Rethinking Misalignment in Vision-Language Model Adaptation from a Causal Perspective

Yanan Zhang[1,2,*], **Jiangmeng Li**[2,*], **Lixiang Liu**[1,2], **Wenwen Qiang**[2,†]
[1]University of Chinese Academy of Sciences
[2]Institute of Software Chinese Academy of Sciences
zhangyanan110199@gmail.com, {jiangmeng2019, lixiang, qiangwenwen}@iscas.ac.cn

Abstract

Foundational Vision-Language models such as CLIP have exhibited impressive generalization in downstream tasks. However, CLIP suffers from a two-level misalignment issue, i.e., task misalignment and data misalignment, when adapting to specific tasks. Soft prompt tuning has mitigated the task misalignment, yet the data misalignment remains a challenge. To analyze the impacts of the data misalignment, we revisit the pre-training and adaptation processes of CLIP and develop a structural causal model. We discover that while we expect to capture task-relevant information for downstream tasks accurately, the task-irrelevant knowledge impacts the prediction results and hampers the modeling of the true relationships between the images and the predicted classes. As task-irrelevant knowledge is unobservable, we leverage the front-door adjustment and propose Causality-Guided Semantic Decoupling and Classification (CDC) to mitigate the interference of task-irrelevant knowledge. Specifically, we decouple semantics contained in the data of downstream tasks and perform classification based on each semantic. Furthermore, we employ the Dempster-Shafer evidence theory to evaluate the uncertainty of each prediction generated by diverse semantics. Experiments conducted in multiple different settings have consistently demonstrated the effectiveness of CDC.

1 Introduction

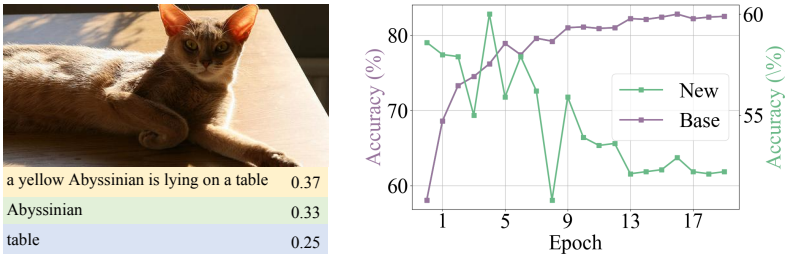
Benefiting from large-scale training data, foundational Vision-Language models such as CLIP [1], ALIGN [2], and Florence [3] have demonstrated remarkable zero-shot generalization capabilities across a wide range of downstream tasks.

Despite the strong generalization of CLIP, there exists a *two-level misalignment* between CLIP and downstream tasks, which hinders its adaptation to these tasks. Specifically, the first *misalignment* is caused by the discrepancy between the pre-training objectives of CLIP and the objectives of downstream tasks, i.e., the *task misalignment*. During pre-training, the texts associated with images often contain rich semantic information, e.g., "a yellow Abyssinian is lying on a table". Accordingly, as shown in Figure 1(a), in the learned embedding space, the distance between the image and the entire textual description can be close, indicating a strong semantic relationship, but the distance between the image and the individual semantic elements such as "Abyssinian" or "table" is not sufficiently close. When performing a downstream task that requires classifying an image as an "Abyssinian", the model encounters a challenge in capturing the specific semantics associated with "Abyssinian" and may fail to complete the task. To address this issue, soft prompt tuning [4, 5, 6] is proposed. By introducing learnable tokens into the model and tuning them with task-specific data, CLIP can identify task-related semantics and adapt to the corresponding task.

[*]Equal contributions.
[†]Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1238



(a) (b)

Figure 1: (a) A motivating example of task misalignment, illustrating the cosine similarities between an image and various text descriptions in the embedding space of CLIP. (b) A motivating experiment on data misalignment, showing the accuracy trends for base and new classes across different training epochs on the DTD dataset.

However, not all downstream tasks have available annotated data for prompt tuning. As a result, inconsistency exists between the training and testing data, constituting the second *misalignment*, i.e., *data misalignment*. The inconsistency in data misalignment can arise due to two main reasons: (1) *Label Inconsistency*: The training and testing classes do not completely overlap. For instance, some classes present in the training data might not appear in the testing data and vice versa. We refer to the classes that appear in both training and testing as base classes and the classes that appear only in testing as new classes. (2) *Distribution Inconsistency*: Even if they share the same class names, the distributions of the classes in the training and testing data may differ, resulting in distribution inconsistency. In such cases, the testing classes are essentially also new classes. Researchers generally agree that prompt tuning potentially leads to the overfitting of the CLIP model to the base classes, deviating CLIP from its original behavior and thus hindering its generalization to new classes [7, 8]. To explore whether the overfitting holds, we conduct prompt tuning based on MaPLe [6] on the DTD dataset [9]. We record the performance of the base classes and the new classes under different training epochs. As shown in Figure 1(b), as the number of training epochs increases, the performance on base classes gradually improves; however, the performance on the new classes first increases and then decreases. The phenomenon is consistent with the definition of overfitting.

To explore why overfitting occurs and how it impacts the recognition of new classes, we revisit the pre-training and adaptation processes of CLIP and develop a Structural Causal Model (SCM) [10, 11] to analyze the causal relationships among variables abstracted from these two processes. From the perspective of data generation, generative factors refer to the elements that control the content of images and texts. We argue that the knowledge embedded in CLIP determines the content that can be obtained from CLIP. Therefore, the knowledge in CLIP is equivalent to the generative factors it contains. Via soft prompt tuning, we expect the learned prompts to capture the semantics determined by the task-relevant generative factors and eliminate the influence of task-irrelevant generative factors. However, due to the data misalignment, the task-relevant generative factors estimated on the base classes potentially being task-irrelevant for the new classes. When estimating the causal relationship between images and the label space of new classes, task-irrelevant generative factors could interfere. In addition, task-irrelevant generative factors cannot be observed in practice, so we cannot remove their effects by adjusting for them. To solve this problem, we introduce task-related semantics as intermediate variables from images to the label space of new classes and apply the front-door adjustment to estimate the true causal relationship between the two.

To implement the front-door adjustment, we propose Causality-Guided Semantic Decoupling and Classification (CDC), consisting of Visual-Language Dual Semantic Decoupling (VSD) and Decoupled Semantic Trusted Classification (DSTC). Specifically, we incorporate multiple prompt templates within the model and encourage distinct templates to represent different semantics through VSD. In addition, decoupled semantics exhibit varying classification uncertainties. Therefore, we propose DSTC, which performs the classification task independently based on each decoupled semantic and estimates the corresponding uncertainties simultaneously. Subsequently, we fuse the results to obtain more accurate and reliable predictions. Through experiments conducted in various settings, our proposed CDC has effectively enhanced the performance of CLIP.

https://doi.org/10.52202/079017-1238

Our **contributions** include: (i) We identify the *two-level misalignment* that exists in the adaptation of CLIP to downstream tasks and illustrate that, due to the data misalignment, the prompt learned through soft prompt tuning becomes overfitted to the base classes; (ii) We develop an SCM to investigate the pre-training and adaptation processes of CLIP, revealing how the overfitting occurs and how it impacts the recognition of new classes. We discover that the task-irrelevant generative factors serve as the confounder when estimating the true causal relationship between the images and the label space of new classes; (iii) To mitigate the impact of task-irrelevant generative factors on downstream tasks, we propose CDC, which implements the front-door adjustment. Through experiments on multiple datasets and various tasks, the effectiveness of CDC has been demonstrated.

