

Multimodal Causal Reasoning Benchmark: Challenging Multimodal Large Language Models to Discern Causal Links Across Modalities

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have showcased exceptional Chain-of-Thought (CoT) reasoning ability in complex textual inference tasks including causal reasoning. However, will these causalities remain straightforward when crucial hints hide in visual details? If not, what factors might influence cross-modal generalization? Whether we can effectively enhance their capacity for robust causal inference across both text and vision? Motivated by these, we introduce **MuCR** - a novel **Multimodal Causal Reasoning** benchmark that leverages synthetic siamese images and text pairs to challenge MLLMs. Additionally, we develop tailored metrics from multiple perspectives, including image-level match, phrase-level understanding, and sentence-level explanation, to comprehensively assess MLLMs' comprehension abilities. Our experiments reveal that current MLLMs fall short in multimodal causal reasoning compared to their performance in purely textual settings. Additionally, we find that identifying visual cues across images is key to effective cross-modal generalization. Finally, we propose a **VcCoT** strategy that better highlights visual cues, and our results confirm its efficacy in enhancing multimodal causal reasoning.

1 Introduction

Causal reasoning is the process of identifying the relationship between a cause and its effect, which is regarded as a fundamental capability of artificial intelligence (Liu et al., 2024c). Recent advancements in CoT reasoning capabilities of MLLMs (OpenAI, 2024b; Guo et al., 2025) have driven significant progress in complex analytical tasks, including causal reasoning within the textual modality (Jin et al., 2023; Bagheri et al., 2024; Ashwani et al., 2024). These developments involve enabling MLLMs to generate coherent explanations (Kiciman et al., 2023), providing multi-step chain-of-thought (CoT) (Bao et al., 2024), or

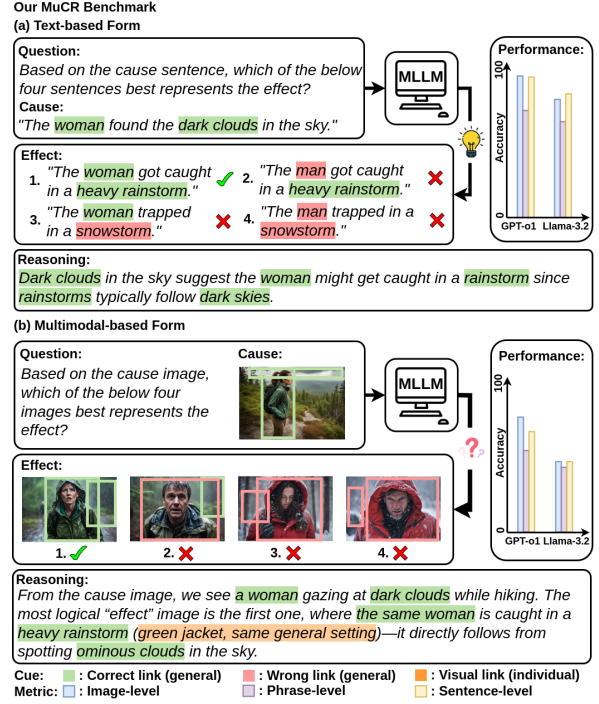


Figure 1: An example from MuCR challenges MLLMs with weather-related causality across two modalities.

even analyzing complex causal relationships that typically demand expert-level topological structure knowledge (Vashishtha et al., 2023). Despite these advancements, existing linguistic benchmarks (Singh et al., 2021; Du et al., 2022; Jin et al., 2023) are beginning to fall short in assessing the more advanced visual capabilities of the latest MLLMs such as GPT-o1 (OpenAI, 2024b), Deepseek-R1 (Guo et al., 2025), Gemini-1.5 (DeepMind, 2024), and Claude-3.5 (ClaudeAI, 2024b), not to mention facilitating cross-modal comparison and analysis (as shown in Figure 1).

Following this, we propose three key questions: Can MLLMs achieve the same level of causal reasoning comprehension as they do in textual modality? If not, what factors might influence cross-modal generalization? How can we enhance their capacity for robust causal inference? We find

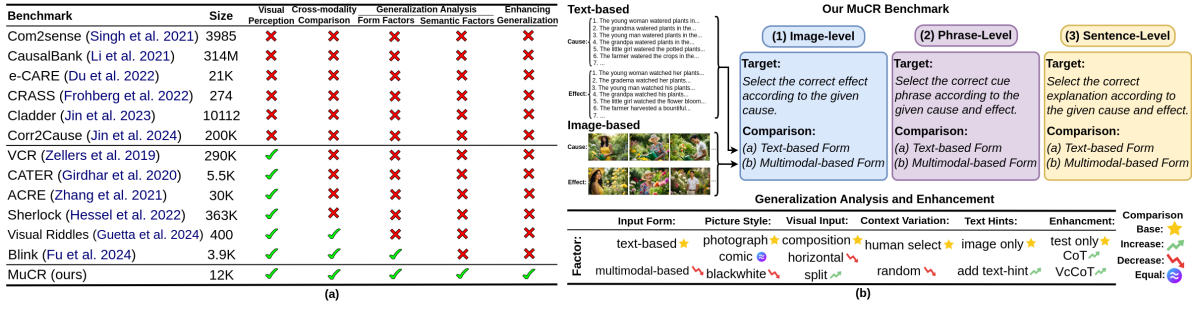


Figure 2: (a) Comparison of our MuCR and related datasets on reasoning tasks. (b) Detailed illustration of our dataset structure and corresponding cross-modal generalization exploration.

that most existing benchmarks fail to address such comparisons or support further exploration in this area. Especially, as shown in Figure 2 (a), we identify two major drawbacks in previous benchmarks: **Absence of visual modality**: Linguistic causal reasoning benchmarks (Singh et al., 2021; Li et al., 2021; Du et al., 2022; Frohberg and Binder, 2022; Jin et al., 2023, 2024) fail to assess visual comprehension ability of MLLMs. **Incomplete of cross-modal analysis**: Most causal reasoning VQA tasks (Zellers et al., 2019a; Girdhar and Ramanan, 2020; Zhang et al., 2021; Hessel et al., 2022) neglect cross-modal comparison. Recently, some benchmarks (Bitton-Guetta et al., 2024; Fu et al., 2024) have begun exploring this domain. For instance, Blink (Fu et al., 2024) examines cross-modal comparisons and conducts basic generalization analyses involving factors like shape and size. As illustrated in Figure 2(b), our proposed MuCR comprehensively evaluates causal reasoning at the image, phrase, and sentence levels and offers a multi-faceted analysis of cross-modal generalization that encompasses both visual form factors and semantic elements. Moreover, we propose a novel VcCoT strategy to further enhance cross-modal generalization by improving visual cue perception.

We evaluate current state-of-the-art (SOTA) MLLMs on our MuCR benchmark. Experiment results indicate that all models fall short of human performance, particularly in multimodal settings. Moreover, they exhibit a pronounced cross-modal gap when discerning causal links across modalities. In addition, we conduct in-depth generalization analysis and demonstrate that visual semantic factors, especially the ability to identify visual cues across siamese images, play a pivotal role.

Our contributions are summarized as follows:

- We identify the limitations of current causal reasoning benchmarks, including failing to evaluate the advanced visual capabilities of

the latest MLLMs and offering incomplete cross-modal analyses.

- We propose the MuCR benchmark, which can comprehensively evaluate MLLMs’ causal reasoning ability across two modalities.
- Our extensive experiments with SOTA MLLMs reveal interesting insights and suggest potential directions for future research.

2 Related Work

2.1 Causal Reasoning

The ability to perform causal reasoning is widely considered a core feature of artificial intelligence. With the development of Large Language Models (LLMs), they have exhibited increasingly robust capabilities in causal reasoning tasks. Previous benchmarks, such as Com2sense (Singh et al., 2021) and CausalBank (Li et al., 2021), are becoming insufficient for evaluating linguistic abilities. To address this, Romanou et al. (2023) introduced the CRAB benchmark, which requires LLMs to capture explicit causal relationships in real-world scenarios. However, these benchmarks focus solely on the text modality, leaving the crucial question of multimodal reasoning unaddressed. Hessel et al. (2022) introduced Sherlock to challenge MLLMs in identifying visual clues scattered throughout a scene and making reasoning inferences combined with commonsense and life experience. More recently, Guetta et al. (2024) and Fu et al. (2024) presented complex visual reasoning challenges to further explore MLLMs’ capabilities. Although these benchmarks have considered the visual modality, they still fail to comprehensively analyze cross-modal generalization capacity. In this paper, we make an early attempt to extensively explore multimodal causal reasoning tasks across modalities.

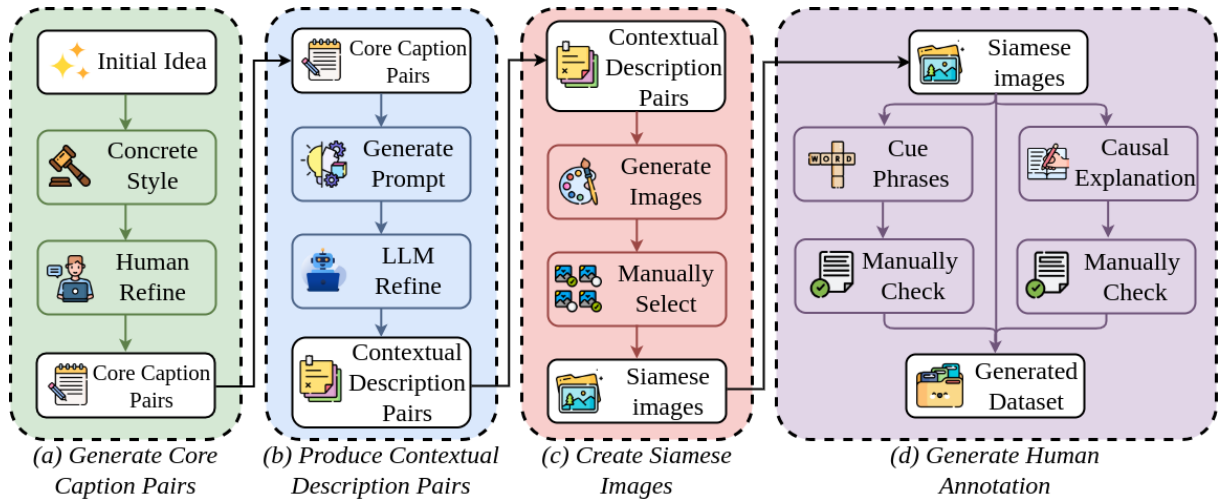


Figure 3: The overview of our MuCR benchmark construction process. It follows synthesis in four core levels: generating core caption pairs, producing contextual description pairs, creating siamese images, and generating human annotations.

