

# Causality-aligned Prompt Learning via Diffusion-based Counterfactual Generation

Xinshu Li

xinshu.li@unsw.edu.au

The University of New South Wales  
Sydney, Australia

Ruoyu Wang

ruoyu.wang5@unsw.edu.au

The University of New South Wales  
Sydney, Australia

Erdun Gao

erdun.gao@adelaide.edu.au  
The University of Adelaide  
Adelaide, Australia

Mingming Gong

mingming.gong@unimelb.edu.au

The University of Melbourne

Melbourne, Australia

MBZUAI

Abu Dhabi, United Arab Emirates

Lina Yao

lina.yao@data61.csiro.au

CSIRO's Data 61

The University of New South Wales  
Sydney, Australia

## Abstract

Prompt learning has garnered attention for its efficiency over traditional model training and fine-tuning. However, existing methods, constrained by inadequate theoretical foundations, encounter difficulties in achieving causally invariant prompts, ultimately falling short of capturing robust features that generalize effectively across categories. To address these challenges, we introduce the *DiCap* model, a theoretically grounded Diffusion-based Counterfactual prompt learning framework, which leverages a diffusion process to iteratively sample gradients from the marginal and conditional distributions of the causal model, guiding the generation of counterfactuals that satisfy the minimal sufficiency criterion. Grounded in rigorous theoretical derivations, this approach guarantees the identifiability of counterfactual outcomes while imposing strict bounds on estimation errors. We further employ a contrastive learning framework that leverages the generated counterfactuals, thereby enabling the refined extraction of prompts that are precisely aligned with the causal features of the data. Extensive experimental results demonstrate that our method performs excellently across tasks such as image classification, image-text retrieval, and visual question answering, with particularly strong advantages in unseen categories.

## CCS Concepts

- Computing methodologies → Computer vision tasks.

## Keywords

Vision Language Models, Prompt Learning, Diffusion Process, Counterfactual Generation

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or 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](mailto:permissions@acm.org).

*Conference acronym 'XX, Woodstock, NY*

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-XXXX-X/2018/06  
<https://doi.org/XXXXXXX.XXXXXXX>

## ACM Reference Format:

Xinshu Li, Ruoyu Wang, Erdun Gao, Mingming Gong, and Lina Yao. 2018. Causality-aligned Prompt Learning via Diffusion-based Counterfactual Generation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 Introduction

Pre-trained vision-language foundation models [28, 29, 48] integrate large-scale multimodal information, effectively breaking down boundaries between different domains and significantly broadening their applications. To facilitate knowledge transfer from these models to downstream tasks, prompt engineering [53, 56] employs optimized templates that guide the model in generating more relevant and accurate outputs. However, the manual creation of prompts is time-consuming and labor-intensive, leading to the emergence of prompt learning [23, 63, 64], which allows for the automatic acquisition of tunable prompts and improves the efficiency of transferring knowledge to downstream tasks.

Current prompt learning methods often suffer from capturing spurious correlations between variables, especially in the absence of additional constraints [37, 38], which can lead to degraded performance under distribution shifts. To mitigate this issue, causal prompt learning methods [10, 15, 30, 40] have been proposed to identify and leverage underlying causal relationships, thereby improving the robustness and generalization of prompts. As shown in Figure 1, prompt embeddings should focus on the causal features of an image, such as the camel's hump, rather than the non-causal factors like the frequently co-occurring desert background or yurts. However, an important challenge remains: how can prompt learning effectively distinguish causal features from spurious correlations in complex visual data?

Counterfactual learning [12, 21, 32–34, 55] offers a promising approach to this challenge, which involves answering the question of how the outcomes of an event would change if its underlying causal factors were modified [45]. Specifically, in the case of Figure 1, the counterfactual speculation might be: “What would the image look like if a cow were standing where the camel originally stood?” By generating counterfactual images as negative samples and aligning the original image as a positive sample, contrastive



**Figure 1: A motivating example: In camel classification tasks, robust prompt embeddings should align with the camel’s causal features, such as physical traits, rather than non-causal features, like desert backgrounds or yurts. Counterfactual learning poses questions like “What if a cow stood here?” to generate counterfactual images, which are used in contrastive learning framework to guide prompts in identifying and aligning with causal features, offering a practical approach to causal-invariant prompt embeddings.**

learning encourages prompt embeddings to capture consistent features across variations in non-causal factors, guiding them to focus on essential causal features for robust visual understanding.

Nonetheless, generating counterfactual samples remains a significant challenge, particularly for high-dimensional image data [44, 59]. Current efforts primarily rely on multimodal information [15, 30, 31], such as identifying semantic similarities between related image prompts or leveraging knowledge graphs to construct counterfactuals. However, these approaches face critical limitations: 1) Their success hinges on sufficient multimodal information, which is often scarce and poses a significant bottleneck. 2) The lack of a robust theoretical foundation to ensure counterfactual identifiability and constrain estimation errors undermines their reliability. 3) The use of low-dimensional vectors to represent counterfactual images limits the model’s ability to capture shared high-dimensional non-causal features and distinguish causal ones. Therefore, a key challenge lies in developing methods that generate high-dimensional counterfactual samples with guaranteed error bounds while reducing dependence on multimodal information to enhance the robustness of prompt learning.

In this work, we leverage recent advancements in diffusion models [9, 16, 51, 54] to design methods for counterfactual prompt learning. Diffusion models offer several key advantages for this task: 1) Their isometric diffusion process preserves high-dimensional image features, minimizing information loss and enabling the possibility of uni-modal reliance; 2) The iterative sampling mechanism [16] naturally extends to interventions on key causal features [6], facilitating the minimal sufficiency of the generated counterfactual images; 3) Diffusion models are underpinned by rigorous mathematical guarantees, ensuring the reliability of counterfactual generation.

Specifically, we propose a novel **Diffusion-based Counterfactual prompt (*DiCap*)** learning framework to align prompts with causal invariant features. First, we invert the diffusion process via denoising [16], approximating the gradient of the log-likelihood of the input distribution. In parallel, we leverage anti-causal predictors [52], which infer causes from observed effects by reversing the natural cause-effect flow. Subsequently, causal interventions are

performed using gradients derived from these predictors, enabling minimal sufficiency of the generated counterfactuals via Markov Chain Monte Carlo sampling. Furthermore, we provide rigorous theoretical derivations to establish the error bounds of counterfactual estimation, ensuring its reliability. Finally, we design a contrastive learning task [7, 25] to train tunable prompt embeddings, enabling them to attract factual samples while repelling counterfactual ones, effectively aligning with the causal features in the images.

