M³CoT: A Novel Benchmark for Multi-Domain Multi-step Multi-modal Chain-of-Thought

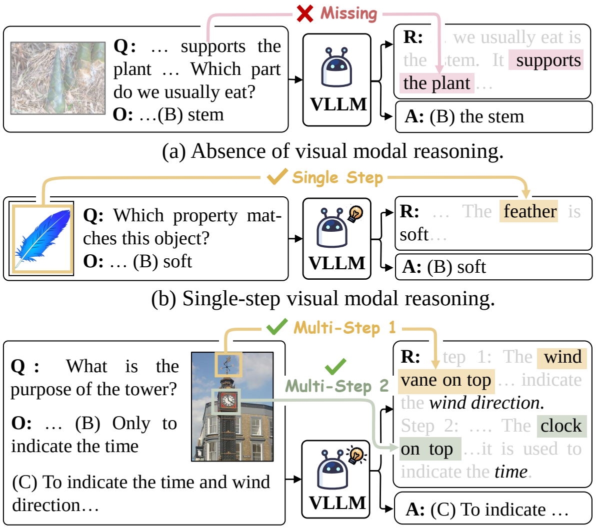
Qiguang Chen[♠] **Libo Qin**[♣*] **Jin Zhang**[♠] **Zhi Chen**[♢]
Xiao Xu[♠] **Wanxiang Che**[♠*]

[♠] Research Center for Social Computing and Information Retrieval
[♠] Harbin Institute of Technology, China
[♣] School of Computer Science and Engineering, Central South University, China
[♢] Shanghai AI Laboratory

`{qgchen,car}@ir.hit.edu.cn`, `lbqin@csu.edu.cn`

Abstract

Multi-modal Chain-of-Thought (MCoT) requires models to leverage knowledge from both textual and visual modalities for step-by-step reasoning, which gains increasing attention. Nevertheless, the current MCoT benchmark still faces some challenges: (1) *absence of visual modal reasoning*, (2) *single-step visual modal reasoning*, and (3) *Domain missing*, thereby hindering the development of MCoT. Motivated by this, we introduce a novel benchmark (M³CoT) to address the above challenges, advancing the multi-domain, multi-step, and multi-modal CoT. Additionally, we conduct a thorough evaluation involving abundant MCoT approaches on Vision Large Language Models (VLLMs). In addition, we highlight that the current VLLMs still struggle to correctly reason in M³CoT and there remains a large gap between existing VLLMs and human performance in M³CoT, despite their superior results on previous MCoT benchmarks. To our knowledge, we take the first meaningful step toward the multi-domain, multi-step, and multi-modal scenario in MCoT. We hope that M³CoT can serve as a valuable resource, providing a pioneering foundation in multi-domain, multi-step, multi-modal chain-of-thought research.



(c) Multi-step visual modal reasoning.

Figure 1: The example of Absence of visual modal reasoning (a), Single-step visual modal reasoning (b), and Multi-step visual modal reasoning (c). **Q**: textual question; **O**: textual options; **R**: generated rationale; **A**: generated answer.

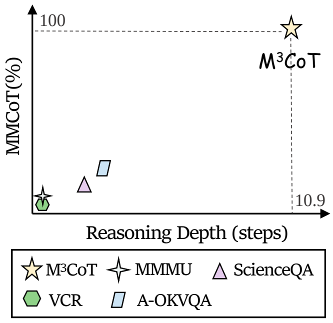
1 Introduction

Recent advancements in Large Language Models (LLMs) have led to notable improvements in Chain-of-Thought (CoT) in textual modality (Wei et al., 2022a,b; Wang et al., 2023b; Hu et al., 2024). In addition, some works begin to extend the textual CoT capabilities to multi-modal CoT reasoning (MCoT). Take Figure 1 (c) as an example, multi-modal CoT requires both the visual and textual features to generate a rationale and a final answer. To this end, Lu et al. (2022a) introduced ScienceQA benchmark and laid the foundation for MCoT. Subsequently, Zhang et al. (2023c) proposed a two-stage approach during multi-modal reasoning for MCoT. Additionally, Wang et al. (2023a) developed T-SciQA framework to distill high-quality rationales from ChatGPT, which attains an average accuracy of 96.2%, surpassing even the human accuracy of 88.4%.

Inspired by the recent remarkable advancements in the MCoT literature (surpassing human performance), we seek to explore an interesting question: *Has MCoT task been solved perfectly?* In our deep analysis, the conclusion is definitely "*NO*". As shown in Figure 2 (a), we observe that current benchmarks are too simple, leading to an overestimation of current progress. Furthermore, we find that the existing benchmarks exhibit three major drawbacks (see Figure 2 (b)): (1) ***Absence of visual modal reasoning***: As shown in Figure 1 (a), the model can successfully produce rationale and answer solely based on the textual modality context of "*supports the plant*", which cannot truly reflect the

* Corresponding Author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8199–8221
August 11-16, 2024 ©2024 Association for Computational Linguistics



(a)

	# Q	# I	Domain			MMCoT	Rationale
			Science	Mathmatic	Commonsense		
Geometry3K (Lu et al., 2021)	3,002	2,342	✘	✔	✘	✘	✘
TQA (Kembhavi et al., 2017)	**26,260**	3,455	✘	✘	✔	✘	✘
MathVista (Lu et al., 2023a)	5,487	6,141	✘	✔	✘	✘	✘
MME (Fu et al., 2023)	2,194	1,097	✘	✘	✔	✘	✘
SeedBench (Li et al., 2023a)	-	19,242	✘	✘	✔	✘	✘
MM-Vet (Yu et al., 2023)	205	187	✘	✔	✔	✘	✘
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., 2022a)	21,208	10,332	✔	✘	✘	∼8%	✔
MMMU (Yue et al., 2023)	11,550	11,264	✔	✔	✘	∼8%	Science Only (<18%)
$\mathtt{M^3CoT}$ (ours)	11,459	11,293	✔	✔	✔	**100%**	✔

(b)

Figure 2: Comparison of $\mathtt{M^3CoT}$ and multi-modal related datasets on (a) MCoT reasoning complexity and (b) detailed diversity. MMCoT: the ratio of samples with multi-step MCoT (MMCoT) in the datasets; **#X**: the size of X, **Q**: Question; **I**: Image. The simplicity of the previous benchmarks lies in its MMCoT, domain, and reasoning depth. We will describe the details of the corresponding statistics in Appendix A.1.

ability of multi-modal CoT model. (2) ***Single-step visual modal reasoning***: As illustrated in Figure 1 (b), the model only requires a single-step "feather" object to predict the correct rationale and answer, which cannot be satisfied in the complex multi-step CoT scenario. (3) ***Domain Missing***: Commonsense and mathematics are important domains for evaluating multi-modal CoT (Wei et al., 2022b; Qin et al., 2023), but the current benchmarks lack these topics, hindering the comprehensive evaluation progress of multi-modal CoT. Nevertheless, in real-world scenarios, multi-step MCoT reasoning is frequently observed in diverse domains. For example, as illustrated in Figure 1 (c), vision large language models (VLLMs) are required to identify the correct options integrating at least two multi-modal reasoning steps (as indicated by the orange and green lines.). Such multi-step MCoT tasks are required to effectively perform multi-step reasoning across multiple modalities, which cannot be achieved by previous single-step multi-modal CoT approaches.

Motivated by these observations and issues, we introduce a novel benchmark about multi-domain multi-step multi-modal chain-of-thought reasoning ($\mathtt{M^3CoT}$) based on ScienceQA (Lu et al., 2022b). Specifically, to address the first issue, we directly remove samples that could infer the final answer without the need for images. To tackle the second issue, we manually annotate and select multi-step multi-modal samples. Specifically, we provide expert annotators with textual context and rationales without images. Experts are required to determine when multi-step reasoning cannot be resolved solely based on textual context. Subsequently, we present the images to experts to ascertain whether multi-step reasoning occurred across textual and visual modalities. To solve the third issue, we explore LLM-guided augmentation to synthesize the multi-step MCoT data for commonsense and mathematics domains. We evaluate abundant representative MCoT approaches on $\mathtt{M^3CoT}$ in extensive scenarios, yielding several **key takeaways**: (1) *VLLM shows CoT emergence phenomenon at the parameter level over 10 billion* ($\geq$ 13B); (2) *Fine-tuning has better hope on multi-step MCoT, compared with the failures of vanilla in-context-learning, tool usage, and prompting strategies*. (3) *$\mathtt{M^3CoT}$ is tough enough and all methods still struggle compared with human performance*.

In conclusion, the primary contributions of our work are summarized as follows:

- We identify the weaknesses of current multi-modal CoT benchmarks that can not handle complex multi-step reasoning scenarios, which motivates researchers to rethink the current progress of multi-modal CoT.
- To the best of our knowledge, we are the first to consider the multi-domain, multi-step, multi-modal CoT scenario and introduce $\mathtt{M^3CoT}$ to this end.
- We evaluate abundant representative MCoT approaches on $\mathtt{M^3CoT}$ and summarize some insightful takeaways, hoping to inspire more breakthroughs in this direction.

To facilitate further research, all data and code are available at https://github.com/LightChen233/M3CoT.