2.2 LLMs’ Generalization

The field of LLMs generalization has gained significant traction in recent years, with numerous tasks proposed to evaluate models’ ability to handle previously unseen contexts and domains. Existing tasks can be broadly divided into compositional, cross-task, cross-lingual, cross-domain, and robustness-based categories. Compositional tasks, such as CFQ (Keysers et al., 2020) and COGS (Kim and Linzen, 2020), test whether models can systematically combine smaller linguistic units to form novel expressions. Cross-task generalization often involves multi-task learning setups, such as DecaNLP (McCann et al., 2018) and BIG-Bench (Srivastava et al., 2022), where models must adapt to tasks with minimal guidance. Cross-lingual benchmarks, like XNLI (Conneau et al., 2018) and XTREME (Hu et al., 2020), measure performance across languages, while cross-domain tasks emphasize shifting between specialized fields (Li et al., 2023; Zhou et al., 2024). Meanwhile, robustness-oriented evaluations such as HellaSwag (Zellers et al., 2019b) and adversarial GLUE (Wang et al., 2021) assess how well models withstand noisy, ambiguous, or adversarial inputs. In this paper, we shift our focus to the generalization in multimodal causal reasoning tasks, conducting a concise but comprehensive analysis of the factors that hinder cross-modal generalization and exploring strategies to enhance it for robust causal reasoning.

3 The MuCR Dataset

In this section, we detail the construction of the MuCR dataset. Figure 3 illustrates the system-

atic workflow of our multimodal cause-and-effect benchmark generation including: generating core caption pairs, producing contextual description pairs, creating siamese images, and generating human annotations (see Appendix A.2 for further examples and details).

3.1 Dataset Creation

Generating Core Caption Pairs. The MuCR benchmark is designed to assess MLLMs’ ability to perform causal inference across modalities. To achieve this, we begin by generating core caption pairs that clearly illustrate cause-and-effect relationships. In order to minimize individual bias, we employ twelve volunteers and group each two as a team: one processes and refines the captions based on initial ideas and iterative feedback, while the other reviews them and offers suggestions for improvement (see Appendix A.3 for an explanation of why we structure the generation process this way, as well as illustrative examples). Through these steps, we create 4,000 cause-and-effect caption pairs.

Producing Contextual Description Pairs. While core caption pairs effectively depict the cause-and-effect relationship, they often lack contextual details such as appearance, clothing, and environmental context that serve as crucial visual cues for high-quality cause-and-effect image synthesis. To address this issue, we leverage the linguistic capabilities of LLMs to enhance core caption pairs by enriching contextual details. By maintaining these elements consistently across images, our approach not only effectively depicts



Figure 4: (a) Examples from our MuCR dataset featuring different categories and styles. The “Mixture” category represents two or more tags involved in the causality. (b) Category distribution overview showing the proportions of human, animal, character, plant, and mixture styles. (c) Style distribution overview illustrating the proportions of comic, photographic, and black-white styles.

causality at a semantic level but also improves visual coherence (see Appendix A.4 for further explanation).

Create Siamese Images. We employ diffusion models with contextual descriptions as prompts to generate cause-and-effect image pairs. Specifically, we utilize DALL-E (Ramesh et al., 2021), DeepAI (DeepAI, 2024), Stability-AI (Stability AI, 2023), and Flux1 (FLUXAI, 2024) for image synthesis, aiming to minimize model bias and enhance the diversity of the generated images. We also incorporate three styles (photograph, comic, and black-white) when creating these images. Specifically, each sentence yields 10 images per style, resulting in 20 images for every cause-and-effect pair in one style (a total of 240k images). Then, volunteers manually select the two representations that best capture the semantic causality and maintain visual consistency. This process produces 12k cause-and-effect image pairs spanning various categories (humans, animals, plants, characters, and mixtures) and three styles (photograph, comic, and black-white). Figure 4 illustrates examples from our MuCR benchmark, showcasing multiple categories and styles alongside an overview of their distribution (see Appendix A.5 for more high-quality samples).

Generate Human Annotation. We require volunteers to create text annotations for each cause-and-effect image pair. As shown in Figure 3, it consists of a phrase-level list (cue phrases) and sentence-level description (cause-and-effect explanations). The cue phrases comprise a list of four options, each being a word or phrase. Among these, only one phrase correctly explains or is highly relevant to the causality, while the other three are striking elements in the images but do not serve as proper cues. The sentence-level annotation is designed to verify whether the MLLMs truly under-

stand multimodal causality and can select reasonable explanations. To achieve this, we require volunteers to structure the explanation by first describing the content of the cause, followed by the content of the effect, and concluding with the causal link connecting between them.

3.2 Evaluation Metrics

Image-level Metric. The image-level metric is called cause-to-effect (C2E) score. It is designed to assess whether the MLLMs can identify cue links and make the correct choice from four potential effects according to the given cause. Given the cause in the form $\mathcal{G}^*(c)$ (* can either be $\mathcal{G}^{text}$ representing text-based form or $\mathcal{G}^{multi}$ representing multimodal-based form), the model is required to select the optimal choice among four potential effects $\{\mathcal{G}^*(e)^{(i)}\}_{i=1}^4$. The C2E score can be computed as follows:

$$S_I^* = F(Q_I, \mathcal{G}^*(c), \{\mathcal{G}^*(e)^{(i)}\}_{i=1}^4), \quad (1)$$

$$f_I(S_I^*) = \begin{cases} 1, & S_I^* = S_I^{*'} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where S_I^* represents the MLLMs’ prediction. F represents MLLM. Q_I represents corresponding question for Image-level. f_I represents the function to calculate the C2E score. $S_I^{*'}$ represents the correct answer.

Phrase-level Metric. The phrase-level metric is called CP score (Cue Phrase), which tests MLLMs’ capability to distinguish the correct cue from a list of fraudulent phrases according to the cause and effect. Given the cause-and-effect pairs $\{\mathcal{G}^*(c), \mathcal{G}^*(e)\}$, the model is required to select the optimal choice among four potential cue phrases $\{T_P^{(i)}\}_{i=1}^4$. The CP score can be computed as fol-

lows:

$$S_P^* = F(Q_P, \mathcal{G}^*(c), \mathcal{G}^*(e), \{T_P^{(i)}\}_{i=1}^4) \quad (3)$$

$$f_P(S_P^*) = \begin{cases} 1, & S_P^* = S_P^{*'} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where S_P^* represents the MLLMs' prediction. F represents MLLM. Q_P represents corresponding question for Phrase-level. f_P represents the function to calculate the **CP** score. $S_P^{*'}$ represents the correct answer.

Sentence-level Metric. Our final metric is designed to evaluate MLLMs' ability to identify the correct explanation according to the cause and effect. The sentence-level metric is called the **explanation (EXP)** score. Specifically, we collect four candidate explanations that share similar causalities but differ in their cues. Only one explanation accurately captures the causal relationship and matches the detailed cues, while the other three do not. Given the condition $\{\mathcal{G}^*(c), \mathcal{G}^*(e)\}$ with the corresponding question Q_S , the model is required to select the optimal choice among four potential explanations $\{T_E^{(i)}\}_{i=1}^4$. The **EXP** score is then computed as follows:

$$S_S^* = F(Q_S, \mathcal{G}^*(c), \mathcal{G}^*(e), \{T_S^{(i)}\}_{i=1}^4) \quad (5)$$

$$f_S(S_S^*) = \begin{cases} 1, & S_S^* = S_S^{*'} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where S_S^* represents the MLLMs' prediction. F represents MLLM. f_S represents the function to calculate the **EXP** score. $S_S^{*'}$ represents the correct answer.

4 Experiments

4.1 Experimental Setup

We evaluated several popular MLLMs on our MuCR benchmark, including GPT-o1 (OpenAI, 2024b), GPT-4o (OpenAI, 2024a), Claude-3.5 (ClaudeAI, 2024b), Claude-3.0 (ClaudeAI, 2024a), Gemini-2.0 (DeepMind, 2025), Gemini-1.5 (DeepMind, 2024), Qwen2.5-VL (Yang et al., 2024), and Llama3.2-Vision (Meta, 2024). For the currently popular models, DeepSeek-R1 (Guo et al., 2025) and DeepSeek-V3 (Liu et al., 2024a), we did not fully evaluate their performance since their image readers currently only support extracting text from images without additional functionality (see Appendix B.1 for a comparison of their text-based

performance). Additionally, we also considered some lightweight open-source models, including LLaVA-NeXT (Li et al., 2024), OpenFlamingo-v2 (Awadalla et al., 2023), LLaVA-v1.6 (Liu et al., 2024b), MiniGPT4-v2 (Zhu et al., 2023), and InstructBLIP (Dai et al., 2023). Since some models only accept a single image input, we provided all of them with a composite image composed of multiple smaller images, as shown in Figure 9 (a). Finally, we established a human performance baseline on the MuCR benchmark using crowd workers for comparison.