Our main contributions are summarized as follows:

- We propose a novel, theoretically grounded **Diffusion-based Counterfactual prompt (*DiCap*)** learning framework that generates robust prompt representations aligned with causal features in images, significantly improving model performance on unseen distributions.
- Through rigorous theoretical derivations, we established the sufficient conditions that guarantee the identifiability of the generated counterfactual samples and bound the counterfactual estimation errors.
- Extensive experimental results on image classification, image-text retrieval, and visual question answering (VQA) consistently confirm the superiority, stability, and generalizability of our approach over existing baselines.

## 2 Related Work

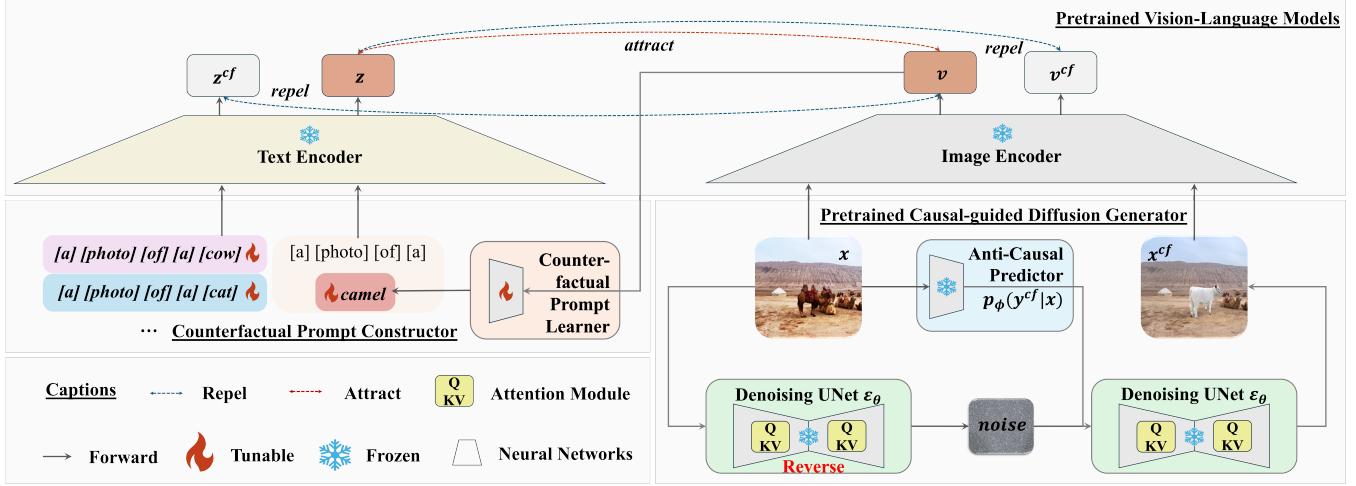
### 2.1 Causal Prompt Learning

Prompt learning has evolved from static prompts in CLIP [48] to learnable soft prompts in CoOp [64], improving task-specific adaptation but struggling with unseen classes due to distribution shifts. CoCoOp [63] mitigates this by conditioning prompts on image features, enhancing generalization further. Recent advancements highlight the potential of causal learning in refining prompt robustness. Li et al. [35] leverages causal contrastive learning to generate instance-dependent soft prefixes for multi-class classification, addressing label ambiguity. Zhang et al. [61] introduces causal prompting with front-door adjustment to reduce biases in large language models. Li et al. [30] presents CPKP, which improves semantic richness and generalization through ontological knowledge graphs. Lyu et al. [40] explores causal relations in sentiment classification using three causal prompts. Li et al. [31] combines counterfactual generation and learnable prompts for better radiology report generation. He et al. [15] proposes a method that constructs counterfactuals by identifying semantically similar images.

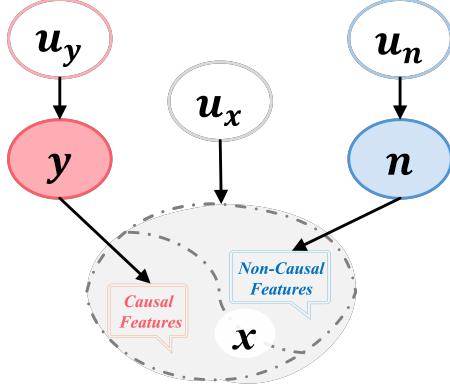
Despite these advancements, existing methods lack theoretical guarantees for counterfactual generation and fail to fully integrate causal reasoning into prompt learning. To bridge this gap, we propose DiCap, a causal diffusion model for prompt learning that establishes theoretical error bounds in counterfactual generation, improving both effectiveness and generalization to unseen classes.

### 2.2 Counterfactual Generation

By constructing hypothetical scenarios with minimal causal modifications, counterfactual generation serves as a powerful tool for understanding causal mechanisms and improving model robustness. Existing counterfactual generation approaches fall into four categories: 1) Multimodal-guided methods: Kim et al. [26], Prabhu et al. [47] leverage language models or textual concepts to reduce



**Figure 2: DiCap Method Overview:** The process begins by inputting the factual image  $x$  into a pretrained denoising model, generating noise as a proxy for exogenous variables. The image is also passed through an anti-causal predictor to obtain the counterfactual label gradient  $\nabla_x p_\phi(y^cf|x)$ , which guides the denoising model to generate the counterfactual image  $x^cf$ . Each factual image is paired with a unique trainable dynamic prompt vector that captures causal features using a dual-contrastive loss. In addition to the standard contrastive loss, a “counterfactual as hardest” strategy aligns the dynamic prompt embedding with the factual image while distancing it from the counterfactual. Fine-tuning is applied only to the trainable counterfactual prompt learner, with the remaining network components frozen for computational efficiency.



**Figure 3: The Data Generation Process of the images:** The features of the image  $x$  can be divided into causal and non-causal components. The label of the image,  $y$ , serves as the cause of the causal features, while the remaining factors,  $n$ , cause the non-causal features. Additionally,  $u_x$ ,  $u_y$ , and  $u_n$  represent the exogenous variables which are the causes of these endogenous variables.

annotation costs but risk semantic misalignment; 2) Object-centric methods: Jacob et al. [18], Jeanneret et al. [20], Zemni et al. [60] enable fine-grained edits via scene priors but require external annotations and exhibit limited generalization; 3) Adversarial attack derivatives: Khorram and Fuxin [24] convert perturbations into semantic edits but prioritize robustness over causal interpretation. 4) Diffusion-based methods: Diffusion models’ iterative sampling process aligns more naturally with causal interventions compared

