What's Missing in Vision-Language Models? Probing Their Struggles with Causal Order Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite the impressive performance of vision-language models (VLMs) on downstream tasks, their ability to understand and reason about causal relationships in visual inputs remains unclear. Robust causal reasoning is fundamental to solving complex high-level reasoning tasks, yet existing benchmarks often include a mixture of reasoning questions, and VLMs can frequently exploit object recognition and activity identification as shortcuts to arrive at the correct answers, making it challenging to truly assess their causal reasoning abilities. To bridge this gap, we introduce VQA-Causal and VCR-Causal, two new benchmarks specifically designed to isolate and rigorously evaluate VLMs' causal reasoning abilities. Our findings reveal that while VLMs excel in object and activity recognition, they perform poorly on causal reasoning tasks, often only marginally surpassing random guessing. Further analysis suggests that this limitation stems from a severe lack of causal expressions in widely used training datasets, where causal relationships are rarely explicitly conveyed. We additionally explore fine-tuning strategies with hard negative cases, showing that targeted fine-tuning can improve model's causal reasoning while maintaining generalization and downstream performance. Our study highlights a key gap in current VLMs and lays the groundwork for future work on causal understanding.

1 Introduction

Pre-trained vision-language models have demonstrated impressive performance across a wide range of tasks, including visual question answering [1, 14], reasoning [35], and object detection [15]. However, strong performance on these benchmarks does not necessarily reflect a rich understanding of visual inputs. Recent studies have revealed that VLMs struggle with tasks demanding high-level visual understanding, such as verb comprehension, spatial reasoning, attribute attachment, and counting [2, 9, 21, 31, 34]. Crucially, whether VLMs possess genuine causal reasoning abilities remains largely unexplored. For instance, can VLMs distinguish between "The woman holding an umbrella is caused by the rain." and "The rain is caused by the woman holding an umbrella."? Robust causal understanding and reasoning are fundamental to tackling complex real-world decision making [10], but this capability in VLMs remains largely unexplored.

Existing benchmarks that aim to assess reasoning in VLMs often conflate causal reasoning with other types of reasoning tasks [1, 35], and many questions can be answered by object recognition or activity understanding alone. For example, our analysis of the Visual Question Answering (VQA) and Visual Commonsense Reasoning (VCR) benchmarks reveals that only 0.92% of questions in the VQA validation set [1] and 35.43% in the VCR validation set [35] involve causal reasoning. Our analysis of 100 randomly selected VCR questions found that 46% could be answered correctly through object detection or activity understanding alone, without requiring genuine causal reasoning.

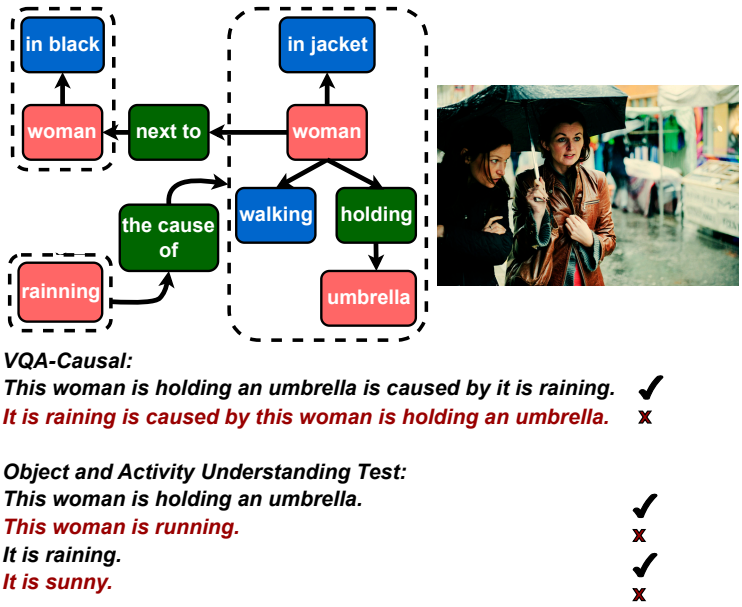

Figure 1: Examples from the VQA-Causal test and the Object and Activity Understanding test. Models tend to focus on low-level visual features such as objects and activities which are represented by the red and blue nodes in the scene graph on the left, but fail to capture high-level visual features such as relationships between activities, especially causal relationships in our case, which are represented by the green nodes in the scene graph.

