

See or Guess: Counterfactually Regularized Image Captioning

Anonymous Authors

ABSTRACT

Image captioning, which generates natural language descriptions of the visual information in an image, is a crucial task in vision-language research. Previous models have typically addressed this task by aligning the generative capabilities of machines with human intelligence through statistical fitting of existing datasets. While these models demonstrate proficiency in describing the content of normal images, they may struggle to accurately describe those where certain parts of the image are obscured or edited. Conversely, humans effortlessly excel at it in this case. The weaknesses these models exhibit, including hallucinations and limited interpretability, often result in performance declines when applied to scenarios involving shifted association patterns. In this paper, we present a generic image captioning framework that leverages causal inference to make existing models more capable of interventional tasks, and counterfactually explainable. Specifically, our approach consists of two variants that utilize either total effect or natural direct effect. We incorporate these concepts into the training process, enabling the models to handle counterfactual scenarios and thereby become more generalizable. Extensive experiments on various datasets have demonstrated that our method can effectively reduce hallucinations and increase the model’s faithfulness to the images, with a high portability for both small-scale and large-scale image-to-text models.

KEYWORDS

Image Captioning, Counterfactual Causal Inference, Object Hallucination, Image-to-text Generation

1 INTRODUCTION

As a fundamental task in vision-language understanding research, image captioning requires models to mimic the human ability to compress huge amounts of visual information into descriptive language [3, 19, 36]. A large amount of image-to-text methods [31] have been developed, among which recent large multimodal models [12, 18, 43] perform surprisingly well in describing an image in details. Despite their good performance in real scenarios, their capabilities still differ from those of humans in interventional scenarios. For example, in Figure 1, the BLIP model [13] can generate a sentence that accurately describes the factual image at left. However, when the bicycle is masked or changed to a tree as shown in the counterfactual images, it generates incorrect descriptions such as “A man sitting on a bench in front of a river.” Such errors reveal

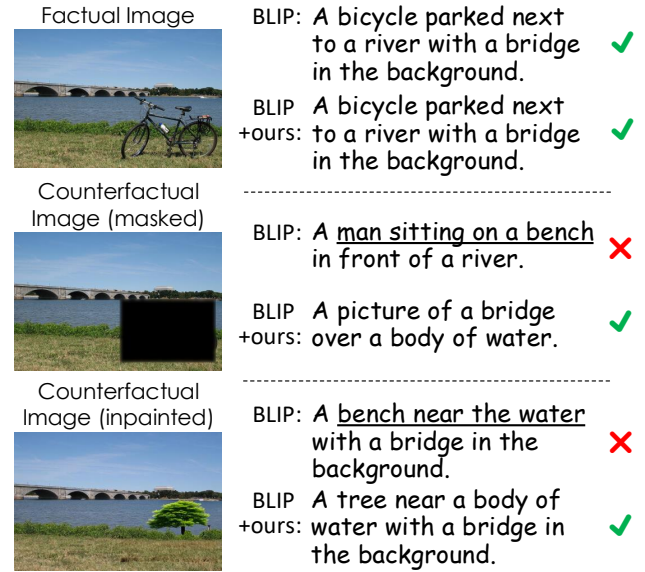


Figure 1: An example of generated captions of different methods in the factual and two counterfactual scenarios (masked and inpainted).

that the model might not have precisely understood the image. Instead, it may make guesses based on the common association patterns in the datasets. For example, the frequent co-occurrence of a river and a man in the dataset may lead the model to form shortcut connections and wrongly generate “man” for most images with a river.

The above analysis suggests that while current models may exhibit impressive performance, it does not necessarily imply their ability to accurately comprehend the contents of an image and generate appropriate descriptions, a capability inherently possessed by humans. Such weaknesses may result in hallucinations and hinder the interpretability of models since people cannot exactly tell which parts of the image correspond to the generated words in the text. Furthermore, when these learned models are applied to other scenarios with shifted association patterns, their performance may suffer a substantial deterioration.

To overcome the above shortcomings, we design a novel framework that integrates causal inference into any image captioning model to mitigate the shortcut correlations. Specifically, we leverage counterfactual concepts to enhance the correspondence between visual and textual characteristics. We hope that when certain regions of an image are removed, the generated text should not include the descriptions of those regions. While this idea is intuitive, it is challenging to implement for the following reasons. First, existing counterfactual models primarily focus on classification tasks [1, 8, 39]. However, we handle a generation task that necessitates the consideration of sequential impacts between words. Moreover, in our

Permission to make digital or hard copies of all or part of this work for personal or professional use, by individuals or small groups, is granted by ACM, provided that the fee of \$12.00 is paid directly to ACM. This permission is granted without fee where organizations register with ACM for a Special Rate Plan. For all other uses, permission should be sought from ACM. Copyright © 2024 ACM. ISBN 978-1-4503-XXXX-X/18/06. <https://doi.org/XXXXXXX.XXXXXXX>

Unpublished working draft. Not for distribution. Distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. MM '24, xxx, xxx

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-XXXX-X/18/06 <https://doi.org/XXXXXXX.XXXXXXX>

multi-modal scenario, the influence of the image on the word can be attributed to two paths: (1) a direct influence from the image to the word, and (2) an indirect influence, where the image first impacts preceding words, which subsequently influence the current word. It presents a nontrivial challenge to distinguish between these two paths and enhance the first one to minimize hallucination while preserving linguistic fidelity.

To address the above challenges, in this paper, we first formalize the image-to-caption generation task as a sequential causal graph, where each word in the generated text is determined by both its previous words and the image. Following this causal graph, we leverage the causal concepts of total effect (TE) and natural direct effect (NDE) to discriminate different reasons for the word generation process. Then we can intervene in the cause, and enhance the correspondences between the image and words while controlling the other influential factors. Finally, we propose a counterfactually regularized image captioning framework. The main contributions of our paper are as follows:

- We propose a generic framework to counterfactually regularize image captioning models and thus make them more human-like, explainable, and robust.
- We propose two causal methods based on total effect and natural direct effect to enhance the correspondence between the visual and textual characteristics.
- We extensively experimented on various models and datasets to demonstrate the high generality and interpretability of our methods, which can effectively reduce object hallucinations and enhance the faithfulness of the model to the images.