2 Related Work

Adaptation in Vision Language models. To enhance the performance of vision-language models such as CLIP on downstream tasks, researchers propose to utilize few-shot adaptation. Generally, two main technical approaches have emerged: 1) adapter-based methods [4, 12, 13], which incorporate adapters into CLIP to fine-tune the features generated by CLIP, enabling better adaptation to specific tasks; 2) prompt-based methods, which primarily inject task information by appending learnable tokens to CLIP. While methods like CoOp [4], CoCoOp [5], and ProGrad [14] add tokens to the text input layer, VPT [15] and CAVPT [16] introduce tokens on the image branch as well. MaPLe [6] and UPT [17] incorporate tokens on both the image encoder and text encoder. In addition, MaPLe adds tokens into the intermediate layers of the network to enhance the model's adaptability to the task. Among all prompt-based approaches, the most similar to our method is ArGue [18], which aims to improve model performance by iterating over the attributes that each category possesses. However, ArGue relies on the assistance of other large language models for attribute generation and requires additional preprocessing steps. In contrast, CDC achieves competitive performance with ArGue without relying on any external knowledge, solely through the end-to-end training.

Causal Inference. In recent years, causal inference [10, 11] has been extensively employed in computer vision and multi-modal learning. Generally, researchers explore the causal inference in two promising directions: counterfactual [19, 20, 21, 22] and intervention [23, 24, 25, 26, 27]. [28] and [21, 22] propose to generate the counterfactual images to improve the robustness of the model. [27] proposes intervening in the distribution of training data to eliminate the interference of the spurious correlation on predictions for domain generalization problems. To the best of our knowledge, we are the first to apply causal inference to analyze CLIP's overfitting in downstream classification. We identify the confounders and eliminate their effects through the front-door adjustment.

3 Problem Formulation and Analysis

3.1 Problem Formulation

CLIP [1] is an impressive vision-language model comprised of an image encoder and a text encoder, capable of embedding images and texts into a shared multi-modal latent space. Through contrastive learning, paired image-text embeddings are optimized for maximum cosine similarity, while unpaired ones are minimized. The intuition behind such design is based on a fundamental assumption: any paired images and texts are expected to form a tight semantic correspondence, as they contain similar content. By pre-training on a large-scale dataset consisting of 400 million image-text pairs, CLIP [1] demonstrates strong generalization in zero-shot and few-shot downstream tasks.

In zero-shot classification with C classes, an image x is transformed into an embedding v via the image encoder. For the text branch, CLIP [1] constructs prompts for each class using a static template, e.g., "a photo of a [CLASS]", where [CLASS] denotes the class name of the $c\text{-}th$ class, $c \in \{1, 2, ..., C\}$. These prompts are inputted into the text encoder to obtain embedding vectors w_c for each class. By calculating the cosine similarity between the image embedding v and the text embeddings w_c, we can obtain the probability of image x belonging to the $c\text{-}th$ class:

$$p(y = c|x) = \frac{exp(sim(v, w_c)/\tau)}{\Sigma_{c'=1}^{C} exp(sim(v, w_{c'})/\tau)}, \tag{1}$$

where $sim(\cdot, \cdot)$ denotes the cosine similarity, and τ is a temperature scalar.

https://doi.org/10.52202/079017-1238

In the few-shot setting, where a certain amount of labeled data is available, it is feasible to learn a template instead of manually specifying one. Initially, researchers construct a learnable template $t = \{p_1, p_2, ..., p_d, l_c\}$ by appending the word embedding l_c of the c-th class to d learnable tokens as the input of the text encoder. In recent years, intermediate layers of both text and image encoders also employ learnable tokens. These tokens are optimized with the cross-entropy loss to adapt to the specific task. The learned template performs the classification task in the same way as the static one.

3.2 Problem Analysis: Causal Perspective

To understand how the knowledge contained in CLIP [1] affects the downstream classification tasks, we propose an SCM shown in Figure 2. The directed acyclic graph $G =< V, E >$ depicts the causal relationships among variables abstracted from the pre-training and adaptation processes. Each node $V_i \in V$ corresponds to a variable. Edge $V_i \rightarrow V_j \in E$ signifies that V_i is the cause of V_j.

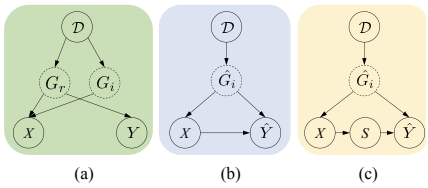


Figure 2: SCMs. Solid and dashed circles indicate the observable and unobservable variables, respectively.

Firstly, we define each variable in Figure 2. $\mathcal{D}$ represents the data used in the pre-training stage. Generative factors, the fundamental semantic elements that control the generation of images and texts, can determine the content of both images and texts. In a specific task, these generative factors can be divided into two subsets: the set of task-relevant factors G_r and the set of task-irrelevant factors G_i. Taking the pet classification task as an example, task-relevant generative factors determine features such as coat colors, body shapes, and ear types, which are highly relevant to identifying pet breeds. Conversely, task-irrelevant generative factors influence semantics which do not provide helpful information for pet classification. For a downstream task, X represents the image space, while Y corresponds to the true labels of these images. $\hat{Y}$ represents the predicted labels obtained through the learning. Additionally, S denotes the semantics within the images X closely relevant to the current task.

By training on diverse data, CLIP acquires relatively comprehensive visual and language knowledge. Therefore, it can be assumed that through the learning of $\mathcal{D}$, we obtain comprehensive generative factors that constitute the entire visual space. Since the learning process of pre-training is not specific to any particular downstream task, these generative factors include not only task-related but also task-irrelevant ones, i.e., $\mathcal{D} \rightarrow G_i$ and $\mathcal{D} \rightarrow G_r$, for a specific task. By combining the generative factors, any image $x \in X$ can be generated, regardless of whether it has appeared in the pre-training dataset $\mathcal{D}$. Images typically include not only content related to their category under a particular task but usually other content as well, hence $G_r \rightarrow X$ and $G_i \rightarrow X$. However, only task-related generative factors can determine the true label Y of images, i.e., $G_r \rightarrow Y$.