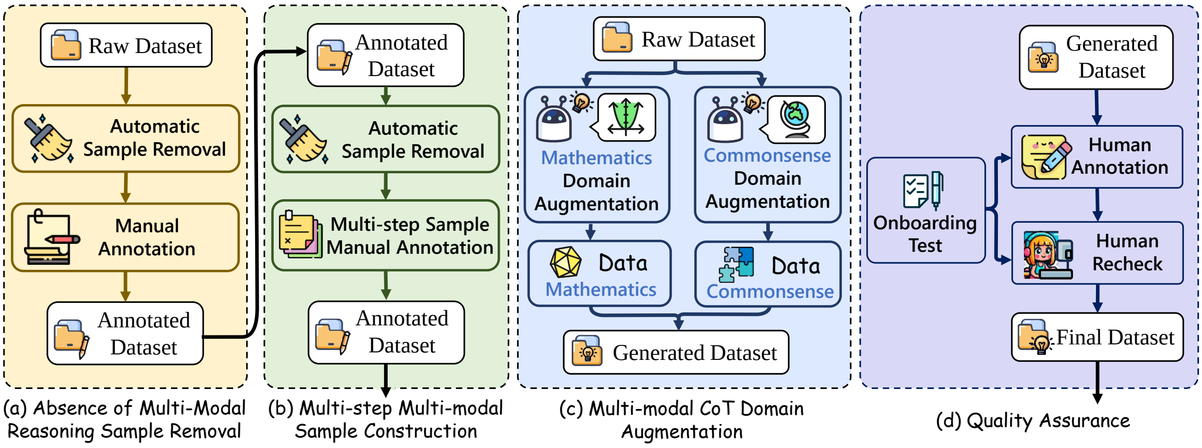


Figure 3: Dataset construction workflow including (a) *Absence of Visual Modal Reasoning Sample Removal* (§ 3.1), (b) *Multi-step Multi-modal Sample Construction* (§ 3.2), (c) *Multi-modal CoT Domain Augmentation* (§ 3.3), and (d) *Quality Assurance* (§ 3.4).

2 Problem Formalization

This section describes the definition of multi-step multi-modal CoT. Specifically, unlike the traditional textual CoT, multi-step multi-modal CoT should consider a scenario involving an image I, a question Q, a context C and a set of n options $\mathcal{O} = \{o_1, ..., o_n\}$. First we construct a textual prompt $\mathcal{T}$:

$$\mathcal{T} = \texttt{Prompt}(Q, C, \mathcal{O}), \tag{1}$$

where $\texttt{Prompt}(\cdot)$ represents any method used to convert textual inputs into an instruction format.

Then, model should generate a step-wise rationale $\mathcal{R}_m = \{S_1, ..., S_m\}$, each step determined by[1]:

$$S_i = \underset{S_i \in \mathcal{R}_m}{\text{argmax}}\ P(S_i|I, \mathcal{T}), \tag{2}$$

$$P(S_i|I,\mathcal{T}) = \begin{cases} P(S_i|\mathcal{T}, \mathcal{R}_{i-1}), & S_i \notin \mathcal{S}; \\ P(S_i|I, \mathcal{T}, \mathcal{R}_{i-1}), & S_i \in \mathcal{S}, \end{cases} \tag{3}$$

where $\mathcal{S}$ indicates steps that require multi-modal reasoning. This reasoning is considered multi-step and multi-modal if $|\mathcal{S}| \geq 2$.

Finally, the model arrives at the final answer $\mathcal{Y}$, which is denoted as:

$$\mathcal{Y} = \underset{o \in \mathcal{O}}{\text{argmax}}\ P(o|\mathcal{R}_m). \tag{4}$$

3 Dataset Annotation

This section describes the annotation process of M^{3}CoT, including: *Absence of Visual Modal Reasoning Sample Removal* (§ 3.1), *Multi-step MCoT Sample Construction* (§ 3.2), *MCoT Domain Augmentation* (§ 3.3), and *Quality Assurance* (§ 3.4). The samples we generate and retain at each stage are detailed in Figure 11.

[1]All step segmentation in this paper follows the ROSCOE (Golovneva et al., 2023).

3.1 Absence of Visual Modal Reasoning Sample Removal

This section focuses on addressing the absence of visual modal reasoning challenge from ScienceQA.

***Automatic Sample Removal*:** First, we directly filter out samples without images, thereby refining the dataset to include only those samples that potentially require multi-modal reasoning.

***Manual Annotation*:** Despite the automatic process, some samples containing images are still irrelevant for multi-modal reasoning (see Figure 1 (a)). Therefore, we further employ manual annotation, requiring experts to verify whether each sample meets the criteria for MCoT. Specifically, our annotation process and instructions are shown in Appendix A.2.

3.2 Multi-step MCoT Sample Construction

This section aims to incorporate multi-step reasoning characteristics from the last processed data.

Automatic Sample Removal: In this step, we first automatically filter out some simple samples with overly simplistic rationales, which comprise less than two steps. By doing this, we can reduce the manual annotation burden and increase the reliability of M^{3}CoT. More details are illustrated in Appendix A.3.

Multi-step Sample Manual Annotation: After automatic sample removal, we further utilize manual annotation to obtain the final multi-step multi-modal reasoning dataset. Specifically, human experts are first provided with textual context and rationales without visual modality. They are focused on determining whether it is necessary to answer the samples multiple times based on the visual content. Once experts find that multi-step reasoning needs multiple times reasoning based on the image, we will provide them with corresponding images to let them finally confirm whether the sample needs to utilize multi-step image and text modalities reasoning to obtain the final multi-step reasoning paths.

3.3 MCoT Domain Augmentation

In order to make up for the missing data in previous work on mathematics and commonsense, we constructed M^{3}CoT based on MATH (Hendrycks et al., 2021) and Sherlock (Hessel et al., 2022) dataset to enhance the dataset within respective domains. More details are illustrated in Appendix A.4.2.

Mathematics Domain Augmentation: It is worth noting that MATH (Hendrycks et al., 2021) is a single modal dataset only with textual questions, rationales and answers, lacking corresponding options and images. To construct the options, we first prompt LLM to generate the related and similar options. Then, for lack of images, we further convert the geometry code and formula code into images, and use HTML framework to splice them together.

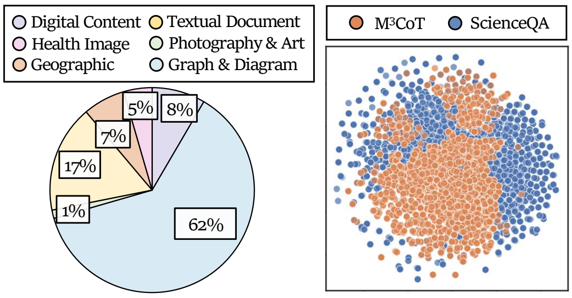
Commonsense Domain Augmentation: In order to expand the field of commonsense, we use the Sherlock (Hessel et al., 2022), which only contains some visual clues and does not contain any specific questions, options, and answers. Therefore, following Zhang et al. (2023b), we require LLM to cautiously generate questions, options, and answers. Specifically, we incorporate the multiple visual clues in Sherlock to LLM and enforce LLM generates models based on multiple image clues instead of single ones to ensure multi-step multi-modal reasoning.

3.4 Quality Assurance

This section aims to improve annotated data quality. More details are shown in Appendix A.5.

Onboarding Test: Annotators are all required to undergo a preliminary test, annotating 100 samples. Their results are evaluated by three experts, and only those achieving at least 80% accuracy proceed to subsequent annotation tasks.

Human Annotation: To address potential hallucinations or logical errors in generated samples, annotators are first asked to review and refine the rationale, ensuring accuracy and coherence.



(a) CLIP-based categorization of our collected images.

(b) Image representation via CLIP encoding and t-SNE dimensionality reduction.

Figure 4: Image diversity analysis (a) and the representation comparison (b) between M^{3}CoT and ScienceQA, where the point area in Figure (b) represents the image semantics coverage in the semantic space.

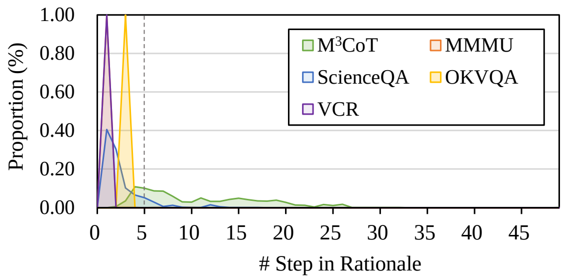


Figure 5: Comparison of the distribution of steps in the rationale for existing benchmarks. Notably, the distributions for MMMU and VCR overlap.

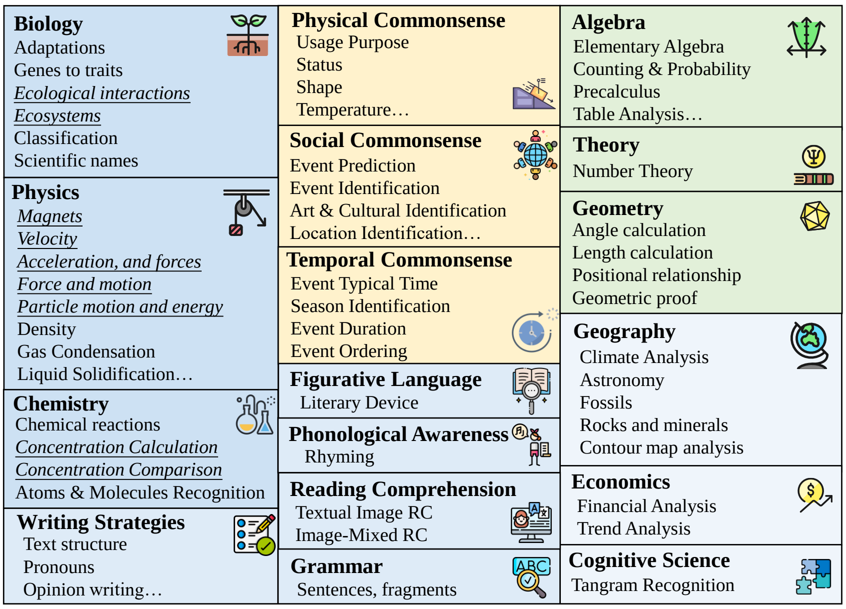
Human Recheck: After that, these annotators are required to recheck all data twice to determine if the data meets multi-step multi-modal reasoning criteria and possesses a coherent logical rationale. All samples in M^{3}CoT for which at least two annotators agree can be accepted. The kappa coefficient between annotators achieves 0.85, which indicates perfect agreement (Landis and Koch, 1977).

4 Data Analysis

This section provides some detailed data analysis to better understand M^{3}CoT.

Basic statistics M^{3}CoT is partitioned randomly into three subsets: train, validation, and test splits, containing 7,863, 1,108, and 2,358 samples, respectively. Compared to ScienceQA, M^{3}CoT demands more intricate reasoning, with an average length of 294, much higher than ScienceQA's 48.