2 RELATED WORK

2.1 Image Captioning

Image captioning, a crucial aspect of image-to-text generation [31], has evolved from convolution neural network (CNN)-based encoders and recurrent neural network (RNN)-based decoders [34, 37] to Transformer architectures [5, 11], and further into the era of vision language pretraining (VLP) models [13, 14, 42]. Recent advancements in Large Vision Language Models [12, 18, 43] have sparked renewed interest in the field. In addition, some methodologies explore the integration of multimodal representation models like CLIP [27] to furnish visual support for language models [19]. Our proposed approach has a model-agnostic nature and flexibility. Considering the notable performance of VLP models, we opt to validate the effectiveness across various architectures (decoder-only and encoder-decoder) and model scales by employing ClipCap [19], BLIP [13], and BLIP2 [12] as backbone models.

2.2 Object Hallucination in Image Captioning

Alleviating hallucination of image captioning models does not solely hinge on improved image perception capability but also on factors like over-reliance on language priors or biases during sequence generation [25, 28], potentially leading to guesswork that is not faithful to the image. Researchers [28] thus propose utilizing the CHAIR metric to quantify hallucination occurrence. Some efforts have been made to reduce model reliance on common or biased co-occurrences by adjusting object label co-occurrence statistics [4].

Other methods maintain semantic consistency to reduce object hallucinations by learning consensus representations through aligning scene and language graphs [40], or by aligning textual tokens and visual objects using masked language modeling [6]. However, these methods may blur semantic and visual alignments and overly rely on dataset co-occurrence patterns, harming interpretability and performance in real scenarios. Our approach considers causality, aiming to establish the correct vision-to-language guiding relationship during the generation process.

2.3 Counterfactual Causal Inference

Causal inference seeks to unravel the causal relationships and underlying mechanisms driving observed outcomes [2, 7, 23]. Moreover, counterfactual causal inference offers a framework to enhance [1, 32] and explain [8, 9] models in counterfactual scenarios. However, the majority of these counterfactual-related works are tailored for classification tasks, such as image classification [1, 8, 39], representations learning [32, 41], or visual question answering [10, 15, 20], rather than for generation tasks. Classification tasks exhibit a deterministic correspondence between input and output, whereas, in the generation process, the counterfactual image and preceding generated tokens collectively influence the subsequent token generation, creating an effect propagation. Some researchers [35] have performed Maximum Likelihood Estimation (MLE) on interventional distribution to reduce spurious correlations learned by models due to observed confounding factors. However, the applicability of their framework is limited to the strong ignorability assumption and lacks causal analysis in multi-modal scenarios. Capturing this causal correspondence [22] is challenging, especially in multimodal scenarios, and has thus received little attention in prior literature. In this paper, we endeavor to leverage counterfactual causal inference to tackle this challenge and gain deeper insights into a model’s generation behavior.

3 PRELIMINARIES: A CAUSAL LOOK AT IMAGE CAPTIONING

This section presents the fundamental concepts and notations of causal inference [7, 22] and how we apply it in image captioning. In the following, capital letters, *i.e.*, cause X , Mediator M and Effect Y , represent random variables. The values or subscripts of these random variables indicate their observed values.

As for image captioning, a model is used to process an input image I and produce a corresponding textual description, *i.e.*, a sequence $S = (s_1, s_2, \dots, s_L)$, where s_i is a token in the sequence and L is the sequence length. The sequence of the preceding tokens of s_j is denoted as $S_{<j} = (s_1, s_2, \dots, s_{j-1})$. We will later present how to treat these variables from a causal perspective.

3.1 Causal Graph

A causal graph describes the causal relations between different random variables in a graph manner [24]. In a causal graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, a node $v \in \mathcal{V}$ represents a variable and a directed edge $e \in \mathcal{E}$ represents a causal relationship between variables. The *direct effect* means that there is an edge between two variables, *e.g.*, in Figure 2(a), X has a *direct effect* on Y . The *indirect effect* means that

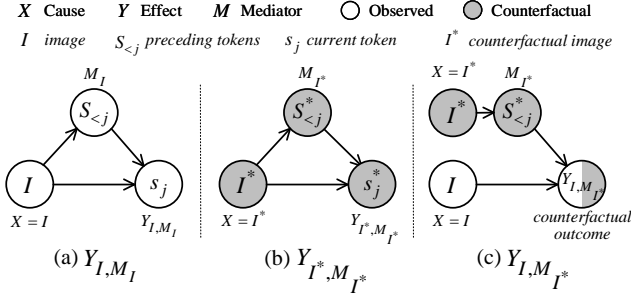


Figure 2: Illustration of causal graphs and counterfactual causal effect notations.

two variables are not directly linked, but are connected via some mediator variables, e.g., X has an indirect effect on Y if $X \rightarrow M \rightarrow Y$.

Considering the process of auto-regressive generation, at each step, the current token s_j is determined by all the preceding tokens $S_{<j}$, and the visual information of the input image I as well. As shown in Figure 2(a), at step j , $S_{<j}$ is influenced by I , and s_j is jointly determined by I and $S_{<j}$. We use $Y_{I, M_j} = Y(X = I, M = M_j)$ to denote the probability of token s_j when the cause X is set to I and the mediator M is set to $M_j = S_{<j}$.

3.2 Counterfactual Causal Effects

In causal inference, counterfactual causal effects compare hypothetical outcomes under factual and counterfactual treatments [2, 23].

As shown in Figure 2(b), the value of the counterfactual of variable X is equal to the counterfactual image I^* , where I^* is created by intervening in the factual image I . The hypothetical outcome of Y is denoted as $Y_{I^*, M_{I^*}} = Y(X = I^*, M = M_{I^*})$, where the mediator $M_{I^*} = S_{<j}^*$. The total effect (TE) is the difference between two hypothetical conditions: one being factual transition where $X = I$ (under treatment, corresponding to Figure 2(a)) and the counterfactual being $X = I^*$ (under no-treatment, corresponding to Figure 2(b)). Mathematically, the total effect can be expressed as

$$\text{TE}_{I, I^*} = Y_{I, M_{I^*}} - Y_{I^*, M_{I^*}}. \quad (1)$$

TE_{I, I^*} measures the effect of all factors (i.e., direct and indirect effects) resulting from changing image I to I^* .

