

ED-SAM: An Efficient Diffusion Sampling Approach to Domain Generalization in Vision-Language Foundation Models

Thanh-Dat Truong¹, Xin Li², Bhiksha Raj^{3,4}, Jackson Cothren⁵, Khoa Luu¹

¹CVIU Lab, University of Arkansas, USA ²University at Albany, Albany NY, USA

³Carnegie Mellon University, USA ⁴Mohammed bin Zayed University of AI, UAE

⁵Dep. of Geosciences, University of Arkansas, USA

{tt032, jcothre, khoaluu}@uark.edu bhiksha@cs.cmu.edu, xli48@albany.edu

Abstract

The Vision-Language Foundation Model has recently shown outstanding performance in various perception learning tasks. The outstanding performance of the vision-language model mainly relies on large-scale pre-training datasets and different data augmentation techniques. However, the domain generalization problem of the vision-language foundation model needs to be addressed. This problem has limited the generalizability of the vision-language foundation model to unknown data distributions. In this paper, we introduce a new simple but efficient Diffusion Sampling approach to Domain Generalization (ED-SAM) to improve the generalizability of the vision-language foundation model. Our theoretical analysis in this work reveals the critical role and relation of the diffusion model to domain generalization in the vision-language foundation model. Then, based on the insightful analysis, we introduce a new simple yet effective Transport Transformation to diffusion sampling method. It can effectively generate adversarial samples to improve the generalizability of the foundation model against unknown data distributions. The experimental results on different scales of vision-language pre-training datasets, including CC3M, CC12M, and LAION400M, have consistently shown State-of-the-Art performance and scalability of the proposed ED-SAM approach compared to the other recent methods.

1 Introduction

The vision-language foundation models trained based on contrastive learning and exemplified by CLIP [60], have gained more attention due to their outstanding performance on various tasks. Although the vision-language foundation models have shown advantages on various downstream visual tasks, limited studies investigate their generalizability. Meanwhile, the generalizability of the foundation models still majorly relies on the large-scale pre-training datasets. While many prior studies [20, 26, 83, 42, 45, 87, 3, 78, 5] have been introduced to domain generalization for classification [34, 83, 75, 2, 41], detection [74, 43], semantic segmentation [88, 35, 10, 29], there are limited studies that address the domain generalization problem in the vision-language founda-

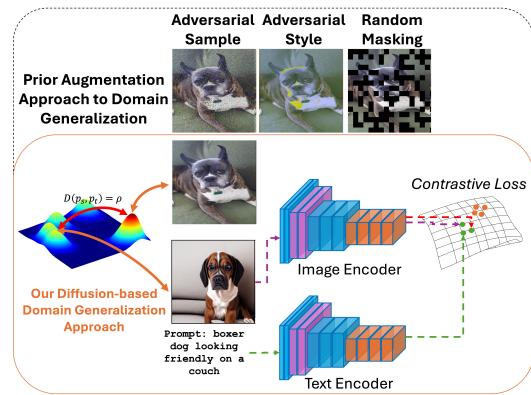


Figure 1: Comparison between Our Proposed Diffusion-based Domain Generalization with Prior Methods [75, 88, 39].

tion model. Despite being trained on a large-scale dataset, the generalizability of the vision-language foundation model has to be considered because it is a key factor in guaranteeing the performance of models against unknown data distributions. The domain generalization approaches are urgently needed for foundation model training to ensure optimal performance and generalizability. The current vision-language foundation models trained using contrastive learning often rely on data augmentations to improve their robustness and prevent overfitting. However, these methods are not effective enough to improve the generalization of the foundation model. In particular, to improve the performance of CLIP models, most of the prior visual foundation models perform the data augmentation on visual inputs [39, 48, 60, 50, 31] to increase the number of training samples and create challenging samples. These augmentation methods aim to increase the diversity of the data, thus enhancing the generalization of the foundation models. However, these visual augmentations concentrate on pixel-level modification like masking, adversarial perturbations, adversarial styles, or color jittering, which have a limited impact on enriching the semantic information of visual concepts. Therefore, the generalizability to unknown data distributions of vision-language models remains limited.