These issues make it difficult for current benchmarks to independently and effectively evaluate the causal reasoning ability of VLMs.

To address this gap, we introduce VQA-Causal and VCR-Causal, the first benchmarks specifically designed to rigorously and independently evaluate VLMs' causal reasoning abilities. Constructed from different sources, VQA [1] and VCR [35], this dual-benchmark setup enables fine-tuning on VQA data and both in-domain (VQA-Causal) and out-of-domain (VCR-Causal) evaluation, thereby providing a robust assessment of models' causal reasoning capabilities and their generalizability across datasets. VQA-Causal consists of 1,947 instances, and VCR-Causal contains 3,511 instances, with each image paired with 12 caption pairs using different causal conjunctions. Each caption pair differs only in the causal relationship between events, as illustrated in Figure 1. This counterfactual approach ensures a comprehensive evaluation of the model's understanding of causal relationships, avoiding potential biases toward specific causal expressions.

We evaluate 10 widely-used VLMs, covering a broad spectrum of architectures and objectives, including score-based and generative models trained with diverse objectives. The evaluated models include CLIP [23], NegCLIP [34], BLIP [11], FLAVA [28] , LLaVA [18] and so on. All models perform poorly on both VQA-Causal and VCR-Causal, with nine out of ten achieving no more than 52% accuracy, which is only marginally above a random guess (50%) and significantly below human performance (98%). These findings highlight a fundamental limitation of existing VLMs in causal reasoning.

To better understand whether the poor performance on causal reasoning tasks stems from a lack of basic visual understanding, we constructed a controlled evaluation set by modifying the VQA-Causal dataset. This modified dataset contains the same 1,947 instances as VQA-Causal, but each image is paired with four captions, two of which correctly describe the image. The incorrect captions differ by altering the object or modifying the described activity, as illustrated in Figure 1. Our results reveal that while VLMs perform well in recognizing objects and activities, they struggle significantly with reasoning about causal relationships between activities, further reinforcing our findings from VQA-Causal and VCR-Causal.

We then investigate why VLMs trained on large-scale image-text corpora fail to learn causal relationships between events in visual inputs. Focusing on LAION-400M [26] (used by OpenCLIP) and MSCOCO [17] (used in FLAVA and NegCLIP) [28, 34], we found that explicit causal expressions

are extremely rare. Quantitatively, only 0.08% of LAION-400M and 0.01% of MSCOCO instances contain explicit causal expressions. This scarcity explains why VLMs excel at object and activity recognition but struggle with causal reasoning.

To mitigate this limitation, we explored fine-tuning strategies incorporating hard negative cases, captions that differ from the correct ones only in the causal order, demonstrating that targeted fine-tuning can significantly enhance causal reasoning. Our fine-tuned model, CausalCLIP, achieves notable improvements on both in-domain and out-of-domain benchmarks while maintaining strong performance on downstream tasks.

Our contributions are as follows:

- We introduce VQA-Causal and VCR-Causal, the first benchmarks specifically designed to isolate and comprehensively evaluate causal reasoning in VLMs, addressing a critical limitation in existing benchmarks. Moreover, this setup allows us to use one dataset (e.g., VQA) as an in-domain source for fine-tuning and evaluate the model's causal reasoning ability on both in-domain (VQA-Causal) and out-of-domain (VCR-Causal) benchmarks to assess generalization.
- Our experimental results reveal that while VLMs excel in object and activity understanding, they perform poorly on causal reasoning tasks, with some only marginally surpassing random guessing. Additionally, our analysis of four widely used datasets for VLM training, fine-tuning, and benchmarking uncovers a severe lack of causal expressions, providing insight into why models fail to learn causal relationships between different activities during training process.
- We explore fine-tuning with hard negative cases and demonstrate that targeted fine-tuning can enhance causal reasoning performance. Our approach achieves notable improvements on both in-domain and out-of-domain benchmarks while maintaining minimal impact on downstream task performance.

2 Benchmarks for Causal Order Reasoning

Existing benchmarks, such as VQA [1], VCR [35], and GQA [7], include questions related to causal reasoning. However, many instances within these datasets involve multiple types of reasoning, making it difficult to isolate and evaluate a model's specific understanding of causal relationships. Additionally, a significant portion of the causal reasoning examples fail to truly assess a model's comprehension of causality.