Multi-modal diversity As shown in Figure 4 (a), M^{3}CoT features diverse image types, which categorized by CLIP (Radford et al., 2021). Furthermore, Figure 4 (b) demonstrates that M^{3}CoT spans



Statistic	Number
Total Samples	11,459
Domain classes	3
Topic classes	17
Category classes	263
Science size	7,973
Mathematic size	1,166
Commonsense size	2,163
Train set size	7,973
Dev set size	1,127
Test set size	2,359
Average question length	14.33
Average choice length	3.60
Average context length	19.35
Average rationale length	293.93

Figure 6: Detailed Analysis of **topic** and categories (partial) in the data set, where the *underlined and italics* mean the data we selected from ScienceQA.

a broader semantic space, suggesting the enhanced semantic richness and coverage.

Rationale diversity In comparison to existing benchmarks, the rationale process in M^{3}CoT is characterized by an increased proportion of steps that are more uniformly distributed, as shown in Figure 5. Specifically, ScienceQA averages 2.5 steps, OKVQA 3.0, MMMU only 1.0, and VCR only 1.0. In M^{3}CoT, the reasoning process involves a significantly higher average of 10.9 steps, highlighting the complexity and challenges presented by M^{3}CoT.

Domain Diversity In M^{3}CoT, questions are categorized into three primary domains: science knowledge, mathematics and commonsense. As illustrated in Figure 6, the dataset encompasses 17 topics and 263 categories, highlighting the extensive variety of the questions. This variety is essential for assessing the generalization abilities of various models and for furthering multi-modal research.

5 Experiments

5.1 Experiments Setting

We evaluate various VLLMs in M^{3}CoT, including *Kosmos-2* (Peng et al., 2023), *InstructBLIP* (Dai et al., 2023), *LLaVA-V1.5* (Liu et al., 2023), *CogVLM* (Wang et al., 2023c), *Gemini* (Google, 2023), *GPT4V* (OpenAI, 2023). In addition, we explore some prompting strategies. Specifically, we utilize `Direct` approach to submitting samples in the VLLMs required format; `CoT` (Kojima et al., 2022) with "Let's think step-by-step!"; `Desp-CoT` (Wu et al., 2023b) with an initial image description prompting; `CCoT` (Mitra et al., 2023) with better description in graph format. Following the settings of Kojima et al. (2022); Qin et al. (2023), we extract the final generated answer through regular expressions.

5.2 Results for M^{3}CoT

Results are presented in Table 1. We have the following observations:

There remains a significant disparity between open source VLLMs and GPT4V. Open source VLLMs still lag behind GPT4V by at least 7.98% on the M^{3}CoT benchmark. It highlights the limitations in the interaction and reasoning capabilities of existing open-source VLLMs, when compared to GPT-4V, especially in advanced tasks.

It exhibits a significant gap between GPT4V and human. Despite GPT4V's impressive results, it substantially trails human performance, demonstrating GPT4V still struggles to M^{3}CoT.

Zero-shot Multi-modal Chain-of-Thought only benefits larger VLLMs. As shown in Table 1 and Table 3, MCoT strategy fails to enhance reasoning abilities in VLLM with fewer than 13B parameters. Therefore, larger VLLMs ($\geq 13B$) can better observe emergent capabilities.

Model	Science			Commonsense			Mathematics			Total
	Lang	Natural	Social	Physical	Social	Temporal	Algebra	Geometry	Theory	
Random	32.70	30.62	26.71	32.97	22.22	20.33	35.71	27.50	23.81	28.56
InstructBLIP-13B (Dai et al., 2023)										
`Direct` (Dai et al., 2023)	38.39	30.52	26.27	76.67	70.66	35.77	30.00	22.50	19.05	35.94
`CoT` (Kojima et al., 2022)	38.39	30.01	27.55	80.00	70.25	33.33	30.71	21.25	19.05	36.07
`Desp-CoT` (Wu et al., 2023b)	16.59	27.84	22.77	54.44	52.89	30.08	27.86	28.75	28.57	29.25
`CCoT` (Mitra et al., 2023)	13.27	26.95	24.84	62.22	67.36	41.46	25.00	25.00	23.81	31.28
LLava-V1.5-13B (Liu et al., 2023)										
`Direct` (Liu et al., 2023)	36.97	27.46	20.22	52.22	23.55	27.64	22.86	45.00	4.76	27.05
`CoT` (Kojima et al., 2022)	46.45	38.31	27.87	67.78	64.05	49.59	26.43	30.00	23.81	39.52
`Desp-CoT` (Wu et al., 2023b)	47.87	29.25	27.23	68.89	59.92	47.15	26.43	36.25	9.52	35.98
`CCoT` (Mitra et al., 2023)	38.86	31.55	28.18	72.22	61.57	39.84	29.29	36.25	28.57	36.45
CogVLM-17B (Wang et al., 2023c)										
`Direct` (Wang et al., 2023c)	52.61	37.42	26.91	55.56	54.13	29.27	29.29	32.50	23.81	37.19
`CoT` (Kojima et al., 2022)	51.18	43.81	29.30	54.44	39.26	31.71	35.71	33.75	33.33	38.91
`Desp-CoT` (Wu et al., 2023b)	46.92	35.63	25.80	48.89	47.52	38.21	27.14	31.25	19.05	35.07
`CCoT` (Mitra et al., 2023)	47.39	34.99	25.80	62.22	46.28	35.77	30.71	37.50	23.81	35.63
Gemini (Google, 2023)										
`Direct` (Google, 2023)	73.93	41.25	31.21	56.67	71.49	62.60	30.71	27.50	28.57	45.17
`CoT` (Kojima et al., 2022)	67.30	49.68	36.31	68.89	60.33	66.67	23.57	21.25	9.52	47.50
`Desp-CoT` (Wu et al., 2023b)	49.29	43.68	27.07	63.33	57.85	70.73	28.57	30.00	28.57	41.85
`CCoT` (Mitra et al., 2023)	36.49	31.16	27.39	71.11	36.78	55.28	20.71	16.25	0.00	32.61
GPT4V (OpenAI, 2023)										
`Direct` (OpenAI, 2023)	80.09	54.66	43.95	87.78	67.77	82.11	42.14	43.75	**42.86**	56.95
`CoT` (Kojima et al., 2022)	**90.52**	**63.09**	**46.97**	83.33	**75.21**	**82.93**	**45.71**	**50.00**	38.10	**62.60**
`Desp-CoT` (Wu et al., 2023b)	79.62	54.66	36.94	**88.89**	74.38	73.98	20.71	32.50	33.33	53.54
`CCoT` (Mitra et al., 2023)	84.83	55.30	39.81	80.00	65.70	81.30	32.86	21.25	28.57	54.44
Human	97.63	91.70	87.92	97.80	94.24	91.87	85.71	90.00	76.19	91.17

Table 1: Main experimental results on selected VLLMs. "Random" and "Human" performance are the average accuracy by three attempts. Detailed descriptions of these performances are shown in Appendix B.1.1. Complete experiments are provided in Table 3.

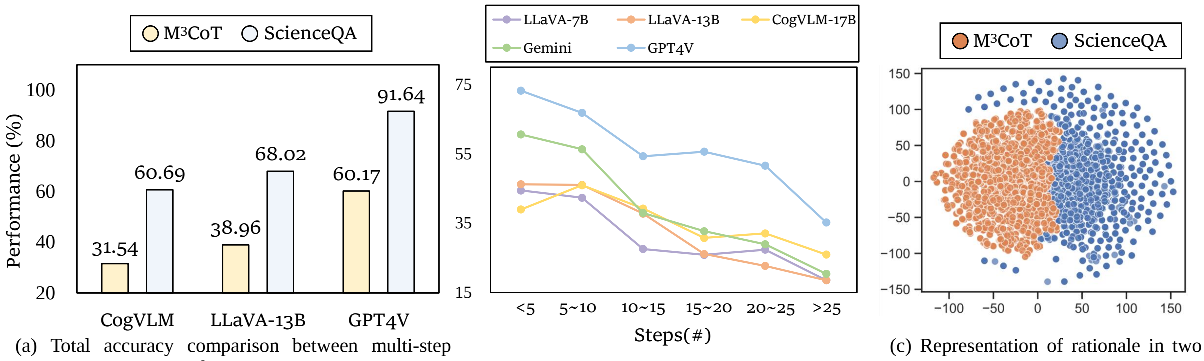
5.3 Analysis

To gain a deeper understanding of why VLLMs fail on $\texttt{M}^3\texttt{CoT}$, we analyze various factors to explore what influence the performance on $\texttt{M}^3\texttt{CoT}$. We provide more analysis details in Appendix B.3.1 to confirm our speculations.

Multi-step MCoT poses a greater challenge than single-step one. As shown in Figure 7 (a), VLLM has achieved amazing performance in single-step reasoning. However, compared with single-step MCoT data in ScienceQA, multi-step MCoT data in $\texttt{M}^3\texttt{CoT}$ maintains at least a 29.06% performance decrease (Figure 7 (a)). In order to further understand the difference in model reasoning with different numbers of steps, we calculated the accuracy of different steps. As illustrated in Figure 7 (b), an increase in the number of reasoning steps is associated with a significant decline in the model's performance. In Figure 7 (c), minimal rationale semantic distribution overlap between datasets further proves that the multi-step MCoT is an Out-of-Distribution (OOD) problem compared with single-step MCoT. For all, we attribute the low performance to the multi-step complexities for $\texttt{M}^3\texttt{CoT}$.

Multi-step MCoT needs higher rationale quality for better performance. We comprehensively assess the predicted rationale quality of various VLLMs based on five dimension criteria. As shown in Figure 8, we observe that rationale quality incrementally improves $\texttt{M}^3\texttt{CoT}$ performance, while it markedly impacts the accuracy in CoT tasks. In the future, we believe that improving rational quality is one of the key challenges to solving $\texttt{M}^3\texttt{CoT}$.

Multi-step MCoT needs more multi-modal interaction. To assess the necessity for more complex multi-modal interaction reasoning in $\texttt{M}^3\texttt{CoT}$, we examine how multi-modal interaction degrees impact performance. Specifically, we measure this by defining the similarity between images and reasoning steps to judge which steps are related to the image, identifying steps to sufficient similarity as



(a) Total accuracy comparison between multi-step MCoT sample in M³CoT and single-step MCoT sample in ScienceQA samples with the similar image.