Further, intervening both X and M allows the total effect to be decomposed into two components, namely the natural direct effect (NDE) and the total indirect effect (TIE). Unlike TIE that focuses on the effect brought by changes in the mediator M , NDE is the effect of X on Y that results solely from changes in X , without any influence from M , which can be denoted as

$$\text{NDE}_{I, I^*} = Y_{I, M_{I^*}} - Y_{I^*, M_{I^*}}. \quad (2)$$

The first term $Y_{I, M_{I^*}}$ corresponds to Figure 2(c), which keeps $X = I$ and conducts intervening on M via I^* to form a counterfactual outcome $Y_{I, M_{I^*}}$. The second term $Y_{I^*, M_{I^*}}$ corresponds to Figure 2(b). Formula 2 describes the variation of Y when X is changed from I to its counterfactual I^* while M is held constant at $M(X = I^*)$.

This paper explores how to use TE or NDE to reduce object hallucination in image captioning and improve interoperability.

4 COUNTERFACTUAL REGULARIZATION

In this section, we introduce how to construct counterfactual data, propose our framework, and design two counterfactual regularization losses that can be universally applied to existing image captioning models.

4.1 Constructing Counterfactual Data

Collecting or generating counterfactual images for the factual ones is difficult. However, by adding a mask, it is easy to achieve minimal changes to the original image when constructing the counterfactual one, which can be regarded as an approximation of the idealized counterfactual image. Specifically, we construct counterfactual images by leveraging datasets that have labeled bounding boxes for corresponding phrases in the image captions. As shown in Figure 3(a), we first select the entity to intervene (\tilde{S}), e.g., “black poodle”. Then we identify its corresponding region to intervene (r) in the image based on the labeled bounding boxes in the datasets. A black mask is used to replace the region r to create a counterfactual image I^* . If the entity to intervene corresponds to more than one bounding box (e.g., when \tilde{S} is “a group of people”), all related regions will be masked. Next, we employ the initial image captioning model to generate a counterfactual caption S^* . The counterfactual captions are used to model $S_{<j}^*$ and s_j^* in the causal graphs (Figure 2), i.e., modeling the generated words for the counterfactual image I^* . Please note that the goal of S^* is to facilitate the estimation of causal effects, instead of being used as a ground-truth caption for counterfactual images. Thus, it is much easier to obtain compared with ground-truth labels for counterfactual images. Accordingly, we have $(I, S, \tilde{S}, I^*, S^*)$ prepared for dataset \mathcal{D} .

4.2 Our Framework

We propose a framework by incorporating negative log-likelihood (NLL) loss \mathcal{L}_{NLL} with TE or NDE regularization loss, i.e., \mathcal{L}_{TE} or \mathcal{L}_{NDE} , which will be described later. Formally, the vanilla negative log-likelihood (NLL) loss is as follows:

$$\mathcal{L}_{\text{NLL}} = - \sum_{(I, S) \in \mathcal{D}} \sum_{i=1}^L \log f_{\theta}(s_i | I, S_{<i}), \quad (3)$$

where $f_{\theta}(\cdot)$ refers to the model that takes the image and preceding text sequence $S_{<i}$ as input and outputs a probability distribution on the vocabulary to generate the next token s_i , with parameters θ .

We add the counterfactual regularization loss to allow the model to learn together with the NLL loss. A hyperparameter α determines the weight of losses, ensuring balanced optimization. The final loss is denoted as:

$$\begin{aligned} \mathcal{L}_1 &= \alpha \mathcal{L}_{\text{NLL}} + (1 - \alpha) \mathcal{L}_{\text{TE}}, \\ \mathcal{L}_2 &= \alpha \mathcal{L}_{\text{NLL}} + (1 - \alpha) \mathcal{L}_{\text{NDE}}. \end{aligned} \quad (4)$$

The whole optimization includes two stages: (1) training the model with vanilla NLL loss (Formula 3); (2) training the model with either \mathcal{L}_1 or \mathcal{L}_2 (Formula 4) using the constructed counterfactual images and their corresponding generated counterfactual captions.

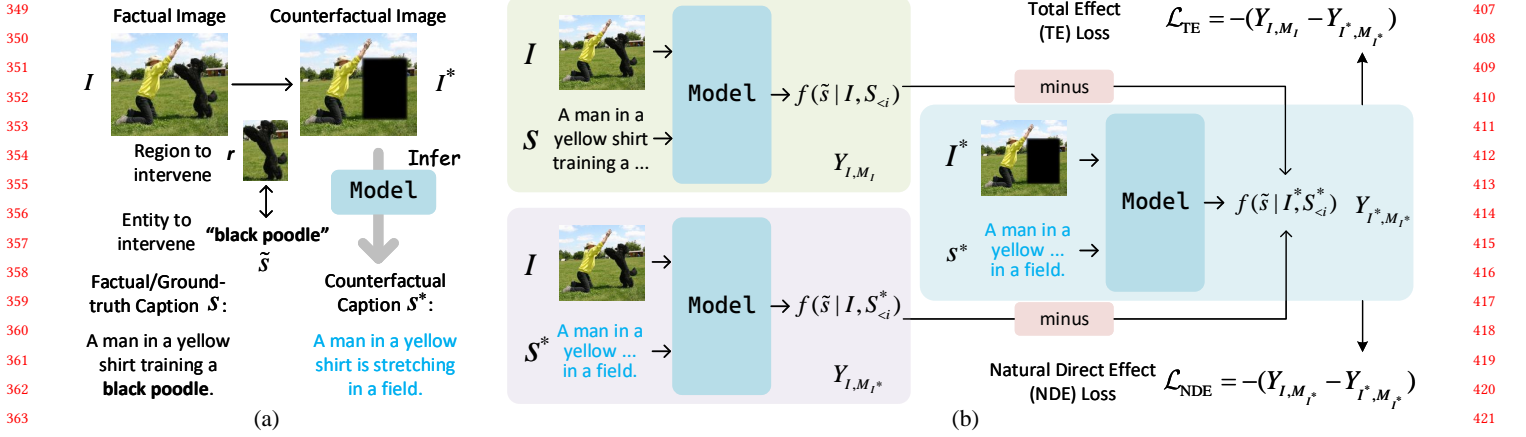


Figure 3: Our framework of counterfactual regularization. (a) shows how to prepare counterfactual images and captions by an example. (b) illustrates how the TE loss and NDE loss are calculated in the example. Counterfactual captions are in blue. The phrase corresponding to the image region in the mask is “black poodle”. Best viewed in color.