Q: Why is the person holding on to a rope?

Answer Choices:
1. The person is climbing over the boat.
2. The person is trying to tie the rope to something.
3. The rope helps the person get to the other side of the train tracks.
4. To keep from being washed away.

Figure 2: The VCR dataset fails to genuinely evaluate a model's causal reasoning ability. In this example, the model can eliminate choice 3 by recognizing that there is no train tracks in the image. It can also rule out captions 2 and 1 by observing that the person is not tying the rope to or climb over anything. As a result, the model can arrive at the correct answer purely through object and activity understanding, without requiring genuine causal reasoning.

For example, as illustrated in Figure 2, the VCR [35] question "Why is the person holding on to a rope?" allows a model to select the correct answer by merely identifying the absence of specific objects, such as train tracks, in the image, thus eliminating an incorrect option. Furthermore, by recognizing that the depicted activity is not "tying the rope to something," or "climbing over the boat" the model can exclude another 2 options. With only basic object and activity recognition, the model can reach the correct answer without demonstrating genuine reasoning about the causal relationships between different entities in the visual input.

In contrast, our proposed datasets place a strong emphasis on requiring models to understand the causal relationships between various events within the visual input. Our newly developed evaluation corpora adopt the format proposed by Kamath et al. [9], featuring an image paired with several captions that vary only in causal order. Specifically, VQA-Causal and VCR-Causal are constructed from the widely-used VQA and VCR datasets [1, 35]. A key contribution of our work is that every instance in our dataset demands models to genuinely

reason about the causal relationships between different events in the visual input, rather than taking shortcuts by merely identifying objects and activities to arrive at the correct answer.

2.1 Dataset Construction

VQA-Causal We constructed the VQA-Causal dataset using the validation set and validation annotation files from the VQA dataset [1]. Specifically, we selected all instances from the validation set where the questions contained the keyword "Why" to form our VQA-Causal dataset. Each original question and answer pair was transformed into two sentences connected by causal conjunctions, differing only in the causal order while keeping everything else identical. For example, given an image with the original VQA question "Why is this woman holding an umbrella?" and the correct answer "It is raining", we retained the image and generated two captions using the causal conjunction *is caused by* : "This woman is holding an umbrella is caused by it is raining." and "It is raining is caused by this woman is holding an umbrella.", as illustrated in Figure 1.

We used 12 causal conjunctions to create 12 groups of caption pairs for each image. These conjunctions were carefully chosen to capture variations in the syntactic ordering of causes and effects, as such variations can potentially influence a model's performance on causal reasoning tasks. Specifically, some conjunctions, such as *is due to*, *is caused by*, *is a result of*, *is the effect of*, *is the consequence of*, *because*, and *owe to*, place the effect before the cause in the sentence structure. In contrast, others such as *result in*, *cause*, *lead to*, *give rise to*, and *bring about to*, place the cause before the effect. Each group contains one caption expressing a correct causal relationship and one expressing an incorrect relationship, differing only in the causal direction.

In total, we extracted 1,947 instances from the VQA dataset, with each image paired with 12 distinct caption pairs. This setup offers several advantages: (1) It enables a rigorous evaluation of the model's ability to reason about causal relationships within visual inputs. (2) The use of diverse causal conjunctions allows us to assess the model's understanding and sensitivity to different causal expressions, while also mitigating potential biases that may arise from over-reliance on any single conjunction during the reasoning process.

VCR-Causal Similarly, we constructed the VCR-Causal dataset using the validation set and annotation files from the VCR dataset. We selected instances containing "Why" in their questions to form the VCR-Causal dataset. The VCR-Causal dataset contains a total of 3,511 instances, with each image associated with 12 caption pairs.

We conducted human verification on a randomly sampled subset of both VQA-Causal and VCR-Causal. Two human annotators with NLP backgrounds were asked to judge whether captions for each instance were (1) semantically coherent given the image context and (2) fluent. Each annotator reviewed 50 image-caption pairs from each dataset. Results show that over 96% of the captions were rated as both fluent and reasonable, indicating that our generation process yields high-quality, interpretable inputs for evaluating causal reasoning.