(b) Performance comparison of different models across different annotated steps in M³CoT.

(c) Representation of rationale in two datasets via RoBERTa encoding and t-SNE dimensionality reduction.

Figure 7: Performance comparison and reason analysis of $\mathtt{M^3CoT}$ and ScienceQA.

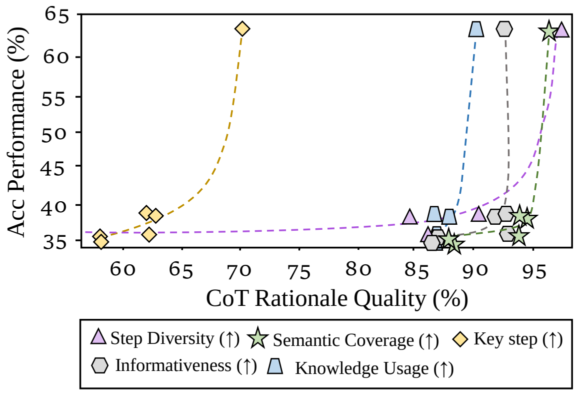


Figure 8: Analysis of the correlation between multi-dimensional qualities for model-generated rationale and final accuracy performance. The rationale qualities are computed by ROSCOE (Golovneva et al., 2023).

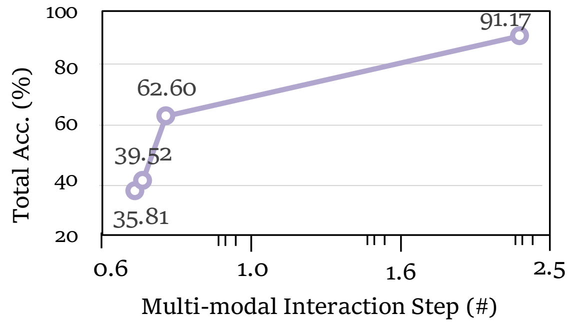


Figure 9: Analysis of the correlation between averaged multi-modal interaction steps and accuracy performance.

multi-modal interaction steps. Figure 9 illustrates a positive correlation between averaged multi-modal interaction steps and reasoning performance, indicating $\mathtt{M^3CoT}$ benefits from more multi-modal reasoning steps for optimal performance.

5.4 Exploration

In addition to the zero-shot CoT evaluation, we further evaluate the models on $\mathtt{M^3CoT}$ under three setups: (1) *Multi-modal Tool Usage*; (2) *Multi-modal In-Context-Learning*; (3) *Fine-tuning*. To our knowledge, these are the first comprehensive exploration of the multi-modal CoT scenarios.

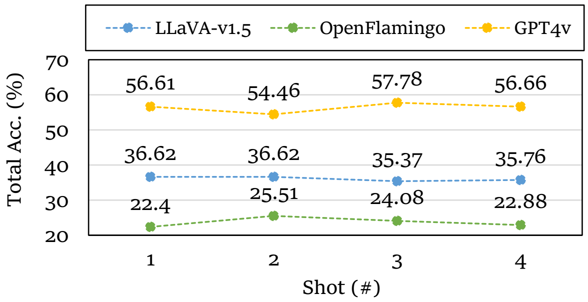
5.4.1 Tool Usage Exploration

Multi-modal tool usage on text-modal LLM fails on M^3CoT. Several studies highlight that ChatGPT can well use external multi-modal tools to help multi-modal reasoning. However, Table 2 reveals that those tool-usage models in single modality are significantly worse than that of GPT4V by 28.21%, and some even are worse than random baseline. We attribute it to the fact that the current tool usage framework cannot observe the visual modal during planning, which caused incorrect tool planning and tool usage, like confusing description and captioning tools. (as shown in Appendix B.3.2) This indicates the necessity for enhanced multi-modal information interaction within $\mathtt{M^3CoT}$. And we will show more implementation details in Appendix B.2.1.

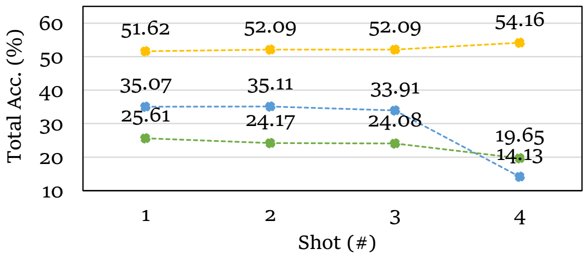
5.4.2 In-Context-Learning Exploration

Performance can not be boosted by text-only examples. Contrasting with textual CoT (Wei et al., 2022b; Shi et al., 2022), we find that ICL, even with chosen in-domain examples, fails to significantly improve multi-modal reasoning, as shown in Figure 10 (a). This suggests a need for more diverse multi-modal examples for $\mathtt{M^3CoT}$.

Performance may even be harmed by image and text interleaving example. In Figure 10 (b), it reveals that LLaVA-13B, untrained in interleaved image-text data, suffers performance degradation with more samples. Surprisingly, despite be-



(a) In-Context-Learning analysis on total accuracy performance with **textual only** demonstrations.



(b) In-Context-Learning analysis on total accuracy performance with **image and text interleaving** demonstrations.

Figure 10: Performance change analysis of In-Context-Learning (ICL) CoT on textual modality and multi-modality demonstrations.

ing trained on interleaved image-text data, OpenFlamingo (Awadalla et al., 2023) still exhibits a slight performance decline. In contrast, GPT4V, which is thoroughly trained on high-quality image-text interleaving examples, improves the performance as the number of shots increases. However, its performance is still lower than direct `CoT`. These indicate the future direction of high-quality interleaved samples and multi-step cross-modal interaction to enhance performance. All implementation details are shown in Appendix B.2.2.

5.4.3 Finetuning Exploration

To further explore the improvement on $\texttt{M}^3\texttt{CoT}$, we conduct finetuning experiments for more effective multi-modal reasoning. We will show more implementation details in Appendix B.2.3.

Finetuning on $\boldsymbol{\texttt{M}^3\texttt{CoT}}$ ***can result better performance.*** Table 2 reveals that our benchmark training set significantly enhances model performance. It enables traditional vision-language models (VLMs) to surpass the zero-shot VLLMs, which is the value of our dataset in boosting VLM effectiveness. Finetuned VLMs (the lowest is 44.85%) outperform most open-source VLLMs with zero-shot prompting (the highest is 38.86%). In addition, some finetuned VLMs have even surpassed Gemini's overall accuracy of 47.50%, which demonstrates that finetuning can effectively boost the performance.

Finetuning on VLLMs tends to be more effective than on Traditional VLM. Further, we found that the performance of VLLMs generally improves as their number of parameters increases. This also proves the importance of utilizing models with sufficient parameters in our $\texttt{M}^3\texttt{CoT}$ to achieve the target performance.

Takeaways: *(1) Visual and textual information should be both considered for tool planning. (2) We should consider better multi-modal interleaving for better ICL in* $\texttt{M}^3\texttt{CoT}$*. (3) Fine-tuning has better hope on multi-step MCoT, compared with the failures of vanilla in-context-learning, tool usage, and prompting strategies.*

6 Related Work

Chain-of-Thought (Wei et al., 2022b) (CoT) is a highly effective step-by-step strategy for enhancing zero-shot and few-shot reasoning in Large Language Models (LLMs) (Kojima et al., 2022; Zhou et al., 2022; Zelikman et al., 2022; Qin et al., 2023, 2024a; Zhuang et al., 2023). In addition, some works begin to extend the textual CoT capabilities to multi-modal CoT reasoning (MCoT) (Wang et al., 2023d; Singh et al., 2023; He et al., 2023). To this end, Lu et al. (2022a) introduce the ScienceQA benchmark, laying the foundation for MCoT. Subsequently, Zhang et al. (2023c) formally formalize the MCoT concept and enhanced its performance using a two-stage approach during multi-modal reasoning. Additionally, Wang et al. (2023a) develop a novel framework to integrate more knowledge with high-quality CoT rationales from larger LLMs for better MCoT reasoning. Further, Mondal et al. (2024) integrate CoT, Knowledge Graphs, and multi-modalities together for better MCoT. Ge et al. (2023); Zheng et al. (2023); Yao et al. (2023) manually decouple the chain-of-thought reasoning steps, integrating better multi-modal interaction. Building upon these works, Tan et al. (2023) introduce the self-consistency mechanism (Wang et al., 2022) into the training process to enable more accurate/reliable reasoning. Wei et al. (2023) propose a novel approach to improve reasoning capabilities in image and text encoders through the integration of multi-hop cross-modal attention and sentence-level contrastive learning. Chen et al. (2023) further extend the MCoT benchmark to generation tasks for

Model	Science			Commonsense			Mathematics			Total
	Lang	Natural	Social	Physical	Social	Temporal	Algebra	Geometry	Theory	
Random	32.70	30.62	26.71	32.97	22.22	20.33	35.71	27.50	23.81	28.56
Tool-Usage										
HuggingGPT (Shen et al., 2023)	17.57	20.93	10.33	8.70	14.75	9.76	11.35	22.50	9.52	14.60
VisualChatGPT (Wu et al., 2023a)	30.09	36.28	7.78	43.48	29.92	33.33	21.99	21.25	28.57	25.92
IdealGPT (You et al., 2023)	31.73	31.63	26.23	56.52	50.00	26.83	20.57	30.00	38.10	32.19
Chameleon (Lu et al., 2023b)	43.87	26.05	25.44	39.13	37.30	48.78	17.73	26.25	23.81	34.29
Finetuning (Traditional VLM)										
MM-CoT$_{base}$ (Zhang et al., 2023d)	41.71	46.49	39.90	59.34	60.91	27.64	48.57	35.00	28.57	44.85
MC-CoT$_{base}$ (Tan et al., 2023)	53.55	63.98	43.56	61.54	69.55	29.27	42.86	33.75	28.57	53.51
MM-CoT$_{large}$ (Zhang et al., 2023d)	45.50	50.19	43.56	63.74	64.61	33.33	40.71	61.25	28.57	48.73
MMR (Wei et al., 2023)	50.24	50.32	43.56	**76.92**	**66.67**	31.71	50.71	**65.00**	**38.10**	50.67
MC-CoT$_{large}$ (Tan et al., 2023)	**42.65**	**67.43**	**50.56**	58.24	60.49	**56.10**	**57.86**	62.50	14.29	**57.69**
Finetuning (VLLM)										
LLaMA-Adaper-7B (Zhang et al., 2023a)	62.56	72.29	30.21	76.92	59.67	72.36	30.71	38.75	38.10	54.89
LLaVA-V1.5-7B (Liu et al., 2023)	65.88	73.44	35.14	80.22	56.79	67.48	32.86	47.50	19.05	56.74
LLaVA-V1.5-13B (Liu et al., 2023)	**68.72**	72.41	**40.86**	**83.52**	64.61	69.11	**35.71**	45.00	38.10	**59.50**
CogVLM-17B (Wang et al., 2023c)	65.88	**77.52**	29.09	81.32	**65.43**	**75.61**	**35.71**	**46.25**	**47.62**	58.25
GPT4V$_{CoT}$ (OpenAI, 2023)	90.52	63.09	46.97	83.33	75.21	82.93	45.71	50.00	38.10	62.60
Human	97.83	92.62	94.31	96.28	92.41	88.71	87.23	88.75	85.71	91.61

Table 2: Fine-tuning results on various VLLMs.

better commonsense reasoning evaluation.