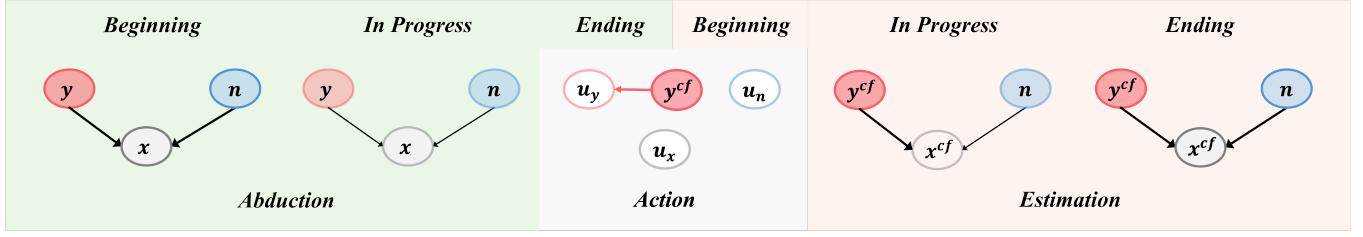
to other generative models. Early works [3, 4, 19, 57] focused on generating high-fidelity samples using gradient guidance or regularization but were primarily designed for image editing, lacking theoretical guarantees for causal consistency. To incorporate counterfactual reasoning, Sanchez and Tsaftaris [51] introduced classifier-guided diffusion models [9], leveraging external classifiers to direct sample generation toward specific counterfactual outcomes. Alternatively, Chao et al. [6] explored classifier-free models [17], adopting a different approach to counterfactual sample generation without explicit classifier constraints. While these methods advance generation quality and scenario adaptability, they fail to address causal identifiability and cross-task generalization.

Our DiCap framework bridges these gaps by integrating counterfactual generation with prompt learning. Grounded in structural causal models, DiCap enables minimal sufficiency of the generated counterfactuals via a denoising process and extracts causally invariant prompts through contrastive learning. By enforcing error-bounded theoretical rigor and enabling multimodal task compatibility, DiCap surpasses existing methods reliant on heuristic designs or domain-specific priors.

### 3 Methodology

#### 3.1 Method Overview

As shown in Figure 2, our DiCap model consists of the following key components: First, denoising sampling is performed using anti-causal gradient guidance from a pretrained diffusion model to generate counterfactual images, which are then used as the hardest negative samples in a dual-contrastive task to guide the generation of prompt embeddings that align with causal features. The following sections detail our framework and its theoretical foundations.



**Figure 4: Diffusion-Based Counterfactual Generation:** During diffusion, noise is progressively added to the factual image  $x$ , transforming endogenous variables governed by causal dependencies into independent exogenous variables (*Abduction*). As noise increases, causal relationships weaken, represented by thinner arrows. By replacing the exogenous variable  $u_y$  which is the cause of the factual label  $y$  with the counterfactual label  $y^{cf}$  at each timestep (*Action*), the denoising process, guided by the anti-causal gradient  $\nabla_x p_\phi(y^{cf}|x)$ , generates counterfactual images  $x^{cf}$  that minimize causal dependencies while enhancing non-causal features  $n$  (*Estimation*).

### 3.2 Data Generation Process

Figure 3 illustrates the general data generation process [65] for the images. The features of the image  $x$  can be classified into causal and non-causal features. The label of the image,  $y$ , causes the causal features, while  $n$  represents other factors leading to non-causal features. Exogenous variables  $u_x$ ,  $u_y$ , and  $u_n$  represent the unobserved causes of  $x$ ,  $y$ , and  $n$ , respectively.

### 3.3 Diffusion-Based Counterfactual Generation

Following Sanchez and Tsaftaris [51], we generate counterfactuals via a classifier-guided diffusion model [9]. However, while Sanchez and Tsaftaris [51] interprets the model through a bivariate framework of images and labels, we adopt a more comprehensive perspective by incorporating non-causal factors into the reasoning process, refining the overall modeling of counterfactual variations.

Counterfactual reasoning involves abduction, action, and estimation. As shown in Figure 4, we show how the diffusion model aligns with these stages to effectively generate counterfactuals, highlighting its unique advantages in counterfactual inference.

**Abduction:** The model identifies exogenous variables associated with nodes in the causal graph, such as  $u_x$ ,  $u_y$ , and  $u_n$ , as shown in Figure 3. This parallels the forward mechanism in diffusion models, where causal dependencies weaken, leading to independent exogenous variables [51]. We use a forward implicit diffusion process [54], reformulating the DDIM process as an ordinary differential equation (ODE) and applying the Euler approximation to derive independent exogenous factors.

$$\begin{aligned} \mathbf{x}_{t+1} \leftarrow & \sqrt{\bar{\alpha}_{t+1}} \left( \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right) \\ & + \sqrt{1 - \bar{\alpha}_{t+1}} \epsilon_\theta(\mathbf{x}_t, t), \quad t = 0, \dots, T-1, \end{aligned} \quad (1)$$

where  $\mathbf{x}_0 = \mathbf{x}$ ,  $\epsilon_\theta$  is a pretrained model to estimate the noise introduced at each time-step  $t$ ,  $\bar{\alpha}_t := \prod_{j=0}^t (1 - \beta_j)$  means a time-dependent variance.

In contrast to the forward process during the pretraining of the diffusion, where independent Gaussian noise is directly added to the image, we input the  $\mathbf{x}_t$  into  $\epsilon_\theta$  to generate noise that aligns with most of the original features in each timestep, which is employed in both the forward noise addition and reverse denoising processes

for counterfactual generation. This structured noise reduces interference from random noise on non-target attributes, ensuring the counterfactuals closely align with the factuals in non-causal features. However, this noise inevitably retains certain causal features related to the label, which will be addressed and adjusted in the **Action** phase.

**Action and Estimation:** In the Langevin dynamics framework, the diffusion process is controlled by gradient guidance, which is particularly effective for counterfactual generation [54]. Following Sanchez and Tsaftaris [51], we guide the sampling using the gradient of the anti-causal predictor  $\nabla_x p_\phi(y^{cf}|x)$ , where  $y^{cf}$  is the counterfactual label. This gradient directs the process at each timestamp, ensuring counterfactual images  $x^{cf}$  reflect the causal effects of label changes as shown in the sampling equation below.

$$\begin{aligned} \epsilon &\leftarrow \epsilon_\theta(\mathbf{x}_t, t) - s\sqrt{1 - \bar{\alpha}_t} \nabla_{\mathbf{x}_t} \log p_\phi(y^{cf} | \mathbf{x}_t) \\ \mathbf{x}_{t-1} &\leftarrow \sqrt{\bar{\alpha}_{t-1}} \left( \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon, \quad (2) \\ t &= T, \dots, 1, \end{aligned}$$

where  $s$  is a hyperparameter [17] controlling the scale of anti-causal gradients. In Section 4.3.2,  $s$  is analyzed to reveal the impact of counterfactual image quality on prompt learning performance.