In recent years, in parallel with the development of vision-language models, the diffusion model has shown its outstanding performance in data distribution modeling and generative AI. The diffusion approach, designed based on the nonequilibrium thermodynamics [21], is able to model the data distribution via the parameterized Markov chain trained using variational inference. Hence, the diffusion models can synthesize novel, high-quality, and complex data. Moreover, the diffusion models are also able to efficiently model the conditional data distributions, e.g., text-to-image diffusion [62]. Inspired by the success of diffusion, this paper fundamentally investigates its role and relation to the generalizability of the vision-language foundation models. In particular, we first model the domain generalization of the vision-language foundation model via the worst-case formula over data distributions that are near the training domain of the latent space. Then, using the Lagrangian relaxation and diffusion properties, we introduce a novel transformation approach that improves the generalizability of the vision-language foundation model by expanding training data distribution via the diffusion model. Our theoretical analysis has shown the proposed approach is robust and well-generalized. It also has a better domain generalization compared to prior methods [75, 88].

Contributions: This paper introduces a novel *Diffusion-based Domain Generalization* approach, a simple yet effective approach to improving the generalizability of the vision-language model, i.e., CLIP, by exploiting the power of the diffusion model (Fig. 1). In particular, first, we form the domain generalization problem of the vision-language model via the worst-case formula over the training data distribution. By modeling the data conditional distribution via the diffusion model, we further provide a complete theoretical analysis of the relation of the diffusion model to adversarial augmentation. Second, we introduce a new simple yet efficient *Transport Transformation* to diffusion sampling that can synthesize adversarial samples to improve the generalizability of the vision-language model. Thanks to our proposed Transport Transformation, our approach efficiently expands the training data distributions, therefore improving the ability to generalize to unseen data distributions of the vision-language model. Finally, our extensive experiments on pre-training vision-language datasets at different scales, including CC3M, CC12M, and LAION400M, have shown the robustness of the proposed approach. Our approach has improved the performance of CLIP significantly on various benchmarks and outperformed other augmentation and domain generalization approaches. The theoretical analysis and empirical results guarantee that the proposed approach is simple yet scalable and contributes to the generalizability improvement of vision-language foundation models.

2 Related Work

Vision-Language Foundation Model The contrastive language-image training [60, 31, 84, 46, 79, 8] has become a prominent approach in developing the large-scale vision-language model [60, 31]. CLIP [60] and ALIGN [31] first introduced contrastive learning to learn strong representations of images and texts for cross-modal alignment. CoCa [84] proposed an additional decoder and generative image captioning. SLIP [48], DeCLIP [40], FLIP [39] further improve the performance by using self-supervised training techniques. LaCLIP [12] improved the performance of CLIP by introducing text augmentation via the large language model. LiT [86] and BASIC [56] improve the zero-shot transfer ability via further fine-tuning the language encoder. SimVLM [80], OFA [77], and BLIP [37] train the vision-language model within an encoder-decoder framework with language generative losses. SigLIP [85] proposed a Sigmoid loss to compute the image-text similarity.

Denoising Diffusion Probabilistic Model (DDPM) has achieved state-of-the-art performance in density estimation and image synthesis [21, 62]. The DDPM model defines a Markov chain of diffusion steps to gradually add random noise to data followed by learning to reverse the diffusion process via the UNet [21] to construct the data sample from noise. Subsequent studies further improved by reweighing the learning objective [33], improving the variance schedule [51], using distillation [47]. Denoising diffusion implicit models (DDIM) [69] was introduced to accelerate the sampling process by generalizing DDPMs. Meanwhile, the Latent Diffusion Model (LDM) [62] proposed a two-stage diffusion model where the diffusion process is performed on the latent space. Other approaches improve the DDPMs by introducing cascaded generation [22], incorporating with GANs [81], using wavelet transformation [57], and introducing momentum-based diffusion [11, 28]. The diffusion model also has an ability of conditional synthesis, e.g., text-to-image [62, 61, 65], image editing [32, 49]. This conditional ability can be implemented as explicit conditions [62], classifier guidance [52, 11, 9], or classifier-free guidance [23]. The later studies further improve diffusion models by introducing a single-step diffusion [24], subject-driven fine-tuning [63].