4.2 Experimental Results

Figure 5 presents the main results of popular MLLMs and human performance on the MuCR benchmark, leading to the following observations: (1) **All models on MuCR lag behind human performance in both settings.** Among these models, GPT-o1 (OpenAI, 2024b) achieves the highest scores, with 94% on C2E score, 75% on CP score, and 93% on EXP score in the text condition, while 87% on C2E, 62% on CP, and 78% on EXP in the multimodal condition. Nevertheless, these results still fall short of human performance, suggesting substantial room for improvement. (2) **All models exhibit a significant cross-modal performance gap.** All models show a noticeable drop in performance when handling multimodal causal inference, whereas humans do not. This discrepancy indicates potential factors restricting cross-modal generalization in MLLMs, likely stemming from the visual component, given that these models already demonstrate robust causal reasoning in text-based cases.

Figure 6 presents the multimodal performance of various lightweight open-source models, revealing that they lag significantly behind GPT-o1. Among these, LLaVA-Next achieves the best results, with 29% on C2E, 17% on CP, and 21% on EXP, which are only around the random selection baseline of 25%. Compared to models like Llama3.2-Vision and Qwen2.5-VL, there is still considerable room for improvement for these models.

5 Cross-modal Generalization Analysis and Enhancement

In this section, we examine the factors that may affect cross-modal generalization. Building on previous findings that attribute these gaps primarily to the visual component, we focus on two main categories: visual format factors and visual semantic

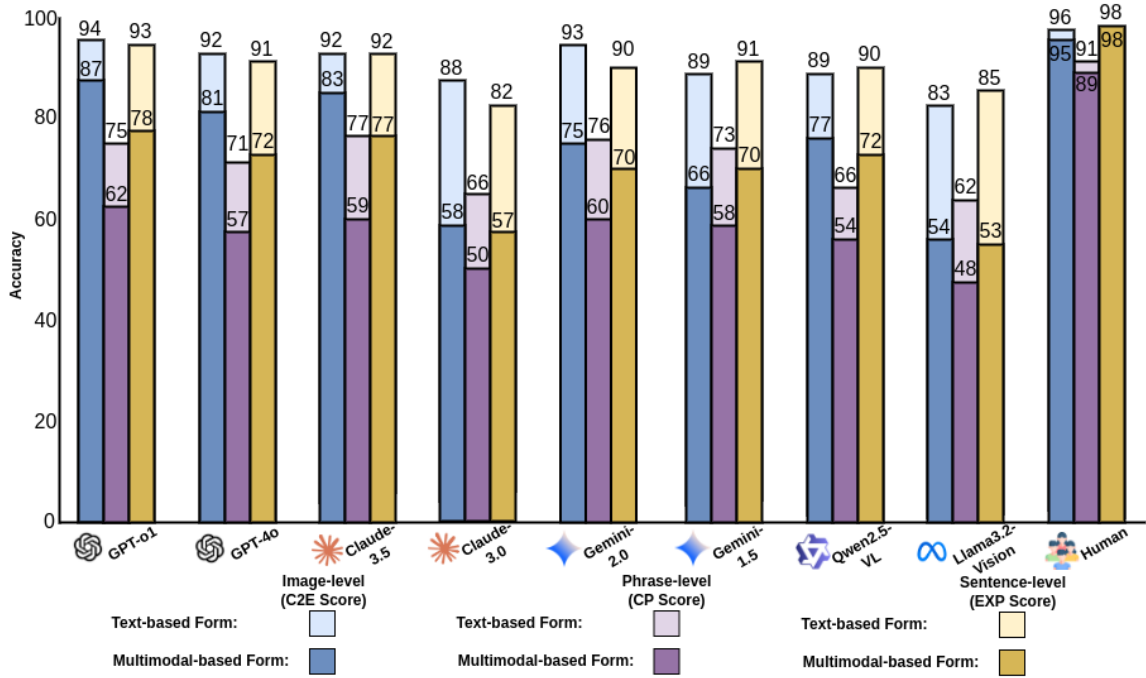


Figure 5: Main experimental results of several popular MLLMs on our MuCR benchmark. “Human” performance is represented by the average accuracy of ten attempts by volunteers.

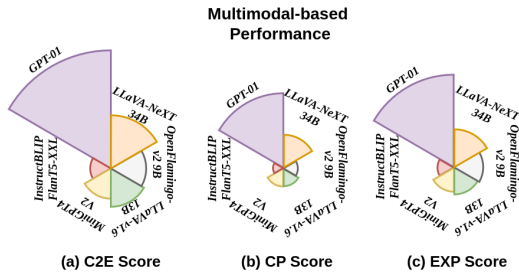


Figure 6: Experimental results of lightweight open-source models on the multimodal-based form. For detailed numbers see Table 6. Best viewed by zooming in.

factors.

- **Visual Format Factors.** These involve cases that share the same underlying semantics but differ in how they are visually presented, such as variations in picture style or the form of the visual input.
- **Visual Semantic Factors.** These involve cases with consistent visual formats but slight semantic differences, such as contextual variations in image details or the inclusion of additional text hints, resulting in richer semantic content.

In addition to investigating these cross-modal generalization factors, we also explore potential enhancement strategies based on our findings.

5.1 Visual Format Factors

Picture Style. We investigate how different picture styles may affect causal reasoning. Figure 7

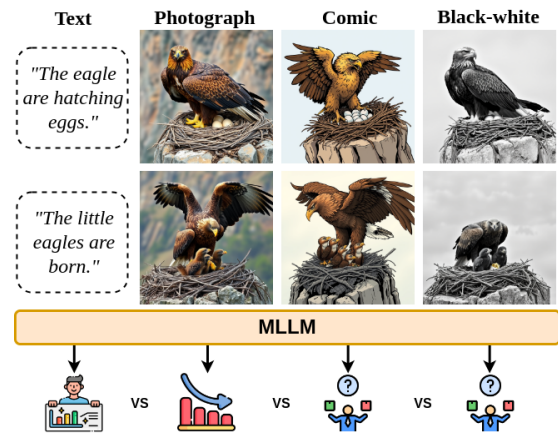


Figure 7: An example of cause and effect showing in three picture styles with the same semantic meanings.

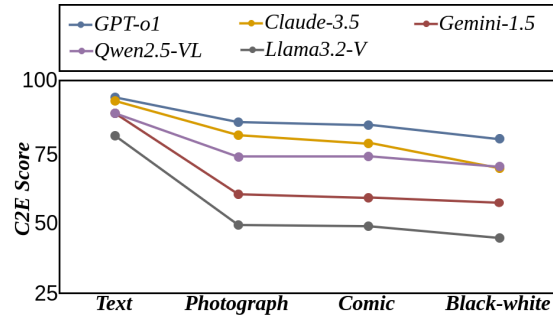


Figure 8: The C2E score of different models tested on three different picture styles.

shows an example of the same cause-and-effect scenario presented in three styles. As indicated by the results in Figure 8, MLLMs perform similarly when presented with photographs and comic images, but with a slight drop for black-white images. Overall, altering the picture style while keeping the

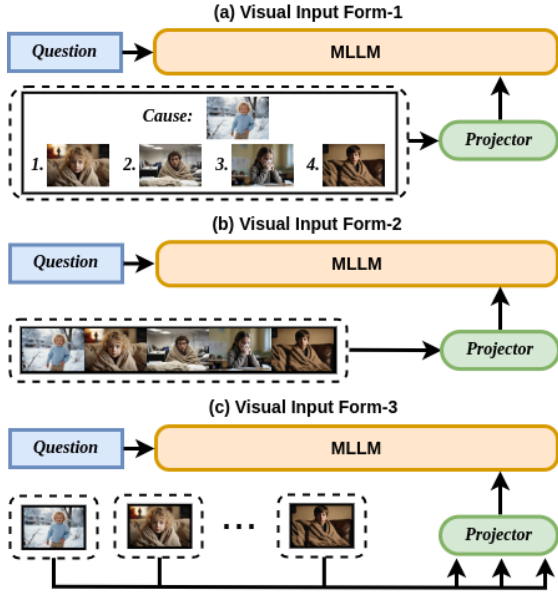


Figure 9: The illustration of three different visual input forms we examined.

Visual Input	Style	C2E	CP	EXP
GPT-o1 (OpenAI, 2024b)				
Form-1		87.50	62.00	78.00
Form-2	Mixture	84.25	60.50	79.00
Form-3		89.00	67.50	86.25
Claude-3.5 (ClaudeAI, 2024b)				
Form-1		83.50	59.75	77.50
Form-2	Mixture	53.50	36.00	68.50
Form-3		85.00	66.75	82.25
Gemini-1.5 (DeepMind, 2024)				
Form-1		66.50	58.50	70.50
Form-2	Mixture	69.50	57.25	63.00
Form-3		83.50	65.25	84.00

Table 1: The performance of different visual input forms on our MuCR benchmark. The mixture means we test on mixture picture style.

same semantic content has only a minimal effect on MLLMs’ performance (see Appendix C.1 for detailed comparison).

Form of Visual Input. We also explore whether the structure of visual inputs affects the final output. Figure 9 illustrates the three types of visual input forms we examined. Table 1 presents the performance of three models on MuCR using these different formats. It indicates that all models get marked performance improvements. Our case analysis suggests that, compared to Form-3, Forms-1 and Form-2 restrict MLLMs’ ability to perceive certain details that could serve as crucial visual cues for enhancing multimodal causal reasoning (see Appendix C.2 for case studies).



Figure 10: Two image pairs illustrate the same cause-and-effect relationship but exhibit different contextual correlations.