4.3 Total Effect Regularization

When the region corresponding to “black poodle” is masked in the image (Figure 3), we hope the model will significantly lower the generation probabilities of the words “black poodle” to reduce hallucination. To achieve this goal from a causal perspective, we maximize the total effect of changing I to I^* on the generation of “black poodle” (i.e., \tilde{S}), which is given in Formula 1. Maximizing this total effect can be fulfilled by minimizing the following total effect (TE) loss:

$$\mathcal{L}_{\text{TE}} = - \sum_{(I, S) \in \mathcal{D}} \sum_{j=1}^{L_{\tilde{S}}} \left[\log f_{\theta}(\tilde{s}_j | I, S_{<p+j}) - \frac{1}{L_{S^*}} \sum_{i=1}^{L_{S^*}} \log f_{\theta}(\tilde{s}_j | I^*, S_{<i}^*) \right], \quad (5)$$

where the first part corresponds to Y_{I, M_I} in Formula 1 and calculates the likelihood of generating \tilde{S} (e.g., “black poodle”) given the factual image I and previously generated words, and the second part estimates Y_{I^*, M_I^*} in Formula 1 with the likelihood of generating \tilde{S} at any position given the counterfactual image I^* and preceding tokens $S_{<i}^*$ that are generated from I^* . Here, \tilde{s}_j denotes the j -th token of the entity to intervene \tilde{S} , of which the length is $L_{\tilde{S}}$. p represents the index of the first position where the entity to intervene \tilde{S} appeared in the ground truth or factual caption (e.g., if “black” and “poodle” are the 9th and 10th words, then $p = 9$). L_{S^*} is the length of counterfactual caption S^* . The specific occurrence position of the entity to intervene \tilde{S} is explicit in the ground truth, while it may not necessarily appear in the counterfactual caption. Therefore, we need to estimate the probability of their occurrence using the average value.

In the example shown in Figure 3, the word “black poodle” is the entity to intervene (\tilde{S}). We estimate the first term by the probability of generating each token in “black poodle” at the position it appeared in the factual caption, i.e. using the preceding tokens “A man in a yellow shirt training a”. The second term is calculated by the average probabilities of generating each token in “black poodle” at any position in the counterfactual caption “A man in a yellow

shirt is stretching in a field.”, where the preceding tokens are those before each step.

4.4 Natural Direct Effect Regularization

To improve the visual perception ability of the model, another option is to maximize the natural direct effect (NDE) rather than the total effect (TE). The natural direct effect is to measure the direct effect resulting from changes in the image. As shown in Formula 2, the first part is to calculate the likelihood of generating each token in the entity to intervene \tilde{S} at any position, from the image I and the preceding tokens $S_{<i}^*$ that have been generated from I^* . Whereas, the second part is the likelihood of generating \tilde{S} at any position from the counterfactual image I^* and preceding tokens $S_{<i}^*$ that are generated from I^* , which is the same as the second part of TE. Formally, we calculate NDE loss as:

$$\mathcal{L}_{\text{NDE}} = - \sum_{(I, S) \in \mathcal{D}} \sum_{j=1}^{L_{\tilde{S}}} \left[\frac{1}{L_{S^*}} \sum_{i=1}^{L_{S^*}} \left(\log f_{\theta}(\tilde{s}_j | I, S_{<i}^*) - \log f_{\theta}(\tilde{s}_j | I^*, S_{<i}^*) \right) \right]. \quad (6)$$

The first part here looks simpler than that in TE loss because both the first part and the second part in NDE loss average the probabilities over any position in S^* whose length is L_{S^*} .

Figure 3(b) presents an example. The first term is estimated by the probability of generating each token in “black poodle” at any position in the counterfactual caption “A man in a yellow shirt is stretching in a field.”, but with the factual image I as input. The second term is again the average probabilities of generating each token in “black poodle” at any position in the counterfactual caption with the counterfactual image I^* as input. By maximizing the NDE effect, the direct influence of the image is enhanced, thereby the model is more inclined to see the image and generate the correct next token, rather than to guess it.

Table 1: Evaluation results on counterfactual images with masks. CH_s (CHAIR_s), P@5 (Precision@5), and nDCG@5 are automatic measures for evaluating hallucination. Faith. (Faithfulness) and Overall denote results given by human judges. The best result is highlighted in bold, while the second best is underlined.