2.2 Causal Order Reasoning Test

Task For the VQA-Causal and VCR-Causal benchmarks, we follow the experimental setup used by Kamath et al. [9]. The input consists of an image paired with two caption options, which differ only in their causal order, as illustrated in Figure 1. Consistent with Kamath et al. [9], we evaluate the models under a zero-shot setting. Our evaluation metric is the proportion of images for which the matching score between the image and the correct caption is higher than the matching score between the image and the incorrect caption.

Models We select both score-based and text-generation based models:

- **Score-based models:** CLIP ViT-B/32, CLIP ViT-L/14 [23], FLAVA [28], BLIP ITM ViT-B, BLIP ITM ViT-L [11], BLIP2 ITM [13], BLIP2 Feature Extractor [13], NegCLIP [34], and RobustCLIP [25]. These models produce matching scores for each image-caption pair independently. Among them, NegCLIP is fine-tuned with hard negatives samples, making it more sensitive to the word order and RobustCLIP is fine-tuned with adversarial augmentations to improve the model's robustness.

- **Text-generation models:** LLaVA1.5 [18], Vicuna1.5 [5, 38]. We include Vicuna to validate that a language model relying solely on text input cannot effectively solve the causal reasoning tasks in our benchmark, thereby demonstrating the benchmark's reliability. By comparing Vicuna with LLaVA, which takes both image and text inputs, we further investigate whether LLaVA is capable of leveraging visual information to support causal reasoning.

For LLaVA, we follow the settings in Kamath et al. [9] and reformulate the task by converting the two captions into two questions. For example:

1. "This woman is holding an umbrella is caused by it is raining. Does it reflect the proper causal relationship?"
2. "It is raining is caused by this woman is holding an umbrella. Does it reflect the proper causal relationship?"

We measure the probabilities of models answering "yes" or "no" to these questions. Correctness is determined based on one of the following criteria:

1. The model assigns the highest "yes" probability to the correct option.
2. If both answers are "no", the lowest "no" probability is assigned to the correct option.

2.3 Benchmarking Results

Table 1 present the performance of nine score-based models and two generation-based models on our VQA-Causal and VCR-Causal benchmarks, respectively. Overall, all models perform near random and far below human estimate, revealing a clear lack of robust causal reasoning ability in current VLMs. Detailed results for each model across the twelve causal conjunctions are provided in Table **??** and Table **??**.

Model	**VQA-Causal**	**VCR-Causal**
	Score-Based Models	
BLIP ITM Base	48.94	50.66
BLIP ITM Large	48.68	47.99
BLIP2 ITM	50.76	49.95
BLIP2 FE	51.51	50.76
CLIP ViT B/32	51.62	50.35
CLIP ViT L/14	50.74	51.66
NegCLIP	50.89	51.30
RobustCLIP	50.66	53.68
FLAVA	48.52	49.99
	Text-Generation Models	
Vicuna 1.5	50.86	56.03
LLaVA 1.5	53.19	52.12
Random	50.00	50.00
Human Estimate	99.17	98.17

Table 1: Accuracy on the causal-order reasoning tests of ten VLMs for the VQA-Causal and VCR-Causal benchmarks. All models perform only marginally above random guessing and remain significantly below human-level performance. "BLIP2 FE" denotes the BLIP2 feature extractor model. Detailed results for each of the twelve causal conjunctions are provided in Table **??** and Table **??** in Appendix.

Causal Conjunctions Performance Across both the VQA-Causal and VCR-Causal benchmarks, the CLIP model family, including CLIP ViT B/32, CLIP ViT L/14, NegCLIP and RobustCLIP [23, 25, 34], demonstrates relatively stronger performance on conjunctions such as is caused by, is due to, is the consequence of, because, owe to, and is the effect of. These conjunctions share a common syntactic structure in which the result precedes the cause. In contrast, these models perform notably worse on conjunctions such as result in, cause, lead to, give rise to, and bring about to, where the cause

appears before the result. However, FLAVA [28] exhibits the opposite trend. On the VQA-Causal benchmark, it performs relatively poorly on conjunctions where the result comes first, but shows stronger performance on those where the cause precedes the result. These observations suggest that the syntactic ordering of cause and effect within a sentence plays a critical role in model performance, and that certain models may be sensitive to specific linguistic patterns of causal expression.