Diffusion-Based Counterfactual Generation offers the significant advantage of enabling the minimal sufficiency of the generated counterfactual images, which are defined as follows:

**Definition 3.1. Minimal sufficiency** of counterfactual images mandates that the counterfactual  $x^{cf}$  is produced by the smallest perturbation to the factual image  $x$  that alters its label from  $y$  to  $y^{cf}$ . Formally,

$$x^{cf} = \arg \min_{\mathbf{x}'} d(\mathbf{x}, \mathbf{x}') \text{ subject to } k(\mathbf{x}') = y^{cf} \text{ and } y^{cf} \neq y, \quad (3)$$

where  $d(\cdot, \cdot)$  is a distance metric and  $k$  is a mapping function:  $\mathcal{X} \rightarrow \mathcal{Y}$ .

**Remark:** Reflecting on our counterfactual generation mechanism in Figure 4, we identify three key aspects that ensure the minimal sufficiency of counterfactual images. 1) Forward noise injection weakens causal dependencies, generating independent noise and plausible latent exogenous variables  $u_x$ ,  $u_n$ , and  $u_y$ . 2) The

anti-causal predictor gradient  $\nabla_x p_\phi(y^{cf} | x)$  guides the generation process, aligning the image with the counterfactual class  $y^{cf}$ . 3) Noise derived from the pretrained diffusion model  $\epsilon_\theta$  governs both forward perturbation and reverse denoising, preserving the relationship between non-causal features and the counterfactual image  $x^{cf}$ . Together, these mechanisms ensure **minimal yet sufficient** changes during the transition.

### 3.4 Bounding Counterfactual Error

We begin by articulating the necessary conditions for recovering the latent exogenous variable via an invertible function, as shown in Theorem 3.2. These conditions are derived through rigorous mathematical reasoning within the structural causal models (SCM) [45] discussed in this paper.

**THEOREM 3.2.** Suppose that for  $x \in \mathcal{X} \subset \mathbb{R}^m$  and continuous exogenous noise  $\mathbf{u}_x \sim \text{Unif}[0, 1]^m$  with  $m \geq 3$ ,  $x$  is governed by the structural equation:

$$x = f(y, \mathbf{n}, \mathbf{u}_x), \quad (4)$$

where  $y \in \mathcal{Y} \subset \mathbb{R}$ ,  $\mathbf{n} \in \mathcal{N} \subset \mathbb{R}^m$ , and  $\mathbf{u}_x \perp y, \mathbf{n}$ . Consider a forward function  $g : \mathcal{X} \rightarrow \mathcal{L}$  and a reverse function  $h : \mathcal{L} \times \mathcal{N} \times \mathcal{Y} \rightarrow \mathcal{X}$ , such that:

$$l := g(x), \quad \hat{x} := h(l, y, \mathbf{n}). \quad (5)$$

Assume the following conditions hold:

- (1) The structural function  $f$  is invertible and differentiable with respect to  $\mathbf{u}_x$ , and its Jacobian  $J_{f_{y,n}}$  is positive definite for all  $y \in \mathcal{Y}$  and  $\mathbf{n} \in \mathcal{N}$ .
- (2) The function  $g$  is invertible and differentiable with respect to  $x$ .
- (3) The recovered latent variable is independent of the parent variables,  $g(x) \perp y, \mathbf{n}$ .
- (4) The transformation  $q_{y,n}(\mathbf{u}_x) := g(f(y, \mathbf{n}, \mathbf{u}_x))$  satisfies  $J_{q_{y,n}|q_{y,n}^{-1}(1)} = c(y, \mathbf{n})A$ , for all  $1 \in \mathcal{L}$  and  $y \in \mathcal{Y}$ ,  $\mathbf{n} \in \mathcal{N}$ , where  $c$  is a scalar function and  $A$  is an orthogonal matrix.

Then,  $g(f(y, \mathbf{n}, \mathbf{u}_x)) = \tilde{q}(\mathbf{u}_x)$  for some invertible function  $\tilde{q}$ .

The first condition in **Theo. 3.2** ensures **identifiability** [39, 41] of causal estimation under additive noise models [14] and post-nonlinear models [62], from which corollaries are derived to identify settings for precise counterfactual estimation and establish error bounds.

**COROLLARY 3.3.** Suppose the function pairs  $(g, h)$  that satisfy  $h(g(x), y, \mathbf{n}) = x$  where  $x := f(y, \mathbf{n}, \mathbf{u}_x)$  and a do-operation  $\text{do}(\mathbf{y} := y^{cf})$ , then counterfactual estimation  $h(g(x), y^{cf}, \mathbf{n})$  will recover true counterfactual outcome  $x^{cf} := f(y^{cf}, \mathbf{n}, \mathbf{u}_x)$  under the conditions in **Theo. 3.2**.

**Corol. 3.3** implies that the counterfactual predictions will align with the true counterfactual outcomes if no information is lost in the forward and reverse steps. Building on this, we further derive the following error bound for classifier-guided diffusion-based counterfactual generation.

**COROLLARY 3.4.** Consider the function pairs  $(g, h)$ , if the reconstruction error is less than  $\delta$ , i.e.,  $d(h(g(x), y, \mathbf{n}), x) \leq \delta$ , then the error between counterfactual estimation  $h(g(x), y^{cf}, \mathbf{n})$  and true

counterfactual outcome  $x^{cf} := f(y^{cf}, \mathbf{n}, \mathbf{u}_x)$  will be at most  $\delta$ , i.e.,  $d(h(g(x), y^{cf}, \mathbf{n}), x^{cf}) \leq \delta$ , under the conditions in **Theo. 3.2**.

**Corol. 3.4** indicates that if the similarity between the reconstructed and factual images is high, the model can yield credible counterfactual samples with substantial precision. **Theo. 3.2**, **Corol. 3.3** and **Corol. 3.4** provide the theoretical guarantees for the counterfactual image generation method based on classifier-guided diffusion models.

### 3.5 Counterfactual Prompt Construction

**Closest as Counterfactual:** We then discuss how to select counterfactual labels to generate images that enhance prompt learning by aligning prompt embeddings with the causal features in the images rather than redundant details. Following Robinson et al. [50], harder-to-distinguish negative samples are more effective as they encourage contrastive learning to capture subtle class distinctions. Therefore, we select the class with the second closest predicted probability to the factual label  $y$  as the counterfactual label  $y^{cf}$ . For instance, for an image labeled as a cat, when input into the anti-causal predictor, the probability of predicting it as a cat should be closest to that of another feline, such as “tiger”, rather than an unrelated animal like “dog”. Thus, the gradients related to “tiger” is used to synthesize the counterfactual image, which subsequently functions as the most challenging negative sample in the contrastive learning task. Experimental results in Section 4.3.4 demonstrate the effectiveness of this sampling strategy.

