We hope that CoMT can inspire more breakthroughs on introducing multi-modal generation into the reasoning process.

Code — <https://github.com/czhzhc/CoMT>

Extended version — <https://arxiv.org/abs/2412.12932>

1 Introduction

Recently, large vision-language models (LVLMs) have achieved remarkable success across various multi-modal tasks (Liu et al. 2024b; Zhu et al. 2023; Qin et al. 2024b; Zhang et al. 2024b; Fei et al. 2024b). In addition, LVLMs have also emerged with amazing capabilities, especially the capability of chain-of-thought (CoT) reasoning, which can perform step-by-step reasoning (Lu et al. 2022; Chen et al. 2024b; Xu et al. 2024; Fei et al. 2023). Specifically, Zhang et al. (2023) first formally introduce the concept of Multimodal-CoT (MCoT) and extend it into a rationalizing-answering stages paradigm. Wang et al. (2024a) propose T-SciQ to distill the advanced large language models (LLMs)

*These authors contributed equally.

†Corresponding Author.

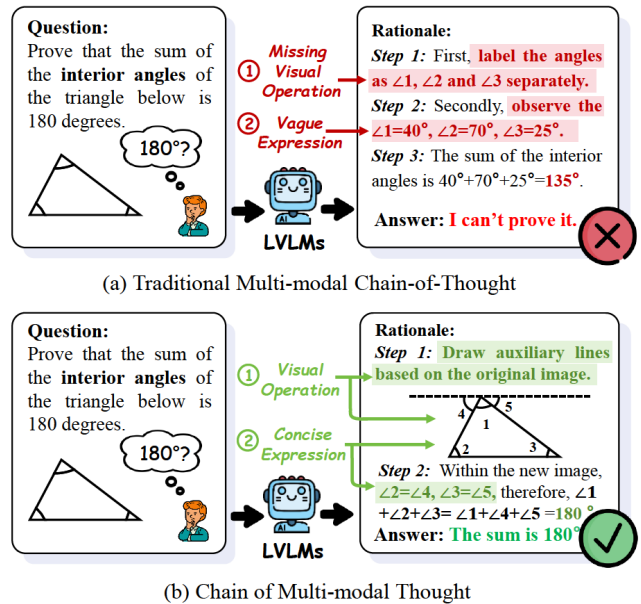


Figure 1: Comparison between (a) traditional multi-modal CoT and (b) chain of multi-modal thought, where images in rationales are needed to be generated from LVLMs to assist textual reasoning in rationale.

to smaller models for better MCoT reasoning. Building on this foundation, Zheng et al. (2024) propose DDCoT, utilizing advanced LLMs to split questions into a series of sub-questions and then answer them by LVLMs. Mondal et al. (2024) further inject the knowledge graph information into the MCoT reasoning process, reducing the hallucinations of LLMs. He et al. (2024) devise a novel latent space learning approach to acquire image features through diffusion processes, achieving more complex CoT reasoning capabilities.

While remarkable success has been witnessed in MCoT, current MCoT benchmarks still follow a traditional paradigm that reads multi-modal input but can only produce single-modal reasoning output. Such a paradigm lacks integrated multi-modal reasoning output, leading to the following issues:

Benchmark	#Question	#Image	#VO	MMCoT	MT	Avg. MT Step	Rationale
VCR (Zellers et al. 2019)	290k	99,904	✗	~4%	✗	✗	✓
A-OKVQA (Schwenk et al. 2022)	24,903	23,692	✗	~21%	✗	✗	✓
KI-VQA (Li et al. 2023b)	4,290	4,189	✗	~17%	✗	✗	✓
ScienceQA (Lu et al. 2022)	21,208	10,332	✗	~8%	✗	✗	✓
MMMU (Yue et al. 2023)	11,550	11,264	✗	~8%	✗	✗	<18%
M ³ CoT (Chen et al. 2024b)	11,459	11,293	✗	100%	✗	✗	✓
CoMT (ours)	3853	14,801	4	100%	✓	3.11	✓

Table 1: Comparison of CoMT and multi-modal related datasets.¹**#X**: the size of X; **VO**: supported visual operations; **MMCoT**: the ratio of samples with multi-step MCoT (MMCoT) in the datasets; **MT**: Multi-modal Thought. **Avg. MT Step**: The average step of Multi-modal Thought. Our benchmark has the following two advantages: (1) abundant rationale containing multi-modal thought, (2) more comprehensive and fine-grained visual operation.

- (1) **Missing Visual Operations.** *Effective multi-modal reasoning often requires visual operations.* However, traditional MCoT paradigms produce only textual reasoning outputs, which greatly hinders the multi-modal reasoning. As shown in Figure 1 (a), traditional methods can express operations in language, such as “label the angles”, but they fail to execute visual operations, omitting the actual image-processing procedure.
- (2) **Vague Expressions.** The adage “a picture is worth a thousand words” highlights the limitations of text in conveying visual reasoning conditions. As shown in Figure 1 (a), phrases like “ $\angle 1=40^\circ$ ” are imprecise in the absence of actual annotations, failing to accurately reflect the mapping relationship between angles and measures, thus leading to ambiguity and loss of visual information.