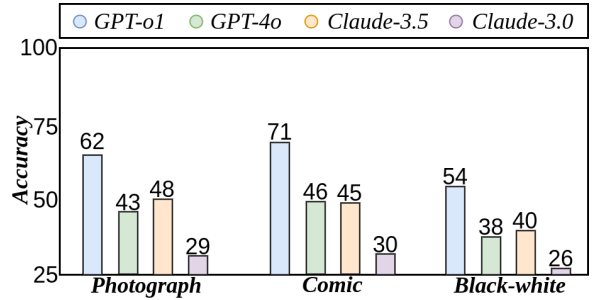


Figure 11: Using human selection as the standard, the models exhibit varying levels of selection accuracy.

5.2 Visual Semantic Factors

Contextual Variation. In addition to examining visual format factors, we also explore whether visual semantics influence MLLMs’ final output. As shown in Figure 1, MLLMs, particularly GPT-o1, can identify visual cues such as action, appearance, and environment, and integrate these details into their causal inference process. Additionally, the case study in the above paragraph also confirms that visual cues are essential for accurate multimodal causal inference. To further investigate, we assess whether the ability to identify visual cues correlates with multimodal causal reasoning performance. For this purpose, we use manually selected siamese image pairs that best capture semantic causality and maintain visual consistency, along with some pairs that exhibit minor contextual variations (see Figure 10). Our challenge is as follows: given a human-selected cause image, the models must identify the corresponding effect image from random 3 samples and 1 correct one. Figure 11 shows that among the four models tested, GPT-o1 excels at identifying visual cues, while Claude-3.0 performs the worst, with GPT-4.0 and Claude-3.5 falling in between (see Appendix C.3 for case stud-

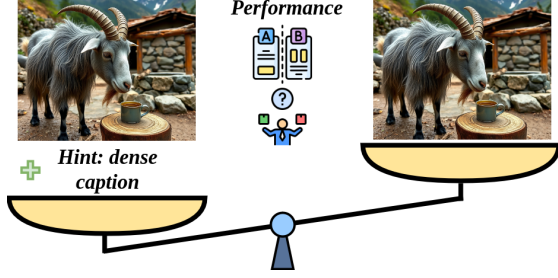


Figure 12: Does adding text hints compensate for shortcomings in visual cue perception?

Add Hints	Style	C2E	CP	EXP
GPT-o1 (OpenAI, 2024b)				
Before		87.50	62.00	78.00
After	Mixture	91.25	69.50	88.50
GPT-4o (OpenAI, 2024a)				
Before		81.25	57.25	72.50
After	Mixture	89.00	66.50	87.50
Claude-3.5 (ClaudeAI, 2024b)				
Before		83.50	59.75	77.50
After	Mixture	87.50	68.50	86.00
Claude-3.0 (ClaudeAI, 2024a)				
Before		58.00	50.25	57.00
After	Mixture	73.00	59.50	77.00

Table 2: The impact of adding text hints on different models.

ies). This finding confirms a positive correlation between an MLLM’s ability to identify visual cues, distinguish contextual variations, and its overall multimodal causal reasoning performance.

Text Hints. Since we verified a positive correlation between multimodal causal reasoning and visual cue perception, the next question is whether text hints can compensate for shortcomings in visual cue perception. To explore this, we use the contextual descriptions generated during dataset creation as dense captions, as they provide detailed raw information while preserving correct semantic meanings. Table 2 shows that adding text hints significantly improves MLLMs’ performance, suggesting that enhancing visual cue identification is a promising avenue for improving cross-modal generalization.

5.3 Generalization Enhancement

Based on our above analysis, the most crucial factor affecting MLLMs’ cross-modal generalization is the ability to identify visual cues. In response, we propose VcCoT, a method designed to enhance visual cue identification for causal inference. Inspired by MMCoT (Zhang et al., 2023)

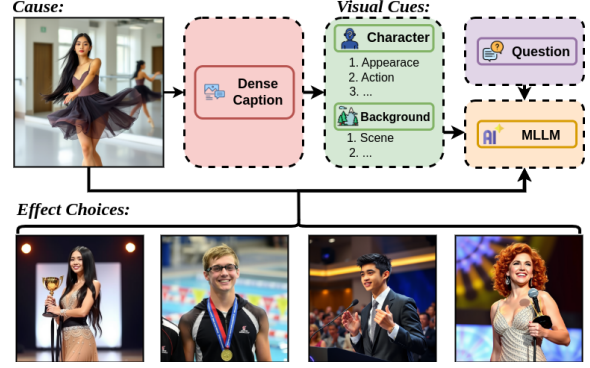


Figure 13: The structure of our VcCoT. Best viewed by zooming in.

Strategy	Style	C2E	CP	EXP
GPT-o1 (OpenAI, 2024b)				
Direct		87.50	62.00	78.00
CoT		86.25	61.50	76.00
CCoT	Mixture	88.00	64.00	79.50
MMCoT		84.25	60.5	86.50
VcCoT		89.75	66.5	83.00

Table 3: The performance of different CoT strategies on MuCR benchmark.

and CCoT (Mitra et al., 2024), our approach first converts images into dense captions, then extracts visual details categorized as *Character* and *Background*. Finally, these cues guide the MLLMs’ reasoning process, as illustrated in Figure 13. Table 3 demonstrates that VcCoT achieves superior performance than others. We also show some qualitative results in Appendix C.4.

6 Conclusion

In this paper, we introduce MuCR, a novel multimodal causal reasoning benchmark that challenges MLLMs to discern causal links across different modalities by leveraging synthetic siamese images and text pairs. We also propose comprehensive metrics to assess MLLMs’ understanding from multiple perspectives, including image-level alignment, phrase comprehension, and sentence-level explanation. Our experimental results reveal that current MLLMs exhibit a cross-modal gap in causal reasoning compared to their strong performance in purely textual settings. In-depth analysis highlights that effective visual cue identification is key to enhancing generalization, as MLLMs often struggle with implicit causal dependencies hidden in visual details. In response, we propose VcCoT, a method designed to improve visual cue identification for causal inference, with experimental results demonstrating its effectiveness.

7 Limitation

Although our research provides a comprehensive analysis of the potential factors affecting generalization from visual components, it has two notable limitations. First, as noted by Wang et al. (2024a), cross-linguistic variations can influence performance and may require transfer learning. Figure 14 presents a simple comparison of transferring the question language from English to Chinese using the C2E score, indicating that cross-linguistic factors affect the final output of the models. However, due to human resource constraints, we did not extend this study to the CP and EXP scores, as these metrics require human reannotation of cue phrases and sentence explanations.

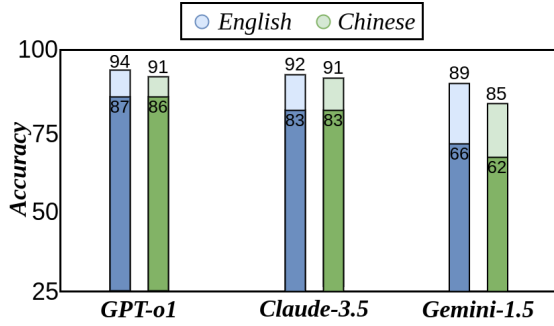


Figure 14: A comparison of different models on the C2E score with cross-linguistic setting.

Fine-tune	Style	C2E	CP	EXP
LLaVA-v1.6 (Liu et al., 2024b)				
Before	Mixture	23.50	11.00	16.50
After		20.00	13.75	15.25
MiniGPT4-v2 (Zhu et al., 2023)				
Before	Mixture	17.75	11.50	15.25
After		19.00	13.50	16.00
InstructBLIP (Dai et al., 2023)				
Before	Mixture	12.25	6.50	9.50
After		7.50	3.25	4.75

Table 4: The impact of direct fine-tuning on different models.

Additionally, we explored fine-tuning a few lightweight open-source models. As shown in Table 4, direct fine-tuning with the correct choices did not improve and in some cases even decreased the performance of these models. Our observations indicate that these models fail to capture the causal links between cause-and-effect images through fine-tuning. Notably, InstructBLIP even lost its ability to caption images accurately, exhibiting severe hallucinations. Due to limited resources, we did not investigate whether reinforcement learning (Guo et al., 2025) or alternative strategies (Niklas et al., 2025) could further address

the generalization problem on larger models such as Qwen2.5-VL (Yang et al., 2024) or LLaMA3.2-Vision (Meta, 2024).

