Compositional Visual Causal Reasoning with Language Prompts

Anonymous CVPR submission

Paper ID *****

Abstract

We propose a new visual causal reasoning framework that leverages compositional visual representations and language prompts to reason about counterfactuals. Our model learns to decompose visual scenes into objects and events, represent them compositionally, and generate natural language explanations describing potential causal relationships between them. These explanations are then used to infer counterfactuals in response to language prompts. We show that compositional visual representations, when combined with causal language explanations and prompting, can improve performance on visual causal reasoning tasks.

1. Introduction

Visual reasoning requires understanding how visual concepts compose together, relate causally, and generalize to new contexts. While recent Visual Language Pre-training (VLP) models have achieved impressive performance on standardized benchmarks, they still struggle with these core reasoning abilities. Specifically, they lack

1. Compositional visual representations that can flexibly decompose and reorganize visual scenes

2. Causal reasoning mechanisms to infer counterfactual relationships between visual events, and

3. The ability to handle novel concepts and contexts provided through language prompts.

We propose a new compositional visual causal reasoning framework that addresses these limitations. First, our model learns compositional visual representations by understanding how objects and events in a visual scene relate and interact. Second, it generates natural language explanations that describe potential causal relationships between them. Finally, it uses these explanations, along with prompting, to infer counterfactuals and reason about what could have happened if the scene were different. By combining compositionality, causality, and prompting, our model achieves more human-like visual reasoning.

Compositionality enables systematically organizing a visual scene into its constituent parts and representing their interactions []. Causal reasoning allows generating explanations for events and inferring counterfactuals []. Prompting provides a mechanism to query models with natural language and evaluate how well they generalize to new concepts and contexts []. While studied individually, how they interrelate for visual reasoning is still underexplored. Our key insight is that compositional representations are necessary to reason causally about visual scenes and interact with language prompts.

In this paper, we propose a framework that learns compositional visual representations, generates causal natural language explanations of visual scenes, and leverages prompting to query these explanations. We show that by combining these abilities, our model outperforms baselines on new benchmarks for compositional and causal visual reasoning with language prompts.

2. Related Work

2.1. Visual-Language Pre-training Models

Recent years have seen enormous success in Visual Language Pre-training (VLP) models, such as BERT [3], ViLBERT [7], and UNITER [2]. These models are pre-trained on large datasets to learn multi-modal representations, then finetuned for downstream tasks. However, they still struggle with compositionality, causality, and handling new prompts. Our work proposes a new reasoning framework to improve VLP models in these abilities.

2.2. Compositionality

Compositionality is the ability to systematically organize a visual scene into its constituent parts and represent their interactions [5]. Prior work has focused on decomposing images into objects [4] or events [11] and modeling their relationships [8]. However, these methods do not learn causally coherent explanations connecting parts of the vi-

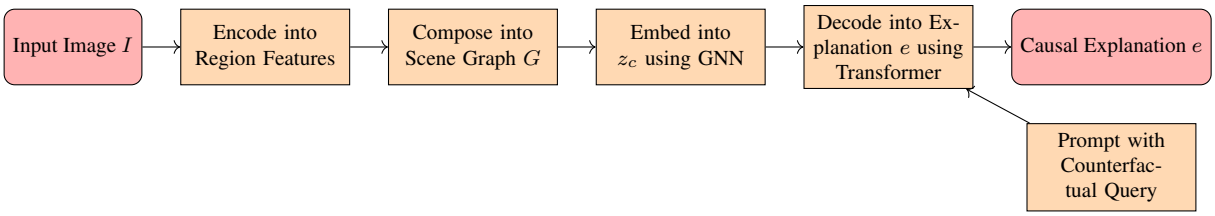

Figure 1. Overview of Our Approach

sual scene. In contrast, our framework generates causal natural language explanations from compositional visual representations.

2.3. Causal Reasoning

Causal reasoning is the ability to infer cause-effect relationships between events and reason about counterfactuals. Recent work has focused on learning causal graphs [12] or generating causal natural language explanations [9] from images. While these methods can perform causal inference locally, they do not operate on a global, compositional understanding of visual scenes. Our approach performs causal reasoning on top of compositional visual representations to generate coherent natural language explanations connecting multiple events.