Given that the training data $\mathcal{D}$ for CLIP is invariant, in Figure 2(a), the path $X \leftarrow G_i \leftarrow \mathcal{D} \rightarrow G_r \rightarrow Y$ is blocked, resulting in X being associated with Y solely through G_r, i.e., $X \leftarrow G_r \rightarrow Y$. When performing downstream classification tasks, we aim to model the relationship between X and $\hat{Y}$ that arises from G_r. As shown in Figure 2(b), $X \rightarrow \hat{Y}$ represents the objective of our modeling. G_i and G_r are unobservable and easily confused in practice. Therefore, we perform soft prompt tuning to help capture the information of G_r and eliminate the influence of G_i. However, the estimated task-related factors are not always accurate. Furthermore, due to the data misalignment, the task-relevant generative factors estimated based on the training data of base classes are not necessarily the task-relevant factors for new classes. $\hat{G}_i$ denotes the set of task-irrelevant generative factors that are incorrectly retained, and the set of factors that are task-relevant for the base classes but task-irrelevant for the new classes. Since we incorrectly assume that $\hat{G}_i$ is task-relevant, $\hat{G}_i$ influences $\hat{Y}$, i.e., $\hat{G}_i \rightarrow \hat{Y}$. When the learned prompt template is overfitted to the base classes, generative factors in $\hat{G}$ will further increase, hindering the estimation of the true causal relationship between X and $\hat{Y}$.

According to the definition of the backdoor path, when estimating the true causal relationship between X and $\hat{Y}$, there exists a backdoor path $X \leftarrow \hat{G}_i \rightarrow \hat{Y}$, where $\hat{G}_i$ serves as a confounder. Given that $\hat{G}_i$ is unobservable, it becomes impracticable to eliminate spurious relationships caused by $\hat{G}_i$ via the back-door adjustment. To address this issue, we introduce an intermediate variable S that represents task-relevant semantics and is derived from X, thus establishing the path $X \rightarrow S$, as

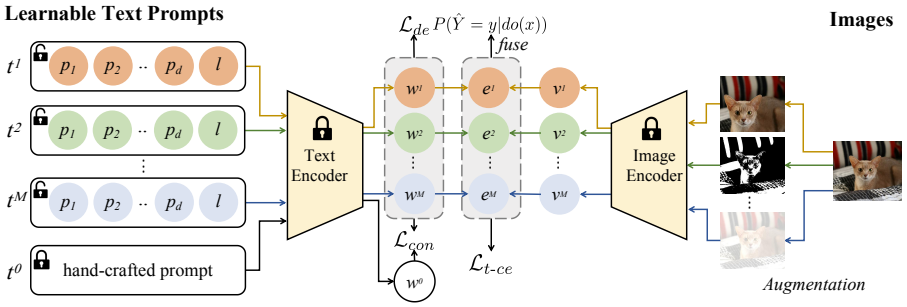


Figure 3: Framework of CDC. t^m denotes a single template, while $p_1, p_2, ..., p_d$ represent tokens in the template. Different colors indicate diverse templates. "fuse" refers to the process of generating the final classification results from multiple template results as shown in Equation (8). The text encoder and the image encoder are frozen, and only the tokens in the prompt templates are learnable.

shown in Figure 2(c). The classification for X can be regarded as predicting based on the semantics S that X possesses, thus $S \to \hat{Y}$. S satisfies the front-door criterion for $(X, \hat{Y})$, and the causal effect from X to $\hat{Y}$ can be calculated using the front-door formula:

$$P(\hat{Y} = y|do(x)) = \sum_s \underbrace{P(s|x)}_{\text{first term}} \underbrace{\sum_{x'} P(y|x', s)P(x')}_{\text{second term}}. \tag{2}$$

To estimate the true causal relationship between X and $\hat{Y}$, we will discuss how to obtain the semantic s and how to estimate the probability of a sample x belonging to class y.

4 Methodology

Guided by the front-door criterion, we propose CDC, comprising two essential components: VSD and DSTC. Specifically, VSD focuses on decoupling the semantics represented by diverse templates. DSTC aims to generate the classification scores via Equation (2) based on the decoupled semantics. We will elaborate on the details of these two components in Section 4.1 and Section 4.2, respectively.

4.1 Visual-Language Dual Semantic Decoupling

To decouple the semantics embedded in the images and texts, the concept of *semantic set* is introduced. A semantic set serves as a collection that encapsulates the semantics related to the same visual attribute. For example, multiple semantic sets can be constructed for the task of pet classification, such as the set of coat colors, the set of body shapes, and the set of ear types. Each semantic set contains some possible values that represent the variations of the corresponding attribute. For instance, the semantic set "coat colors" includes semantics like red, white, purple, etc. The semantic sets enable us to understand and analyze the semantic information embedded in pet images.

To represent different semantic sets, we employ multiple templates in the CLIP model, i.e., $t^m \in \{t^1, t^2, ..., t^M\}$ in Figure 3, where M denotes the number of templates. Each template aims to characterize a unique semantic set. To achieve decoupling of the semantics represented by these templates, we propose VSD, which considers both text and image branches. On the image branch, we leverage diverse augmentation methods to generate the image inputs and employ the corresponding augmentation method for each individual template. The intuition behind this is to help different templates capture distinct feature invariance against different augmentations, e.g., performing random rotation augmentation enables the template to capture features with rotation invariance. Since defining the precious semantic sets, such as coat colors and ear types, remains challenging without the help of external knowledge, we group semantics that exhibit the same type of feature invariance into a semantic set in practice.

On the text branch, we expect to maximize the diversity of embeddings corresponding to different templates, ideally achieving orthogonality. To achieve this objective, we aim to ensure that the embedding of the c-th class in the m-th template, denoted as w_c^m, cannot be accurately classified by

https://doi.org/10.52202/079017-1238

embeddings from other templates, i.e., using any $w^{m'}$ to classify w_c^m, where $m' \neq m$, there is no bias towards the results. Specifically, the objective function can be formalized as:

$$\mathcal{L}_{de} = \frac{1}{M-1}\frac{1}{C}\sum_{m'=1,m'\neq m}^{M}\sum_{c=1}^{C}\sum_{\bar{c}=1}^{C} P(\bar{c}|w_c^m, w^{m'})\log P(\bar{c}|w_c^m, w^{m'}), \tag{3}$$

where $P(\bar{c}|w_c^m, w^{m'})$ denotes the probability that w_c^m is predicted as the $\bar{c}$-th class by the embedding of m'-th template. Specifically, $P(\bar{c}|w_c^m, w^{m'})$ can be calculated as:

$$P(\bar{c}|w_c^m, w^{m'}) = \frac{exp(sim(w_c^m, w_{\bar{c}}^{m'})/\tau)}{\sum_{c'=1}^{C} exp(sim(w_c^m, w_{c'}^{m'})/\tau)}. \tag{4}$$

Intuitively, the embedding vector w_c^m from the m-th template can be regarded as a sample to be classified. The objective of Equation (3) is to maximize the entropy of the classification results, resulting in a uniform distribution. The classification results are obtained using the embedding of all other templates. $\mathcal{L}_{de}$ aims to enhance the semantic distinctions among different templates, thereby encouraging each template to maintain a unique semantic.