Domain Generalization aims to learn a robust model from single or multiple source data so that the model can later be well generalized to unseen data domains. One stream of the domain generalization approach focuses on using data augmentation to improve the generalizability of the model [39, 25]. Recent studies adopted image masking [39] or the image-editing technique or style transfer via the diffusion model to improve the performance of object classification [72, 18, 16, 30], object detection [14], or 3D classification [68]. Another stream of domain generalization focuses on learning the invariant feature space by jointly optimizing a multi-domain autoencoder [19, 36], removing domain specific via normalization [73, 13, 70], learning in the frequency domain [76, 27, 44]. Adversarial training has been introduced to learn the robust model by forming novel domains via the generated adversarial samples. Adversarial Data Augmentation (ADA) [75] first introduced an approach to generate adversarial samples via max-min iterative training. Later, M-ADA [59] further improved ADA by training an additional autoencoder. Other approaches learn the domain-invariant features with adversarial samples via meta-learning [17, 58], or image-style adversarial learning [88]. In addition, another domain generalization approach improves the generalization ability by re-designing the deep neural network [38, 54] or using an ensemble of expert models [1]. To the best of our knowledge, *these prior studies have not fully investigated the fundamentals of diffusion to domain generalization of the vision-language foundation models. Therefore, in this paper, we provide a theoretical analysis of diffusion to the generalizability of the vision-foundation model, followed by proposing a new simple yet efficient diffusion-based domain generalization approach.*

3 Theoretical Analysis of Generalizability in Foundation Model

3.1 Preliminary

Diffusion Model formulates the data distribution $p(\mathbf{x})$ by gradually denoising a normally distributed variable via the reverse process of a fixed Markov Chain of length T , i.e., $p(\mathbf{x}_0) = \int p(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$, with a Gaussian transition starting at $p(\mathbf{x}_T) = p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. The diffusion model includes the forward and backward processes. The forward diffusion process, i.e., $q(\mathbf{x}_i|\mathbf{x}_{i-1})$ is defined as:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{i=1}^T q(\mathbf{x}_i|\mathbf{x}_{i-1}) \quad q(\mathbf{x}_i|\mathbf{x}_{i-1}) = \mathcal{N}(\mathbf{x}_i, \sqrt{1 - \beta_i} \mathbf{x}_{t-1}, \beta_i \mathbf{I}) \quad (1)$$

where β_i is a variance schedule. Then, the backward process, i.e., $p(x_{k-1}|x_{k-1})$, is defined as:

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{i=1}^T p(\mathbf{x}_{i-1}|\mathbf{x}_i) \quad p(\mathbf{x}_{i-1}|\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_{i-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_i, i), \boldsymbol{\Sigma}_\theta(\mathbf{x}_i, i)) \quad (2)$$

The backward process adopts a denoising model ϵ_θ to predict the denoised variant from \mathbf{x}_i . Then, the model is learned via the usual variational bound on negative log-likelihood is as follows:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{\mathbf{x}, \epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I}), i} [\|\epsilon - \epsilon_\theta(\mathbf{x}_i, i)\|_2^2] \quad (3)$$

where θ is the parameter of ϵ , $\mathbf{x}_i = \sqrt{\alpha_i} \mathbf{x} + \sqrt{1 - \alpha_i} \epsilon$, $\alpha_i = 1 - \beta_i$, $\bar{\alpha}_i = \prod_{s=1}^i \alpha_s$, and i is uniformly sampled from 1 to T , i.e., $i \in \mathcal{U}(1, T)$. The diffusion model is capable of modeling the conditional distribution, i.e., $p(\mathbf{x}|\mathbf{p})$ where \mathbf{p} is the condition (e.g., a text prompt). This ability can be done by implementing a conditional denoising model $\epsilon_\theta(\mathbf{x}_i, i, \mathbf{p})$.