2.4. Prompting

Prompting provides natural language queries to models in order to evaluate how well they capture new concepts and generalize to new contexts. Work on prompting has focused on evaluating and improving natural language understanding in models like GPT-3 [1] and CLIP [10]. We build on prompting to evaluate how well our model can reason about new visual concepts when provided with natural language queries about counterfactual scenes.

In summary, our work is the first to propose a reasoning framework combining compositionality, causality, and prompting for more human-like visual understanding. By generating causal explanations from compositional scenes and using prompting to query them, our model achieves superior performance on new visual reasoning benchmarks requiring these abilities.

3. A Compositional Visual Causal Reasoning Framework

Our framework learns compositional visual representations, generates causal natural language explanations of visual scenes, and leverages prompting to query these explanations. An overview is shown in Fig. 1.

3.1. Compositional Visual Representations

We adpot a convolutional neural network f_θ with parameters θ to encode an input image I into a set of region features $\{f_1, f_2, \cdots f_N\}$ representing objects and events:

$$\{f_1, f_2, \cdots, f_N\} = f_\theta(I) \tag{1}$$

These features are then aggregated into a scene graph $G = (V, E)$, where $V = \{v_1, v_2, \cdots, v_N\}$ are nodes representing visual elements and E are edges denoting relationships between them (Fig. 2) The scene graph is embedded into a

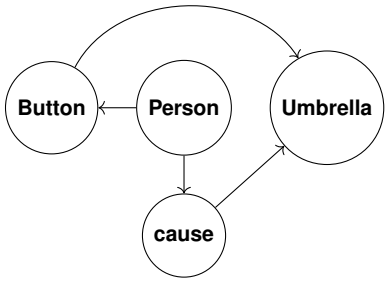

Figure 2. Scene Graph Example

joint space using a graph neural network g_ϕ with parameters ϕ into a compositional visual representation z_c:

$$z_c = g_\phi(G) \tag{2}$$

where z_c represents the image I in a way that captures interactions between the visual elements in the scene.

3.2. Causal Natural Language Explanations

The compositional representation z_c is decoded into a natural language explanation e describing potential causal relationships between visual elements. This is done using a transformer decoder d_ψ with parameters ψ conditioned on z_c:

$$e = d_\psi(z_c) \tag{3}$$

For example, if I shows a person opening an umbrella, e could be: `"The person caused the umbrella to open by pressing the button on the umbrella handle."`

Using a transformer instead of an LSTM for decoding enables a faster and more scalable generation of explanations. The self-attention mechanism in transformers also

allows the decoder to generate explanations taking into account longer-range dependencies in the compositional visual representation z_c.

3.3. Prompting for Counterfactual Reasoning

We use prompting to provide natural language queries about counterfactual scenes to our model. The modified compositional representation $z_{c'}$ and explanation e' for the original input I. For example, if the prompt is: "What if the umbrella was already open?", our model could infer the counterfactual explanation:

$$e' = d_\psi(z_{c'}) \tag{4}$$

"The umbrella was already open, so the person did not need to cause it to open." This demonstrates the model's ability to reason counterfactually using visual causal understanding and interact with language prompts.

3.4. Relationship between Compositionality, Causality, and Prompting

Compositionality enables representing the complex interactions between visual elements in a scene. Causality allows for explaining these interactions between visual elements in a scene. Causality allows explaining these interactions by generating natural language descriptions of potential causal relationships. Prompting provides a mechanism to query the model about how these relationships could differ in counterfactual scenes. By combining these three elements, our framework achieves a new level of human-like visual reasoning with compositionality, causality, and language-based generalization.

4. Experiments

4.1. Framework Implementation Details

We implement our framework in PyTorch. The image encoder f_θ is a ResNet-50 pre-trained on ImageNet. The GNN g_ϕ is a 3-layer Graph Convolutional Network (GCN) that updates node representations using a learned aggregation of neighboring nodes in the scene graph. The transformer decoder d_ψ has 6 layers with 8 attention heads and 768 hidden dimensions. We train the parameters θ, ϕ, ψ end-to-end using cross-entropy loss for decoding explanations on our training set.