Actually, when humans perform reasoning, they naturally integrate images into the process: using visual thought for concrete, detailed reasoning while using textual thought for abstract, logical reasoning (Lehmann et al. 2010; Lin et al. 2024; Wu et al. 2024b). Take Figure 1 (b) as an example, LVLMs can accurately locate the specific angle by generating an annotated image. By labeling the angles and drawing auxiliary lines, LVLMs can perform clearer expressions and better multi-modal reasoning. Inspired by this, in this paper, we aim to explore a new MCoT paradigm that requires generating multi-modal reasoning outputs.

To fill this gap, we introduce a novel Chain of Multi-modal Thought benchmark (CoMT). Unlike the traditional MCoT benchmarks, CoMT requires both multi-modal input and multi-modal reasoning output, aiming to enhance LVLMs’ performance in concise expression and complex visual operations in real-world scenarios. Specifically, CoMT contains four categories to comprehensively assess the ability of LVLMs to use multi-modal thought processes: (1) **Visual Creation** assesses the ability to generate images from scratch, thereby visualizing abstract problems; (2) **Visual Deletion** evaluates the removal of irrelevant information from given images; (3) **Visual Update** examines the integration of updated images while retaining prior information; (4) **Visual Selection** tests the selection of specific visual fea-

tures for improved image comparison. The detailed comparisons and analyses are shown in Table 1.

We evaluate abundant representative LVLMs and prompting strategies on CoMT in extensive scenarios, yielding several **key takeaways**: (1) *CoMT presents a significant challenge; nearly all zero-shot methods perform only marginally better than random, which demonstrates huge gaps compared with human performance.* (2) *In-context learning (ICL) has better hope on triggering LVLMs for better multi-modal thought in CoMT.* (3) *Future advancements in CoMT should focus on integrating multi-modal generation, logical reasoning and visual operations into MCoT more effectively.*

Our main contributions are as follows:

- To our knowledge, this is the first work to establish a benchmark for chain of multi-modal thought (CoMT) in LVLMs, which encompasses four fundamental operations for comprehensive evaluation.
- We evaluate various representative LVLMs and prompting strategies, revealing a huge performance gap between LVLMs and humans. Except for Gemini, nearly all LVLMs perform at random chance levels.
- We explore in-context learning to enhance performance and highlight some future directions for integrating multi-modality into MCoT reasoning, hoping to provide insights for further research.

2 Benchmark Construction

We introduce CoMT², which aims to assess the ability of multi-modal thought, consisting of four types: *Visual Creation* (§2.1), *Visual Deletion* (§2.2), *Visual Update* (§2.3), and *Visual Selection* (§2.4). Specially, we design a specified question-answering template, which involves *question*, *options*, *image*, *rationale*, and *answer*, to standardize the format for all tasks within CoMT. More annotation details are shown in Technical Appendix C.

2.1 Visual Creation

An image is worth a thousand words. As shown in Figure 2 (a), visual creation tasks emphasize generating images from textual descriptions to improve multi-modal reasoning.

¹The background of traditional MCoT and our CoMT can be found in our Technical Appendix A.

²The quality assurance of CoMT can be found in Technical Appendix B.