Contrastive Language-Image Pretraining (CLIP) [60] has shown its outstanding performance in the training vision-language foundation model using language supervision. Formally, let $\mathbf{x}, \mathbf{p} \sim p(\mathbf{x}, \mathbf{p})$ be the source training data of the CLIP model where \mathbf{x} is the image, and \mathbf{p} is the corresponding prompt, $F^{\mathbf{x}}$ and $F^{\mathbf{p}}$ be the vision and language encoder, and $\mathbf{f}^{\mathbf{x}}$ and $\mathbf{f}^{\mathbf{p}}$ be the features extracted by the vision and language encoder, respectively, i.e., $\mathbf{f}^{\mathbf{x}} = F^{\mathbf{x}}(\mathbf{x})$ and $\mathbf{f}^{\mathbf{p}} = F^{\mathbf{p}}(\mathbf{p})$. The CLIP model is learned via contrastive loss, where the pairs of images and corresponding texts are the positive pairs. The CLIP model can be formulated as follows:

$$\theta_{F^{\mathbf{x}}}^*, \theta_{F^{\mathbf{p}}}^* = \arg \min_{\theta_{F^{\mathbf{x}}}, \theta_{F^{\mathbf{p}}}} \mathbb{E}_{\mathbf{x}, \mathbf{p}_t \sim p(\mathbf{x}, \mathbf{p})} - \log \frac{\exp(\text{sim}(F^{\mathbf{x}}(\mathbf{x}), F^{\mathbf{p}}(\mathbf{p})) / \tau)}{\sum_k \exp(\text{sim}(F^{\mathbf{x}}(\mathbf{x}), F^{\mathbf{p}}(\mathbf{p}^k)) / \tau)} \quad (4)$$

where $\theta_{F^{\mathbf{x}}}, \theta_{F^{\mathbf{p}}}$ are parameters of $F^{\mathbf{x}}$ and $F^{\mathbf{p}}$, \mathbf{p}^k is the negative text sample of \mathbf{x} , τ is the temperature to scale logits, sim is the dot product to measure distance between features. For simplicity, Eqn. (4) only illustrates the contrastive loss over images. In practice, a symmetrical loss over texts is also applied, and the loss is the average of the contrastive loss over images and texts.

3.2 Domain Generalization of Contrastive Language-Image Pre-Training

In our paper, we aim to develop a domain generalization approach to CLIP that is able to better generalize to new unknown data distributions. In this work, we consider the training data of CLIP drawn from a single source data [75], i.e., $\mathbf{x}_s, \mathbf{p}_s \in p(\mathbf{x}_s, \mathbf{p}_s)$. Inspired by prior work in robust optimization, we propose to model the domain generalization of CLIP via the worst-case problem around the source data distribution $p(\mathbf{x}_s, \mathbf{p}_s)$ as follows:

$$\theta_{F^{\mathbf{x}}}^*, \theta_{F^{\mathbf{p}}}^* = \arg \min_{\theta_{F^{\mathbf{x}}}, \theta_{F^{\mathbf{p}}}} \sup_{p_t: \mathcal{D}(p_t, p_s) \leq \rho} \mathbb{E}_{\mathbf{x}_t, \mathbf{p}_t \sim p_t(\mathbf{x}_t, \mathbf{p}_t)} - \log \frac{\exp(\text{sim}(F^{\mathbf{x}}(\mathbf{x}_t), F^{\mathbf{p}}(\mathbf{p}_t)) / \tau)}{\sum_k \exp(\text{sim}(F^{\mathbf{x}}(\mathbf{x}_t), F^{\mathbf{p}}(\mathbf{p}^k)) / \tau)} \quad (5)$$

where $\mathbf{x}_t, \mathbf{p}_t$ are images and prompt sampled from p_t , $\mathcal{D}(p_s, p_t)$ is the Wasserstein metric measure the distance between two data distributions p_s and p_t , ρ is the distance constraint, p_t is ρ -away unknown data distributions from p_s , i.e., $\mathcal{D}(p_s, p_t) \leq \rho$. Eqn. (5) aims to guarantee good performance of the CLIP model against the unknown data distribution.