While boosting the diversity of the semantics in different templates, $\mathcal{L}_{de}$ may hinder the learning of the task-relevant semantics. To alleviate this problem, we propose introducing a consistency loss to assist each template in capturing information that is close to the hand-crafted template. The proposed consistency loss can be formulated as follows:

$$\mathcal{L}_{con} = -\frac{1}{C}\sum_{m=1}^{M}\sum_{c=1}^{C}\log P(c|w_c^m, w^0), \tag{5}$$

where w^0 denotes the embeddings of the hand-crafted template t^0 obtained via the original CLIP model. See **Appendix A** for the detailed templates.

Via VSD, we obtain distinct semantic sets that contain decoupled semantics. Different classes have varying values on the same semantic set, so we consider the values of s in Equation (2) to be the text embeddings of each class on all semantic sets, i.e, $s \in \{w_c^m\}_{m=1,c=1}^{M,C}$.

4.2 Decoupled Semantic Trusted Classification

To implement Equation (2), we need to clarify how to estimate the $P(\hat{Y}|do(x))$ based on the decoupled semantics. Firstly, the probability $P(s|x)$ that a sample x has semantic s is obtained by calculating the cosine similarity between the representation v of x and s, i.e., $P(s|x) = N(sim(s, v))$, where N denotes the normalization operation.

After determining the possible values of s, we clarify how to estimate the second term in Equation (2). Let x' represent the samples from the training set and $s = w_c^m$, then: (i) if $c \neq y$, $P(y|x', s) = 0$. By updating the template, we expect each class c to carry unique semantics w_c^m. Therefore, it cannot be assumed that the specific semantics of c is attributable to another class; (ii) if $c = y$ but x' does not belong to the c-th class, $P(y|x', s) = 0$; (iii) if $c = y$ and x' belong to the c-th class, we maximize the similarity between s and x' by prompt tuning so that the semantics of s and x' are as consistent as possible. As x' contains both task-relevant and task-irrelevant semantics, and its task-relevant semantic is decoupled to $w_c^1, w_c^2, ..., w_c^M$ via VSD, s is the subset of x'. Therefore, the second term in Equation (2) shrinks to $P(y|s)$.

Ideally, if s contains and only contains the category-specific semantic information that explicitly supports the classification of sample x' into c, $P(y = c|s) = 1$, and $P(y \neq c|x', s) = 0$. However, because of the data misalignment, while applying the learned templates to the new classes, this ideal scenario does not happen. In reality, the qualities of the learned semantics vary for different templates and classes, allowing the value of $P(y|s)$ to alter, thereby resulting in significant differences in the uncertainty of classification results derived from different templates. To solve this issue, we propose to estimate the uncertainty [29, 30] of each prediction from diverse templates via Dempster-Shafer theory [31, 32]. Specifically, we consider $e_c^m = h(sim(w_c^m, v))$ as the evidence of classifying the feature v as belonging to the c-th class under the m-th semantic set, where h represents a function that maps the cosine similarity to a non-negative value. We assume that there exists a Dirichlet distribution $\alpha^m = [\alpha_1^m, \alpha_2^m, ..., \alpha_C^m]$, where $\alpha_c^m = e_c^m + 1$. This formulation allows us to quantify

the belief that the feature v belongs to class c as well as the uncertainty associated with the prediction as:

$$b_c^m = \frac{e_c^m}{A^m}, u^m = \frac{C}{A^m}, \tag{6}$$

where $A^m = \sum_{c'=1}^{C} e_{c'}^m + 1$. Once we have obtained the classification beliefs and uncertainties for all semantic sets, we proceed to perform evidence fusion to obtain the final classification results. Specifically, for the m-th and m'-th semantic set, after evidence fusion, we obtain the belief of the sample x belonging to class c, as well as the uncertainty of the classification as:

$$b_c^{mm'} = \frac{1}{1-C}(b_c^m b_c^{m'} + b_c^m u^{m'} + b_c^{m'} u^m), u^{mm'} = \frac{1}{1-C} u^m u^{m'}. \tag{7}$$

The results from all prompt templates can be iteratively fused, as shown in the following equation:

$$\begin{aligned} B_c^m &= \begin{cases} b_c^1, & \text{if } m = 1 \\ \frac{1}{1-C}(B_c^{m-1} b_c^m + B_c^{m-1} u^m + b_c^m U^{m-1}), & \text{if } 1 < m \leq M \end{cases}, \\ U^m &= \begin{cases} u^1, & \text{if } m = 1 \\ \frac{1}{1-C} U^{m-1} u^m, & \text{if } 1 < m \leq M \end{cases}. \end{aligned} \tag{8}$$

In this formulation, B_c^m represents the belief after fusing the results from the first m templates, while U^m denotes the corresponding uncertainty. The process of evidence fusion can be understood as the summation of results over all semantics s in Equation (2). Consequently, the result in Equation (2) can be reformulated as:

$$P(\hat{Y} = c | do(x)) = \frac{B_c^M}{\Sigma_{c'=1}^{C} B_{c'}^M}. \tag{9}$$

To calculate the probability that sample x belongs to class c during the testing phase, we employ Equation (9). Accordingly, during the training stage, the cross-entropy loss which is used for learning the prompts is adjusted to the trusted cross-entropy loss to model the uncertainty information. For each training sample, the trusted cross-entropy can be formalized as:

$$\mathcal{L}_{t\text{-}ce} = \sum_{m=1}^{M} (\psi(A) - \psi(\alpha_y^m)), \tag{10}$$

where y denotes the index of the ground-truth class and $\psi(\cdot)$ denotes the digamma function.

4.3 Overall Architecture

We demonstrate the architecture of our proposed method in Figure 3. While training, the overall loss function can be formalized as:

$$\mathcal{L}_{CDC} = \mathcal{L}_{t\text{-}ce} + \beta \mathcal{L}_{de} + \gamma \mathcal{L}_{con}, \tag{11}$$

where β and γ denote the hyper-parameters to tune the influence of $\mathcal{L}_{de}$ and $\mathcal{L}_{con}$, respectively. Refer to **Algorithm 1** and **Algorithm 2** for the detailed training and testing pipeline.

5 Experiments

Following previous works [5, 6], we conduct experiments to evaluate our proposed method with three different settings. These settings encompass the base-to-new setting as well as two out-of-distribution (OOD) settings, i.e., the cross-dataset setting and the cross-domain setting. Refer to **Appendix B** for a detailed overview of the evaluation protocol.