Methods	Flickr30k Entities					MSCOCO				
	CH _s ↓	P@5	nDCG@5	Faith.	Overall	CH _s ↓	P@5	nDCG@5	Faith.	Overall
ClipCap	20.45	80.08	79.97	0.320	0.447	64.05	36.02	36.01	0.687	0.713
+ObjL [4]	21.18	79.16	79.07	0.353	0.453	64.64	35.62	35.58	0.653	0.727
+ObjMLM [6]	25.37	75.07	74.97	0.140	0.313	70.07	29.89	29.91	0.533	0.547
+TE (ours)	<u>19.78</u>	<u>80.48</u>	<u>80.39</u>	<u>0.373</u>	<u>0.467</u>	<u>63.58</u>	<u>36.32</u>	<u>36.35</u>	<u>0.753</u>	<u>0.807</u>
+NDE (ours)	19.64	80.53	80.51	0.400	0.493	63.04	36.55	36.65	0.760	0.820
BLIP	12.14	88.00	88.00	0.740	0.793	33.70	66.17	66.22	1.167	1.200
+ObjL [4]	10.61	89.17	89.19	0.613	0.767	33.07	67.12	67.08	1.187	1.107
+ObjMLM [6]	<u>10.11</u>	<u>89.31</u>	<u>89.45</u>	<u>0.687</u>	<u>0.787</u>	<u>33.90</u>	<u>65.67</u>	<u>65.77</u>	<u>1.113</u>	<u>1.180</u>
+TE (ours)	<u>10.23</u>	<u>89.63</u>	<u>89.68</u>	<u>0.767</u>	<u>0.827</u>	<u>31.10</u>	<u>68.49</u>	<u>68.58</u>	<u>1.213</u>	<u>1.267</u>
+NDE (ours)	9.53	89.83	89.93	0.873	0.913	30.43	69.24	69.33	1.247	1.273
BLIP2	8.01	91.95	91.96	0.807	0.847	30.28	69.91	69.88	1.227	1.233
+ObjL [4]	8.02	91.90	91.96	0.847	<u>0.887</u>	30.26	70.19	70.13	1.133	0.947
+ObjMLM [6]	8.12	92.00	92.01	0.800	0.867	34.84	65.23	65.19	1.140	1.100
+TE (ours)	<u>7.61</u>	<u>92.09</u>	<u>92.14</u>	0.867	0.913	<u>29.60</u>	<u>70.54</u>	<u>70.49</u>	<u>1.340</u>	<u>1.273</u>
+NDE (ours)	7.51	92.21	92.24	<u>0.860</u>	0.880	29.26	70.70	70.68	1.353	1.280

5 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the capability of our model for alleviating object hallucination, reducing biases in training data, and interpreting the correspondence between captions and image regions.

5.1 Experiment Setup

5.1.1 Datasets. To evaluate the effectiveness of our model, we construct counterfactual images and captions as mentioned in Section 4.1. We choose Flickr30k Entities [26] and MSCOCO [17] as our datasets, which have high-quality image annotations for constructing masked counterfactual images.

Flickr30k Entities (Flickr) is built upon the existing Flickr30k dataset [38] that contains 31,783 images. The dataset provides 244k coreference chains and 276k manually annotated bounding boxes within the images. We use the original split of this dataset. Entities that occur more than once within a caption are removed to avoid confusion. After pre-processing, the final dataset consists of 29k/1k/1k samples for training/validation/test, respectively.

MSCOCO (COCO) consists of more than 328k images with annotated objects, phrases, and relationships. We adopt the Karpathy split of the MSCOCO dataset and coreference relationships in the annotations are utilized to establish correspondences between the image regions and phrases (entities) to intervene in the captions.

Both datasets are composed of diverse phrase categories, where the Flickr dataset covers over 1,000 categories, while the COCO dataset is more concentrated on 80 categories.

5.1.2 Backbones and Baselines. Our proposed counterfactual regularization losses are model-agnostic and can be applied to various models. We conduct experiments on three backbones: ClipCap [19], BLIP [13] and BLIP2 [12], which respectively serve as

representative models for decoder-only, encoder-decoder, and multimodal large language model architectures. In addition to the above three image captioning baselines, we compare our methods with another two methods that aim to alleviate the object hallucination in image captioning: (1) **ObjL** [4] utilizes object labels as training augmentation to diminish models’ object bias on hallucination; (2) **ObjMLM** [6] conducts a whole object mask to mitigate object hallucination in masked language modeling. More implementation details are in supplementary pages.

5.1.3 Evaluation Methodology. Compared to baselines, our methods are expected to significantly reduce object hallucination on counterfactual test sets while maintaining the generation ability on factual test sets (it is not trivial due to different distributions between training and test sets). We employ both automatic and human evaluation in our experiments for convincing conclusions.

Automatic Evaluation: We evaluate hallucination by using:

- **CHAIR_s** [28]: It measures whether models generate a masked phrase, *i.e.*, phrases whose corresponding regions have been masked in the counterfactual image:

$$\text{CHAIR}_s = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}, \quad (7)$$

where a lower CHAIR_s score indicates a reduced presence of hallucination or increased faithfulness.

- **Ranking-based Metrics:** We generate a set of five candidate captions with the highest probability of being generated for a given counterfactual image, among which the ones without the masked phrase are regarded as positive while those with the masked phrase are as negative. **Precision@5** and **nDCG@5** are employed¹ to assess the object hallucination in fine-grained.

¹<https://github.com/microsoft/rankerEval>

Table 2: Evaluation results on factual images without masks. BLEU-4, ROUGE-L, CIDEr are automatic measures for evaluating generation quality. Faith. (Faithfulness) and Overall denote content accuracy and overall caption quality given by human judges. The best result is highlighted in bold, while the second best is underlined.

Methods	Flickr30k Entities					MSCOCO				
	BLEU-4	ROUGE-L	CIDEr	Faith.	Overall	BLEU-4	ROUGE-L	CIDEr	Faith.	Overall
ClipCap	<u>23.38</u>	<u>48.33</u>	57.09	0.947	0.793	<u>28.79</u>	52.75	<u>126.92</u>	1.360	1.253
+ObjL [4]	23.58	48.17	58.13	0.927	0.787	26.67	49.40	115.15	1.373	1.173
+ObjMLM [6]	17.01	43.64	32.42	0.793	0.813	19.25	46.76	73.82	1.040	1.013
+TE (ours)	24.03	48.92	59.08	0.973	0.860	28.94	52.98	127.27	<u>1.413</u>	1.300
+NDE (ours)	23.32	48.05	<u>58.69</u>	<u>0.960</u>	<u>0.840</u>	28.77	<u>52.89</u>	125.70	1.433	<u>1.280</u>
BLIP	<u>37.14</u>	56.67	<u>95.40</u>	1.433	1.260	34.63	56.83	<u>153.39</u>	1.873	1.593
+ObjL [4]	36.93	56.37	92.72	1.420	1.147	33.00	56.66	148.00	1.840	1.300
+ObjMLM [6]	35.61	56.56	94.88	1.420	1.253	31.71	65.55	133.67	1.713	1.540
+TE (ours)	37.28	56.77	<u>95.40</u>	<u>1.447</u>	<u>1.300</u>	<u>34.65</u>	<u>56.82</u>	153.37	<u>1.900</u>	<u>1.620</u>
+NDE (ours)	37.00	<u>56.68</u>	95.45	1.473	1.307	34.66	56.83	153.73	1.920	1.640
BLIP2	<u>37.61</u>	<u>58.11</u>	<u>103.41</u>	1.473	1.373	34.72	58.13	154.17	1.880	1.553
+ObjL [4]	34.30	56.61	93.43	1.473	1.240	29.81	55.13	135.42	1.820	1.340
+ObjMLM [6]	36.04	56.94	94.76	1.480	1.220	34.06	57.39	151.51	1.800	1.453
+TE (ours)	37.64	58.24	103.68	<u>1.520</u>	<u>1.393</u>	<u>34.80</u>	<u>58.24</u>	<u>154.77</u>	<u>1.933</u>	1.647
+NDE (ours)	37.56	<u>58.11</u>	102.63	1.533	1.427	34.88	58.34	155.14	1.947	<u>1.613</u>