Domain Generalization of CLIP In our paper, we are interested in the problem of domain generalization of CLIP where we aim to improve the performance of CLIP, especially when using the CLIP model for downstream tasks, e.g., zero-shot classification, linear probing, or fine-tuning. In this learning scenario, since the target data distribution p_t is completely unknown, the hyper-parameter ρ plays an important role since it will indicate the generalizability of the CLIP model to new data (or test domains). To solve the Eqn (5), the Lagrange multiplier can be adopted to reform Eqn (5) as:

$$\mathbf{x}_t^* = \arg \max_{\mathbf{x}_t} \left\{ \mathcal{L}_{CLIP}(\mathbf{x}_t, \mathbf{p}_s) - \lambda \mathcal{D}(p_t(\mathbf{x}_t, \mathbf{p}_s), p_s(\mathbf{x}_s, \mathbf{p}_s)) \right\} \quad (6)$$

where \mathbf{x}_t^* is the **adversarial sample** (corresponding to prompt \mathbf{p}_s) to improve the generalization and robustness of the CLIP model, \mathcal{L}_{CLIP} is the contrastive language-image pretraining loss defined in Eqn. (4), λ is the hyper-parameter that is inverse proportional of ρ , $\mathcal{D}(p_t(\mathbf{x}_t, \mathbf{p}_s), p_s(\mathbf{x}_s, \mathbf{p}_s))$ is the transportation cost that moving from the $\mathbf{x}_s, \mathbf{p}_s \sim p_s(\mathbf{x}_s, \mathbf{p}_s)$ to the distribution p_t . Since our paper aims to improve the generalizability of the CLIP model on the downstream vision tasks, the scope of this work focuses on the adversarial sample in the vision domain. Eqn. (6) aims to create augmented samples so that the distribution of augmented samples is ρ -away from the original one and increases contrastive learning loss. Then, using these augmented samples will potentially improve the generalizability of CLIP.

Limitation of Prior Work Prior work adopts adversarial training [75], augmentation methods [39], or adversarial style augmentation [88] to generate adversarial/augmented samples to improve the generalizability. Although prior results have shown the potential performance improvement, these approaches remain limited in terms of expanding their generalizability to unknown distributions. Indeed, adversarial learning [75, 88] tries to add the perturbation via maximizing loss or adversarial styles into images. Meanwhile, the augmentation methods create different variations of images by performing heuristic pixel-wise image operations (e.g., masking, cropping, color jittering, etc.). However, the data distributions of augmented samples generated by prior methods [75, 88, 39] remain unchanged or little changed compared to the original data distribution. This can be explained since, despite the different variations of augmented samples, the content information, e.g., object appearances, shapes, etc., and the semantic background information remained the same. For example, as shown in Fig. 2, the target object of augmented samples created by [75, 88, 39] remains unchanged. The semantic background, in general, is similar to the original image, with noise added.

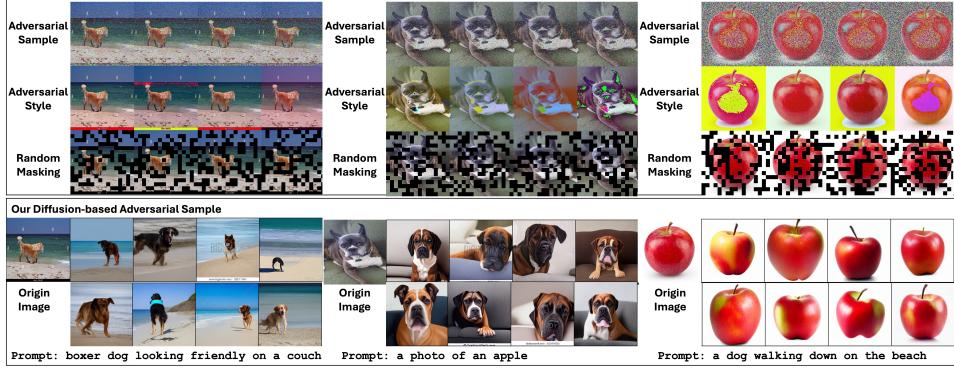


Figure 2: The Comparison of Our Diffusion-based Adversarial Sample and Prior Augmentations (Adversarial Sample [75], Adversarial Style [88], Masking Sample [39]).