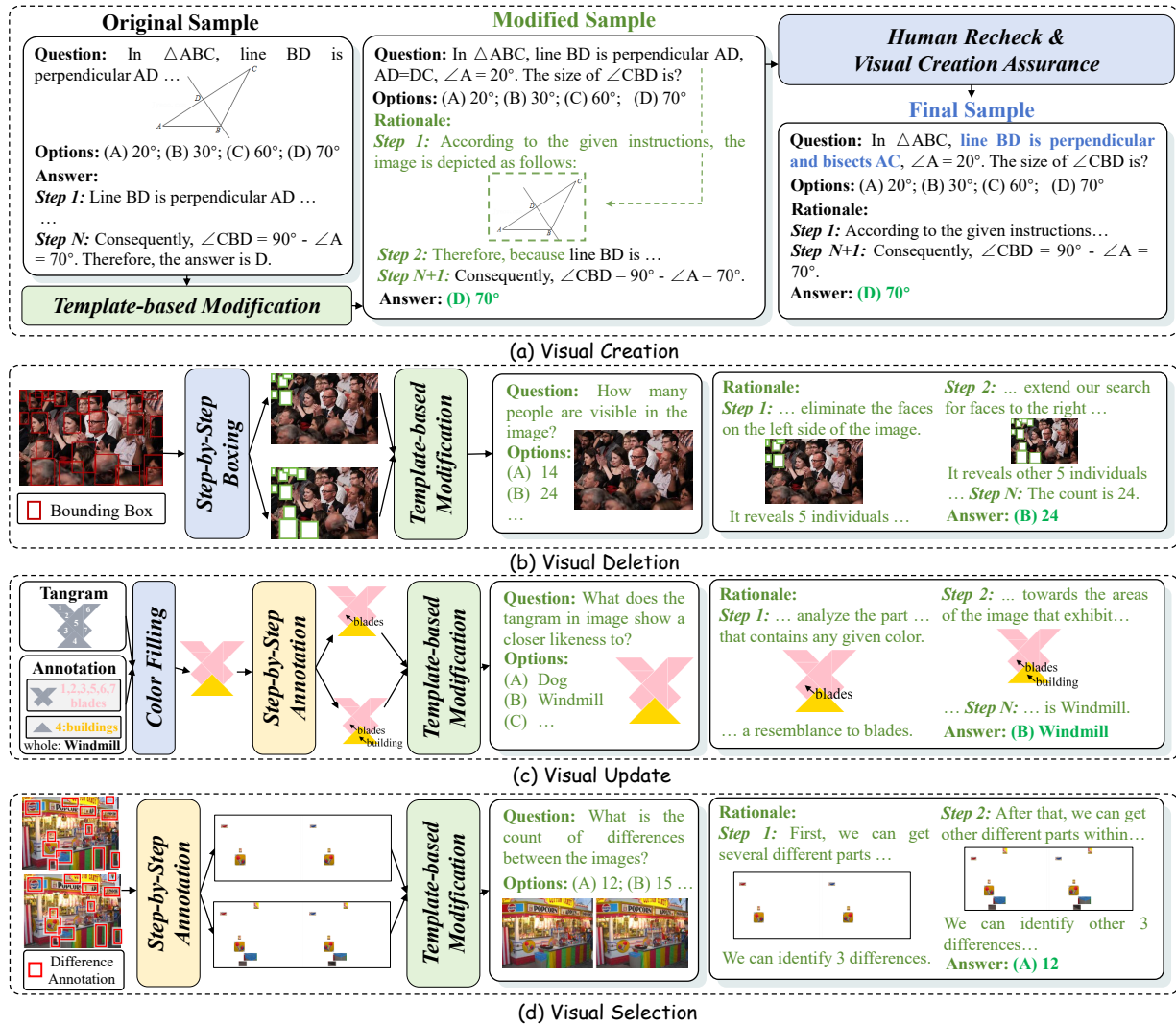


Figure 2: The overall annotation process for four tasks of CoMT, which consists of (a)visual creation, (b)visual deletion, (c)visual update, and (d)visual selection.

- **Original Dataset:** We develop visual creation tasks based on the GeoQA+ dataset (Cao and Xiao 2022), which includes geometric images and textual questions as input, with textual rationales as output.
- **Template-based Modification:** We first follow the template to modify the visual creation data. Specifically, we maintain the original question and option part from GeoQA+ and split the whole response into rationale and the final answer. Furthermore, we reposition the image from question to the output rationale as visual thought, with step information supplemented.
- **Human Recheck:** To ensure the accurate reproduction of images, we manually augment the geometric description within the question by aligning with the image details.

2.2 Visual Deletion

In logical reasoning, it is crucial to eliminate redundant information and clarify the logical chain. By progressively removing visual features, LVLMs experience reduced confusion, enabling step-by-step reasoning for the final answer, as illustrated in Figure 2 (b).

- **Original Dataset:** We utilize the crowd-counting task from the JHU-CROWD++ dataset (Sindagi, Yasarla, and Patel 2020), which includes images with numerous faces and corresponding boxing.
- **Step-by-Step Boxing:** The most crucial aspect of crowd-counting is identifying human individuals where faces serve as a significant visual feature. To demonstrate the marking and removal of redundant visual features, we batch-mask faces based on the boxing provided, preparing for the next operation.
- **Template-based Modification:** We construct the com-

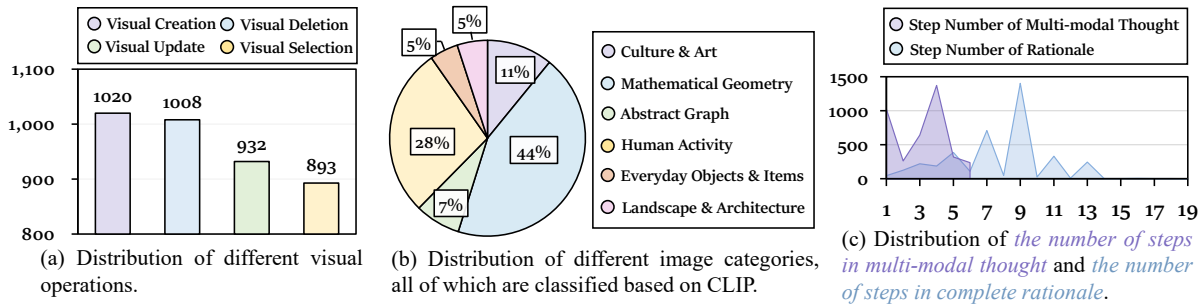


Figure 3: Distribution of CoMT tasks across four types of image processing.

plete sample by following the CoMT template, involving inquiries about the people count in the image (*question*) and clarifications of the identified count (*rationale*), etc. The prepared images serve as the visual thought within the rationale.

2.3 Visual Update

Marking can help sort out the logic. LVLMs often make mistakes in reasoning due to forgetting visual features, while humans mitigate this by annotating images. Inspired by this, as illustrated in Figure 2 (c), we propose the *Visual Update* task to annotate the images step-by-step.

- **Original Dataset:** We leverage the KILOGRAM (Ji et al. 2022) dataset to implement tangram recognition, including the tangram image and labels of both individual pieces and the whole shape.
- **Tangram Annotation:** For accurate assessments, we enhance the original tangram by applying different colors to each label category which consists of multiple individual pieces. After coloring, we explicitly annotate each category with label texts.
- **Template-based Modification:** Finally, we follow the CoMT template to construct the whole sample and combine the enhanced images with the textual rationales to represent the multi-modal thoughts.

2.4 Visual Selection

Text cannot indicate the location intuitively. Accurately selecting among similar objects using text alone is challenging due to the inherent difficulty in precise location and difference descriptions. Following this intuition, we construct the *Visual Selection* task, as shown in Figure 2 (d).