5.1 Experimental Settings

Datasets. In the base-to-new setting, we conduct experiments based on 11 datasets: ImageNet [33], Caltech101 [34], Oxford Pets [35], Stanford Cars [36], Flowers102 [37], Food101 [38], FGVC Aircraft [39], SUN397 [40], DTD [9], EuroSAT [41], and UCF-101 [42]. In the following, we

 https://doi.org/10.52202/079017-1238

Algorithm 1 The training pipeline of CDC

Input: 16-shot dataset X, the learning rate ℓ, two hyper-parameter β and γ, the number of templates M, and the list of augmentation methods $\{\mathcal{A}^1, \mathcal{A}^2, ..., \mathcal{A}^M\}$.
Initialize the parameters of CLIP with the parameters of the pre-trained model.
Randomly initialize the learnable tokens $\{\theta^1, \theta^2, ..., \theta^M\}$.
repeat
  for i-th training iteration **do**
    Iteratively sample a minibatch $\mathcal{X}'$ from $\mathcal{X}$.
    for m-th template **do**
      $\mathcal{X}'^m = \mathcal{A}^m(\mathcal{X}')$.
      Generate the image features v^m of $\mathcal{X}'^m$ with the image encoder of CLIP.
      Generate the text features w^m with the text encoder.
      Generate the classification evidence e^m with v^m and w^m: $e_c^m = h(sim(w_c^m, v^m))$.
      Using Equation (10) to calculate $\mathcal{L}_{t\text{-}ce}$.
      Using Equation (3) and Equation (5) to calculate $\mathcal{L}_{de}$ and $\mathcal{L}_{con}$.
      Using Equation (11) to calculate $\mathcal{L}_{CDC}$.
      $\theta^m = \theta^m - \ell\nabla_\theta\mathcal{L}_{CDC}$.
    end for
  end for
until θ converge.

Algorithm 2 The testing pipeline of CDC

Input: The testing dataset X, the number of templates M, and $\{\theta^1, \theta^2, ..., \theta^M\}$.
Sample a test data x from X
for m-th template **do**
  Generate the image features v^m of x.
  Generate the text features w^m with the text encoder.
  Generate the classification evidence e^m with v^m and w^m: $e_c^m = h(sim(w_c^m, v^m))$.
end for
Using Equation (8) and Equation (9) to generate the final classification results.

abbreviate these datasets as ImageNet, Caltech, Pets, Cars, Flowers, Food, Aircraft, SUN, DTD, EuroSAT, and UCF. For the out-of-generalization task, we adopt ImageNet as the source dataset. The remaining 10 are the target datasets in the cross-dataset setting, and ImageNetV2 [43], ImageNet-S [44], ImageNet-A [45], and ImageNet-R [46] are the target datasets in the cross-domain setting.

Experimental Details. We follow the experimental settings of the baseline method MaPLe. Specifically, we utilize a pre-trained CLIP with ViT-B/16 as the visual encoder. The number of learnable tokens is fixed at 2, whereas the prompt depth varies, being 9 for the base-to-new setting and 3 for the OOD setting. The learning rate is 0.035, and the batch size is 4. All models are trained using an SGD optimizer on an NVIDIA 3090 GPU. Our proposed CDC introduces three additional hyperparameters: β and γ, which represent the weights for $\mathcal{L}_{de}$ and $\mathcal{L}_{con}$, respectively, and M, which denotes the number of prompts. Furthermore, different augmentation methods have a notable impact on the model's performance. We set $\beta = 5$, $\gamma = 0.01$, and $M = 4$ in the base-to-new setting, and $\beta = 3$, $\gamma = 0.01$, and $M = 8$ in the OOD setting. For a detailed analysis of the influence of varying hyper-parameters and augmentation methods, please refer to **Appendix C**. We report the accuracies of base classes and novel classes, along with their harmonic mean (HM), averaged over 3 runs.

5.2 Base-to-New Generalization

As shown in Table 1, we compare our method with three recent works, CoOp, CoCoOp, and MaPLe. For a fair comparison, we do not include methods that utilize external knowledge, such as ArGue [18], HPT [47], and CoPrompt [8], although our method demonstrates competitive performance.

Compared to the baseline method MaPLe, our approach has achieved remarkable improvements in the accuracy of base classes by 1.06%, and the accuracy of new classes by 2.24%, which leads to an improvement of 1.70% in HM. Specifically, across all 11 datasets, our method outperforms MaPLe

Table 1: The comparison with baseline methods on base-to-novel generalization setting.

Dataset	CoOp [4]			CoCoOp [5]			MaPLe [6]			CDC			
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM	Δ
Avg	82.69	63.22	71.66	80.47	71.69	75.83	82.28	75.14	78.55	**83.34**	**77.38**	**80.25**	+1.70
ImageNet	76.47	67.88	71.92	75.98	70.43	73.10	76.66	70.54	73.47	**77.50**	**71.73**	**74.51**	+1.04
Caltech	98.00	89.91	93.73	97.96	93.81	95.84	97.74	94.36	96.02	**98.20**	**94.37**	**96.25**	+0.23
Pets	93.67	95.29	94.47	95.20	97.69	96.43	95.43	97.76	96.58	**96.07**	**98.00**	**97.02**	+0.44
Cars	**78.12**	60.40	68.13	70.49	73.59	72.01	72.94	**74.00**	73.47	73.80	73.97	**73.88**	+0.41
Flowers	**97.60**	59.67	74.06	94.87	71.75	81.71	95.92	72.46	82.56	96.93	**75.07**	**84.61**	+2.05
Food	88.33	82.26	85.19	90.70	91.29	90.99	90.71	92.05	91.38	**90.87**	**92.33**	**91.59**	+0.21
Aircraft	**40.44**	22.30	28.75	33.41	23.71	27.74	37.44	35.61	36.50	37.47	**37.50**	**37.48**	+0.98
SUN	80.60	65.89	72.51	79.74	76.86	78.27	80.82	78.70	79.75	**82.37**	**80.03**	**81.18**	+1.43
DTD	79.44	41.18	54.24	77.01	56.00	64.85	80.36	59.18	68.16	**82.70**	**64.10**	**72.22**	+4.06
SAT	92.19	54.74	68.90	87.49	60.04	71.21	94.07	73.23	82.35	**95.10**	**82.33**	**88.26**	+5.91
UCF	84.69	56.05	67.46	82.33	73.45	77.64	83.00	78.66	80.77	**85.70**	**81.73**	**83.67**	+2.90

Table 2: Comparison of CDC with recent approaches on cross-dataset evaluation.

	Source	**Target**									
	ImageNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	SAT	UCF	Avg
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	**24.74**	67.01	46.49	48.06	**68.69**	66.30
CDC	**71.76**	**94.47**	**90.77**	**66.27**	**72.67**	**86.27**	24.50	**68.07**	**46.60**	**49.13**	68.60	**66.73**

in all 11 datasets for base classes and in 10 datasets for novel classes. Notably, on the ImageNet dataset, our approach achieves an accuracy increase of 0.84% in base classes, 1.19% in new classes, and 1.04% in HM. For the challenging datasets SAT, DTD, and UCF, our method delivers significant HM improvements of 5.91%, 4.06%, and 2.90%, respectively. In summary, our proposed method significantly enhances the generalization of the CLIP model for unseen classes, while ensuring stable improvements in the performance of base classes, demonstrating its superiority on a wide range of datasets.