3.3 The Relation of Diffusion to Adversarial Augmentation

As aforementioned, the goal of the adversarial sample in Eqn. (6) is to move the data sample from the source training \mathbf{x}_s to the \mathbf{x}_t^* in the ρ -away distribution so that maximize the contrastive language-image pretraining loss $\mathcal{L}_{CLIP}(\mathbf{x}_t, p_s)$. As shown in Eqn. (6), the sample \mathbf{x}_t^* is depending on the source training sample \mathbf{x}_s , the text prompt p_s , and the distance between two distributions ρ . Therefore, in our work, we consider the adversarial sample \mathbf{x}_t^* is draw from a ρ -away distribution conditioned on \mathbf{x}_s , p_s , and ρ , i.e., $\mathbf{x}_t^* \in p(\mathbf{x}_t^* | \mathbf{x}_s, p_s, \rho)$.

The Source Data Distribution Since the image \mathbf{x}_s and p_s is a pair of image and text, without a strict argument, we could assume that the image \mathbf{x}_s is conditioned on the text prompt p_s , i.e., $\mathbf{x}_s \in p(\mathbf{x}_s | p_s)$. As presented in Sec. 3.1, the conditional distribution $p(\mathbf{x}_s | p_s)$ could be efficiently modeled by the diffusion. Let $\mathbf{z}_s \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ be the latent variable of image \mathbf{x}_s . Then, the image \mathbf{x}_s can be modeled via the backward process of diffusion conditioned on p_s and \mathbf{z}_s as in Eqn. (2) For simplicity, we rewrite the data distribution $\mathbf{x}_s \sim p(\mathbf{x}_{0:T} | p_s)$ via the latent variable \mathbf{z}_s as $\mathbf{x}_s \sim p(\mathbf{x}_s | \mathbf{z}_s, p_s)$.

The Diffusion-based Adversarial Augmentation Fig. 3 illustrates the graphical model that define the relation among \mathbf{x}_s , p_s , \mathbf{z}_s , \mathbf{x}_t^* , and ρ . The relation in this graphical model is established based on two conditions of the adversarial sample $p(\mathbf{x}_t^* | \mathbf{x}_s, p_s, \rho)$ and the conditional diffusion model $p(\mathbf{x}_s | \mathbf{z}_s, p_s)$. As shown in the graphical model, we have observed that the adversarial sample \mathbf{x}_t^* depends on $(\mathbf{x}_s, p_s, \rho)$ while the image \mathbf{x}_s is conditioned on (\mathbf{z}_s, p_s) . Therefore, for simplicity, without a strict argument, we assume that the adversarial sample \mathbf{x}_t^* is equivalently depending on $(\mathbf{z}_s, p_s, \rho)$ defined as follows

$$\mathbf{x}_t^* \sim p(\mathbf{x}_t^* | \mathbf{x}_s, p_s, \rho) \Rightarrow \mathbf{x}_t^* \sim p(\mathbf{x}_t^* | \mathbf{z}_s, p_s, \rho) \Rightarrow \mathbf{x}_t^* \sim p(\mathbf{x}_t^* | \mathbf{z}_s^*, p_s) \text{ where } \mathbf{z}_s^* = \mathcal{T}(\mathbf{z}_s, \rho) \quad (7)$$

where \mathcal{T} is the transport transformation on the latent space. Intuitively, instead of moving the image \mathbf{x}_s in the image space to \mathbf{x}_t^* in the new distribution with a transportation cost of $\mathcal{D}(p_t(\mathbf{x}_t^*, p_s), p_s(\mathbf{x}_s, p_s))$ as in Eqn. (6) which is a challenging problem, we are going to move the latent variable \mathbf{z}_s to \mathbf{z}_s^* via the transport function \mathcal{T} controlled by ρ . Since the latent space of the diffusion model is tractable (as it is a Gaussian distribution), moving \mathbf{z}_s to \mathbf{z}_s^* on latent space is controllable and easier than moving samples on the image space. Then, the adversarial sample of \mathbf{x}_t^* can be achieved by the reverse process of the diffusion model. With our proposed diffusion-based augmentation approach, thanks to the power of the diffusion model [62], our approach is able to synthesize novel adversarial samples that still maintain the semantic conditions on the prompt p_s while being effectively used to improve the generalizability in training CLIP model. As shown in Fig. 2, our proposed approach can generalize a new sample with the sample condition prompt, but the content and semantic background of the image have been changed significantly. This helps to strongly expand the data distribution during training to improve the generalizability of unknown data distribution.