- **Original Dataset:** We construct the task from the spotdiff³ dataset. This dataset provides pairs of similar images and corresponding difference annotations, requiring precise identification of differences between two images.
- **Step-by-Step Annotation:** According to the annotations, we extract the distinct areas of image pairs in batches, keeping the same position and size as the original images.
- **Template-based Modification:** We then supplement the textual section within the template and integrate corresponding images to construct a multi-modal rationale.

³<https://www.allstarpuzzles.com/spotdiff/index.html>

Statistics	Number
Total Sample	3,853
Total Image	14,801
Average Question Length	22.66
Average Choice Length	1.33
Average Rationale Length	104.74
Average Multi-modal Thought Step	3.11
Average Rationale Step	7.71

Table 2: Basic statistics of CoMT, including sample numbers, steps of rationale, length of rationale, and image number generated in CoT.

3 Benchmark Analysis

Basic statistics As shown in Table 2, CoMT comprises 3,853 samples and 14,801 images. CoMT encompasses two primary domains within M³CoT (Chen et al. 2024b) and four visual operations (illustrated in Figure 3 (a)) for comprehensive evaluation. Additionally, CoMT requires more intricate reasoning, with an average length of 104.7 words and 7.7 steps per sample, significantly higher than ScienceQA’s 48 words and 2.5 steps.

Multi-modal diversity CoMT includes a diverse array of multi-modal tasks (*visual creation*, *visual deletion*, *visual update* and *visual selection*), ranging from mathematical problems to commonsense challenges, such as geometry and recognition. Furthermore, as depicted in Figure 3 (b), CoMT features a wide range of image types encompassing “Culture & Art”, and “Abstract Graph”, etc, classified by CLIP (Radford et al. 2021).

Rationale diversity As illustrated in Figure 3 (c), CoMT exhibits a *broad range* in the number of reasoning steps. Additionally, the multi-modal thought steps also show both diversity and sufficient volume. This allows for a comprehensive evaluation across different steps within CoMT.

4 Experiments

4.1 Experiments Setting

In our experiments, we select a range of LVLMs as backbones, including those trained on image generation tasks as well as those that are not, including *Gemini-Pro* (Team et al. 2023), *Qwen-VL* (Bai et al. 2023), *LLaVA-NeXT* (Liu

Model	Visual Creation		Visual Deletion		Visual Update		Visual Selection		Average	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Random	27.10	26.75	25.17	25.15	24.06	24.05	25.59	25.55	25.48	25.37
<i>Qwen-VL-7B</i> (Bai et al. 2023)										
Direct (Bai et al. 2023)	21.49	12.78	<u>26.35</u>	<u>18.29</u>	<u>37.64</u>	<u>30.34</u>	22.08	13.80	<u>26.89</u>	<u>18.80</u>
CoT (Kojima et al. 2022)	<u>23.96</u>	<u>19.22</u>	12.63	11.81	33.62	26.13	23.22	<u>18.00</u>	23.26	18.79
Desp-CoT (Wu et al. 2023)	19.90	13.23	20.94	7.73	30.59	23.85	<u>26.05</u>	10.48	24.37	13.82
VoT (Wu et al. 2024b)	22.08	17.51	14.43	11.71	28.52	21.02	22.08	12.47	21.78	15.68
<i>LLaVA-NeXT-13B</i> (Liu et al. 2024a)										
Direct (Liu et al. 2024a)	<u>26.34</u>	<u>19.72</u>	20.64	<u>20.06</u>	<u>35.47</u>	<u>34.26</u>	22.76	19.60	<u>26.30</u>	<u>23.41</u>
CoT (Kojima et al. 2022)	22.18	12.33	21.44	15.21	26.36	18.99	24.92	19.91	23.73	16.61
Desp-CoT (Wu et al. 2023)	19.90	12.82	23.45	17.47	27.01	18.82	25.59	<u>20.77</u>	23.99	17.47
VoT (Wu et al. 2024b)	20.79	15.58	<u>25.55</u>	18.55	27.55	18.95	<u>26.61</u>	17.23	25.13	17.58
<i>GILL</i> (Koh, Fried, and Salakhutdinov 2023)										
Direct (Koh, Fried, and Salakhutdinov 2023)	<u>16.93</u>	<u>15.75</u>	<u>22.65</u>	<u>13.90</u>	<u>23.43</u>	<u>12.62</u>	18.12	<u>10.16</u>	<u>20.28</u>	<u>13.11</u>
CoT (Kojima et al. 2022)	8.61	9.96	12.63	8.62	18.11	8.20	17.21	8.34	14.14	8.78
Desp-CoT (Wu et al. 2023)	6.83	7.93	20.74	9.60	21.69	10.90	<u>20.95</u>	9.12	17.55	9.39
VoT (Wu et al. 2024b)	5.94	7.01	17.94	11.81	21.04	11.51	14.27	9.23	14.80	9.89
<i>NExT-GPT</i> (Wu et al. 2024a)										
Direct (Wu et al. 2024a)	<u>24.26</u>	<u>19.00</u>	<u>25.75</u>	<u>19.15</u>	24.30	<u>18.04</u>	22.42	16.24	<u>24.18</u>	<u>18.11</u>
CoT (Kojima et al. 2022)	20.20	13.88	23.85	17.25	22.78	17.95	21.52	<u>18.39</u>	22.09	16.87
Desp-CoT (Wu et al. 2023)	17.52	13.93	23.95	14.13	<u>25.38</u>	17.91	<u>22.99</u>	16.90	22.46	15.72
VoT (Wu et al. 2024b)	13.17	12.91	22.85	14.38	25.05	16.28	22.88	18.32	20.99	15.47
<i>AnyGPT</i> (Zhan et al. 2024)										
Direct (Zhan et al. 2024)	19.11	12.18	17.43	11.92	23.10	17.85	<u>27.63</u>	16.91	21.82	14.72
CoT (Kojima et al. 2022)	10.10	10.36	21.74	11.96	24.08	18.37	22.20	15.77	19.53	14.12
Desp-CoT (Wu et al. 2023)	<u>19.31</u>	<u>14.15</u>	22.75	<u>12.22</u>	24.84	18.72	25.59	16.63	<u>23.12</u>	<u>15.43</u>
VoT (Wu et al. 2024b)	11.78	10.22	<u>23.45</u>	11.45	<u>26.36</u>	<u>19.44</u>	25.59	<u>18.43</u>	21.80	14.89
<i>Gemini</i> (Team et al. 2023)										
Direct (Team et al. 2023)	28.91	25.43	30.86	22.28	46.36	46.26	27.63	20.69	33.44	28.67
CoT (Kojima et al. 2022)	27.92	23.07	28.76	22.73	40.24	40.02	27.39	23.60	31.08	27.36
Desp-CoT (Wu et al. 2023)	18.04	14.61	29.36	21.43	31.05	23.20	25.14	11.32	25.90	17.64
VoT (Wu et al. 2024b)	33.27	26.48	27.05	20.79	35.36	27.83	24.92	19.38	30.15	23.62