Compared with previous works, we first propose M^{3}CoT to explore the multi-step MCoT and extend their application across a broader range of domains. In addition, we conduct comprehensive experiments on M^{3}CoT and highlight some takeaways, facilitating future research.

7 Conclusion

In this work, we introduce a novel benchmark (M^{3}CoT), toward multi-domain, multi-step, and multi-modal chain-of-thought scenarios, which is developed through a detailed and comprehensive process. In addition, we conduct a comprehensive analysis involving abundant multi-modal CoT methodologies to understand the limitation of existing frameworks on M^{3}CoT. We sincerely aspire that our work can reassess existing advancements and inspire future research by highlighting new challenges and opportunities.

Limitations

We introduce a new benchmark for multi-domain, multi-step, and multi-modal Chain of Thought (M^{3}CoT) reasoning, performing in-depth analysis and exploring various CoT methodologies to better understand existing frameworks. However, limited by unavoidable human subjectivity, manual annotation may introduce potential biases that may affect the reliability of the data. Furthermore, with the advent of globalization, multilingualism has become increasingly important (Qin et al., 2024b). However, due to regional and cost restrictions, the data does not take into account multilingual backgrounds and was developed only in the single language of English. Moreover, due to the possibility of discontinued or retired models, we will pay more attention to the open-source models in the future.

Ethical Considerations

Data Access We sourced our data from the ScienceQA (Lu et al., 2022b), MATH (Hendrycks et al., 2021), TabMWP (Lu et al., 2022c), KiloGram (Ji et al., 2022), and Sherlock (Hessel et al., 2022). These are open-source and freely available for academic research, aligning with the commitment to ethical data use.

Participant Recruitment We recruit participants from universities and require all participants to pass the CET-6 exam or IELTS score of 6 or above. In addition, all participants come from all over China and may have some national biases. We blurred national differences in the data set as much as possible, limiting it to common human commonsense. All annotators gave informed consent and were compensated in excess of the local minimum wage. In addition, the site does not require IRB review.

Dataset Collection Process Our annotation process begins with an onboarding test, introducing the task through 100 example questions. Participants are compensated $20 for this initial phase, aimed at acquainting them with the task. Subsequently, annotators receive $15 per hour for their contributions, accumulating approximately 450 human-hours for manual annotations. Following this, a recheck process to ensure correct labeling is added an additional 60 hours of work. Overall, six experts and three students are engaged to fulfill the annotation and recheck tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) via grant 62306342, 62236004 and 62441603. This work was also sponsored by CCF-Baidu Open Fund and Excellent Young Scientists Fund in Hunan Province (2024JJ4070). We are grateful for resources from the High Performance Computing Center of Central South University.

References

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhan. 2023. Chain of thought prompt tuning in vision language models. *arXiv preprint arXiv:2304.07919*.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Google. 2023. Gemini: A family of highly capable multimodal models.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2023. Multi-modal latent space learning for chain-of-thought reasoning in language models. *arXiv preprint arXiv:2312.08762*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mengkang Hu, Yao Mu, Xinmiao Chelsey Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. 2024. Tree-planner: Efficient close-loop task planning with large language models. In *The Twelfth International Conference on Learning Representations*.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. 2022. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 582–601.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, and Min Zhang. 2023b. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023b. Chameleon: Plug-and-play compositional reasoning with large language models. *CoRR*, abs/2304.09842.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022c. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2023. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2311.17076*.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. *arXiv preprint arXiv:2401.12863*.

OpenAI. 2023. Gpt-4 technical report.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2024a. Large language models meet nlp: A survey.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024b. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, and Gust Verbruggen. 2023. Assessing gpt4-v on structured reasoning tasks. *arXiv preprint arXiv:2312.11524*.

Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Xihong Yang, and Stan Z Li. 2023. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. *arXiv preprint arXiv:2311.14109*.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023a. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *arXiv preprint arXiv:2305.03453*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023c. Cogvlm: Visual expert for pretrained language models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zefeng Wang, Zhen Han, Shuo Chen, Volker Tresp, and Jindong Gu. 2023d. Towards the adversarial robustness of vision-language model with chain-of-thought reasoning.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jingxuan Wei, Cheng Tan, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z Li. 2023. Enhancing human-like multi-modal reasoning: A new challenging dataset and comprehensive framework. *arXiv preprint arXiv:2307.12626*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C Gee, and Yixin Nie. 2023b. The role of chain-of-thought in complex vision-language reasoning task. *arXiv preprint arXiv:2311.09193*.

Fanglong Yao, Changyuan Tian, Jintao Liu, Zequn Zhang, Qing Liu, Li Jin, Shuchao Li, Xiaoyu Li, and Xian Sun. 2023. Thinking like an expert: Multimodal hypergraph-of-thought (hot) reasoning to boost foundation modals. *arXiv preprint arXiv:2308.06207*.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A. Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

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

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhehao Zhang, Xitao Li, Yan Gao, and Jian-Guang Lou. 2023b. CRT-QA: A dataset of complex reasoning question answering over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2131–2153, Singapore. Association for Computational Linguistics.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023d. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Ziyu Zhuang, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Zixian Feng, Weinan Zhang, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. *arXiv preprint arXiv:2308.07902*.

Appendix

A Dataset Annotation Details

A.1 Statistical Analysis of Existing Datasets

In this study, we conduct a comprehensive analysis of the prevalence of multi-step reasoning within existing datasets. The analysis focuses on the "MM-CoT" column in Figure 2(b), which represents the proportion of multi-step multi-modal CoT (Chain of Thought) data. This examination is critical as it reveals a significant deficiency in the current datasets regarding multi-step reasoning capabilities. Our findings indicate that at least 79% of the data across all the benchmarks we examined lack sufficient multi-step reasoning elements. This highlights a pervasive issue in the design and utilization of these datasets, where the absence of complex reasoning processes could impede the development of more sophisticated multi-modal models.

The term %MMCoT, as used in this context, is consistent with the representation in Figure 2(b). Both denote the proportion of multi-step multi-modal CoT data. To ensure the accuracy of our analysis, we employed a rigorous sampling method. We selected a stratified random sample consisting of 20% of the dataset, evenly distributed, for manual inspection. This approach allowed us to precisely determine the proportion of multi-step multi-modal CoT data within the existing datasets.

The manual inspection involved a detailed review of each sampled data point to identify and categorize the presence of multi-step reasoning. This meticulous process ensured that our findings were not only statistically significant but also reflective of the true nature of the datasets. The results underscore the necessity for enhanced dataset designs that incorporate more multi-step reasoning tasks, thereby facilitating the development of advanced multi-modal models capable of handling complex reasoning scenarios.

In conclusion, our statistical analysis sheds light on a critical gap in existing datasets, emphasizing the need for more comprehensive data that can better support the advancement of multi-modal reasoning capabilities.

A.2 The Details of Absence of Visual Modal Reasoning Sample Removal

We develop the annotation interface based on the open-source Gradio framework. We segment the dataset, distribute the scripts, and deploy them to local computers. In addition, we have designed some manual guidelines for annotators. These guidelines need to be followed by annotation experts. Specifically, our method flow is as follows:

1. We will first mask the image so that the expert can only see the text modal questions, options, rationale, and answers.
2. The expert needs to directly judge whether the question can directly infer the rationale. If it cannot be inferred, it may be necessary for the visual modal information for the rationale generation.
3. And then we ask the experts to check the image to confirm this.

For each step, the guideline instructions are as follows:

[Instruction$_1$]
Firstly, we will conceal all visual content, allowing you to concentrate exclusively on the textual elements, including the questions, available options, correct answers, and rationales.
You need to follow these instructions for annotation:

- You should determine whether the text alone can convey the full context and solution logic of the question.
- You should evaluate whether the textual information is sufficient for a comprehensive understanding and rationale formulation.

[Instruction$_2$]
In cases where the text-based evaluation indicates a need for visual information, please click the “Image Display” button.
Then, the image will be reintroduced for further recheck, confirming the requirement for a multi-modal approach in certain scenarios.

We used three experts to conduct majority voting to judge this matter. This part removes at least 30% of ScienceQA’s multi-modal sample, which also illustrates the limitations of the existing data.