References

- Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: Can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, pages 2–9.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.
- Abdolmahdi Bagheri, Matin Alinejad, Kevin Bello, and Alireza Akhondi-Asl. 2024. C2p: Featuring large language models with causal reasoning. *arXiv preprint arXiv:2407.18069*.
- Guangsheng Bao, Hongbo Zhang, Linyi Yang, Cunxiang Wang, and Yue Zhang. 2024. Llm with chain-of-thought are non-causal reasoners. *arXiv preprint arXiv:2402.16048*.
- Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. 2024. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. *arXiv preprint arXiv:2407.19474*.
- ClaudeAI. 2024a. Claude 3: Anthropic’s large language model. <https://www.anthropic.com/claude>.
- ClaudeAI. 2024b. Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

565	DeepAI. 2024. Deepai: Image generation. https://deepai.org/machine-learning-model/text2img .	619
566		620
567		621
568	DeepMind. 2024. Gemini 1.5 models. https://deepmind.google/technologies/gemini/ .	622
569		623
570	DeepMind. 2025. Gemini 2.0 models. https://deepmind.google/technologies/gemini/ .	624
571		625
572	Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning . In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 432–446. Association for Computational Linguistics.	626
573		627
574		628
575		629
576		630
577		631
578		632
579	FLUXAI. 2024. Introducing flux.1 tools. https://blackforestlabs.ai/flux-1-tools/ .	633
580		634
581	Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In <i>Proceedings of the Thirteenth Language Resources and Evaluation Conference</i> , pages 2126–2140, Marseille, France. European Language Resources Association.	635
582		636
583		637
584		638
585		639
586		640
587	Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In <i>European Conference on Computer Vision</i> , pages 148–166. Springer.	641
588		642
589		643
590		644
591		645
592		646
593	Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2024. Lcm-lookahead for encoder-based text-to-image personalization. In <i>European Conference on Computer Vision</i> , pages 322–340. Springer.	647
594		648
595		649
596		650
597		651
598	Rohit Girdhar and Deva Ramanan. 2020. CATER: A diagnostic dataset for Compositional Actions and Temporal Reasoning. In <i>ICLR</i> .	652
599		653
600		654
601	Nitzan Bitton Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szepktor, Amir Globerson, and Yuval Elovici. 2024. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. In <i>The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track</i> .	655
602		656
603		657
604		658
605		659
606		660
607		661
608	Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. <i>arXiv preprint arXiv:2501.12948</i> .	662
609		663
610		664
611		665
612		666
613	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 <i>European Conference on Computer Vision</i> , pages 558–575. Springer.	667
614		668
615		669
616		670
617		671
618		672
	Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In <i>International Conference on Machine Learning</i> , pages 4411–4421. PMLR.	673
		674
	Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models. In <i>Thirty-seventh conference on neural information processing systems</i> .	675
		676
		677
		678
		679
		680
		681
		682
		683
		684
		685
		686
		687
		688
		689
		690
		691
		692
		693
		694
		695
		696
		697
		698
		699
		700
		701
		702
		703
		704
		705
		706
		707
		708
		709
		710
		711
		712
		713
		714
		715
		716
		717
		718
		719
		720
		721
		722
		723
		724
		725
		726
		727
		728
		729
		730
		731
		732
		733
		734
		735
		736
		737
		738
		739
		740
		741
		742
		743
		744
		745
		746
		747
		748
		749
		750
		751
		752
		753
		754
		755
		756
		757
		758
		759
		760
		761
		762
		763
		764
		765
		766
		767
		768
		769
		770
		771
		772
		773
		774
		775
		776
		777
		778
		779
		780
		781
		782
		783
		784
		785
		786
		787
		788
		789
		790
		791
		792
		793
		794
		795
		796
		797
		798
		799
		800

675	Deng, Chenyu Zhang, Chong Ruan, et al. 2024a.	Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao,	727
676	Deepseek-v3 technical report. <i>arXiv preprint</i>	Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch,	728
677	<i>arXiv:2412.19437</i> .	Adam R Brown, Adam Santoro, Aditya Gupta,	729
678	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae	Adrià Garriga-Alonso, et al. 2022. Beyond the	730
679	Lee. 2024b. Visual instruction tuning. <i>Advances in</i>	imitation game: Quantifying and extrapolating the	731
680	<i>neural information processing systems</i> , 36.	capabilities of language models. <i>arXiv preprint</i>	732
681	Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan	<i>arXiv:2206.04615</i> .	733
682	Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Hao-	Stability AI. 2023. Stability ai: Image generation.	734
683	liang Wang, Tong Yu, et al. 2024c. Large language	https://stability.ai .	735
684	models and causal inference in collaboration: A com-	Aniket Vashishtha, Abbavaram Gowtham Reddy, Ab-	736
685	prehensive survey. <i>arXiv preprint arXiv:2403.09606</i> .	hinav Kumar, Saketh Bachu, Vineeth N Balasub-	737
686	Bryan McCann, Nitish Shirish Keskar, Caiming Xiong,	ramanian, and Amit Sharma. 2023. Causal infer-	738
687	and Richard Socher. 2018. The natural language	ence using llm-guided discovery. <i>arXiv preprint</i>	739
688	decathlon: Multitask learning as question answering.	<i>arXiv:2310.15117</i> .	740
689	<i>arXiv preprint arXiv:1806.08730</i> .	Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan,	741
690	Meta. 2024. Llama 3.2: Revolutionizing edge	Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadal-	742
691	ai and vision with open, customizable mod-	lah, and Bo Li. 2021. Adversarial glue: A multi-	743
692	els. https://ai.meta.com/blog/llama-3-2-connect-	task benchmark for robustness evaluation of language	744
693	2024-vision-edge-mobile-devices/ .	models. <i>arXiv preprint arXiv:2111.02840</i> .	745
694	Chancharik Mitra, Brandon Huang, Trevor Darrell, and	Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan	746
695	Roei Herzig. 2024. Compositional chain-of-thought	Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-	747
696	prompting for large multimodal models. In <i>Proceed-</i>	lingual knowledge editing in large language models.	748
697	<i>ings of the IEEE/CVF Conference on Computer Vi-</i>	In <i>Proceedings of the 62nd Annual Meeting of the</i>	749
698	<i>sion and Pattern Recognition</i> , pages 14420–14431.	<i>Association for Computational Linguistics (Volume 1:</i>	750
699	Muennighoff Niklas, Yang Zitong, Shi Weijia, Li Xi-	<i>Long Papers)</i> , pages 11676–11686, Bangkok, Thai-	751
700	ang Lisa, Fei-Fei Li, Hajishirzi Hannaneh, Zettle-	land. Association for Computational Linguistics.	752
701	moyer Luke, Liang Percy, Candès Emmanuel, and	Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and	753
702	Hashimoto Tatsunori. 2025. s1: Simple test-time	Anthony Chen. 2024b. Instantid: Zero-shot identity-	754
703	scaling. <i>arXiv preprint arXiv:2501.19393</i> .	preserving generation in seconds. <i>arXiv preprint</i>	755
704	OpenAI. 2024a. Hello gpt-4o. https://openai.com/index	<i>arXiv:2401.07519</i> .	756
705	/hello-gpt-4o/ .	An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui,	757
706	OpenAI. 2024b. Introducing openai o1-preview.	Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu,	758
707	https://openai.com/index/introducing-openai-o1-	Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 tech-	759
708	preview .	nical report. <i>arXiv preprint arXiv:2412.15115</i> .	760
709	Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott	Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang.	761
710	Gray, Chelsea Voss, Alec Radford, Mark Chen, and	2023. Ip-adapter: Text compatible image prompt	762
711	Ilya Sutskever. 2021. Dall-e: Creating images from	adapter for text-to-image diffusion models. <i>arXiv</i>	763
712	text. https://www.openai.com/blog/dall-e .	<i>preprint arXiv:2308.06721</i> .	764
713	Angelika Romanou, Syrielle Montariol, Debjit Paul,	Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin	765
714	Leo Laugier, Karl Aberer, and Antoine Bosselut.	Choi. 2019a. From recognition to cognition: Vi-	766
715	2023. Crab: Assessing the strength of causal rela-	visual commonsense reasoning. In <i>Proceedings of the</i>	767
716	tionships between real-world events. <i>arXiv preprint</i>	<i>IEEE/CVF conference on computer vision and pat-</i>	768
717	<i>arXiv:2311.04284</i> .	<i>tern recognition</i> , pages 6720–6731.	769
718	Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoor-	Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali	770
719	molabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun	Farhadi, and Yejin Choi. 2019b. Hellaswag: Can a	771
720	Peng. 2021. COM2SENSE: A commonsense reason-	machine really finish your sentence? <i>arXiv preprint</i>	772
721	ing benchmark with complementary sentences . In	<i>arXiv:1905.07830</i> .	773
722	<i>Findings of the Association for Computational Lin-</i>	Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun	774
723	<i>guistics: ACL/IJCNLP 2021, Online Event, August</i>	Zhu, and Yixin Zhu. 2021. Acre: Abstract causal	775
724	<i>1-6, 2021</i> , volume ACL/IJCNLP 2021 of <i>Findings of</i>	reasoning beyond covariation. In <i>Proceedings of the</i>	776
725	<i>ACL</i> , pages 883–898. Association for Computational	<i>ieee/cvf conference on computer vision and pattern</i>	777
726	Linguistics.	<i>recognition</i> , pages 10643–10653.	778
		Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao,	779
		George Karypis, and Alex Smola. 2023. Multi-	780
		modal chain-of-thought reasoning in language mod-	781
		els. <i>arXiv preprint arXiv:2302.00923</i> .	782

Xiaomao Zhou, Qingmin Jia, Yujiao Hu, Renchao Xie,
Tao Huang, and F Richard Yu. 2024. Geng: An llm-
based generic time series data generation approach
for edge intelligence via cross-domain collaboration.
In *IEEE INFOCOM 2024-IEEE Conference on Com-
puter Communications Workshops (INFOCOM WK-
SHPS)*, pages 1–6. IEEE.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and
Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing
vision-language understanding with advanced large
language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A The MuCR Dataset

A.1 Task Formulation

As shown in Figure 2 (b), our dataset $\mathcal{D} := \{(Q, \mathcal{G}^*(A), \{B^{(i)}\}_{i=1}^4)^{(k)}\}_{k=1}^N$ consisting of N triples, each contains a question Q , an input $\mathcal{G}^*(A)$ (where $*$ represents input form), and four potential choices $\{B^{(i)}\}_{i=1}^4$. The MLLMs are required to according to the question Q and an input $\mathcal{G}^*(A)$ to select the correct answer from four potential choices $\{B^{(i)}\}_{i=1}^4$. The goal of this benchmark is to determine whether the input form ($*$) affects the MLLMs' prediction accuracy. To this end, the biggest challenge is defined as follows:

$$\mathcal{G}^{text}(A) \stackrel{\text{semantic}}{\approx} \mathcal{G}^{multi}(A) \quad (7)$$

where $\stackrel{\text{semantic}}{\approx}$ means $\mathcal{G}^*(A)$ retains identical or closely aligned semantic meaning across different modalities. To address this, we propose a novel transfer strategy that harnesses the linguistic capabilities of LLMs alongside the image generation abilities of diffusion models, effectively preserving semantic content while altering the input form.