Table 3: Main results on various LVLMs. The **bold content** indicates the best performance across all models and all prompting methods, while the underlined content signifies the best performance within a single model across all methods. See Table 4 in Technical Appendix F for complete results.

et al. 2024a), *GILL* (Koh, Fried, and Salakhutdinov 2023), *NExT-GPT* (Wu et al. 2024a), *AnyGPT* (Zhan et al. 2024). Additionally, we explore various prompting strategies: (1) *Direct* prompts the model to directly generate the answer. (2) *CoT* (Kojima et al. 2022) is a widely used prompt method to stimulate LLMs to generate steps with “Let’s think step-by-step!”. (3) *Desp-CoT* (Wu et al. 2023) enhances reasoning quality by instructing the model to generate a description before answering. (4) *VoT* (Wu et al. 2024b) utilizes “Visualize the state after each reasoning step.” to imagine the reasoning path with text-modal. Following Qin et al. (2023) and Chen et al. (2024b), we extract the final generated answers using regular expressions. See Technical Appendix D for further experimental details.

4.2 Main Results

Table 3 presents the main results, from which we derive the following key findings:

All LVLMs perform poorly on the CoMT. Despite Gemini achieving a 28.67% F1 score across four tasks, this performance is marginally better than the random baseline by

3.3%, indicating significant room for improvement. Additionally, except for Gemini, most models perform at or below random levels. We attribute these to the lack of multi-modal reasoning in current LVLMs.

Traditional Multimodal CoT almost completely fails on CoMT. We observe that pure text-modal CoT does not attain improvement in addressing the CoMT problem and even degrades the performance of most models to near-random levels. We attribute it to the fact that the inability of the model to execute specific visual logic expressions and operations results in poor performance.

All models fail to visualize thought in textual words. As demonstrated in Table 3, all LVLMs fail to utilize VoT effectively to improve performance. Specifically, VoT prompts LVLMs to visualize states through *textual* representation and results in an average accuracy decrease of 12.28%. This finding suggests that although textual representation can convey visual features, the inherent differences between modalities still constrain the expression of multi-modal thought.

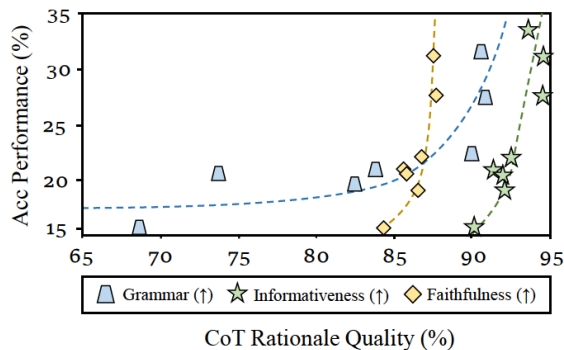


Figure 4: Analysis of the correlation between the model performance and the quality of rationale for different LVLMs based on ROSCOE (Golovneva et al. 2023).

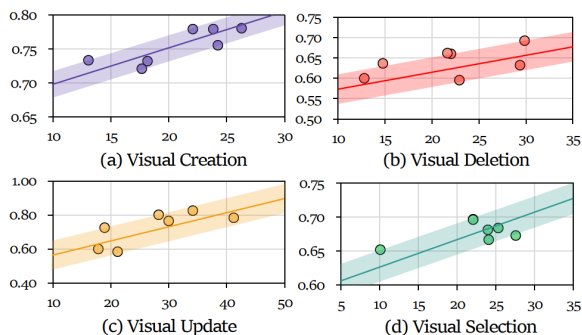


Figure 5: CLIPScore of LVLMs on 4 tasks within CoMT. The x-axis represents the CLIPScore, and the y-axis represents the accuracy.

4.3 Analysis

This section will conduct a further analysis on CoMT. See Technical Appendix E for more implementation details.