**Counterfactual as Hardest:** Our prompt learning method is built upon dual contrastive learning tasks. The first adopts the vision-language contrastive paradigm from Radford et al. [48], implementing the Conditional Context Optimization loss (Eq. (6)) where image features anchor a similarity space that attracts their paired text prompts while repelling counterfactual class prompts.

$$\mathcal{L}_{\text{basic}} = \sum_{i=1}^m -\log \frac{\exp(\mathbf{v}^i \cdot g(\mathbf{w}^{y^i}(\mathbf{v}^i))/\tau)}{\sum_{j=1}^m \exp(\mathbf{v}^i \cdot g(\mathbf{w}^{y^j}(\mathbf{v}^i))/\tau)}, \quad (6)$$

where  $\mathbf{v}$  denotes the image embeddings learned by the image encoder of pretrained vision language model (e.g., CLIP).  $i, m$  and  $\tau$  indicate  $i_{th}$  image in the training dataset, training data size and temperature parameter.  $g(\cdot)$  denotes the text encoder and  $\mathbf{w}$  represents the tunable prompt embeddings, which are learned by the Counterfactual Prompt Constructor.

The second contrastive task uses generated counterfactuals: prompt embeddings serve as anchors, factual image embeddings as positive samples, and counterfactual embeddings as negative samples. The objective function is:

$$\begin{aligned} \mathcal{L}_{cf} = & \sum_{i=1}^m -\log \exp(\mathbf{v}^i \cdot g(\mathbf{w}^{y^i}(\mathbf{v}^i))/\tau) \\ & + \log(\exp(\mathbf{v}^i \cdot g(\mathbf{w}^{y^i}(\mathbf{v}^i))/\tau) + \exp((\mathbf{v}^{cf})^i \cdot g(\mathbf{w}^{y^i}(\mathbf{v}^i))/\tau)), \end{aligned} \quad (7)$$

where  $(\mathbf{v}^{cf})^i$  refers to the embedding of counterfactual image related to factual image  $\mathbf{v}^i$ .

In summary, the overall objective function for the **Diffusion-based Counterfactual prompt learning (DiCap)** method is formulated as follows.

$$\mathcal{L}_{total} = \mathcal{L}_{basic} + \lambda \mathcal{L}_{cf}, \quad (8)$$

where  $\lambda$  is an adjustable hyperparameter. Pseudo-code of **Dicap** is provided in the Algorithm 1.

---

**Algorithm 1** Diffusion-based Counterfactual Prompt Learning

---

**Input:** Training data size  $m$ , Training dataset  $\mathcal{D} = \{(v^i, y^i) | i \in 1, 2, \dots, m\}$ , Counterfactual dataset  $\mathcal{D}^{cf} = \{(v^i, (v^{cf})^i) | i \in 1, 2, \dots, m\}$ , CLIP pretrained text encoder  $g(\cdot)$ , tunable prompt embedding  $w$ , temperature  $\tau$ , loss-re-weight hyper-parameter  $\lambda$ .  
**Output:** Prompt embedding  $w$  after trained.  
Initialize  $w$   
**for**  $i = 0$  **to**  $m - 1$  **do**  
  # Calculate basic loss:  
   $\mathcal{L}_{basic+} = -\log \frac{\exp(v^i \cdot g(w^{y^i}(v^i)) / \tau)}{\sum_{j=1}^m \exp(v^i \cdot g(w^{y^j}(v^i)) / \tau)}$   
  # Calculate counterfactual loss:  
   $\mathcal{L}_{cf+} = \sum_{i=1}^m -\log \exp(v^i \cdot g(w^{y^i}(v^i)) / \tau) + \log(\exp(v^i \cdot g(w^{y^i}(v^i)) / \tau) + \exp((v^{cf})^i \cdot g(w^{y^i}(v^i)) / \tau))$   
  # Calculate total loss:  
   $\mathcal{L}_{total} = \mathcal{L}_{basic+} + \lambda \mathcal{L}_{cf+}$   
  Update  $w$  according to  $\mathcal{L}_{total}$   
**end for**  
return  $w$

---

## 4 Experiments

### 4.1 Task and Datasets

**Image Classification:** Seven publicly available image classification datasets are used: Caltech101 [13], OxfordPets [43], Flowers102 [42], Food101 [5], StanfordCars [27], SUN397 [58] and ImageNet [8]. The datasets are split into seen and unseen classes for generalisation assessment, with training conducted on the seen classes. Following the CLIP few-shot protocol, 16 training examples per class are used, and the full test set is employed for evaluation. The performance of the task is measured using the accuracy score.

**Image-Text Retrieval:** We evaluate image-text retrieval on MSCOCO [36] and Flickr30K [46] using the Karpathy split [22]: MSCOCO has 113K/5K/5K train/val/test images, and Flickr30K has 29K/1K/1K. For few-shot evaluation, training subsets use 0.5%, 1%, and 1.5% of the data, with performance assessed on the whole test set. Results are measured using Recall at 1 (R@1).

**Visual Question Answering:** VQAv2 [11], an extension of the VQA dataset [2], includes questions categorized as Number, Yes/No, and Other. Following Anderson et al. [1], VQA is treated as a classification task, where the model selects the correct answer from predefined options. Questions are converted into masked templates using a pre-trained T5 model [49], forming prompts that link questions and answers. The model predicts whether the prompt-image pair matches. For few-shot evaluation, 0.1%, 0.2%,

and 0.5% of instances are used for training. The performance of the task is measured using the accuracy score.

### 4.2 Baselines

We compare **DiCap** with: 1) the Zero-shot CLIP model [48], 2) classical prompt learning methods CoOp and CoCoOp [63, 64], and 3) the state-of-the-art causal prompt learning method CPL [15]. All methods use CLIP as the pretrained vision-language model. The results for learning-based models are reported as the average over five independent runs to avoid randomness.

### 4.3 Results

**4.3.1 Accuracy Comparison.** This section presents the main results. Performance is measured relative to the zero-shot CLIP, with green representing the percentage of performance exceeding the CLIP baseline and red indicating the percentage falling below it.

**Image Classification:** From the experimental results in Table 1, we have the following observations: 1) All prompt learning methods outperform zero-shot learning on seen classes; however, methods without causal learning, such as CoOp and CoCoOp, exhibit performance degradation on unseen classes. This suggests that these methods may have learned spurious correlations, underscoring the necessity of learning prompt representations with causal invariance. 2) Compared to CPL, another causality-based prompt learning method, our algorithm outperforms it across almost all seen classes and all unseen classes, demonstrating more robust predictive performance. This further validates that our counterfactual generation method better empowers prompt learning to align causal features in the images. 3) Our method achieves near-optimal performance across all datasets, with an average improvement of 17.6% over CLIP on seen classes and 3.87% on unseen classes, strongly affirming the superiority of our approach.