A.2 Overall structure

Section 3 only illustrates the simplified process of our MuCR benchmark generation. Here, we delve into more details about the generation process and the corresponding prompts. Figure 15 showcases the detailed generation process of a weather-related causal case in our MuCR dataset. Our process begins with generating core caption pairs, each consisting of one caption describing the cause and the other stating the effect. We then leverage the language capabilities of LLMs to entail these paired captions into contextually relevant descriptions, enhancing the consistency of sentences to facilitate the creation of cause-and-effect image pairs. Then, we employ diffusion models to generate numerous Siamese images based on these descriptions. Finally, we annotate cue phrases and causality explanations for each pair.

A.3 Generating Core Caption Pairs

Our MuCR benchmark begins with the creation of core caption pairs, where one caption outlines the cause and the other describes the effect. These pairs maintain semantic causality and serve two roles. First, they function as textual causal inference cases to challenge MLLMs' textual reasoning

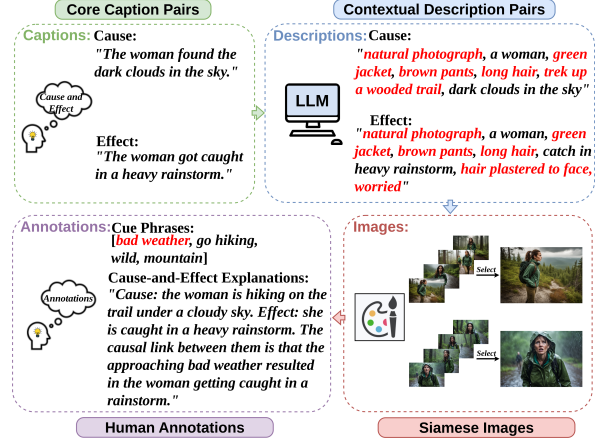


Figure 15: A detailed example of generating our MuCR dataset. Best viewed by zooming in.

ability. Second, they guide the subsequent synthesis of Siamese images. As shown in Figure 16, we employ a structured refinement loop that transforms initial brainstorming ideas into precise caption pairs, clearly depicting the cause-and-effect relationships. This process is guided by the principle: "Whether the expression is concrete and can be effectively represented through visual means". Here, we discuss the rationale behind this rule and explain why volunteers are instructed to create core caption pairs in accordance with it.

Figure 17 compares the initial spark and core caption pairs in image synthesis. The comparison reveals that the initial spark often contains semantically ambiguous elements, leading to visual gaps in the generated images. For instance, the phrase "the baker left the cake in the oven" might result in an image depicting only a cake in the oven, as the diffusion model may struggle to interpret or visually represent the action "left". Another issue is subject conflict. For example, the phrase "the food became inedible" might simply produce an image of unappealing food on a plate. However, within a cause-and-effect scenario, a human would easily infer that "food" refers specifically to the "cake." In contrast, our core caption pairs resolve these ambiguities by translating them into more concrete actions, such as replacing "careless" with "played his phone." This refinement significantly improves the quality of the generated images and the semantic causality between the pairs.

We ask the volunteers to design four paired captions as a group, each sharing similar causalities but containing different visual cues. These groups are intended to explore the capability of distinguishing

Cause	Effect	Category
The man drove his car at an excessive speed .	The man got a speeding ticket .	Person
The woman drove her car at an excessive speed .	The woman was pulled over by the policeman .	Person
The old man drove his car at an excessive speed .	The old man was pulled over by the policeman .	Person
The old woman drove her car at an excessive speed .	The old woman got a speeding ticket .	Person
The shark is attacking the fish.	The fish got wounded on its back.	Animal
The shark is attacking the seal.	The seal got wounded on its back.	Animal
The shark is attacking the manta ray.	The manta ray got wounded and bleeding .	Animal
The shark is attacking the penguin.	The penguin got wounded and bleeding .	Animal
The chrysanthemum is blooming .	The chrysanthemum attracting bees to collect nectar .	Plant
The tulip is blooming .	The tulip attracting bees to collect nectar .	Plant
The rose is blooming .	The rose attracting bees to collect nectar .	Plant
The jasmine is blooming .	The jasmine attracting bees to collect nectar .	Plant
The rabbit worked hard .	The rabbit earn much money .	Character
The monkey worked hard .	The monkey earn much money .	Character
The bear worked hard .	The bear earn much money .	Character
The fox worked hard .	The fox earn much money .	Character
The gardener planted a tree .	The tree grew tall .	Mixture
The farmer planted seeds .	The seeds sprouted into crops .	Mixture
The child planted flowers .	The flowers bloomed in the garden .	Mixture
The woman planted herbs .	The herbs grew in the pot .	Mixture

Table 5: Case studies for the paired caption generation process.

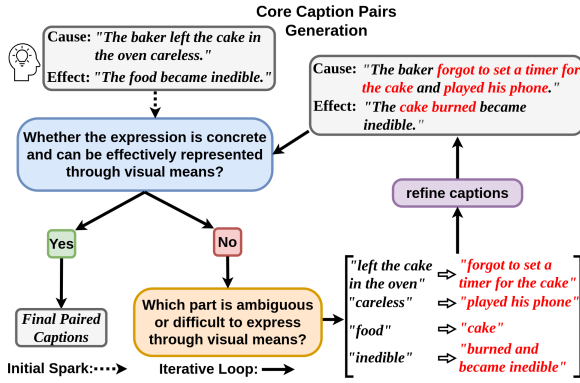


Figure 16: The process of generating paired captions through refinement loops, ensuring the final captions are precise and can be effectively represented through visual means.

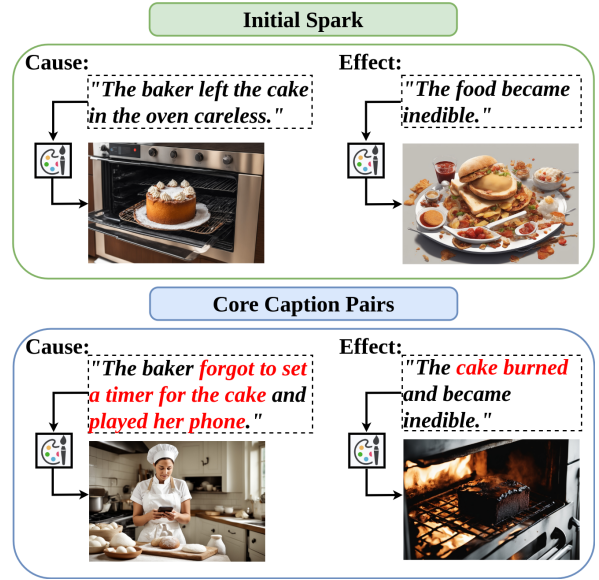


Figure 17: A comparison of directly using initial spark and our core caption pairs to generate cause-and-effect images through diffusion model.

similar causalities occurring in different subjects across various scenarios. Furthermore, to maintain the diversity of our dataset, we include a portion of non-human cases. While many causality scenarios feature humans as subjects, we also incorporate cases involving animals, plants, comic characters, and their interactions. Table 5 shows generated paired-caption examples (i.e., four captions sharing similar causalities but involving different visual cues are paired as a group) for various scenarios (i.e., cases involving humans, animals, plants, comic characters, and mixtures). Abstract expressions are concretized during the paired-caption generation process according to the causality. For instance, the scenario “driving at excessive speed”

is rephrased in terms of its potential outcomes, such as “getting a speeding ticket” or “being pulled over by a police officer”. Similarly, the concept of “blooming” is illustrated through its possible consequence, “attracting bees to gather nectar”. This process leverages causal reasoning to ground abstract ideas in real-world outcomes, thereby enhancing the intelligibility and reproducibility of the generated captions.

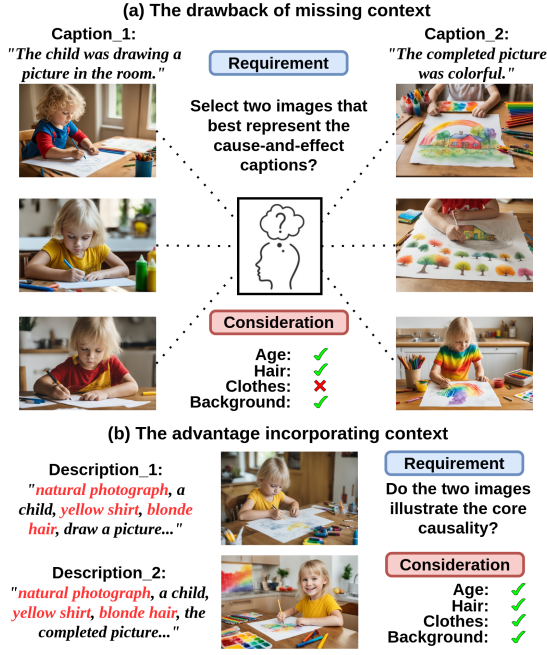


Figure 18: An example of core captions vs contextual descriptions in cause-and-effect image synthesis.