A.3 The Details of Multi-step MCoT Sample Construction

Automatic Sample Removal: In this step, we automatically filter out samples with overly simplistic rationales comprising fewer than two steps. This

Model	Science			Commonsense			Mathematics			Total
	Lang	Natural	Social	Physical	Social	Temporal	Algebra	Geometry	Theory	
Random	32.70	30.62	26.71	32.97	22.22	20.33	35.71	27.50	23.81	28.56
Kosmos-2-2B (Peng et al., 2023)										
`Direct` (Peng et al., 2023)	10.43	28.61	21.18	33.33	17.77	28.46	21.43	21.25	14.29	23.17
`CoT` (Kojima et al., 2022)	18.48	24.14	14.65	30.00	14.46	9.76	17.14	18.75	0.00	18.68
`Desp-CoT` (Wu et al., 2023b)	0.00	0.00	0.00	1.11	0.00	0.00	0.00	0.00	0.00	0.04
`CCoT` (Mitra et al., 2023)	0.00	0.00	0.16	2.22	7.44	1.63	0.00	0.00	0.00	0.99
InstructBLIP-7B (Dai et al., 2023)										
`Direct` (Dai et al., 2023)	30.81	32.31	27.55	60.00	66.94	39.02	35.71	31.25	33.33	36.11
`CoT` (Kojima et al., 2022)	38.39	30.01	26.43	80.00	70.25	33.33	30.71	21.25	19.05	35.76
InstructBLIP-13B (Dai et al., 2023)										
`Direct` (Dai et al., 2023)	38.39	30.52	26.27	76.67	70.66	35.77	30.00	22.50	19.05	35.94
`CoT` (Kojima et al., 2022)	38.39	30.01	27.55	80.00	70.25	33.33	30.71	21.25	19.05	36.07
`Desp-CoT` (Wu et al., 2023b)	16.59	27.84	22.77	54.44	52.89	30.08	27.86	28.75	28.57	29.25
`CCoT` (Mitra et al., 2023)	13.27	26.95	24.84	62.22	67.36	41.46	25.00	25.00	23.81	31.28
LLava-V1.5-7B (Liu et al., 2023)										
`Direct` (Liu et al., 2023)	43.13	37.16	26.43	66.67	58.26	30.89	22.14	35.00	14.29	36.63
`CoT` (Kojima et al., 2022)	38.86	33.59	25.48	71.11	65.29	39.02	29.29	16.25	4.76	35.81
`Desp-CoT` (Wu et al., 2023b)	34.12	32.18	25.32	65.56	57.85	41.46	24.29	31.25	28.57	34.43
`CCoT` (Mitra et al., 2023)	26.54	35.50	28.66	62.22	55.79	44.72	29.29	31.25	9.52	35.72
LLava-V1.5-13B (Liu et al., 2023)										
`Direct` (Liu et al., 2023)	36.97	27.46	20.22	52.22	23.55	27.64	22.86	45.00	4.76	27.05
`CoT` (Kojima et al., 2022)	46.45	38.31	27.87	67.78	64.05	49.59	26.43	30.00	23.81	39.52
`Desp-CoT` (Wu et al., 2023b)	47.87	29.25	27.23	68.89	59.92	47.15	26.43	36.25	9.52	35.98
`CCoT` (Mitra et al., 2023)	38.86	31.55	28.18	72.22	61.57	39.84	29.29	36.25	28.57	36.45
CogVLM-17B (Wang et al., 2023c)										
`Direct` (Wang et al., 2023c)	52.61	37.42	26.91	55.56	54.13	29.27	29.29	32.50	23.81	37.19
`CoT` (Kojima et al., 2022)	51.18	43.81	29.30	54.44	39.26	31.71	35.71	33.75	33.33	38.91
`Desp-CoT` (Wu et al., 2023b)	46.92	35.63	25.80	48.89	47.52	38.21	27.14	31.25	19.05	35.07
`CCoT` (Mitra et al., 2023)	47.39	34.99	25.80	62.22	46.28	35.77	30.71	37.50	23.81	35.63
Gemini (Google, 2023)										
`Direct` (Google, 2023)	73.93	41.25	31.21	56.67	71.49	62.60	30.71	27.50	28.57	45.17
`CoT` (Kojima et al., 2022)	67.30	49.68	36.31	68.89	60.33	66.67	23.57	21.25	9.52	47.50
`Desp-CoT` (Wu et al., 2023b)	49.29	43.68	27.07	63.33	57.85	70.73	28.57	30.00	28.57	41.85
`CCoT` (Mitra et al., 2023)	36.49	31.16	27.39	71.11	36.78	55.28	20.71	16.25	0.00	32.61
GPT4V (OpenAI, 2023)										
`Direct` (OpenAI, 2023)	80.09	54.66	43.95	87.78	67.77	82.11	42.14	43.75	**42.86**	56.95
`CoT` (Kojima et al., 2022)	**90.52**	**63.09**	**46.97**	83.33	**75.21**	**82.93**	**45.71**	**50.00**	38.10	**62.60**
`Desp-CoT` (Wu et al., 2023b)	79.62	54.66	36.94	**88.89**	74.38	73.98	20.71	32.50	33.33	53.54
`CCoT` (Mitra et al., 2023)	84.83	55.30	39.81	80.00	65.70	81.30	32.86	21.25	28.57	54.44
Human	97.63	91.70	87.92	97.80	94.24	91.87	85.71	90.00	76.19	91.17

Table 3: The overall experimental results using selected VLLMs by zero-shot prompting. The "`Direct`" approach refers to submitting samples in the VLLM required format. "Human" is the average accuracy achieved by three college students who have successfully completed a relevant assessment. Complete experiments are provided in the Appendix.

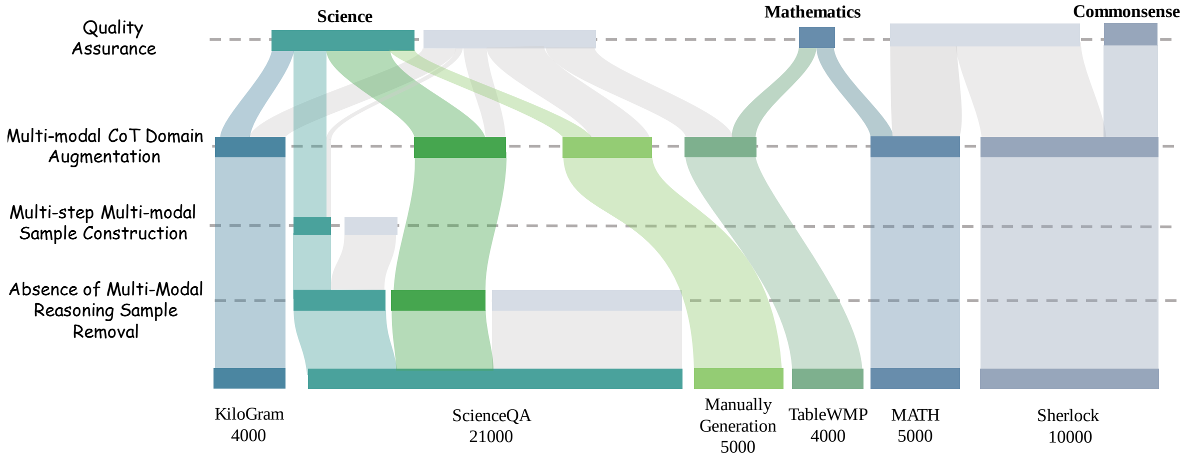


Figure 11: Sample distribution flow chart retained, generated, discarded at different stages

process reduces the manual annotation burden and enhances the reliability of $\mathtt{M^3CoT}$. Since the rationale in ScienceQA includes at least one conclusion and one step of reasoning, samples with fewer than two steps indicate that multiple visual cues were not used for MMCoT reasoning. Thus, this filtering step minimizes annotation workload and costs. Notably, samples with multi-step and single-step reasoning still require manual evaluation in our study.

A.4 Domain Augmentation Details

A.4.1 Mathematics Domain Augmentation Details

Firstly, the MATH (Hendrycks et al., 2021) dataset consists solely of textual mathematical questions, complete with detailed rationales and answers. However, it does not include multiple-choice options and illustrative images, limiting its utility in $\mathtt{M^3CoT}$.

To address this, we employ `gpt-3.5-turbo` to generate relevant and similar multiple-choice options for each question, enhancing the dataset's versatility. Specifically, for the option generation phase, we prompt the LLM with specific questions from the dataset, followed by instructions to generate four plausible options, one correct and three distractors. The prompt included guidelines for ensuring the options are closely related to the question's content, challenging yet not misleading. Specifically, the prompt is defined as follows:

[Input]
Here is the mathematical question:
<Context, Question, Answer, Option>

[Instruction]
Generate four multiple-choice options: one correct answer and three plausible but incorrect distractors.
Ensure the distractors are relevant and challenging without being misleading.
The options should closely relate to the question's subject matter and provide a meaningful test of the reader's understanding.

To compensate for the absence of visual content, we developed a method to translate mathematical expressions and geometric figures described in the questions into visual representations. This process involved generating PNG images from the mathematical and geometric codes. Subsequently, we used a combination of HTML and CSS to integrate these images with the textual content, creating a cohesive multi-modal dataset.

A.4.2 Commonsense Domain Augmentation Details

In our study, we aim to enhance commonsense reasoning domain for $\mathtt{M^3CoT}$ by leveraging visual clues from the Sherlock (Hessel et al., 2022) dataset. Unlike traditional vision question answering datasets, Sherlock (Hessel et al., 2022) does not include predefined questions, options, or answers, focusing instead on visual information to stimulate inference and deduction.

Recent advancements in LLMs have showcased their potential in generating diverse and contextually relevant data (Zhang et al., 2023b). Building on this, our approach involves prompting LLMs with visual clues from Sherlock, tasking them to generate coherent and contextually appropriate questions, multiple-choice options, and corresponding answers. This process demands careful design to ensure the prompts effectively communicate the visual information and desired output format to the LLM.

To ensure comprehensive multi-step, multi-modal reasoning, we develop a prompting methodology to trigger LLMs to consider multiple visual clues simultaneously. Specifically, our experimental setup includes detailed prompting strategies that describe the visual clues from a structured manner to natural language description, allowing the LLM to understand and interpret the information accurately. The prompt is defined as follows:

> **[Instruction]**
> Given the visual clues, generate a question that requires commonsense reasoning to answer. Then, provide four options (A, B, C, D), one of which is the correct answer, and the others are plausible distractors.
>
> **[Highlighting]**
> Ensure that the question and options leverage insights from at least two visual clues for deeper multi-modal interactions.
>
> **<One-shot Example>**
>
> **[Visual Clue]**
> <*multiple detailed description of the visual clues*>

where **[One-shot Example]** denotes we used one-shot in-context-learning to allow the model to better learn the generation of related samples.