**Image Text Retrieval:** Table 2a shows the performance of diverse methods on the image text retrieval task. Our method performs optimally on unseen classes across three different few-shot learning settings. The vanilla instance-conditional prompt learning method, CoCoOp, experiences performance degradation on unseen classes of Flickr30k compared to zero-shot CLIP, further echoing our concerns. On the other hand, our DiCap model outperforms CPL in this task, demonstrating its ability to better focus on causal features and mitigate spurious correlations.

**Vision Question Answering:** In the VQA task, as shown in Table 2b, the zero-shot CLIP model performs poorly, with an average accuracy rate of only 16.8% across both seen and unseen classes. The poor performance is likely attributable to the limited exposure of the pre-trained model to the specific prompt templates used in the VQA task, in contrast to the prompt templates frequently encountered by the pre-trained vision-language models in the first two tasks. Nevertheless, we observe that all prompt learning methods significantly enhance the performance of the pre-trained language model on the VQA task. Notably, in this more challenging vision task, the DiCap method continues to lead, improving the CLIP model's predictive performance by an average of 34.6%.

**4.3.2 Counterfactual Quality Evaluation.** We conduct experiments on the ImageNet dataset, adjusting the scale parameter  $s$  in Eq. 2 to control the quality of counterfactuals and analyze its impact on

<i>Seen</i>	Caltech101	OxfordPets	Flowers102	Food101	StanfordCars	Sun397	ImageNet	Average
<b>CLIP</b>	90.64	91.12	69.80	83.10	55.45	66.72	70.30	75.30
<b>CoOp</b>	<b>97.93</b> <b>8.04</b>	92.82 <b>1.87</b>	<b>96.15</b> <b>37.8</b>	90.27 <b>8.63</b>	<b>74.76</b> <b>34.8</b>	81.24 <b>21.8</b>	86.25 <b>22.7</b>	88.49 <b>17.5</b>
<b>CoCoOp</b>	97.74 <b>7.83</b>	95.43 <b>4.73</b>	94.78 <b>35.8</b>	90.76 <b>9.22</b>	71.14 <b>28.3</b>	79.73 <b>19.5</b>	86.13 <b>22.5</b>	87.96 <b>16.8</b>
<b>CPL</b>	97.55 <b>7.62</b>	95.27 <b>4.55</b>	94.11 <b>34.8</b>	90.73 <b>9.18</b>	70.54 <b>27.2</b>	79.47 <b>19.1</b>	<b>86.35</b> <b>22.8</b>	87.72 <b>16.5</b>
<b>Dicap</b>	97.87 <b>7.98</b>	<b>95.66</b> <b>4.98</b>	94.12 <b>34.8</b>	<b>90.91</b> <b>9.40</b>	73.07 <b>31.8</b>	<b>82.38</b> <b>23.5</b>	86.07 <b>22.4</b>	<b>88.58</b> <b>17.6</b>
<i>Unseen</i>	Caltech101	OxfordPets	Flowers102	Food101	StanfordCars	Sun397	ImageNet	Average
<b>CLIP</b>	92.16	94.13	<b>74.26</b>	91.20	73.65	70.52	77.97	81.98
<b>CoOp</b>	89.74 <b>2.60</b>	95.41 <b>1.36</b>	72.00 <b>3.04</b>	91.00 <b>0.22</b>	73.68 <b>0.04</b>	73.45 <b>4.16</b>	79.92 <b>2.50</b>	82.17 <b>0.23</b>
<b>CoCoOp</b>	90.90 <b>1.37</b>	94.87 <b>0.79</b>	66.21 <b>10.8</b>	91.61 <b>0.45</b>	72.74 <b>1.24</b>	72.48 <b>2.78</b>	82.98 <b>6.43</b>	81.68 <b>0.37</b>
<b>CPL</b>	92.90 <b>0.80</b>	97.93 <b>4.04</b>	72.62 <b>2.21</b>	91.62 <b>0.46</b>	73.04 <b>0.83</b>	76.77 <b>8.86</b>	84.02 <b>7.76</b>	84.13 <b>2.62</b>
<b>Dicap</b>	<b>95.09</b> <b>3.18</b>	<b>98.01</b> <b>4.12</b>	73.61 <b>0.88</b>	<b>91.84</b> <b>0.70</b>	<b>74.92</b> <b>1.72</b>	<b>77.88</b> <b>10.4</b>	<b>84.77</b> <b>7.95</b>	<b>85.07</b> <b>3.87</b>

**Table 1:** Performance comparison of accuracy rate between *DiCap* and the SOTA baselines on the image classification task. Bold indicates the method with the best performance. The green numbers represent the performance improvement over the CLIP model, while the red numbers indicate the performance degradation relative to CLIP. The same applies to the Table 2.

Training Data Used	Methods	Flickr30k	MSCOCO	Average
0.5%	CLIP	78.90	67.70	73.30
	CoCoOp	76.80 <b>2.66</b>	73.50 <b>8.57</b>	75.15 <b>2.52</b>
	CPL	84.00 <b>6.46</b>	74.30 <b>9.75</b>	79.15 <b>7.98</b>
1%	DiCap	<b>85.80</b> <b>8.75</b>	<b>75.10</b> <b>10.9</b>	<b>80.45</b> <b>9.75</b>
	CoCoOp	84.60 <b>7.22</b>	74.33 <b>9.79</b>	79.47 <b>8.41</b>
	CPL	86.60 <b>9.76</b>	75.42 <b>11.4</b>	81.01 <b>10.5</b>
1.5%	DiCap	<b>87.20</b> <b>10.5</b>	<b>75.82</b> <b>12.0</b>	<b>81.51</b> <b>11.2</b>
	CoCoOp	84.90 <b>7.60</b>	75.25 <b>11.2</b>	80.08 <b>9.24</b>
	CPL	85.90 <b>8.87</b>	73.85 <b>9.08</b>	79.88 <b>8.97</b>
	DiCap	<b>86.20</b> <b>9.25</b>	<b>75.95</b> <b>12.2</b>	<b>81.08</b> <b>10.6</b>