5.3 Out-of-Distribution Generalization

Cross-dataset. We evaluate the cross-dataset generalization of CDC by learning prompts on ImageNet and then transferring them to the remaining 10 datasets. From Table 2, for the source dataset ImageNet, our proposed method achieves performance improvement by 1.04% compared to MaPLe, outperforming all previous methods. Compared to MaPLe on the 10 target datasets, our method achieves better performance on 8 datasets. Considering all target datasets, our method achieves an average performance improvement of 0.43%.

Cross-domain. We transfer the prompts learning in ImageNet to its four invariants to evaluate the coss-domain generalization of CDC. From Table 3, compared to MaPLe, our proposed method improves the performance of ImageNetV2, ImageNet-S, and ImageNet-R by 0.80%, 1.18%, and 1.12%, respectively, and brings a slight performance decrease of 0.5% for ImageNet-A, thus leading to an average performance improvement of 0.65%. Generally, CDC enhances the generalization of the model when dealing with data from different domains.

5.4 Ablative Experiments

Table 4 presents the experimental results obtained after adding some of the components to the baseline method MaPLe. The performance of MaPLe is shown in the first row.

https://doi.org/10.52202/079017-1238

Table 3: Comparison of CDC with recent approaches in the cross-domain setting.

	Source	**Target**				
	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R	Avg.
CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
Co-CoOp	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe	70.72	64.07	49.15	**50.90**	76.98	60.28
CDC	**71.76**	**64.87**	**50.33**	50.40	**78.10**	**60.93**

Table 4: The results of the ablative experiments on base-to-novel generalization setting.

M	DSTC	VSD		Base	New	HM
		Image	Text			
1	×	×	×	82.28	75.14	78.55
4	×	×	×	82.96	76.06	79.36
4	✓	×	×	82.65	76.78	79.61
4	✓	✓	×	**83.35**	76.72	79.90
4	✓	×	✓	82.93	76.83	79.77
4	✓	✓	✓	83.34	**77.38**	**80.25**

The effectiveness of multiple templates and DSTC. In Table 4, all results except the baseline method are obtained by increasing the number of templates to four. In the second row, we train the learnable tokens within the templates using the conventional cross-entropy loss and average the predictions from different templates to generate the final classification results. Compared to the baseline, multiple templates lead to a 0.68% improvement in the performance of base classes and a 0.92% enhancement in the performance of new classes, resulting in an overall HM gain of 0.81%. While modeling the uncertainty of each template, as shown in the third line, the performance for new classes achieves an improvement of 0.72%, and HM increases by 0.25%. Without explicitly decoupling the semantics within different templates, increasing the number of templates effectively boosts the performance. We argue that the semantics acquired by distinct templates, each with varied initialization, is inherently independent (See analysis in **Appendix C**). Therefore, simply applying multiple templates guided by the front-door adjustment could enhance the performance.

The effectiveness of VSD. The fourth and fifth rows of Table 4 present the experimental results obtained by decoupling semantics solely within the image and text branches, respectively. When compared to the third row, it becomes evident that decoupling semantics at either branch alone does not yield a substantial improvement in the accuracy of new classes. However, upon comparing the sixth row with the third row, it is observed that when both decoupling methods are employed concurrently, the model achieves a 0.69% increase in base class performance and a 0.60% improvement in novel class performance, leading to an overall HM enhancement of 0.64%. Despite the inherent decoupling of semantics among different templates, the utilization of VSD further promotes the learning of distinct semantics for each template, thereby enhancing the model's overall performance.

6 Conclusion and Future Discussion

Through motivating experiments, we identify the two-level misalignment that exists when CLIP is applied to downstream tasks. We further show that soft prompt tuning may worsen the second misalignment due to overfitting to the base classes, which impairs the generalization of CLIP. To analyze this problem, we propose an SCM and discover that the task-irrelevant generative factors serve as the confounder while estimating the true causal relationship between images and label space of new classes. To address this issue, we propose to implement front-door adjustment via CDC. Specifically, we introduce multiple templates to represent the decoupled semantics, then leverage VSD to further facilitate the decoupling of semantics, and finally fuse predictions from different templates via DSTC. The results under multiple experimental settings demonstrate that our proposed approach effectively improves the generalization of CLIP when adapted to downstream tasks.

Limitations and broader impacts. CDC incurs more time consumption, as applying different templates in the image branch requires encoding an image multiple times during both training and testing. The analysis of two-level misalignment in CLIP and the modeling of SCM is inspiring for the community.

https://doi.org/10.52202/079017-1238

Acknowledgments

The authors gratefully acknowledge the valuable feedback provided by the anonymous reviewers. This work is supported by the Fundamental Research Program, China, Grant No.JCKY2022130C020.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021.

[3] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021.

[4] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV*, volume 13695 of *Lecture Notes in Computer Science*, pages 493–510. Springer, 2022.

[5] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE, 2022.

[6] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19113–19122. IEEE, 2023.

[7] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15144–15154. IEEE, 2023.

[8] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. 2024.

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3606–3613. IEEE Computer Society, 2014.

[10] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, pages 96–146, 2009.

[11] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132(2):581–595, 2024.

 https://doi.org/10.52202/079017-1238

[13] Fang Peng, Xiaoshan Yang, Linhui Xiao, Yaowei Wang, and Changsheng Xu. Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *IEEE Trans. Multim.*, 26:3469–3480, 2024.

[14] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15613–15623. IEEE, 2023.

[15] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, volume 13693 of *Lecture Notes in Computer Science*, pages 709–727. Springer, 2022.

[16] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. *CoRR*, abs/2208.08340, 2022.

[17] Sanjoy Chowdhury, Sayan Nag, and Dinesh Manocha. Apollo : Unified adapter and prompt learning for vision language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10173–10187. Association for Computational Linguistics, 2023.

[18] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. *CoRR*, abs/2311.16494, 2023.

[19] Chun-Hao Chang, George-Alexandru Adam, and Anna Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15212–15221. Computer Vision Foundation / IEEE, 2021.

[20] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15404–15414. Computer Vision Foundation / IEEE, 2021.

[21] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10797–10806. Computer Vision Foundation / IEEE, 2020.

[22] Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. Counterfactual samples synthesizing and training for robust visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13218–13234, 2023.

[23] Yinjie Jiang, Zhengyu Chen, Kun Kuang, Luotian Yuan, Xinhai Ye, Zhihua Wang, Fei Wu, and Ying Wei. The role of deconfounding in meta-learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10161–10176. PMLR, 2022.

[24] Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Bing Su, and Hui Xiong. Interventional contrastive learning with meta semantic regularizer. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 18018–18030. PMLR, 2022.

[25] Jiangmeng Li, Yanan Zhang, Wenwen Qiang, Lingyu Si, Chengbo Jiao, Xiaohui Hu, Changwen Zheng, and Fuchun Sun. Disentangle and remerge: Interventional knowledge distillation for few-shot object detection from A conditional causal perspective. 2023.