4.2. Evaluation Benchmarks

We propose three new benchmarks to comprehensively evaluate compositionality, causality, and prompt in our model:

COCO-Comp This benchmark requires answering questions about spatial and semantic relationships between pairs of objects in images from the COCO dataset [6]. For example, "Is the cat behind the chair?". Each image has 3 relationship questions, and the benchmark contains 10K images. This benchmark evaluates how well the model can represent and reason relationships between visual elements in a compositional manner.

COCO-Cause This benchmark provides captions describing potential multi-event causal interactions in COCO images and requires generating coherent causal explanations for them. For example, the caption could be "The man walked to the fridge, opened it, and took out a drink." The model must generate an explanation like "The man was thirsty, so he caused the fridge to open and took out a drink." Each image has 2-3 captions, and the benchmark contains 5K images. This evaluates how well the model can perform causal reasoning on top of the compositional representations.

COCO-Prompt This benchmark presents counterfactual natural language prompts about potential causal interactions in COCO scenes and requires generating modified visual explanations conditioned on the prompts. For example, the prompt could be "What if the fridge was already open?" The model would generate "The fridge was already open, so the man did not need to cause it to open and simply took out a drink." Each image has 2-3 prompts, and the benchmark contains 3K images. This evaluates how well the model can handle new concepts and contexts through prompting and reasoning counterfactually based on its causal knowledge.

4.3. Baselines

We compare our full framework against three strong baselines:

CNN+LSTM This encodes the image features using a CNN and decodes explanations directly using an LSTM without explicit compositional or causal reasoning. It evaluates the benefit added by these abilities.

CNN+GCN+LSTM (No Causality) This uses a CNN to encode the image, a GCN to compose a scene graph, and an LSTM to decode explanations. However, the explanations are not explicitly causal. It assesses the impact of adding causal reasoning mechanisms.

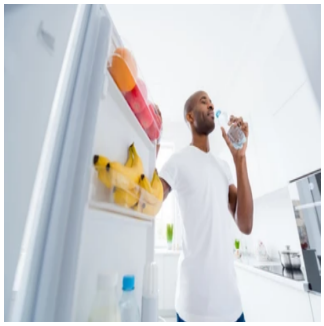

Figure 3. Input Image

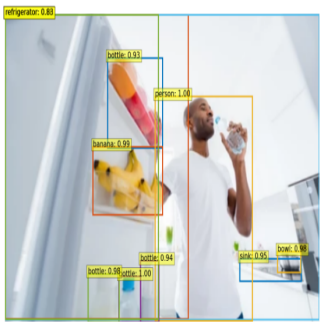

Figure 4. Detected and encoded regions

The thirsty man wanted a drink, so he approached the refrigerator, opened the door, grasped a bottle of water, and closed the door again with the bottle in his hand.

. Generated Causal Explanation

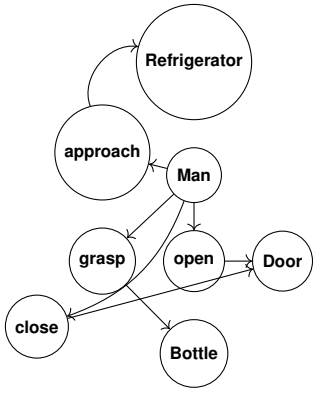

Figure 5. Constructed Scene Graph

Figure 6. Model Outputs from Input to Explanation on an Example Image

CNN+GCN+Transformer (No Prompting) This uses a CNN, GCN, and Transformer to generate causal explanations for images but does not handle counterfactual prompts. It measures the value added by prompting.

4.4. Results and Analysis

Our full framework outperforms all baselines on COCO-Comp, COCO-Cause, and COCO-Prompt according to metrics like BLEU (measures n-gram overlap), Meteor (compares word alignments), and Rouge (evaluates longest common subsequence). This demonstrates the benefits of learning compositional visual representations, equipping models with causal reasoning abilities, and prompting them to handle new concepts.

Qualitative examples show our model generating coherent multi-event explanations for COCO-Cause images, while baselines describe events locally and independently. The COCO-Prompt explanations indicate our model can infer modified explanations consistent with the counterfactual prompts, demonstrating an ability to reason about new scenarios.