(a) Image-Text Retrieval

Training Data Used	Methods	VQAv2	
		Seen	Unseen
0.1%	CLIP	16.01	17.60
	CoCoOp	<b>23.14</b> <b>44.5</b>	23.10 <b>31.3</b>
	CPL	22.00 <b>37.4</b>	20.97 <b>19.1</b>
0.2%	DiCap	22.97 <b>43.5</b>	<b>23.87</b> <b>35.6</b>
	CoCoOp	35.60 <b>122</b>	22.20 <b>26.1</b>
	CPL	35.08 <b>119</b>	21.60 <b>22.7</b>
0.5%	DiCap	<b>36.39</b> <b>127</b>	<b>23.10</b> <b>31.3</b>
	CoCoOp	43.46 <b>171</b>	22.60 <b>28.4</b>
	CPL	43.46 <b>171</b>	20.40 <b>15.9</b>
	DiCap	<b>44.50</b> <b>178</b>	<b>25.50</b> <b>44.9</b>

(b) Visual Question Answering

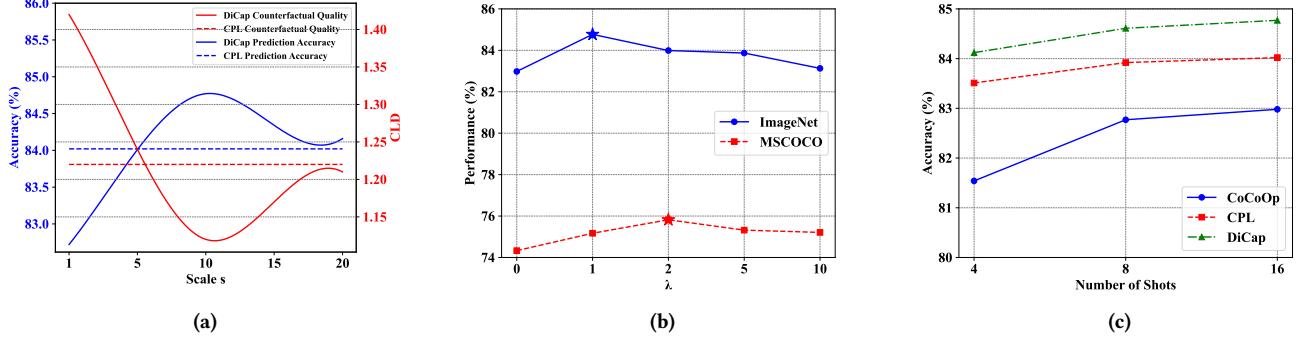
**Table 2:** Performance comparison of Recall@1 (a) and accuracy rate (b) between *DiCap* and the SOTA baselines on the image-text retrieval (a) and visual question answering tasks (b).

prompt learning performance. According to Sanchez and Tsaftaris [51], a smaller scale results in images nearly identical to the factual ones, while a larger scale causes the loss of important non-causal information. Therefore, both excessively large and small  $s$  affect the quality of counterfactuals. To quantify the impact of  $s$  on counterfactual quality, we adopt the CLD metric defined in Sanchez and Tsaftaris [51] (**lower CLD indicates higher counterfactual quality**) and compare the CLD scores and predictive performance of the DiCap and CPL methods under optimal settings.

The results, shown in Figure 5a, reveal that lower CLD scores consistently correlate with improved prompt learning performance,

highlighting that high-quality counterfactuals are critical for performance optimization. Furthermore, within a broad range ( $5 < s < 20$ ), the DiCap method consistently achieves lower CLD scores than CPL, demonstrating its ability to generate higher-quality counterfactuals, which is the fundamental reason for its superior performance compared to CPL.

**4.3.3 Ablation Study. Stability Analysis:** We further examine the impact of the sole hyperparameter  $\lambda$  in Eq. 8 on model performance. As shown in Figure 5b, we present results for unseen classes on the ImageNet and MSCOCO datasets (using 1% of samples from the training set). It is evident that DiCap’s predictive performance remains highly stable across both datasets despite variations in  $\lambda$ ,



**Figure 5:** (a) Counterfactuals enhance prompt learning. As the scale parameter  $s$  varies, model performance (left y-axis) and CLD scores (right y-axis) exhibit an inverse relationship. Since lower CLD scores indicate higher counterfactual quality, this highlights a strong positive correlation between counterfactual quality and model performance. Our DiCap method generates higher-quality counterfactuals within a broad range of  $5 < \text{scale} < 20$ , which directly leads to improved predictive performance. (b) Hyper-parameters analysis of  $\lambda$  on ImageNet and MSCOCO datasets. The blue and red stars represent the best parameters for the datasets; (c) three different shots between DiCap and two baselines on ImageNet unseen classes.

Prompt	ImageNet		MSCOCO(1.5%)		VQAv2(0.5%)	
Length	Seen	Unseen	Seen	Unseen	Seen	Unseen
4	<b>86.07</b>	84.77	63.55	75.95	44.50	25.50
8	85.38	82.94	62.74	74.81	44.08	24.56
16	84.98	80.13	61.92	74.02	43.23	23.84

**Table 3: Ablation Study of Prompt Length.**

Sampling	ImageNet		MSCOCO(1.5%)		VQAv2(0.5%)	
Strategy	Seen	Unseen	Seen	Unseen	Seen	Unseen
Random	85.56	84.08	62.98	74.22	43.83	23.13
Similarity	<b>86.07</b>	<b>84.77</b>	<b>63.55</b>	<b>75.95</b>	<b>44.50</b>	<b>25.50</b>

**Table 4: Performance comparison between two sampling strategies of DiCap on three visual tasks.**

consistently outperforming the vanilla instance-conditional prompt learning method ( $\lambda = 0$ ). This indicates that our approach is robust to changes of  $\lambda$ , demonstrating strong stability.

**Few-shot Evaluation:** Increasing the number of shots in the training data generally enhances predictive accuracy, a trend corroborated by our experiments. As illustrated in Figure 5c, we assess the performance of CoCoOp, CPL, and our DiCap method on unseen classes of ImageNet with shots set to 4, 8, and 16. The results reveal a steady improvement in DiCap’s accuracy as the number of shots increases, with our approach consistently surpassing the other two methods across all shot configurations.

**Prompt Length Analysis:** Table 3 reveals that shorter prompts consistently lead to better generalization across all three benchmarks, particularly under unseen settings. We hypothesize that longer prompts may introduce redundant or task-specific cues that the model can exploit spuriously, hindering its ability to generalize beyond the training distribution. In contrast, shorter prompts act as a form of regularization, encouraging the model to focus on invariant causal features.

**4.3.4 Sampling Strategy.** As shown in Table 4, we compare two negative label sampling methods: one based on classification probability similarity, as described in Section 3.5, and the other involving random sampling of negative labels. We evaluate prompt learning

with counterfactual images generated using each sampling strategy across ImageNet, MSCOCO with 1.5% of training samples, and VQAv2 with 0.5% of training samples. The results demonstrate that learning with more challenging samples, specifically, similarity-based sampling, further enhances DiCap’s predictive accuracy on both seen and unseen classes, validating the effectiveness of our sampling strategy.