A.4 Producing Contextual Description Pairs

The absence of crucial visual cues could introduce randomness in image creation, which may lead to inconsistencies and potentially undermine the perceived causality between the siamese images. Figure 18 highlights the drawbacks of missing context and the advantages of incorporating context. As shown in Figure 18 (a), although the two columns of images accurately represent the core caption, mismatched clothing disrupts the sense of causality, making it difficult to form coherent pairs. In contrast, the example in Figure 18 (b) demonstrates that incorporating contextual information and transforming core captions into contextual descriptions effectively resolves this issue and reduces randomness in image synthesis. To achieve this, we leverage the linguistic capabilities of LLMs to enhance core caption pairs by enriching contextual details such as appearance, clothing, environment, and atmosphere. Additionally, we introduce subtle changes, such as variations in facial expressions, within the contextual description pairs to reflect the passage of time. These detailed variations emphasize the impact of causality over time, making the connection between siamese images more natural and coherent.

we also compare identity-preserving techniques with our prompt-guidance method (Figure 19). Traditional identity-preserving image synthesis methods (e.g., LCM (Gal et al., 2024) and IP-

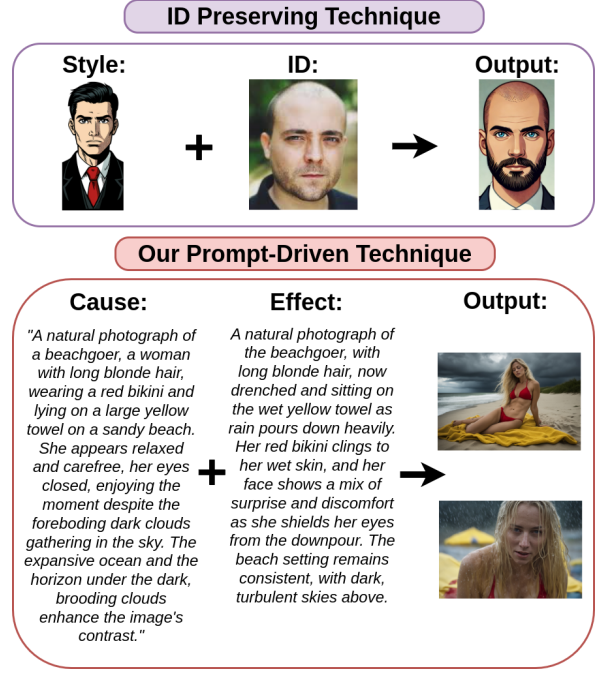


Figure 19: A comparison of identity-preserving technique and our prompt-driven technique on image synthesis.

Adapter (Ye et al., 2023)) focus on image personalization by retaining identity details through a region encoder during the generation process (Wang et al., 2024b). However, this approach leads to two major issues. First, most existing identity-preserving techniques rely heavily on guided images, which limits their capacity for semantically-driven image generation and requires finding a suitable ID image for each causal scenario. Second, as the name suggests, identity-preserving methods focus primarily on maintaining facial identity (appearance) but struggle to incorporate cause-and-effect relationships across images. In contrast, our causal-and-effect image synthesis approach leverages the linguistic capabilities of large language models (LLMs) to integrate a richer spectrum of contextual information. Not only does it preserve human facial identity (appearance), but it also accounts for additional details (e.g., clothing, environment, and overall atmosphere). This ensures that images remain coherent even when modifications are introduced through causal reasoning.

A.5 Siamese Images and Annotations

In this section, we show some high-quality examples as follows:

In the plant category, as shown in Figure 20, take the jasmine flower pair: the cause image shows a blooming jasmine flower, while the effect image

	Cue Phrases bloom bee flower sunshine	Explanation <i>“Cause: the flower is in full bloom. Effect: the flower attracts numerous bees. The causal link between them is that the flower's blooming attracts bees for pollination.”</i>
Category: Plant Style: Photograph		
	Cue Phrases bloom bee flower sunshine	Explanation <i>“Cause: the red tulips are shown in full bloom. Effect: the tulips attract numerous bees. The causal link between them is that the blooming tulips draw the bees.”</i>
Category: Plant Style: Photograph		
	Cue Phrases bloom bee flower sunshine	Explanation <i>“Cause: the red rose is blooming in the garden. Effect: the red rose is attracting bees that are gathering around it. The causal link between them is that the rose's blooming has attracted the bees.”</i>
Category: Plant Style: Photograph		
	Cue Phrases bloom bee flower sunshine	Explanation <i>“Cause: the jasmine flower is in full bloom. Effect: the bees are seen gathering around the blooming jasmine flowers. The causal link between them is that the flower's blooming has attracted the bees.”</i>
Category: Plant Style: Photograph		

Figure 20: Example 1 - Plant

 Category Style Character Comic	Cue Phrases fitness gym muscle dumbbells	Explanation <i>“Cause: the rabbit is lifting dumbbells in a gym. Effect: the rabbit gaining strength. The causal link between them is that the rabbit’s workout in the gym leads to it gaining strength and showing off its muscles.”</i>
 Category Style Character Comic	Cue Phrases fitness gym muscle dumbbells	Explanation <i>“Cause: the cat is lifting weights at the gym. Effect: the cat gains strength and muscle. The causal link between them is that the cat’s consistent weightlifting and workout routine at the gym resulted in gaining muscle strength.”</i>
 Category Style Character Comic	Cue Phrases fitness gym muscle dumbbells	Explanation <i>“Cause: the dog is at the gym, preparing for a workout. Effect: the dog is shown with well-developed muscles. The causal link between them is that the dog’s consistent workout routine at the gym resulted in significant muscle growth and increased strength.”</i>
 Category Style Character Comic	Cue Phrases fitness gym muscle dumbbells	Explanation <i>“Cause: the eagle is lifting a dumbbell at the gym. Effect: the eagle is flexing its muscles. The causal link between them is that the eagle’s consistent weightlifting and workout routine at the gym resulted in gaining significant muscle strength and mass.”</i>

Figure 21: Example 2 - Character

 	Cue Phrases plant grow green land	Explanation <i>“Cause: the man is planting a tree in a garden. Effect: the tree is fully grown with vibrant pink flowers. The causal link between them is that the man's efforts in planting and nurturing the tree led to its growth and blossoming.”</i>
Category: Mixture Style: Photograph		
 	Cue Phrases plant grow green land	Explanation <i>“Cause: the farmer planting seeds in a field. Effect: the crops fill the lush, green field with rows of growth. The causal link between them is that the farmer's planting efforts resulted in the successful growth of the crops.”</i>
Category: Mixture Style: Photograph		
 	Cue Phrases plant grow green land	Explanation <i>“Cause: the young child carefully planting flowers in a garden. Effect: this child standing proudly among a vibrant array of blooming flowers. The causal link between them is that the child's planting efforts resulted in the beautiful, blossoming garden.”</i>
Category: Mixture Style: Photograph		
 	Cue Phrases plant grow green land	Explanation <i>“Cause: the woman carefully planting seedlings in a garden. Effect: this woman smiling and holding a large pot with flourishing plants. The causal link between them is that the woman's diligent planting and care for the seedlings led to their healthy growth, resulting in the flourishing plants she holds with pride.”</i>
Category: Mixture Style: Photograph		

Figure 22: Example 3 - Mixture

Model	Text-based Form			Multimodal-based Form		
	C2E	CP	Exp	C2E	CP	Exp
Popular MLLMs						
GPT-o1 (OpenAI, 2024b)	94.00	75.50	93.00	87.50	62.00	78.00
GPT-4o (OpenAI, 2024a)	92.75	71.75	91.50	81.25	57.25	72.50
Claude-3.5 (ClaudeAI, 2024b)	92.50	77.00	92.75	83.50	59.75	77.5
Claude-3.0 (ClaudeAI, 2024a)	88.25	66.75	82.00	58.00	50.25	57.00
Gemini-2.0 (DeepMind, 2025)	93.00	76.00	90.50	75.50	60.75	70.25
Gemini-1.5 (DeepMind, 2024)	89.00	73.25	91.50	66.50	58.50	70.75
Qwen2.5-VL (Yang et al., 2024)	89.00	66.00	90.00	77.00	54.50	72.00
Llama3.2-Vision (Meta, 2024)	83.50	62.50	86.00	54.00	48.25	53.25
Lightweight Open-source Models						
LLaVA-NeXT (Li et al., 2024)	54.50	37.50	48.00	29.00	17.00	21.00
OpenFlamingo-v2 (Awadalla et al., 2023)	23.00	16.00	17.25	20.00	9.75	18.00
LLaVA-v1.6 (Liu et al., 2024b)	25.25	17.25	18.00	23.50	11.00	16.50
MiniGPT4-v2 (Zhu et al., 2023)	13.50	18.50	16.75	17.75	11.50	15.25
InstructBLIP (Dai et al., 2023)	14.50	10.00	8.50	12.25	6.50	9.50
Human	96.75	91.00	98.50	95.50	89.50	98.50

Table 6: Main experimental results of different models on our MuCR benchmark.

features a group of bees swarming around it. For this pair, we select “bloom” as the positive cue phrase and “bee”, “flower”, and “sunshine” as the negative ones, aligning with the visual information. The annotation emphasizes the connection between the flower’s blooming and the attraction of bees.

In the character category, as shown in Figure 21, consider the cat pair: the first image shows a cat lifting weights at the gym, while the second image depicts the cat gaining strength and muscle. For this, “fitness” is used as the positive cue phrase, with “gym”, “muscle”, and “dumbbells” as the negative ones, matching the visual content. The annotation focuses on the connection between consistent workouts and muscle gains.

In the mixture category, as shown in Figure 22, take the female planting pair: the cause image shows a woman planting seedlings in a garden, while the effect image displays the same woman smiling and holding a large pot of flourishing plants. Here, “plant” is the positive cue phrase, and “grow”, “green”, and “land” are the negative ones, aligning with the visual information. The annotation emphasizes the relationship between her nurturing care and the plant’s growth, along with her pride.