We adopt BLEU [21], ROUGE-L [16], and CIDEr [33] to measure the quality of generated captions on factual image test sets. We do not evaluate the generation quality of counterfactual images using automatic evaluation due to the lack of ground-truth captions. To compensate for this, the quality of captions generated for counterfactual images is analyzed by using human evaluation.

Human Evaluation: To verify whether the automatic measurements are consistent with human experiences, we further conduct a user study. First, we randomly sample 50 factual images from the Flickr and 50 from the COCO dataset. We then create counterfactual images for the 100 factual images and collect top-generated captions from all methods for both factual and counterfactual images. We conduct human evaluations on the 100 factual images and 100 counterfactual images. For each image, we shuffle the generated captions and make the methods anonymous when presented with an image to ensure a fair comparison. Three human assessors majored in English with the age range from 23 to 25, are hired to rate the captions on a 3-level Likert scale from 0 to 2 in two aspects:

- **Faithfulness** measures the degree to which a caption accurately represents the content of the image;

- **Overall** means the overall quality of a caption.

Finally, we calculate the Fleiss' Kappa among their assessments which results in 0.43, meaning a moderate level of agreement. We use their average values as the results.

5.2 Evaluation Results on Counterfactual Images

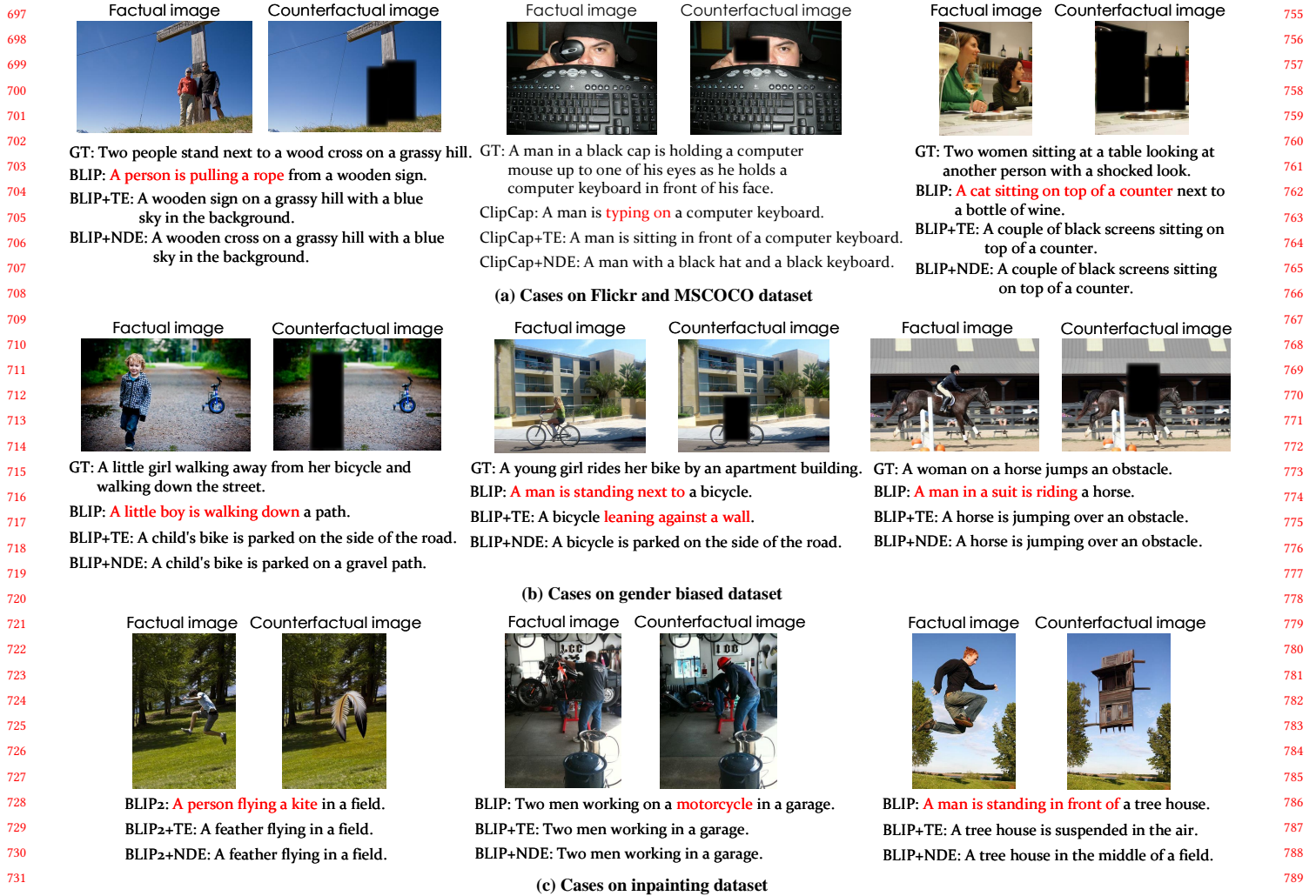
We first compare our proposed models with all baselines on counterfactual test sets in terms of both hallucination and overall generation quality. The results are shown in Table 1, where all methods with the same backbone are grouped for clarity.