[26] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[27] Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, and William Yang Wang. Causal balancing for domain generalization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[28] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[29] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[30] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(2):2551–2566, 2023.

[31] Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[32] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 57–72. Springer, 2008.

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[34] Li Fei-Fei, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007.

[35] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3498–3505. IEEE Computer Society, 2012.

[36] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society, 2013.

[37] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729. IEEE Computer Society, 2008.

[38] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer, 2014.

[39] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

[40] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492. IEEE Computer Society, 2010.

[41] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019.

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[43] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June*

 https://doi.org/10.52202/079017-1238

2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019.

[44] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10506–10518, 2019.

[45] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15262–15271. Computer Vision Foundation / IEEE, 2021.

[46] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8320–8329. IEEE, 2021.

[47] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 5749–5757. AAAI Press, 2024.

A Hand-Crafted Prompts

To prevent the text embeddings obtained via the learned prompts from deviating from the semantics of their respective categories, we introduce a consistency loss $\mathcal{L}_{con}$ to apply constraints on the learned prompts. For each dataset, we borrow from CLIP a hand-crafted prompt t^0 to generate w^0. The prompts employed in this process are detailed below.

*"**Pets**": "a photo of a {}, a type of pet."*

*"**Flowers**": "a photo of a {}, a type of flower."*

*"**Aircraft**": "a photo of a {}, a type of aircraft."*

*"**DTD**": "a photo of a {}, a type of texture."*

*"**SAT**": "a centered satellite photo of {}."*

*"**Cars**": "a photo of a {}."*

*"**Food**": "a photo of {}, a type of food."*

*"**SUN**": "a photo of a {}."*

*"**Caltech**": "a photo of a {}."*

*"**UCF**": "a photo of a person doing {}."*

*"**ImageNet**": "a photo of {}."*

B Evaluation Protocol

Base-to-New Setting. Following CoCoOp and MaPLe, we split the classes into two un-overlapping subsets, i.e., base classes and new classes, for each dataset. In training, base classes are leveraged for learning the prompts under the 16-shot setting, where each class contains 16 annotated samples. Subsequently, the prompts are transferred to new classes which are unseen during the training. In this setting, the evaluation metrics include accuracy on base classes, new classes, and their corresponding harmonic mean (HM), which can be calculated by:

$$HM = \frac{1}{\frac{1}{2}(\frac{1}{\text{Base}} + \frac{1}{\text{New}})}, \tag{12}$$

where Base and New denote the accuracy of base classes and new classes, respectively.

Out-of-Distribution Setting. In the OOD setting, we initially train the model using the source dataset and then transfer the learned prompts to the target dataset to evaluate the robustness of the method. In our experiments, following COOP, CoCoOp and MaPLe, the source dataset is ImageNet, employing all 1000 of its classes for prompt learning, where each class comprises 16 annotated samples. In the cross-dataset setting, 10 datasets, Caltech, Pets, Cars, Flowers, Food, Aircraft, SUN, DTD, EuroSAT, and UCF, are target datasets. In the cross-domain setting, the target datasets include ImageNetV2, ImageNet-S, ImageNet-A, and ImageNet-R, all sharing the same classes as ImageNet but differing in distributions.

C Additional Experiments

C.1 Analysis of the Hyper-Parameters

The number of templates M. Table 5 presents the performance of the model in the base-to-new setting and the corresponding computational complexity when the number of templates changes from 1 to 8. The results in Table 5 are obtained solely using $\mathcal{L}_{t\text{-}ce}$, without the help of VSD. As the number of templates increases, the performance of the model on the base classes and the new classes increases consistently, while at the same time, the computational overhead increases. Although better performance can be achieved when M is further increased to 8, we select $M = 4$ in order to balance the accuracy of classification with the computational overhead.

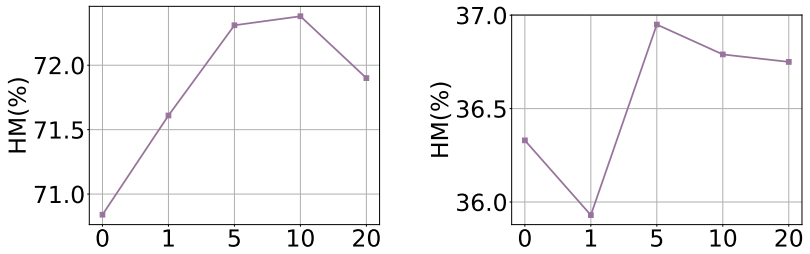
Weight β for $\mathcal{L}_{de}$. Figure 4 gives the HM on both DTD and Aircraft datasets when varying the weights β of $\mathcal{L}_{de}$. From the figure, it is evident that using $\beta = 5$ and $\beta = 10$ effectively improves

https://doi.org/10.52202/079017-1238

Table 5: Comparison of the performance under different template numbers.

M	Base	New	HM	Params	FPS
1	80.87	73.49	77.00	3.55M	600
2	82.80	76.59	79.57	7.10M	280.00
4	82.65	76.78	79.61	14.20M	185.71
8	83.40	77.29	80.23	28.40M	65.63

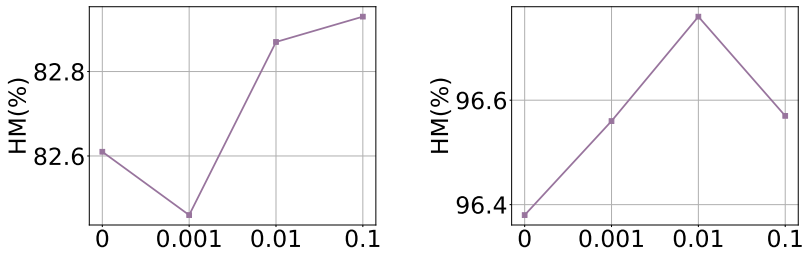
the performance of the model on these two datasets as compared to training without $\mathcal{L}_{de}$, i.e., $\beta = 0$. When β is further increased to 20, the performance of the model begins to decline. Considering the performance on each dataset, in the base-to-new setting, we set $\alpha = 5$.



(a) DTD (b) FGVC Aircraft

Figure 4: The impact of β on performance.

Weight γ for $\mathcal{L}_{con}$. In Figure 5, we provide the experimental results obtained by varying γ when keeping $\beta = 5$. From the results, we observe that setting $\gamma = 0.01$ enables the performance of both datasets to obtain an enhancement compared to $\gamma = 0.0$. When γ is further increased to 0.1, the performance on the Caltech101 dataset shows a significant decrease. Considering all results together, we set $\gamma = 0.01$ in all base-to-new experiments.



(a) UCF101 (b) Caltech101

Figure 5: The impact of γ on performance.