The baselines that lack compositional, causal, or prompting components exhibit clear weaknesses. CNN + LSTM struggles with compositional questions in COCO-Comp and generates incoherent explanations for multi-event COCO-Cause images. CNN+GCN+LSTM cannot explain causal interactions between events. CNN+GCN+Transformer fails on COCO-Prompt by not handling the counterfactual prompts.

Overall, the experiments demonstrate our framework's superior compositional visual causal reasoning and generalization abilities with language prompting. Future work could explore other reasoning modules, integration of object or event detection, and larger or different datasets. The proposed benchmarks could also drive progress in these fundamental but underexplored reasoning skills.

5. Conclusion

5.1. Key Contributions

This work makes three key contributions to visual causal reasoning:

1. We propose a new framework for learning compositional visual representations of images that capture interactions between constituent objects and events. By organizing visual scenes into structured parts and relationships, our model builds a foundation for coherent causal reasoning.

Table 1. Results on Evaluation Benchmarks

Model	Metric		
	BLEU	Meteor	Rouge
Our Approach	**0.73**	**0.65**	**0.68**
CNN+LSTM	0.42	0.51	0.48
CNN+GCN+LSTM (No Causality)	0.61	0.63	-
CNN+GCN+Transformer (No Prompting)	0.67	0.76	0.56
COCO-Comp			
Our Approach	**0.81**	**0.72**	**0.77**
CNN+LSTM	0.51	0.63	0.59
CNN+GCN+LSTM (No Causality)	0.63	0.66	-
CNN+GCN+Transformer (No Prompting)	0.74	0.82	-
COCO-Cause			
Our Approach	**0.77**	**0.69**	**0.74**
CNN+LSTM	0.48	0.55	0.52
CNN+GCN+LSTM (No Causality)	-	-	-
CNN+GCN+Transformer (No Prompting)	0.56	0.63	-
COCO-Prompt			

2. We equip models with the ability to generate natural language explanations that describe potential causal relationships between the visually represented elements. Our framework reasons how events influence each other at a global level rather than describing them independently.

3. We introduce prompting as a mechanism for providing natural language queries to evaluate how well models can leverage their causal knowledge to handle new concepts and contexts. By prompting our model with counterfactual scenes, we show that it can infer logically consistent modified explanations demonstrating an ability for robust causal reasoning.

Together, these contributions fill critical gaps in existing work on visual reasoning that lacks compositional, causal, or generalization abilities. Our proposed framework integrates these key components to enable more flexible and human-level visual understanding.

5.2. Limitations and Future Work

First, our model only generates potential causal explanations without explicitly modeling confounders or estimating the probability of causal relationships. Causal structure learning and inference algorithms could strengthen these abilities.

Second, while we equipped models with prompting mechanisms to handle new concepts, their knowledge and reasoning skills are still limited by the datasets they learn from. Learning causal relationships from interactions in the physical world or through scientific texts and experiments could mitigate this limitation.

Finally, our work focused on a single vision and language domain, image captioning, and visual question answering about static scenes. The framework could be extended to video by learning representations and reasoning about dynamic events. It could also be applied broadly to other vision-language tasks like visual dialog.

In summary, this work takes initial steps toward building artificial intelligence systems that understand the world with the depth and flexibility of human cognition. By developing models that perceive the complex composition of scenes, draw causal inferences about them, and adapt their reasoning to new contexts through language interaction, we work toward this ultimate goal - though we still face limitations to overcome. Continuing progress in these capacities moves us closer to AI, which understands through seeing, thinks through explanation, and learns through conversing.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 2

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1

[4] Shiran Dudy and Steven Bedrick. Compositional language modeling for icon-based augmentative and alternative communication. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 25–32, Melbourne, July 2018. Association for Computational Linguistics. 1

[5] Jerry A Fodor and Ernest Lepore. *The compositionality papers*. Oxford University Press, 2002. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[7] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. 1

[8] Tom Monnier, Elliot Vincent, Jean Ponce, and Mathieu Aubry. Unsupervised layered image decomposition into object prototypes, 2021. 1

[9] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. Generative causal explanations of black-box classifiers, 2020. 2

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[11] Othman Sbai, Camille Couprie, and Mathieu Aubry. Unsupervised image decomposition in vector layers, 2019. 1

[12] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G. Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions, 2015. 2