Automatic evaluation. In terms of the automatic metrics of measuring object hallucination in Table 1 (CH_s, P@5, nDCG@5), our proposed counterfactually regularized methods consistently exhibit superior performance over all baselines on both datasets, demonstrating their effectiveness in mitigating object hallucination. Our NDE regularization performs better than the TE ones. This indicates that maximizing the direct effect of image content on the generated tokens helps build a more precise alignment between visual regions and their corresponding entity phrases. Baselines ObjL and ObjMLM do not always alleviate hallucination effectively, *e.g.*, they exhibit more hallucinations on the ClipCap backbone on the two datasets. In comparison, our methods that regularize the causal effect consistently reduce hallucination in terms of different backbones, datasets, and measures. This demonstrates the effectiveness of adopting a causal perspective when handling hallucinations. Further experiments confirm this will not be affected by different decoding strategies (please see Section A in supplementary pages).

Human evaluation. Human evaluation results in Table 1 show that our methods perform the best regarding both reducing hallucination (Faith.) and overall generation quality (Overall). Moreover, NDE wins more times than TE, and these two methods that we propose are consistently better than baselines ObjL and ObjMLM. Overall, the human evaluation results are consistent with the automatic evaluation results in terms of hallucination and additionally reveal the good generation quality of our methods.

5.3 Evaluation Results on Factual Images

We compare all methods on factual test images to investigate 1) whether our regularization methods compromise any generation capability and 2) whether our method can reduce hallucination on factual images. As shown in Table 2, our proposed methods



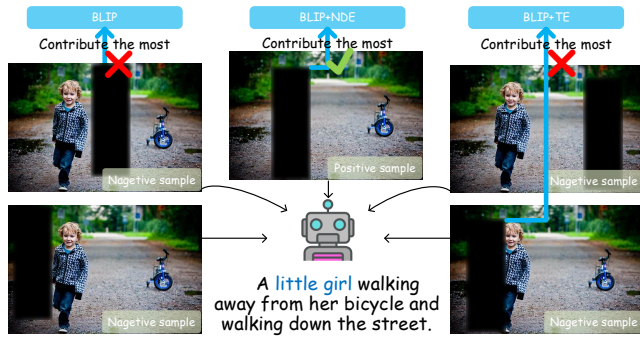


Figure 5: Illustration of the experiment design to evaluate interpretability.

recipe as the test set. Finally, we train a BLIP model and evaluate its performance on the biased test set.

When examining entities related to gender only, the error rates are presented in Table 3. The results indicate that BLIP+NDE model outperforms BLIP in terms of lower error rates for both factual and counterfactual images. The results imply that models with our proposed methods are more robust in handling biased datasets.

5.5 Quantitative Analysis

We present some examples of the generated captions on counterfactual images in Figure 4(a), and more in supplementary pages. Overall, our methods perform better in understanding counterfactual images, avoiding generating captions containing masked information. Instead, they describe what are indeed presented in the images, such as “a wooden cross” and “a couple of black screens”. Conversely, the baseline model without counterfactual regularization often guesses incorrectly. We also present some examples of generated captions for counterfactual images in the biased dataset in Figure 4(b). The baseline model often incorrectly guesses a “man” or “boy” behind the mask, whereas our models describe other objects that are present in the image, such as “a child’s bicycle” and “a horse”.

We further utilize a Latent Diffusion Model [29] to inpaint the masked region with a counterfactual object. As shown in Figure 4(c), an intriguing observation is that the baseline model occasionally hallucinates “a person” in the inpainted image, despite the absence of any human presence in the image. This may be caused by the shortcut connections it learned from the training data, where our methods can robustly avoid this and correctly describe “a feather” or “a tree house” that can be seen in the inpainted images.

5.6 Evaluation of Interpretability

In this experiment, we compare different image captioning models in terms of interpretability. The basic idea is that an interpretable model should generate a noun phrase by using its corresponding region. For example, when generating the phrase “little girl”, the region that contains a little girl should contribute the most to the model generation compared with other regions. The contribution of a region is measured by using an efficient and effective explanation method CXPlain [30]. More specially, for each noun phrase, we first identify its corresponding region (positive sample) and then

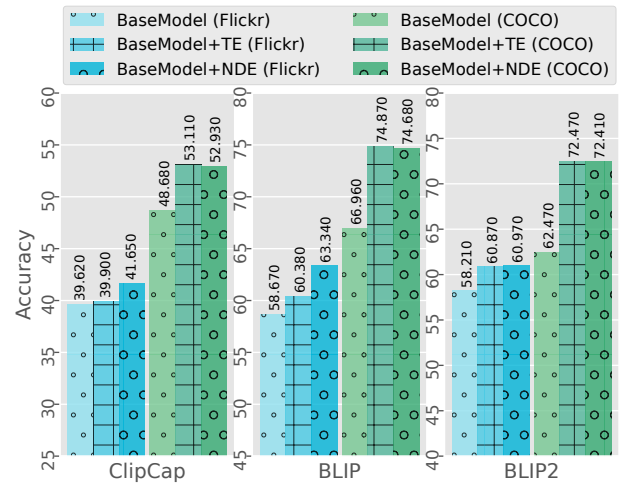


Figure 6: The interpretability performance of different models by identifying the correct masked counterfactual image.

randomly generate four incorrect regions with the same size (negative samples). Next, we rank the five regions by using CXPlain [30], which assigns a higher contribution score to a region if removing this region results in a larger change in the model’s loss. A model is considered interpretable if the positive sample has the highest contribution. Figure 5 shows an example, where the correct region (the region with a little girl) contributes the most in generating the phrase “little girl” for the Blip+NDE model, demonstrating the interpretability of the model. In comparison, for other models (Blip and Blip+TE), the correct region does not have the highest contribution to generating “little girl”.