Impact of Prior Fine-Tuning Strategies Fine-tuning for caption order improves a model's sensitivity to word order, thereby improving its performance on certain causal order tests. For instance, NegCLIP outperforms CLIP models when tested on conjunctions like is due to and is caused by in most cases, demonstrating substantial improvements. However, for conjunctions like result in, cause, and lead to, NegCLIP underperforms compared to CLIP models. This suggests that fine-tuning for word order amplifies the model's strengths for specific conjunctions but also exacerbates its weaknesses for others, particularly those it initially struggled with. Moreover, adversarial robustness fine-tuning, as implemented in RobustCLIP, does not lead to significant improvements in causal order reasoning performance.

3 Activity and Object Understanding Test

To further investigate whether the poor causal reasoning performance of VLMs arises from a lack of understanding of entities in visual inputs, we conducted the Activity and Object Understanding Test. The results show that VLMs exhibit strong capabilities in recognizing objects and activities within images. This suggests that VLMs tend to focus on learning low-level visual features such as objects and activities recognition but fail to capture high-level features like causal relationships between activities.

Data Construction We extended the VQA-Causal dataset to construct this evaluation. For each original instance, we generated four captions: two correct captions that preserve the original causal event but decompose it into independent factual statements, and two incorrect captions, which were carefully crafted by modifying the object or the activity from the correct captions to make them factually inaccurate. This setup allows us to isolate the model's understanding of objects and activities from its ability to reason about causal relationships. An illustration is provided in Figure 1.

Object and Activity Understanding Test We conducted experiments with all score-based VLMs mentioned in Section 2.2 to assess their understanding of objects and activities within the input images. The input to each model consisted of an image paired with four captions described in the last paragraph, as illustrated in Figure 1. We considered the model's response correct if the two captions with the highest scores were the correct ones. This task setup requires models to accurately understand both the objects and activities depicted in the input image to achieve a correct response.

Results and Analysis As shown in Table 2, all models achieve strong performance on the object and activity understanding task. In Figure 1, the red nodes represent objects, the blue nodes indicate the attributes of these objects, and the green nodes depict the relationships between different objects and activities. VLMs tend to focus on learning low-level features which is the red and blue node in the scene graph but fail to capture high-level features which is the green node representing the relationships between objects and activities. This limitation in capturing structured visual relationships may explain why VLMs perform close to random on high-level reasoning tasks, including causal reasoning, as well as on other complex reasoning tasks like spatial reasoning [9] and verb understanding [31], which have been highlighted in previous studies.

4 Why Struggling with Causal Reasoning? A Data-Level Exploration

All evaluated VLMs were pretrained and fine-tuned on large-scale image-text corpora and have shown strong performance on traditional benchmarks. To explore why they fail to learn causal relationships, we investigate this limitation from a data-level perspective.

We selected four widely-used datasets for VLM pretraining and benchmark: LAION-400M [26], which was used to train OpenCLIP [4, 8, 23, 27], and MSCOCO [17], which was used in FLAVA's training and NegCLIP's fine-tuning. For benchmark datasets, we analyzed VQA and VCR [1, 35], two standard datasets commonly used to evaluate the reasoning capabilities of VLMs.

Model	VQA-Causal	O&A Test
BLIP ITM Base	48.94	94.61
BLIP ITM Large	48.68	95.53
BLIP2 ITM	50.76	92.24
BLIP2 FE	51.51	83.98
CLIP ViT B/32	51.62	76.53
CLIP ViT L/14	50.74	85.31
NegCLIP	50.89	87.62
RobustCLIP	50.66	83.26
FLAVA	48.52	71.85

Table 2: Accuracy on the Object and Activity Understanding Test (*O&A Test*) versus the Causal Order Reasoning Test (*VQA-Causal*). All models exhibit strong performance on the *O&A Test*, indicating that while VLMs effectively recognize objects and activities, they struggle with causal reasoning task. "BLIP2 FE" denotes the BLIP2 feature extractor model.

Pre-Training Datasets We randomly sampled about 5,200,000 captions from the LAION-400M dataset to examine the prevalence of causal expressions. Specifically, we looked for captions containing any of the following causal-related terms: *because*, *cause*, *lead to*, *reason*, *is the reason why*, *is the effect of*, *owe to*, *give rise to*, *bring about to*, *result in*. Among the sampled captions, only 4,026 captions ($\sim$0.08%) included causal expressions. Similarly, in the MSCOCO dataset, where we analyzed 415,795 captions, only 53 captions ($\sim$0.01%) contained causal relationships.