Improving the quality of rationale is essential for CoMT. As illustrated in Figure 4, the quality of CoT rationale significantly impacts the CoMT performance. Poor rationale quality constrains the logical coherence of LVLMs, limiting their reasoning capacities, which aligns with Chen et al. (2024b). Consequently, enhancing reasoning quality in LVLMs is a crucial area for further exploration.

CoMT benefits from improved multi-modal thought. To assess the impact of multi-modal thought on performance within CoMT, we calculate the CLIPScore (Hessel et al. 2021) to reflect the similarity between model output and each image within the ideal rationale pre-defined. Averaging these scores yields a multi-modal alignment score for each reasoning chain generated. As shown in Figure 5, there is a significant positive correlation between performance and multi-modal alignment scores across four tasks, which indicates that CoMT benefits from more multi-modal thought.

The performance relies more on the quality of multi-modal alignment than on parameter size. As shown in Table 4 in Technical Appendix F, the IDEFICS2-8B, with fine-grained multi-modal alignment, surpasses the 13B models, even approaching the performance of the Gemini-Pro

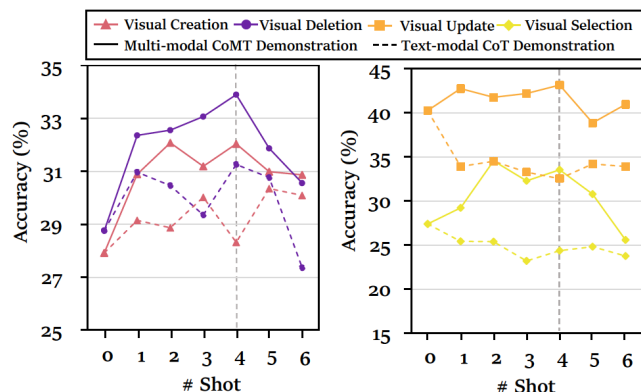


Figure 6: Analysis on In-context Learning of Gemini-Pro (Team et al. 2023) in CoMT.

(>100B, Team et al. (2023)). We think that CoMT performance depends more on multi-modal alignment quality rather than parameter size.

4.4 In-context Learning Explorations

In-context Learning with multi-modal input and output can effectively promote the performance in CoMT. As shown in Figure 6, using in-context learning (ICL) (Li et al. 2023a; Qin et al. 2024a) with multi-modal input and multi-modal output demonstrations significantly improves performance. It not only surpasses zero-shot prompting but also outperforms ICL with text-modal output. This approach can be successful due to the fact that LVLMs can learn to effectively facilitate multi-modal thought through such demonstrations, even though Gemini is limited to producing rationales in the textual modality alone.

Not more demonstrations means better performance in CoMT. As shown in Figure 6, the model exhibits a significant downward trend in performance when the number of demonstrations exceeds four. It shows that more demonstrations are not necessarily better, as multimodal demonstrations often require the consumption of a substantial number of tokens, which can also lead to more complex challenges associated with longer contexts.

4.5 Error Analysis

Insufficient Multi-modal Thought. When dealing with multi-modal problems, models struggle to integrate multi-modal thought most of the time. As illustrated in Figure 7, we observe that despite certain models (e.g., GILL, NEXTGPT, AnyGPT) being trained on image generation tasks, at least 48% of their reasoning processes do not incorporate image generation. This occurs even when image generation is crucial for accurate outcomes, indicating a disjunction between image generation and text processing.

Inaccurate Textual Reasoning. When logical errors occur in textual reasoning, they hinder the advancement towards the correct answer. For example, Figure 10 in Technical Appendix reveals that the model demonstrates poor reasoning logic, with significant logical errors, such as calcula-

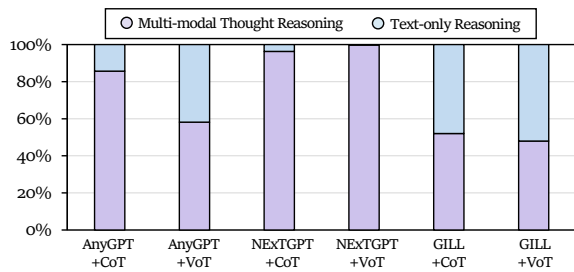


Figure 7: Image generation frequency during reasoning.

tion mistakes (like “ $2*5*5=2*10$ ”). These inaccurate textual reasoning significantly impedes progress in this field.

Incoherent Visual Reasoning. Although certain models generate images when reasoning, not all image contents align with the reasoning path, revealing an immature interaction between modalities. We manually evaluate the generated images, with results shown in Figure 8. The distribution reveals that current LVLMs often generate *irrelevant* images during reasoning (an average of 43%, represented by *score 0*) and fail to perform effective visual logic (on average 45% of images exhibit *logical mistake*, represented by *score 1,2*). The judgment criteria can be found in Technical Appendix C.3. To be specific, Figure 11 in Technical Appendix G shows instances with irrelevant text and image logic.

4.6 Future Directions

Based on the above analysis, we summarize the future directions for current LVLMs tackling CoMT.

How can we effectively integrate multi-modal thought reasoning? The absence of visual thought significantly increases the difficulty when addressing certain multi-modal tasks, such as CoMT. How to enable models to integrate multi-modal reasoning is an intriguing research topic. Furthermore, given the inherent differences between textual and visual modalities, exploring how to align these two modalities during reasoning presents another valuable challenge.

How can we enhance logical reasoning capabilities for textual reasoning? The inadequacies in textual reasoning logic lead to inaccurate conclusions during inference, such as calculation mistakes. Therefore, how to enable models with better textual logic to perform effective text reasoning is a critical topic to explore.