3.4 The Proposed Transport Transformation

It is important to design a transformation \mathcal{T} that satisfies the condition of domain generalization, i.e., $\mathcal{D}(p_s(\mathbf{x}_s, p_s), p_t(\mathbf{x}_t, p_t)) \leq \rho$ in Eqn. (5), to guarantee the generalizability defined in Eqn. (5).

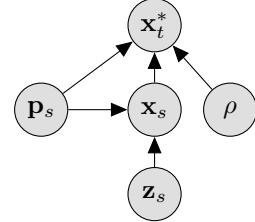


Figure 3: The Relation Between Adversarial Sample and Source Data.

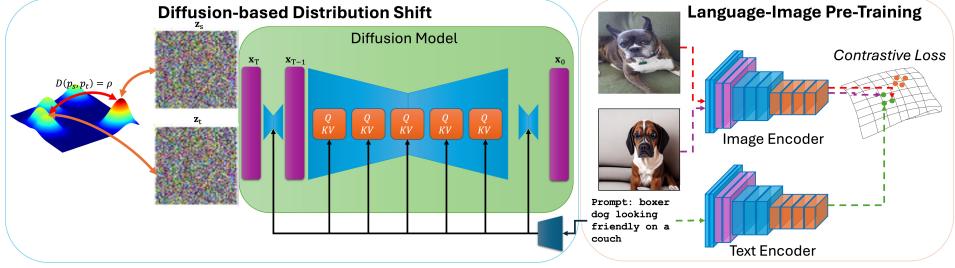


Figure 4: The Proposed Diffusion-based Domain Generalization Framework

Since the data distribution in our approach is displaced in the latent space of \mathbf{z}_s , with a strict argument, the condition of domain generalization via the latent space could be written as follows:

$$\mathcal{D}(p_s(\mathbf{x}_s, \mathbf{p}_s), p_t(\mathbf{x}_t^*, \mathbf{p}_t)) \propto \mathcal{D}(p_s(\mathbf{z}_s), p_t(\mathbf{z}_t^*)) \leq \rho \quad (8)$$

In our proposed approach, in order to meet the requirement as defined in Eqn. (8), the transport transformation \mathcal{T} can be defined as follows:

$$\mathbf{z}_t^* = \mathcal{T}(\mathbf{z}_s, \rho) = \frac{\mathbf{z}_s + \mathcal{N}(\alpha\sqrt{2}, \mathbf{I})}{\sqrt{2}} \quad \text{where } \alpha \sim \mathcal{U}(-\rho, \rho) \quad (9)$$

where α is controllable hyper-parameter uniformly sampled from $\mathcal{U}(-\rho, \rho)$.

Proposition 1: Given $\mathbf{z}_s \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\alpha (-\rho \leq \alpha \leq \rho)$, the condition of distance between distributions $\mathcal{D}(p_s(\mathbf{z}_s), p_t(\mathbf{z}_t^*)) \leq \rho$ holds if the transport transformation \mathcal{T} is defined as $\mathbf{z}_t^* = \mathcal{T}(\mathbf{z}_s, \rho) = \frac{\mathbf{z}_s + \mathcal{N}(\alpha\sqrt{2}, \mathbf{I})}{\sqrt{2}}$. The proof is provided in the appendix.

While there could be multiple transport transformations that satisfy the condition of the distance between two distributions, i.e., $\mathcal{D}(p_s, p_t) \leq \rho$, we have observed that our proposed metric in Eqn. (9) provides a better mechanism to move the sample on the latent spaces. This could be explained since our metric is able to expand the training data distribution by moving the original latent vectors in the latent space while still maintaining the important property as mentioned in **Proposition 1**. In addition, by moving the latent vector \mathbf{z}_s in the latent space with a controlled parameter ρ , our metric can guarantee the semantic content information compared to the original one while creating the diverse semantic variations of the images. This also encourages the diffusion model to avoid synthesizing useless random images with uncontrolled latent vectors.