These findings reveal that causal relationships are exceedingly rare in the training datasets, with less than 0.1% of captions involving causal reasoning, making it difficult for VLMs to learn and generalize causal understanding from visual inputs.

Benchmark Datasets We then examined the causal reasoning content of two commonly used VLM benchmarks: VQA and VCR [1, 35]. In the VQA validation set, only 1,962 out of 214,354 questions ($\sim$0.92%) involved causal reasoning related to visual inputs; In the VCR validation set, 9,401 out of 26,534 questions ($\sim$35.43%) involved causal reasoning.

To further analyze the VCR dataset, we randomly selected 100 questions from the subset involving causal reasoning and conducted a detailed human annotation. We found that 46% of these questions could be answered correctly by relying solely on object detection or activity understanding without requiring any genuine understanding of causal relationships. As shown in Figure 2, a model could eliminate one incorrect option by recognizing the absence of objects such as "train tracks" in the image. Furthermore, the model could identify the actions of the person in the image (e.g., not tying a rope to something or climbing over the boat) to select the correct answer. In such cases, models rely on object detection and activity recognition to arrive at the correct answer without reasoning about the causal relationships between events in the image and thus have good performance on such benchmarks.

5 Data-Level Improvement

We extract a subset of data from the VQA [1] training set and, for each instance, generate 10 caption pairs, each consisting of one correct caption reflecting a valid causal relationship and one incorrect caption serving as a hard negative example. These examples are used to fine-tune the models. This fine-tuning strategy significantly improves the models' causal reasoning performance on both in-domain and out-of-domain datasets, while preserving downstream performance.

Dataset Following the VQA-Causal construction methodology, we extracted all "why" questions from the VQA training set along with their corresponding correct answers, resulting in a total of 4,891 instances. For each instance, we constructed 10 caption pairs using ten different causal conjunctions: *is due to*, *is caused by*, *is a result of*, *is the effect of*, *because*, *result in*, *cause*, *lead to*, *give rise to*, and *bring about to*. Each pair consists of two captions that differ only in the direction of the causal relationship, with all other elements remaining identical.

Finetuning We adopt the fine-tuning setup from NegCLIP [34] and extend CLIP's [23] contrastive learning objective to better support causal reasoning. For each image-caption pair, we introduce hard

Model	Avg	CW1	CW2	CW3	CW4	CW5	CW6	CW7	CW8	CW9	CW10	CW11	CW12
	VQA-Causal (In-Domain)												
NegCLIP	50.89	62.51	62.25	64.77	67.33	66.97	59.94	69.75	28.45	39.39	33.85	29.33	26.14
CausalCLIP	61.46	70.57	71.70	69.54	73.81	72.16	61.43	56.45	55.68	48.84	57.22	53.72	46.38
	VCR-Causal (Zero-Shot)												
Model	Avg	CW1	CW2	CW3	CW4	CW5	CW6	CW7	CW8	CW9	CW10	CW11	CW12
NegCLIP	51.30	52.21	53.89	54.80	55.23	57.11	50.04	50.87	47.22	50.19	51.32	45.43	47.34
CausalCLIP	57.37	59.10	62.49	61.29	63.69	62.92	54.97	50.16	58.96	53.00	58.33	51.38	52.15

Table 3: CausalCLIP demonstrates strong generalization on both VQA-Causal (in-domain) and VCR-Causal (zero-shot) benchmarks. CW1–CW12 correspond to the following twelve causal conjunctions: *is due to*, *is caused by*, *is a result of*, *is the effect of*, *is the consequence of*, *because*, *owe to*, *result in*, *cause*, *lead to*, *give rise to*, and *bring about to*.