How can we achieve effective vision logic for visual reasoning? Since some generated images fail to perform effective visual logic or even be irrelevant, not all visual thoughts generated have a positive influence on the reasoning. How to enable models to develop better visual logic to produce images that are relevant and consistent with the progression of rationale is a topic worth exploring.

5 Related Work

The emergence of Multi-modal Chain-of-Thought (MCoT) techniques elicits the step-by-step zero-shot and few-shot multi-modal reasoning capabilities of Large Vision-Language Models (LVLMs) (Wang et al. 2024c,b; Chen

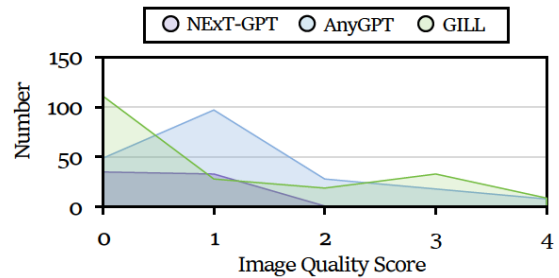


Figure 8: Distribution of human-evaluated image quality scores (\uparrow) which are mainly determined based on *Relevance* and *Logical Correctness*. See Technical Appendix C.3 for evaluation details.

et al. 2024a,c; Liu et al. 2023; He et al. 2024; Qin et al. 2024a; Fei et al. 2024a,c). Pioneering work introduces the ScienceQA benchmark (Lu et al. 2022), involving multi-modal scientific questions. Subsequently, Zhang et al. (2023) formally propose the concept of MCoT and introduce a two-stage framework encompassing both reasoning and answering. Additionally, Tan et al. (2024); Wang et al. (2024a); Zhang et al. (2024a); Mondal et al. (2024); Lee et al. (2024) introduce more knowledge to improve the performance and reduce hallucinations in MCoT reasoning. Following this, Zheng et al. (2024) propose DDCoT, which breaks down the question into a series of sub-questions and solves them using LVLMs. Building upon this, Chen et al. (2024b) further introduce a multi-domain multi-step multi-modal benchmark to fully evaluate the complex MCoT capabilities. Based on traditional MCoT, some works begin preliminary exploration integrating the diffusion model or retriever model as a tool for better MCoT. Meng et al. (2023) propose CoI to generate images as intermediate reasoning steps in single modal tasks, outperforming purely textual CoT. Wu et al. (2024b) propose VoT, requiring text-only LLMs to imagine their vision reasoning paths, which increases the spatial reasoning abilities.

In contrast to our work, their strategies rely solely on textual modalities for reasoning, lacking visual operation or detailed visual expression in reasoning. To fill this gap, we propose CoMT to comprehensively reveal diverse multi-modal thought capabilities. We hope CoMT will inspire research on promoting better multi-modal reasoning.

6 Conclusion

In this work, we introduce a Chain of Multi-modal Thought (CoMT) benchmark to evaluate and improve the multi-modal reasoning capabilities of Large Vision-Language Models (LVLMs). Through extensive experiments, our findings reveal a significant performance gap between LVLMs and human, with models generally not outperforming random chance in zero-shot scenarios. In-context Learning with multi-modal rationale emerges as a promising approach to better integrate visual and textual reasoning in LVLMs. We hope this research lays the groundwork for future enhancements in multi-modal reasoning technologies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) via grant 62306342, 62236004, 62441603 and 62476073. This work was sponsored by the Excellent Young Scientists Fund in Hunan Province (2024JJ4070), the Science and Technology Innovation Program of Hunan Province under Grant 2024RC3024 and CCF-Zhipu Large Model Innovation Fund (NO.CCF-Zhipu202406)). This work was also supported by Key Laboratory of Data Intelligence and Advanced Computing in Provincial Universities, Soochow University. We are grateful for resources from the High Performance Computing Center of Central South University. Libo Qin is the corresponding author.