## 5 Conclusions

This work delves into leveraging causality-driven diffusion models to refine prompt representations by aligning them with the causal characteristics inherent in images, thereby fostering the predictive performance of vision language models, notably on unseen classes. Our approach is underpinned by a robust theoretical foundation, providing stringent error bounds, an aspect not yet captured by existing literature. Extensive experimental results substantiate the robustness of our method across diverse visual tasks, highlighting the pivotal role of counterfactual generation in enhancing the performance of prompt learning. While our current model is primarily tailored for uni-modal prompt learning, it exhibits significant potential for seamless adaptation to multi-modal settings. We leave these avenues for future exploration. We also emphasize the importance of responsible deployment of generating counterfactuals, especially in sensitive domains such as healthcare.

## References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [3] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. 2022. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems* 35 (2022), 364–377.
- [4] Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. 2024. DiG-IN: Diffusion Guidance for Investigating Networks—Uncovering Classifier Differences Neuron Visualisations and Visual Counterfactual Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11093–11103.
- [5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101-mining discriminative components with random forests. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI 13*. Springer, 446–461.
- [6] Patrick Chao, Patrick Blöbaum, and Shiva Prasad Kasiviswanathan. 2023. Interventional and counterfactual inference with diffusion models. *arXiv preprint arXiv:2302.00860* (2023).
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [9] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- [10] Tsu-Jui Fu, Xin Eric Wang, Matthew F Peterson, Scott T Grafton, Miguel P Eckstein, and William Yang Wang. 2020. Counterfactual vision-and-language navigation via adversarial path sampler. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 71–86.
- [11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.
- [12] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*. PMLR, 2376–2384.
- [13] Gregory Griffin, Alex Holub, Pietro Perona, et al. 2007. *Caltech-256 object category dataset*. Technical Report 7694, California Institute of Technology Pasadena.
- [14] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*. PMLR, 1414–1423.
- [15] Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. 2022. Cpl: Counterfactual prompt learning for vision and language models. *arXiv preprint arXiv:2210.10362* (2022).
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [17] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- [18] Paul Jacob, Éloi Zablocki, Hedi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. 2022. STEEX: steering counterfactual explanations with semantics. In *European Conference on Computer Vision*. Springer, 387–403.
- [19] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. 2022. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*. 858–876.
- [20] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. 2023. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16425–16435.
- [21] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. PMLR, 3020–3029.
- [22] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.
- [23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19113–19122.
- [24] Saeed Khorram and Li Fuxin. 2022. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10203–10212.
- [25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.
- [26] Siwon Kim, Jinh Oh, Sungjin Lee, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. 2023. Grounding counterfactual explanation of image classifiers to textual concept space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10942–10950.
- [27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*. 554–561.
- [28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [29] 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.
- [30] Jiangmeng Li, Wenyi Mo, Wenwen Qiang, Bing Su, and Changwen Zheng. 2022. Supporting vision-language model inference with causality-pruning knowledge prompt. *arXiv preprint arXiv:2205.11100* (2022).
- [31] Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmoteab Elsaddik, and Xiaojun Chang. 2025. Contrastive learning with counterfactual explanations for radiology report generation. In *European Conference on Computer Vision*. Springer, 162–180.
- [32] Xinshu Li, Mingming Gong, and Lina Yao. 2024. Self-distilled disentangled learning for counterfactual prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1667–1678.
- [33] Xinshu Li and Lina Yao. 2022. Contrastive individual treatment effects estimation. In *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1053–1058.
- [34] Xinshu Li and Lina Yao. 2024. Distribution-conditioned adversarial variational autoencoder for valid instrumental variable generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 13664–13672.
- [35] Yang Li, Canran Xu, Guodong Long, Tao Shen, Chongyang Tao, and Jing Jiang. 2024. CCPrefix: Counterfactual Contrastive Prefix-Tuning for Many-Class Classification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2977–2988.
- [36] 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.
- [37] Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. 2023. Kept: Knowledge enhanced prompt tuning for event causality identification. *Knowledge-Based Systems* 259 (2023), 110064.
- [38] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.
- [39] Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. 2020. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092* (2020).
- [40] Zhiheng Lyu, Zhijing Jin, Justus Mattern, Rada Mihalcea, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. Psychologically-inspired causal prompts. *arXiv preprint arXiv:2305.01764* (2023).
- [41] Arash Nasr-Esfahany and Emre Kiciman. 2023. Counterfactual (non-) identifiability of learned structural causal models. *arXiv preprint arXiv:2301.09031* (2023).
- [42] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 722–729.
- [43] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3498–3505.
- [44] Nick Pawłowski, Daniel Coelho de Castro, and Ben Glocker. 2020. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems* 33 (2020), 857–869.
- [45] Judea Pearl. 2009. *Causality*. Cambridge university press.
- [46] 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.
- [47] Viraj Prabhu, Sriram Venamandra, Prithvijit Chattopadhyay, and Judy Hoffman. 2023. Lance: Stress-testing visual models by generating language-guided counterfactual images. *Advances in Neural Information Processing Systems* 36 (2023),

- 25165–25184.
- [48] 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.
  - [49] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG] <https://arxiv.org/abs/1910.10683>
  - [50] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592* (2020).
  - [51] Pedro Sanchez and Sotirios A Tsaftaris. 2022. Diffusion Causal Models for Counterfactual Estimation. In *Causal Learning and Reasoning 2022*.
  - [52] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni SGouritsa, Kun Zhang, and Joris Mooij. 2012. On Causal and Anticausal Learning. In *29th International Conference on Machine Learning (ICML 2012)*. International Conference on Machine Learning.
  - [53] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020).
  - [54] Jiaming Song, Chenlin Meng, and Stefano Ermon. [n. d.]. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
  - [55] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems* 28 (2015).
  - [56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
  - [57] Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. 2025. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*. Springer, 338–357.
  - [58] Jianxiong Xiao, James Hays, Krista Ahinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 3485–3492.
  - [59] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2021. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9593–9602.
  - [60] Mehdi Zemni, Mickaël Chen, Éloï Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. 2023. Octet: Object-aware counterfactual explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15062–15071.
  - [61] Congzhi Zhang, Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. 2024. Causal prompting: Debiasing large language model prompting based on front-door adjustment. *arXiv preprint arXiv:2403.02738* (2024).
  - [62] Kun Zhang and Aapo Hyvärinen. 2012. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599* (2012).
  - [63] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16816–16825.
  - [64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.
  - [65] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. 2021. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*. PMLR, 12979–12990.