Model	M1	M2	M3	M4	M5	M6	M7	M8	M1	M2	M3	M4	M5	M6	M7	M8
	COCO								**Flickr30K**							
OpenCLIP	0.30	0.56	0.50	0.75	0.30	0.56	0.10	0.34	0.59	0.84	0.79	0.95	0.59	0.84	0.16	0.57
NegCLIP	0.41	0.68	0.56	0.80	0.41	0.68	0.11	0.39	0.67	0.89	0.79	0.95	0.67	0.89	0.16	0.62
CausalCLIP	0.38	0.64	0.54	0.78	0.38	0.64	0.11	0.38	0.64	0.87	0.78	0.94	0.64	0.87	0.16	0.60

Table 4: CausalCLIP exhibits minimal performance loss compared to NegCLIP and even outperforms OpenCLIP on retrieval tasks across both MSCOCO and Flickr30K datasets. Metrics M1–M8 correspond to: ImagePrec@1, ImagePrec@5, TextPrec@1, TextPrec@5, ImageRecall@1, ImageRecall@5, TextRecall@1, and TextRecall@5, respectively.

negative captions, including (1) the incorrect causal order caption for the same image, and (2) three randomly sampled negative captions from other instances in the dataset. Additionally, we randomly sample one alternative image per instance to serve as a negative image, helping ensure generalization and reduce overfitting.

We conduct fine-tuning experiments using NegCLIP [34], a ViT-B/32 variant of CLIP. For each batch of N images I_N, we concatenate the N corresponding correct captions and N incorrect captions to form a $2N$ caption batch. We then compute a similarity matrix between all images and all captions. Following Yuksekgonul et al. [34], we obtain both row-wise and column-wise cross-entropy losses, while ignoring the loss terms from negative captions in the column-wise direction.

Baseline Since we fine-tune on the NegCLIP model, its original performance on the causal reasoning benchmarks serves as our baseline. It is worth noting that our fine-tuning only uses 10 causal conjunctions and is performed exclusively on data from the VQA training set. However, we evaluate the model on all 12 causal conjunctions using both VQA-Causal (as the in-domain benchmark) and VCR-Causal (as the out-of-domain benchmark). Notably, VCR-Causal serves as a zero-shot test set. This setup allows us to evaluate the model's generalization in two ways: (1) to unseen causal conjunctions not present during fine-tuning, and (2) to out-of-domain dataset, thereby providing a more comprehensive assessment of its causal reasoning abilities.

Evaluation As shown in Table 3, our fine-tuned model *CausalCLIP* achieves strong causal reasoning performance on both in-domain and out-of-domain benchmarks. Furthermore, Table 4 shows that this fine-tuning strategy preserves downstream performance and even outperforms OpenCLIP on retrieval tasks over MSCOCO [17] and Flickr30k [33], following the setup of Yuksekgonul et al. [34].

6 Related Work

VLMs have excelled across a wide range of multimodal tasks, including object detection [15, 36], image-text retrieval [11, 13, 23] , visual question answering [1, 14, 18] and commonsense reasoning [35]. However, many recent benchmarks have been proposed to test specific visual understanding capabilities of VLMs and revealed that VLMs perform poorly on tasks requiring fine-grained reasoning skills, such as counting [20, 21], spatial reasoning [3, 9, 30], verb understanding [6, 31], attribute composition[29, 34, 37]. These suggest that models fail to possess high-level visual understanding beyond low-level recognition.

Among these reasoning abilities, causal reasoning is one of the most foundamental abilities, as it allows models to plan interventions and infer underlying mechanisms crucial for complex real-world decision making tasks [10], but remains largely underexplored. Many existing benchmarks for evaluating reasoning ability often focus on video-language models [12, 22, 24, 32] and frequently conflate causal reasoning with other forms of reasoning [1, 7, 19, 35]. Moreover, Li et al. [16] introduced MuCR by generating images conditioned on given causes and designing evaluation tasks such as selecting the correct effect from multiple candidates. However, many of these benchmarks can still be solved through shortcut strategies—for instance, by detecting salient objects or identifying specific activities—as illustrated in Figure 2. This makes it difficult to determine whether models truly capture the causal order between causes and effects.

Our work addresses this critical gap by introducing two dedicated benchmarks—VQA-Causal and VCR-Causal—that explicitly evaluate whether VLMs can distinguish between alternative causal interpretations of the same visual scene, thus enabling rigorous causal reasoning evaluation in multimodal models.