References

- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.
- Cao, J.; and Xiao, J. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1511–1520.
- Chen, Q.; Qin, L.; Jiaqi, W.; Jinxuan, Z.; and Che, W. 2024a. Unlocking the Boundaries of Thought: A Reasoning Granularity Framework to Quantify and Optimize Chain-of-Thought. In *Proc. of NeurIPS*.
- Chen, Q.; Qin, L.; Zhang, J.; Chen, Z.; Xu, X.; and Che, W. 2024b. M³CoT: A Novel Benchmark for Multi-Domain Multi-step Multi-modal Chain-of-Thought. *arXiv preprint arXiv:2405.16473*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.
- Fei, H.; Li, B.; Liu, Q.; Bing, L.; Li, F.; and Chua, T.-S. 2023. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1171–1182.
- Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Zhang, M.; Lee, M.-L.; and Hsu, W. 2024a. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*.
- Fei, H.; Wu, S.; Zhang, H.; Chua, T.-S.; and Shuicheng, Y. 2024b. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Fei, H.; Wu, S.; Zhang, M.; Zhang, M.; Chua, T.-S.; and Yan, S. 2024c. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Golovneva, O.; Chen, M. P.; Poff, S.; Corredor, M.; Zettlemoyer, L.; Fazel-Zarandi, M.; and Celikyilmaz, A. 2023. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. In *The Eleventh International Conference on Learning Representations*.
- He, L.; Li, Z.; Cai, X.; and Wang, P. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18180–18187.
- Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.
- Ji, A.; Kojima, N.; Rush, N.; Suhr, A.; Vong, W. K.; Hawkins, R. D.; and Artzi, Y. 2022. Abstract visual reasoning with tangram shapes. *arXiv preprint arXiv:2211.16492*.
- Koh, J. Y.; Fried, D.; and Salakhutdinov, R. 2023. Generating Images with Multimodal Language Models. *NeurIPS*.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Proc. of NeurIPS*, 35: 22199–22213.
- Lee, J.; Wang, Y.; Li, J.; and Zhang, M. 2024. Multimodal Reasoning with Multimodal Knowledge Graph. *arXiv preprint arXiv:2406.02030*.
- Lehmann, D.; Pascual-Marqui, R. D.; Strik, W. K.; and Koenig, T. 2010. Core networks for visual-concrete and abstract thought content: a brain electric microstate analysis. *Neuroimage*, 49(1): 1073–1079.
- Li, X.; Lv, K.; Yan, H.; Lin, T.; Zhu, W.; Ni, Y.; Xie, G.; Wang, X.; and Qiu, X. 2023a. Unified Demonstration Retriever for In-Context Learning. *arXiv:2305.04320*.
- Li, Y.; Wang, L.; Hu, B.; Chen, X.; Zhong, W.; Lyu, C.; and Zhang, M. 2023b. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*.
- Lin, W.; Wei, X.; An, R.; Gao, P.; Zou, B.; Luo, Y.; Huang, S.; Zhang, S.; and Li, H. 2024. Draw-and-Understand: Leveraging Visual Prompts to Enable MLLMs to Comprehend What You Want. *arXiv preprint arXiv:2403.20271*.
- Liu, B.; Lyu, C.; Min, Z.; Wang, Z.; Su, J.; and Wang, L. 2023. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. *arXiv preprint arXiv:2312.01714*.
- Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Proc. of NeurIPS*, 36.
- Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Proc. of NeurIPS*, 35: 2507–2521.
- Meng, F.; Yang, H.; Wang, Y.; and Zhang, M. 2023. Chain of Images for Intuitively Reasoning. *arXiv preprint arXiv:2311.09241*.

- Mondal, D.; Modi, S.; Panda, S.; Singh, R.; and Rao, G. S. 2024. KAM-CoT: Knowledge Augmented Multimodal Chain-of-Thoughts Reasoning. *arXiv preprint arXiv:2401.12863*.
- Qin, L.; Chen, Q.; Fei, H.; Chen, Z.; Li, M.; and Che, W. 2024a. What Factors Affect Multi-Modal In-Context Learning? An In-Depth Exploration. *arXiv preprint arXiv:2410.20482*.
- Qin, L.; Chen, Q.; Feng, X.; Wu, Y.; Zhang, Y.; Li, Y.; Li, M.; Che, W.; and Yu, P. S. 2024b. Large Language Models Meet NLP: A Survey. *arXiv preprint arXiv:2405.12819*.
- Qin, L.; Chen, Q.; Wei, F.; Huang, S.; and Che, W. 2023. Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2695–2709.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 146–162. Springer.
- Sindagi, V. A.; Yasarla, R.; and Patel, V. M. 2020. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2594–2609.
- Tan, C.; Wei, J.; Sun, L.; Gao, Z.; Li, S.; Yu, B.; Guo, R.; and Li, S. Z. 2024. Retrieval Meets Reasoning: Even High-school Textbook Knowledge Benefits Multimodal Reasoning. *arXiv preprint arXiv:2405.20834*.
- Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Wang, L.; Hu, Y.; He, J.; Xu, X.; Liu, N.; Liu, H.; and Shen, H. T. 2024a. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19162–19170.
- Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; et al. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.
- Wang, Z.; Han, Z.; Chen, S.; Xue, F.; Ding, Z.; Xiao, X.; Tresp, V.; Torr, P.; and Gu, J. 2024c. Stop Reasoning! When Multimodal LLM with Chain-of-Thought Reasoning Meets Adversarial Image. In *Proc. of CoLM*.
- Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024a. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, 53366–53397.
- Wu, W.; Mao, S.; Zhang, Y.; Xia, Y.; Dong, L.; Cui, L.; and Wei, F. 2024b. Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models. *arXiv preprint arXiv:2404.03622*.
- Wu, Y.; Zhang, P.; Xiong, W.; Oguz, B.; Gee, J. C.; and Nie, Y. 2023. The role of chain-of-thought in complex vision-language reasoning task. *arXiv preprint arXiv:2311.09193*.
- Xu, J.; Fei, H.; Pan, L.; Liu, Q.; Lee, M.-L.; and Hsu, W. 2024. Faithful Logical Reasoning via Symbolic Chain-of-Thought. *arXiv preprint arXiv:2405.18357*.
- Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.
- Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6720–6731.
- Zhan, J.; Dai, J.; Ye, J.; Zhou, Y.; Zhang, D.; Liu, Z.; Zhang, X.; Yuan, R.; Zhang, G.; Li, L.; et al. 2024. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. *arXiv preprint arXiv:2402.12226*.
- Zhang, D.; Yang, J.; Lyu, H.; Jin, Z.; Yao, Y.; Chen, M.; and Luo, J. 2024a. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*.
- Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024b. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
- Zheng, G.; Yang, B.; Tang, J.; Zhou, H.-Y.; and Yang, S. 2024. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Proc. of NeurIPS*, 36.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.