Input Form	Style	C2E	CP	EXP
GPT-o1 (OpenAI, 2024b)				
Text	\	94.00	75.50	93.00
Image	Mixture	87.50	62.00	78.00
DeepSeek-R1 (Guo et al., 2025)				
Text	\	96.00	73.50	95.00
DeepSeek-V3 (Li et al., 2014)				
Text	\	91.50	72.25	92.00

Table 7: The performance comparison between GPT-o1 and DeepSeek models in text domain on MuCR.

B Experiments

In this section, we delve into extended experiments and provide supplementary details that were not included in the main paper for the sake of clarity and brevity.

B.1 Experimental Results

As discussed in Section 4.1, we did not include the currently popular models DeepSeek-R1 (Guo et al., 2025) and DeepSeek-V3 (Liu et al., 2024a) in the main paper. Here, we provide a brief comparison of their text-based performance against GPT-o1 (OpenAI, 2024b). Table 7 shows that DeepSeek-R1 achieves results comparable to GPT-o1 in the text domain, while DeepSeek-V3 performs slightly less effectively.

In addition, we provide a detailed breakdown of each model’s performance on our MuCR bench-

mark. Table 6 presents these results. We observe that all popular MLLMs significantly outperform random chance, whereas most lightweight open-source models perform below the random baseline of 25%. This indicates that the latter group lacks robust causal reasoning capabilities.

C Cross-modal Generalization Analysis and Enhancement

C.1 Picture Style

Here, we present a detailed case analysis comparing the influence of picture style on Claude-3.5’s predictions, as illustrated in Figure 23.

In the black-white images, Image 1 shows a warthog bending down to drink water, placing it in a vulnerable position. The cause is clear—the warthog’s need to drink compels it to lower its head, thus reducing its awareness of potential threats. Among the follow-up images, Image 5 best represents the effect: it shows a crocodile emerging from the water, poised to attack a drinking animal, maintaining consistent compositional elements such as the animal at the water’s edge and the predator’s emergence. While Images 2, 3, and 4 depict similar scenarios with different animals, Image 5 most directly mirrors the cause-and-effect relationship suggested by Image 1. However, the analysis in this style tends to lack detail in some of the incorrect answers, which could potentially influence the model’s predictive accuracy in nuanced cases.

In contrast, the comic style analysis also begins with Image 1, where a warthog is depicted looking down at ripples in the water, seemingly unaware of any lurking danger. The potential effects are illustrated across multiple images: Image 2 shows a wildebeest encountering a crocodile, Image 3 depicts a zebra facing a crocodile, Image 4 features a gazelle or antelope in a similar scenario, and Image 5 shows another warthog confronting a crocodile. Here, Image 5 stands out as the best representation of the effect because it features the same animal as in the cause image in a comparable setting, now facing the implied threat signaled by the ripples. The consistent composition and environmental context reinforce the direct cause-and-effect relationship.

The comic style analysis provides a richer context and more detailed narrative for the causal relationship, whereas the black-white analysis, although accurate in identifying the correct image, offers less detailed reasoning for some incorrect options.

C.2 Form of Visual Input

Our case analysis demonstrates that, compared to Form-3, Forms 1 and 2 impose limitations on MLLMs’ ability to recognize and leverage critical visual cues necessary for multimodal causal reasoning. As shown in Figure 24, Form-3 provides GPT-4o with direct visual information, enabling it to successfully identify essential details, such as the continuity in a person’s appearance across cause-and-effect images. This was evident in GPT-4o’s output, where it correctly determined that the woman in the cause image, overwhelmed by paperwork, was the same individual in the effect image, now engaged in a serious discussion about work. This recognition of visual consistency is crucial for establishing causal relationships. However, when using Form-1, GPT-4o was unable to incorporate this specific visual cue and instead selected a different effect image (a generic team meeting) based on a more abstract textual interpretation rather than a direct visual correlation.

The key issue with Forms 1 and 2 is that they rely on structured textual descriptions that predefine categories of reasoning, which may inadvertently filter out implicit but important visual details. These formats encourage MLLMs to focus on generalized textual patterns rather than independently deriving causal relationships from visual features like facial expressions, body language, and scene continuity. By contrast, Form-3 allows GPT-4o to analyze raw visual inputs more freely, enhancing its ability to establish causal links based on direct visual observation rather than abstracted textual hints. This distinction highlights the potential shortcomings of rigid textual input structures in multimodal causal reasoning tasks. While textual guidance can be helpful, it may also constrain the model’s reasoning process, making it less sensitive to nuanced visual cues. Ensuring that MLLMs receive input formats that preserve rich visual information is therefore essential for improving their ability to perform causal inference in multimodal settings.

C.3 Contextual Variation

Visual cues are crucial for accurate multimodal causal inference because they provide a consistent framework for linking cause and effect. Taking Figure 25 as an example:

- **Consistency:** Shared elements like the hiking setting, outfit, and subject positioning help the model recognize that the images belong to the

same sequence.

- **Clear Transitions:** Changes in lighting, perspective, and mood signal the progression from cause (a clear, well-lit forest) to effect (a foggy, atmospheric scene), reinforcing the narrative flow.
- **Disambiguation:** Detailed cues identify Image 2 as the best continuation among similar options, ensuring the causal relationship is accurately maintained.

The analysis shows that visual cues—ranging from consistent environmental context and subject details to nuanced transitions in lighting, perspective, and mood—are crucial for establishing a clear and coherent narrative. These cues allow the model to accurately determine the causal links between images, ensuring that the inferred relationships are both logical and contextually grounded. Without such detailed visual information, the model would face challenges in differentiating between similar scenarios, potentially leading to inaccurate or incomplete causal inferences.

C.4 Qualitative Results of VcCOT

To prove the efficiency of our VcCoT, we provide some qualitative results, as shown in Figure 26.



Figure 23: Case study for picture style influence. Best viewed by zooming in.

Question:

Based on the cause image, which of the below four images best represents the effect?

Case:



Effects:

(1):



(3):



(2):



(4):



GPT-4o

Visual Input Form-1

From the provided image, the **cause** is represented by a woman in an office setting appearing overwhelmed with paperwork. This suggests a scenario of being overworked, stressed, or under pressure at work.

Now, let's analyze the possible **effects** depicted in the four images:

- Image 1:** A woman having a discussion with a colleague, possibly about work-related matters.
- Image 2:** A group meeting with individuals in conversation, possibly discussing workload or solutions.
- Image 3:** Two people talking, one appearing to provide guidance or support.
- Image 4:** Two people holding a certificate, which could indicate recognition, achievement, or resolution.

Best Choice: Image 2

Reason: If the cause is an overwhelmed employee, a likely effect is a team meeting or discussion to address workload issues, making **Image 2** the most logical representation of the effect.

Visual Input Form-3

Let's analyze the best causal match.

- Cause:** A woman looking exhausted at her desk, overwhelmed by paperwork.
- Effect 1:** The same woman (with same appearance) in a serious discussion with another woman, possibly about work.
- Effect 2:** A tense discussion or argument between men in a work setting.
- Effect 3:** A positive business handshake.
- Effect 4:** A certificate being awarded, symbolizing success.

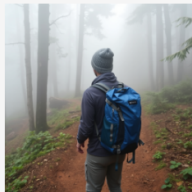
Best Answer:

Effect 1 seems to be the best match because it directly follows from the overwhelmed state in the cause image. It suggests she is now addressing work-related issues, possibly discussing a problem that arose due to the workload.

Figure 24: Case study for visual input form influence.

Question:

Based on the cause image, which of the below four images best represents the effect while maintaining visual consistency?

Case:**Effects:****(1):****(3):****(2):****(4):****Claude-3.5**

Looking at the visual sequence, Image 2 best represents the effect while maintaining visual consistency with Image 1. Here's why

1. Image 2 maintains the same:

- Hiking setting
- Outfit (gray beanie, same jacket, blue backpack)
- Subject positioning (clear view of the backpack)
- Professional photo quality and composition

2. It creates a clear cause-and-effect relationship:

- Image 1 shows the hiker starting their journey in a clear, well-lit forest
- Image 2 shows them continuing into foggy, atmospheric conditions, suggesting progression deeper into the trail

3. Visual Transition:

- The lighting changes naturally from clear to misty
- The perspective shifts from frontal to profile view, creating a narrative flow
- The mood transitions from preparation to contemplation

While Images 3, 4, and 5 are similar, they either show different clothing (Image 4 with an orange jacket), different compositions, or don't flow as naturally from the initial image. Image 2 provides the most seamless visual continuation from Image 1 while maintaining all key visual elements.

Figure 25: Case study for Contextual Variation.

Question:

Based on the cause image, which of the below four images best represents the effect?

Effects:

(1): 

(2): 

(3): 

(4): 

Case:



GPT-o1

VcCoT

Consider the following visual hint

Character:

- chef
- white hat
- white coat
- wear glass
- watering plants

Background:

- gardern

Results

Based on the given visual hints and character consistency, the best effect image should maintain the following elements from the cause image:

- Character:** A chef wearing a white hat, white coat, and glasses, who was previously watering plants.
- Background:** A garden setting.

Analysis of the Images:

- The second image (chef picking herbs in the garden) closely aligns with the cause image. It maintains:**
 - The same character (chef).
 - Consistency in attire (white hat, white coat, and glasses).
 - The same garden environment.
 - A logical cause-and-effect sequence: after watering plants, the chef is now picking herbs, likely for cooking.
- Other images do not align as well:**
 - The third and fourth images depict scientists or researchers in lab coats, working in a controlled indoor environment, which does not match the chef's attire or setting.
 - The fifth image shows elderly gardeners in an outdoor setting, but they are not chefs and do not maintain the white uniform or glasses.

Conclusion:

The second image (chef picking herbs in the garden) best represents the effect, as it maintains character consistency and follows a logical cause-and-effect progression.

Figure 26: Qualitative result for VcCoT.

25