7 Conclusion

We introduce VQA-Causal and VCR-Causal, the first benchmarks designed to comprehensively evaluate VLMs' causal reasoning abilities across 12 causal conjunctions. Despite strong performance in object and activity recognition, all ten evaluated models perform poorly on causal reasoning tasks—nine achieving no more than 53% accuracy, barely above chance. To understand this limitation, we analyze four commonly used datasets including LAION-400M, MSCOCO, VQA, and VCR, and find that explicit causal expressions are exceedingly rare in LAION-400M and MSCOCO datasets, with fewer than 0.1% of instances involving causal relationships. Moreover, only 0.92% of VQA and 35.43% of VCR questions require causal reasoning, and 46% of sampled VCR questions can be solved using shortcuts without genuine causal reasoning. Finally, we extract 4,891 causality-related instances from the VQA training set and construct contrastive training data by pairing correct captions with hard negative examples that differ only in causal direction. Fine-tuning with this data significantly improves causal reasoning performance on both in-domain and out-of-domain benchmarks, while maintaining downstream task performance.

8 Limitations

Our work uncovers the weaknesses of current VLMs on causal reasoning tasks. By analyzing both their training data and benchmark datasets, we proposed a data-level fine-tuning strategy that significantly enhances causal reasoning ability with minimal impact on downstream performance. However, this approach mainly focuns on data-level and does not address the underlying model architecture. A promising direction for future research is to improve causal and fine-grained reasoning at the model architectural level. For example, researchers could adjust attention weights to guide the model foucs more on fine-grained visual features or implement broadly generalizable modifications to specific VLM components to improve both causal reasoning ability and fine-grained visual understanding. Finally, although our study focuses on vision–language models, causal reasoning and fine-grained visual reasoning in other multimodal settings, such as video–language models, remains an important direction for further investigation.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[3] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024.

[4] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[5] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

[6] L. A. Hendricks and A. Nematzadeh. Probing image-language transformers for verb understanding. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.318. URL `https://aclanthology.org/2021.findings-acl.318/`.

[7] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[8] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

[9] A. Kamath, J. Hessel, and K.-W. Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, 2023.

[10] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[11] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[12] J. Li, L. Niu, and L. Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282, 2022.

[13] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[14] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[15] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

[16] Z. Li, H. Wang, D. Liu, C. Zhang, A. Ma, J. Long, and W. Cai. Multimodal causal reasoning benchmark: Challenging multimodal large language models to discern causal links across modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5509–5533, 2025.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[19] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[20] R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, and T. Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180, 2023.

[21] L. Parcalabescu, A. Gatt, A. Frank, and I. Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In L. Donatelli, N. Krishnaswamy, K. Lai, and J. Pustejovsky, editors, *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online), June 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.mmsr-1.4/`.

[22] P. Parmar, E. Peh, R. Chen, T. E. Lam, Y. Chen, E. Tan, and B. Fernando. Causalchaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. *Advances in Neural Information Processing Systems*, 37:92769–92802, 2024.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[24] I. S. Rawal, A. Matyasko, S. Jaiswal, B. Fernando, and C. Tan. Dissecting multimodality in videoqa transformer models by impairing modality fusion. In *International Conference on Machine Learning*, pages 42213–42244. PMLR, 2024.

[25] C. Schlarmann, N. D. Singh, F. Croce, and M. Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *International Conference on Machine Learning*, pages 43685–43704. PMLR, 2024.

[26] C. Schuhmann, R. Kaczmarczyk, A. Komatsuzaki, A. Katta, R. Vencu, R. Beaumont, J. Jitsev, T. Coombes, and C. Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021.

[27] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=M3Y74vmsMcY`.

[28] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022.

[29] Y. Tang, Y. Yamada, Y. Zhang, and I. Yildirim. When are lemons purple? the concept association bias of vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14333–14348, 2023.

[30] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, S. Li, and N. Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024.

[31] Z. Wang, A. Blume, S. Li, G. Liu, J. Cho, Z. Tang, M. Bansal, and H. Ji. Paxion: Patching action knowledge in video-language foundation models. *Advances in Neural Information Processing Systems*, 36:20729–20749, 2023.

[32] J. Xiao, X. Shang, A. Yao, and T.-S. Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[33] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

[34] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, J. Zou, et al. When and why vision-language models behave like bags-of-words, and what to do about it? In *11th International Conference on Learning Representations, ICLR 2023*. International Conference on Learning Representations, ICLR, 2023.

[35] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[36] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022.

[37] T. Zhao, T. Zhang, M. Zhu, H. Shen, K. Lee, X. Lu, and J. Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

[38] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.