To show the average results over our test set, we use accuracy as the measure. Accuracy refers to the percentage of cases in which the positive sample has the highest contribution. Figure 6 shows that the TE method consistently performs better than the backbone model without regularization. The NDE method outperforms the TE method across backbone models on Flickr by a significant margin, while they perform comparably on MSCOCO. This suggests that our proposed counterfactual regularization is effective in enhancing interpretability and our proposed NDE method is the most effective.

6 CONCLUSION

This paper proposes using counterfactual causal effects to model the relationship between vision and language. We employ two counterfactual regularization methods based on the concepts of total effect (TE) and natural direct effect (NDE) to improve image captioning models. Experimental results consistently show the superiority of our methods over baselines in terms of alleviating hallucination across different backbones and datasets. The NDE method performs the best in generating faithful captions for counterfactual images and accurately interpreting the most relevant image regions corresponding to a phrase in a caption. In the future, we plan to integrate the counterfactual regularization methods into more complicated multimodal generation scenarios with both image and text as input, such as visual question answering and multimodal dialogue.

REFERENCES

- [1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10044–10054.
- [2] Alexander Balke and Judea Pearl. 1995. Counterfactuals and Policy Analysis in Structural Models. In *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995*, Philippe Besnard and Steve Hanks (Eds.). Morgan Kaufmann, 11–18. https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=4144&proceeding_id=11
- [3] Manuele Barraco, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. With a Little Help from Your Own Past: Prototypical Memory Networks for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3021–3031.
- [4] Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1381–1390.
- [5] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10578–10587.
- [6] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, 2128–2140. <https://doi.org/10.18653/v1/2023.eacl-main.156>
- [7] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
- [8] Yash Goyal, Ziyang Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*. PMLR, 2376–2384.
- [9] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating Counterfactual Explanations with Natural Language. *CoRR* abs/1806.09809 (2018). arXiv:1806.09809 <http://arxiv.org/abs/1806.09809>
- [10] Yusuke Hirota, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, Ittetsu Taniguchi, and Takao Onoye. 2021. Visual Question Answering with Textual Representations for Images. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*. IEEE, 3147–3150. <https://doi.org/10.1109/ICCVW54120.2021.00353>
- [11] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8928–8937.
- [12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICLR 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19730–19742. <https://proceedings.mlr.press/v202/li23q.html>
- [13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.
- [14] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 121–137.
- [15] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. 3285–3292.
- [16] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.
- [18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *CoRR* abs/2304.08485 (2023). <https://doi.org/10.48550/ARXIV.2304.08485> arXiv:2304.08485
- [19] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. *CoRR* abs/2111.09734 (2021). arXiv:2111.09734 <https://arxiv.org/abs/2111.09734>
- [20] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 12700–12710. <https://doi.org/10.1109/CVPR46437.2021.01251>
- [21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [22] Judea Pearl. 2001. Direct and Indirect Effects. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, Jack S. Breese and Daphne Koller (Eds.). Morgan Kaufmann, 411–420. https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=126&proceeding_id=17
- [23] Judea Pearl. 2010. Causal inference. *Causality: objectives and assessment* (2010), 39–58.
- [24] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press* 19, 2 (2000).
- [25] Suzanne Petryk, Spencer Whitehead, Joseph E. Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2024. Simple Token-Level Confidence Improves Caption Correctness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 5742–5752.
- [26] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [28] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4035–4045. <https://doi.org/10.18653/v1/d18-1437>
- [29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. <https://doi.org/10.1109/CVPR52688.2022.01042>
- [30] Patrick Schwab and Walter Karlen. 2019. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 10220–10230. <https://proceedings.neurips.cc/paper/2019/hash/3ab6b46e1d6b21d59a3c3a0b9d0f6ef-Abstract.html>
- [31] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiamenti, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 539–559.
- [32] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 580–599.
- [33] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.
- [34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.
- [35] Xinyi Wang, Wenhui Chen, Michael Saxon, and William Yang Wang. 2021. Counterfactual Maximum Likelihood Estimation for Training Deep Networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 25072–25085. <https://proceedings.neurips.cc/paper/2021/hash/d30d0f522a86b3665d8e3a9a91472e28-Abstract.html>
- [36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.
- [37] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*. 684–699.
- [38] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

1045	[39] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. 2021. Counterfactual Zero-Shot and Open-Set Visual Recognition. In <i>IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021</i> . Computer Vision Foundation / IEEE, 15404–15414. https://doi.org/10.1109/CVPR46437.2021.01515	1103
1046		
1047		
1048		
1049	[40] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. 2021. Consensus Graph Representation Learning for Better Grounded Image Captioning. In <i>Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021</i> . AAAI Press, 3394–3402.	1104
1050		
1051		
1052		
1053		
1054		
1055		
1056		
1057		
1058		
1059		
1060		
1061		
1062		
1063		
1064		
1065		
1066		
1067		
1068		
1069		
1070		
1071		
1072		
1073		
1074		
1075		
1076		
1077		
1078		
1079		
1080		
1081		
1082		
1083		
1084		
1085		
1086		
1087		
1088		
1089		
1090		
1091		
1092		
1093		
1094		
1095		
1096		
1097		
1098		
1099		
1100		
1101		
1102		
	[41] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. <i>Advances in Neural Information Processing Systems</i> 33 (2020), 18123–18134.	1105
	[42] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In <i>Proceedings of the AAAI conference on artificial intelligence</i> , Vol. 34. 13041–13049.	1106
	[43] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. <i>CoRR</i> abs/2304.10592 (2023). https://doi.org/10.48550/ARXIV.2304.10592 arXiv:2304.10592	1107
		1108
		1109
		1110
		1111
		1112
		1113
		1114
		1115
		1116
		1117
		1118
		1119
		1120
		1121
		1122
		1123
		1124
		1125
		1126
		1127
		1128
		1129
		1130
		1131
		1132
		1133
		1134
		1135
		1136
		1137
		1138
		1139
		1140
		1141
		1142
		1143
		1144
		1145
		1146
		1147
		1148
		1149
		1150
		1151
		1152
		1153
		1154
		1155
		1156
		1157
		1158
		1159
		1160