In addition, for some topics in ScienceQA where the data is too sparse due to the last "absence of multi-modal reasoning sample removal" and "multi-step multi-modal sample construction", we used a similar method to synthesize data in ScienceQA or used other open source data sets, like TabMWP (Lu et al., 2022c), KiloGram (Ji et al., 2022), for data augmentation. Moreover, we also manually synthesized some geographical images through Matplotlib, constructed some samples using rules, and polished and modified them using ChatGPT.

A.5 Quality Assurance Details

Due to space limitations, we only describe the rationale rewriting (§A.5.3) section in the appendix. This part is mainly to improve the quality of the rationale data set.

A.5.1 Human Annotation Details

In order to better mark the correctness of the logical chain of reasoning, first, we divide the steps according to the ROSCOE (Golovneva et al., 2023) settings to obtain a clearer step-by-step rationale visualization. Secondly, we provide the corresponding sample image, question, options, answer and step-segmented rationale for experts to annotate each time. During annotation, we allow experts to discard samples that are of poor quality and cannot be modified to ensure the quality of the data set.

A.5.2 Human Recheck Details

In assessing the capability of a given sample to fulfill the criteria for multi-step multi-modal reasoning, this process employs a structured approach. Initially, we decompose the reasoning rationale into discrete steps by ROSCOE (Golovneva et al., 2023). Following this segmentation, experts are required to focus more on ascertaining which step needs image modality in rationale. After that, it requires further verification that the sample has considered integration of image modality for a minimum of two distinct steps. This methodological framework ensures a thorough recheck of the sample with the specified requirements of multi-step multi-modal reasoning. Specifically, the instructions given by our experts are as follows:

> **[Input]**
> Here is an example: <*EXAMPLE*>
>
> **[Instruction]**
> You need to judge whether a given sample meets the requirements of multi-step multi-modal reasoning.
>
> 1. First, we have broken down the steps for you.
> 2. Secondly, please determine which steps in rationale require image modality.
> 3. Finally, please confirm whether the sample requires image modality for at least two steps.

where "*[EXAMPLE]*" represents an example containing an image, question, options, answer, step-segmented rationale and annotation detail information.

Furthermore, the human recheck process actually has two rounds. In the second round, the sample discard rate is less than 5%.

A.5.3 Rationale Rewriting

The rationale quality within the ScienceQA dataset has been found poor expression, with some explanations not adequately addressing the posed questions. To mitigate this issue, we have employed `gpt-3.5-turbo` to perform rationale rewriting, aiming to elevate the overall quality of M^{3}CoT before human annotation. Specifically, to achieve this, we designed a specific prompting strategy for the LLM, which prompt is defined as follows:

[Instruction]
Improve the quality of ScienceQA dataset rationales by rewriting them for enhanced relevance, accuracy, and clarity.

1. Read the question and the provided rationale from the ScienceQA dataset.
2. Evaluate the existing rationale for its relevance and accuracy in answering the question.
3. Rewrite the rationale to better answer the question, ensuring the new version is clear, concise, and directly related to the question's core topic. Maintain scientific accuracy and use accessible language suitable for the intended audience.

[One-shot Example]
Question: <*Question*>
Original Rationale: <*Rationale*>
Rewritten Rationale: <*Rewritten Rationale*>

[Input]
Question: <*Question*>
Original Rationale: <*Rationale*>
Rewritten Rationale:

This approach ensures that the rewritten rationales are not only relevant but also adhere to a high standard of clarity and coherence. Each rationale is assessed both automatically and manually to confirm its relevance and quality improvement over the original version.

A.6 Image Redundancy Removal

Additionally, we observed a significant number of highly similar samples in the ScienceQA. To reduce redundancy and maintain diversity for image, we remove samples where the questions are identical, and the grayscale image similarity exceeded 99%.

B Experiment Details

B.1 Main Result Details

Due to space limitations, we only show some of the LLM test results in the main table. The specific experimental results are shown in Table 3.

B.1.1 Heuristic baselines

This study employs two heuristic baselines. The first is a random selection method, where an answer is chosen randomly, with its accuracy determined by averaging three random seeds. The second baseline evaluates human performance through participants who must successfully complete preliminary qualification tasks. The "Human" accuracy is the average accuracy achieved by three participants.

B.2 Exploration Details

B.2.1 Tool Usage Details

Model Selection In this section, we introduce a suite of tool-augmented LLMs, including *HuggingGPT* (Shen et al., 2023), *VisualChatGPT* (Wu et al., 2023a), *IdealGPT* (You et al., 2023), and *Chameleon* (Lu et al., 2023b). Specifically, *VisualChatGPT* (Wu et al., 2023a) and *IdealGPT* (You et al., 2023) are engineered to tackle complex issues through iterative problem-solving processes. Conversely, *HuggingGPT* (Shen et al., 2023), and *Chameleon* (Lu et al., 2023b) employ LLMs to decompose complicated challenges into a series of manageable sub-problems, addressing them in a sequential manner. This array of approaches highlights the diverse capabilities and potential of LLM-aided visual reasoning systems in executing sophisticated problem-solving strategies.

B.2.2 In-Context-Learning Details

Model Selection In this section, we explore three notable models: *LLaVA-V1.5-13B* (Liu et al., 2023), *OpenFlamingo-7B* (Awadalla et al., 2023), and *GPT4V* (OpenAI, 2023), each demonstrating unique capabilities in the context of In-Context Learning (ICL). LLaVA-V1.5-13B (Liu et al., 2023) is built upon a foundation of instruction-following data of high quality without any specific image-text interleaving training. *OpenFlamingo* (Awadalla et al., 2023) is a VLLM, optimized for tasks involving complex image-text interleaving sequences. GPT4V (OpenAI, 2023) is the state-of-the-art VLLM that can learn efficiently from limited image-text interleaving demonstrations. Therefore, we assume it is well-trained in image-text interleaving scenarios.

Exemplar Selection In order to complete in-domain sample selection as much as possible, we only randomly select samples under the same categories from the development set.

B.2.3 Finetuning Details

Model Selection Our fine-tuning section incorporates a carefully curated selection of models, which includes a series of traditional Vision-Language Models (VLMs) and Vision Large Language Models (VLLMs). Specifically, VLMs contain MM-CoT (Zhang et al., 2023c), MC-CoT (Tan et al., 2023), and MMR (Wei et al., 2023). VLLM include LLaMA-Adapter (Zhang et al., 2023a), LLaVA-V1.5 (Liu et al., 2023), CogVLM (Wang et al., 2023c). This selection was strategically made to encompass a wide range of parameter sizes, architectures, and functionalities. This diversity ensures a comprehensive evaluation of the state-of-the-art in multi-modal capability on finetuning settings.

Experiment Setting In the context of finetuning VLLMs, parameter efficient tuning always achieves better performance compared with full-parameter tuning and offers a training cost reduction (Hu et al., 2021; Liu et al., 2023). To leverage these benefits, we employ LoRA (Hu et al., 2021) for parameter-efficient tuning across our experiments. However, for specific cases like the LLaMA-Adapter, we integrate a compact adapter module for fine-tuning, adding minimal additional parameters to the model's architecture.

Our training configurations include a selection of batch sizes from $\{2, 4, 8\}$ and learning rates ranging from $[1e-6, 8e-5]$. We standardize the maximum sequence length to 512 tokens for uniformity across all model trainings. The experiments are conducted on NVIDIA A100 and A800 GPUs to ensure optimal performance and efficiency. For all experiments, model selection is based on the best performance on the development set, which is then validated on the test set for final evaluation.

B.3 Error Analysis

B.3.1 Zero-shot Chain-of-Thought Error Analysis

In order to further analyze the typical errors in the data set, we analyzed the cases of GPT4V on different domains (as shown in Figure 12, Figure 13, and Figure 14). We can find that all cases have logical errors and visual information interaction errors or deficiencies. This view is also consistent with Section 5.3 mutual confirmation. Therefore, we believe that the lack of high-quality logical reasoning capabilities and complex multi-modal interactions of large models leads to the failure of multi-step multi-modal reasoning of the model.

B.3.2 Tool Usage Error Analysis

In the context of tool-usage methodologies, an initial challenge emerges from the image information mistake and logical mistake, as highlighted in Appendix B.3.1. This limitation becomes particularly problematic in complex scenarios involving multiple tools and steps, as seen in multi-modal task planning. These scenarios demand precise tool selection and sequencing; however, the lack of visual interaction during tool planning leads to frequent errors in tool selection (as shown in Figure 16) and tool-chain redundancy(as shown in Figure 15). Incorrect tool planning or selection can cascade through the process, culminating in complete failure of the intended task (as shown in Figure 16). This issue underscores the need for enhanced model capabilities in processing and integrating visual modalities to accurately navigate multi-step, multi-tool workflows on $\texttt{M}^3\texttt{CoT}$.

Question:

What can you infer about the airplane?

Options:

(A) It is a commercial plane (B) It is a military plane

(C) It is a stunt plane (D) It is a cargo plane

GPT4V CoT

Sure, let's break it down step by step!

Step 1: Appearance and Design:

The airplane in the image has a distinct design that is not typical of commercial planes. It appears to be a single-seater with a narrow fuselage and wing design.

The markings and paint job suggest it is not a typical commercial or cargo plane.

Step 2: Purpose:

Commercial planes are usually larger and designed to carry passengers, but this plane doesn't seem to have multiple windows or the size typical of commercial aircraft.

Cargo planes are also generally larger and are built for carrying goods, but this plane lacks the bulk and structure one would expect from a cargo plane.

Stunt planes are typically agile, lightweight, and sometimes have distinctive designs. They are often smaller and built for maneuverability.

Step 3: Markings:

[Image Information Mistake] The markings on the plane seem to be more in line with military insignias or symbols. However, it's also possible for stunt planes to have unique designs or markings for aesthetic or sponsorship reasons.

[Logical Mistake] Based on the provided information and the appearance of the plane in the image, the closest match would be: (B) It is a military plane

However, without more specific information about the plane's origin or its intended use, this is an educated guess based on its visual appearance.