Augmentation methods. Setting $M = 4$, $\beta = 5$, $\gamma = 0.01$, we vary the augmentation methods to explore the effect of different augmentations on model performance and demonstrate the results in Table 6. We set four sets of augmentations for 4 templates. Specifically, for **Augmentation 1**, the augmentation methods include: (i) random crop and random flip; (ii) color jitter and random flip; (iii) random translation and random flip; (iv) random augmentation. For **Augmentation 2**, the four augmentation sets are: (i) random crop and random flip; (ii) random crop, random flip, and color jitter; (iii) random crop, random translation, and random flip; (iv) random crop and other random augmentations. Augmentation 2 adds the random crop to each augmentation set based on Augmentation 1. From Table 6, we notice that for most of the datasets such as ImageNet, Caltech101, etc., there is no great difference in the performance obtained by Augmentation 1 and Augmentation 2.

https://doi.org/10.52202/079017-1238

Table 6: The impact of augmentation methods on the performance.

Dataset	Augmentation 1			Augmentation 2		
	Base	New	HM	Base	New	HM
ImageNet	77.50	71.73	**74.51**	77.30	71.57	74.32
Caltech	98.63	93.67	96.09	98.20	94.37	**96.25**
Pets	95.97	97.67	96.81	96.07	98.00	**97.02**
Cars	72.33	71.63	71.98	73.80	73.97	**73.88**
Flowers	96.93	75.07	**84.61**	96.93	74.67	84.36
Food	90.93	92.23	91.58	90.87	92.33	**91.59**
Aircraft	36.83	36.23	36.53	37.47	37.50	**37.48**
SUN	82.37	80.03	**81.18**	82.00	79.77	80.87
DTD	82.70	64.10	**72.22**	82.13	61.73	70.49
SAT	95.10	82.33	**88.26**	92.87	75.83	83.49
UCF	85.70	81.73	**83.67**	85.40	81.23	83.26

However, for *Cars* and *Aircraft*, Augmentation 2 performs significantly better than Augmentation 1. Capturing discriminative information is challenging for fine-grained classification. We argue that the translation invariance and invariance to the object scale brought by random crop may be critical, so Augmentation 2 performs significantly better than Augmentation1 on these two datasets.

C.2 Template Study

In Table 7, we provide the full results of our method under all templates, as well as the final performance after performing the evidence fusion which is denoted by CDC. From Table 7, CDC performs the best in 10 out of 11 datasets. Among them, several datasets gain essential improvements compared to the results under a single template, including ImageNet, Flowers, SUN, and DTD. This suggests that the essence of our approach is not to search for better discriminative semantics by increasing the number of templates but to eliminate the impacts of task-independent generative factors by combining multiple sets of decoupled semantics. Despite the moderate performance of any individual template, our method achieves a substantial performance enhancement by fusing them.

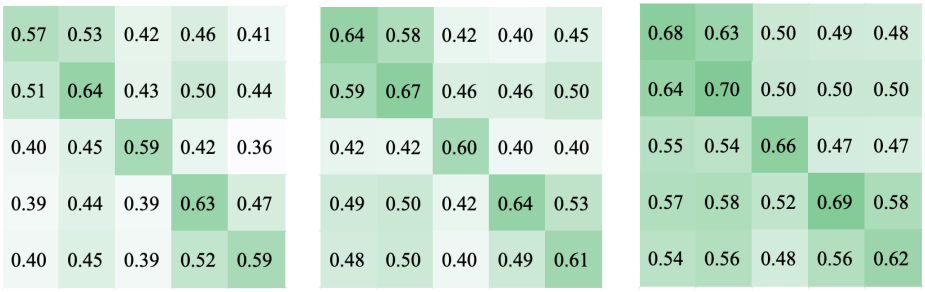
Table 7: The comparison with the results from different templates.

Dataset	S^1	S^2	S^3	S^4	CDC
ImageNet	70.10	70.10	69.73	70.33	**71.73**
Caltech	94.10	**94.73**	94.57	94.17	94.37
Pets	97.60	97.60	97.37	97.87	**98.00**
Cars	72.93	73.57	70.87	72.63	**73.97**
Flowers	72.87	73.23	74.10	71.43	**75.07**
Food	91.60	91.40	91.73	91.73	**92.33**
Aircraft	34.77	35.33	35.70	35.60	**37.50**
SUN	78.53	77.87	77.73	78.17	**80.03**
DTD	59.60	57.83	57.87	60.43	**64.10**
SAT	73.73	75.67	74.57	73.73	**82.33**
UCF	77.83	76.00	77.60	78.13	**81.73**

C.3 Intrinsic Decoupling of Semantics from Different Templates

In **Section 5.4**, we demonstrate that increasing the number of templates can get a performance boost even without explicitly decoupling the semantics contained in different templates. We attribute this to the fact that templates based on different initializations do not learn exactly the same semantics, i.e., the semantics contained in different templates are intrinsically decoupled. To verify this suspicion, we collect the text embeddings obtained based on different templates without applying VSD, and compute the similarities between the embedding obtained from different templates. We plot the results of the DTD dataset in Figure 6. As can be seen from the figure, even for the same category,

https://doi.org/10.52202/079017-1238



(a) (b) (c)

Figure 6: The cosine similarities of the text embeddings under different templates based on the new classes of DTD. (a) The cosine similarities between the text embeddings generated by the 0-th template and the 1-th template. (b) The cosine similarities between the text embeddings generated by the 1-th template and the 2-th template. (c) The cosine similarities between the text embeddings generated by the 2-th template and the 3-th template.

the similarities between the text embedding obtained by different templates are still much less than 1 (i.e., the value on the diagonal line in the figure). This indicates that the text embedding obtained by different templates has a large discrepancy, which confirms our idea that the semantics learned by different templates are inherently decoupled. Therefore, better performance can be obtained through fusing results generated by these decoupled features.

https://doi.org/10.52202/079017-1238

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main contributions include: (i) we identify the problem of two-level misalignment in Vision-Language Model adaptation; (ii) we propose an SCM to analyze the spurious correlations between images and the label space caused by the data misalignment; (iii) we propose CDC to mitigate the spurious relationships while estimating the true causal relationships. These contributions are covered in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in terms of computational efficiency in **Section 6**.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

https://doi.org/10.52202/079017-1238

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not include theorems and formulas in the paper. We analyze the misalignment problem from a causal perspective, and the main assumptions and corresponding analysis are provided in **Section 3.2**.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all the details of our method in **Section 4** and present the necessary hyper-parameters in **Section 5.1** to ensure the reproducibility of our method. Furthermore, we provide the code of our proposed method in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We include the code of our proposed method in the supplementary material. The necessary environments and data preparation procedures are provided in the GitHub repository of our baseline method MaPLe.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the necessary hyper-parameters in **Section 5.1**, and analyze the hyper-parameters in **Appendix C.1**.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We don't report the error bar in our experimental results because the previous works do not report the error bar and we follow them. All of our experimental results are averaged over 3 runs of 3 different seeds.

https://doi.org/10.52202/079017-1238

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the compute workers in **Section 5.1**, and list the params and time of execution in Table 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conduct the research with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We provide the broader impacts of our work in **Section 6**.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific sets), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not pose such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets, models, and code involved in our paper are open source.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

https://doi.org/10.52202/079017-1238

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1238