Figure 12: Response from GPT4V with CoT prompting on commonsense domain.

Question:

Let's analyze the options given for the Geometry problem in the image.

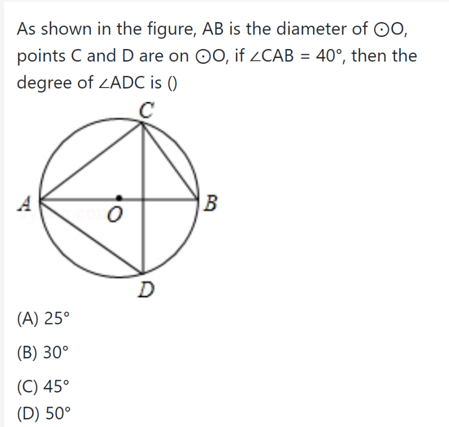
Options:

(A) the answer is option A in the image

(B) the answer is option D in the image

(C) All of the answer choices given in the text are wrong.

(D) the correct choice is option C in the image.



GPT4V CoT

Alright, let's break this problem down.

Step 1:

From the figure:

AB is the diameter of circle O. This means ∠AOB is 180° because it's a straight angle.

Given: ∠CAB = 40°

Step 2:

Recall a property of a circle:

The angle subtended by a semicircle (an angle subtended by the diameter) is always a right angle. This means: ∠CBA = 90°

Step 3:

From the above information:

∠CAB + ∠CBA + ∠CBO = 180°

[Image Information Mistake] Given ∠CAB = 40°, and we know ∠CBA = 90°

So, ∠CBO = 180° - 90° - 40° = 50°

Step 4:

[Image Information Mistake / Logical Mistake] Now, note that in circle O, the arc CD subtends both ∠CAB and ∠ADC at the circumference, making them angles subtended by the same arc.

A property of angles subtended by the same arc is:

Angle subtended by an arc at the center is double the angle subtended by it at any point on the remaining part of the circle.

This means:

2 * ∠CAB = ∠CBO

2 * 40° = 80°

But, ∠CBO = 50° (as we've calculated above)

Thus, the given information is inconsistent.

Given our calculations:

(A) 25° is incorrect.

(B) 30° is incorrect.

(C) 45° is incorrect.

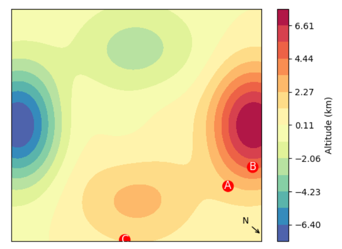
(D) 50° is incorrect.

Based on the provided information and the options, the correct answer is:

(C) All of the answer choices given in the text are wrong.

Figure 13: Response from GPT4V with CoT prompting on mathematics domain.

Context:
The photograph reveals a terrain chart depicting a particular area. The picture shows the terrain map of a certain place. The illustration displays both a contour map paired with a directional symbol.
Question:
What are the altitudes of points Point B and Point A respectively?
Options:
(A) 4.6, 2.1 (B) 4.6, 3.0
(C) 2.1, 4.6 (D) 4.6, 3.9
(E) 3.9, 4.6



GPT4V CoT
Let's analyze the image step by step!
Step 1: to determine the altitudes of points Point B and Point A:
Step 1.0: Understand the Scale:
The color scale on the right provides the altitude in kilometers. The darker red represents a higher altitude while the lighter blue signifies a lower altitude.
Step 1.1: Point B
[Image Information Mistake] Point B is located within the darkest red region.
Referring to the color scale, this darkest red corresponds to the highest altitude on the scale, which is approximately 6.61 km.
Step 1.2: Point A
Point A is located within a lighter shade of red but not the lightest.
[Image Information Mistake] Referring to the color scale, this shade of red seems to lie between 4.44 km and 2.27 km. Given the gradient of the colors, it's closer to 4.44 km than 2.27 km, so it's safe to assume it's around 4.44 km.
Step 2: Based on the analysis
Altitude of Point B is approximately 6.61 km.
Altitude of Point A is approximately 4.44 km.
Step 3:
However, none of the choices provided exactly match these values. Let's find the closest match:
(A) 4.6, 2.1 (B) 4.6, 3.0
(C) 2.1, 4.6 (D) 4.6, 3.9
(E) 3.9, 4.6
Step 4:
From the options, the closest match to our analysis is:
Point B: 4.6 km (although our analysis showed 6.61 km, none of the options matches this, so we have to go with the closest one)
Point A: 3.9 km (close to our estimation of 4.44 km)
[Logical Mistake] Therefore, the best choice would be (D) 4.6, 3.9.

Figure 14: Response from GPT4V with CoT prompting on science domain.

Question:

What is the most probable function of the building with country flags and a tall building nearby?

Options:

(A) A convention center

(B) An embassy

(C) A government building

(D) An global office building

IdealGPT

Tool Track:

Sub-question: What is the architectural style of the building with country flags? Answer: modern
Sub-question: What are the countries represented by the flags? Answer: united states
Sub-question: What is the architectural style of the tall building nearby? Answer: modern
Sub-question: What is the location of the building with country flags and the tall building? Answer: in new york city
Sub-question: What is the surrounding environment of the building with country flags and the tall building? Answer: street
[Image Information Mistake] Sub-question: Are there any signs or logos on the building with country flags that could indicate its function? Answer: no
Sub-question: Are there any security measures visible around the building with country flags that could suggest it is an embassy or government building? Answer: no
Sub-question: Are there any nearby landmarks or attractions that could suggest the function of the building with country flags and the tall building nearby? Answer: no
Sub-question: What is the size of the building with country flags compared to the tall building nearby? Answer: the tall building is taller
Sub-question: Are there any vehicles or transportation modes visible around the building with country flags and the tall building nearby? Answer: no
Sub-question: What is the time of day in the image? Answer: morning
Sub-question: Are there any people visible around the building with country flags and the tall building nearby? Answer: no
Sub-question: What is the overall design of the building with country flags and the tall building nearby? Answer: modern
Sub-question: Are there any events or activities happening around the building with country flags and the tall building nearby? Answer: no
Sub-question: What is the shape and size of the building with country flags? Answer: square
Sub-question: What is the distance between the building with country flags and the tall building nearby? Answer: 2 blocks
Sub-question: What is the color scheme of the building with country flags and the tall building nearby? Answer: white
Sub-question: Are there any other buildings or structures visible in the background that could provide context for the function of the building with country flags and the tall building nearby? Answer: no
Sub-question: What is the overall atmosphere or vibe of the surrounding area? Answer: urban
Sub-question: Are there any other visual cues or details that could suggest the function of the building with country flags and the tall building nearby? Answer: no
Sub-question: What is the surrounding neighborhood of the building with country flags and the tall building nearby? Answer: new york city
Sub-question: What is the history of the building with country flags and the tall building nearby? Answer: the building is the united nations building
Sub-question: What is the view from the top of the tall building nearby? Answer: a view of the city
Sub-question: What is the purpose of the nearby buildings and structures? Answer: to serve as a headquarters
Sub-question: Are there any nearby businesses or organizations that could suggest the function of the building with country flags and the tall building nearby? Answer: no
Sub-question: What is the age of the building with country flags and the tall building nearby? Answer: 20th century
Sub-question: What is the material used to construct the building with country flags and the tall building nearby? Answer: glass
Sub-question: What is the overall shape and design of the tall building nearby? Answer: square
[Image Information Mistake] Sub-question: Are there any notable features or unique characteristics of the building with country flags and the tall building nearby? Answer: no
Sub-question: What is the view of the building with country flags and the tall building nearby from different angles? Answer: different angles

Please follow the above-mentioned instruction to list the Analysis and More Likely Answer.

Reasoner(GPT):

Analysis: Based on the sub-questions and answers provided, we can infer that the building with country flags and the tall building nearby is the United Nations building in New York City. The architectural style of both buildings is modern, and the building with country flags is square in shape and made of glass. The surrounding environment is urban, and there are no visible signs or logos indicating the function of the building with country flags. There are no security measures visible, and there are no nearby landmarks or attractions that could suggest the function of the buildings. The overall atmosphere of the surrounding area is busy, and there are no events or activities happening around the buildings. The tall building nearby is taller than the building with country flags, and the distance between them is two blocks.
[Logical Mistake] The building with country flags and the tall building nearby serve as the headquarters of the United Nations.

More Likely Answer: A government building.

Figure 15: Response from IdealGPT on commonsense domain.

Question:

Which movie set is represented in the image?

Options:

(A) Back to the Future

(B) Indiana Jones and the Last Crusade

(C) Ferris Bueller's Day Off

(D) National Lampoon's Vacation

HuggingGPT

Tool Track:

[Tool Usage Mistake]

IMAGE-CLASSIFICATION(image)

```
QUESTION-ANSWERING(
  question = "Which movie set is represented in the image?",
  choices = [
        "Back to the Future",
        "Indiana Jones and the Last Crusade",
        "Ferris Bueller's Day Off",
        "National Lampoon's Vacation"
  ])
```

It is not possible to accurately determine which movie set is represented in the image without actually viewing the image.

(a) Single Tool Usage Mistake from HuggingGPT

Context:

Question:

Based on the image, which of the following statements is more likely to be true about the person in this picture?

Options:

(A) The woman hate cats.

(B) The woman prefers green over pink

(C) All options are incorrect

(D) The woman is fond of floral decorations

HuggingGPT

Tool Track:

[Tool Usage Mistake]

var_1 = IMAGE-CLASSIFICATION(image)

[Tool Usage Mistake]

var_2 = TEXT-CLASSIFICATION("[Question]Based on the image, which of the following statements is more likely to be true about the person in this picture?[Choices](A) The woman hate cats.(B) The woman prefers green over pink(C) All options are incorrect (D) The woman is fond of floral decorations")

TEXT-Generation(var_2)

Without being able to view the image itself, it is not possible to determine which of the statements is more likely to be true about the person in the picture.

(b) Multi-Tool Chain Usage Mistake from HuggingGPT

Figure 16: Response from HuggingGPT on commonsense domain.