4 The Proposed Diffusion-based Domain Generalization Training Approach

Large Scale Diffusion-based Augmentation Sample Generation As shown in our theoretical analysis, generating our diffusion-based adversarial samples does not require alternative training steps with the CLIP training procedure. We have empirically observed that retraining the text-to-image diffusion model is unnecessary because the pre-trained diffusion model has been well learned on extreme-scale datasets, can model the data distribution well, and generates diverse synthetic data. Therefore, in our approach, we adopt the pretrained Latent Diffusion model [62] to generate the adversarial samples in advance to save the training time of CLIP. Formally, for each image \mathbf{x}_s and its corresponding prompt \mathbf{p}_s , we are going to generate M different augmented samples \mathbf{x}_t^* via the latent diffusion model by the following process:

$$\mathbf{z}_s = \text{LDMForward}(\mathbf{x}_s); \quad \mathbf{z}_t^* = \frac{\mathbf{z}_s + \mathcal{N}(\alpha\sqrt{2}, \mathbf{I})}{\sqrt{2}} \quad \text{where } \alpha \sim \mathcal{U}(-\rho, \rho); \quad \mathbf{x}_t^* = \text{LDMBackward}(\mathbf{z}_t^*) \quad (10)$$

where LDMForward and LDMBackward are the forward and backward processes of the latent diffusion model. Generating the adversarial samples during training will result in a longer training time for CLIP, which is unnecessary. Therefore, we propose to generate the adversarial samples via diffusion in advance, followed by using them to train the CLIP model, which is more time-efficient.

The Diffusion-based Domain Generalization Training Fig. 4 illustrates our proposed domain generalization framework. After the generation steps, each real image has M different adversarial samples. Then, we are able to improve the generability of the CLIP model by training on the real and adversarial samples together. Formally, learning the CLIP model can be re-written as follows:

$$\theta_{F\mathbf{x}}^*, \theta_{FP}^* = \arg \min_{\theta_{F\mathbf{x}}, \theta_{FP}} \mathbb{E}_{\mathbf{x}_s, \mathbf{p}_s, \mathbf{x}_t^*} [\mathcal{L}_{CLIP}(\mathbf{x}_s, \mathbf{p}_s) + \mathcal{L}_{CLIP}(\mathbf{x}_t^*, \mathbf{p}_s)] \quad (11)$$

5 Experiments

5.1 Datasets, Implementations, and Evaluations

Dataset We trained our foundation model on three different image-text datasets at different scales: Conceptual Captions 3M (CC3M) [67], Conceptual Captions 12M (CC12M) [4], and LAION400M [66]. Due to the hardware constraints, our ablation studies are mainly conducted on CC3M and CC12M. We evaluate our models on ImageNet 1K [64] and six common datasets, including, STL-10 [7], Country-211 [71], Caltech-101 [15], Flowers [53], Pets [55], and SUN-397 [82].

Implementation We adopt the implementation of OpenCLIP [6] and Latent Diffusion [62] in our experiments. For the CLIP model, we use the ViT-B/16 architecture. The results of other network backbones are reported in the appendix. For a fair comparison, our model is trained for 32 epochs with a similar hyper-parameter setting as [60, 6]. We utilize 32 NVIDIA A100 GPUs (40GB), and the batch size of our experiments is set to 320 per GPU. For image synthesis, we use the text-to-image latent diffusion model [62] to generate images at the resolution of 256×256 with 10 DDIM steps. For each real image, we generate $M = 10$ different synthetic images. The controlling hyper-parameter of the distance between distributions ρ is set to 0.5 in our experiments. Due to time and hardware constraints, we choose to use only 10 DDIM generation steps. This offers image quality that meets acceptable standards [62] while maintaining efficient data generation time on large-scale datasets (e.g., approximately 7.5 hours to generate 12M adversarial samples of CC12M on 32 GPUs).

Table 1: The Effectiveness of Distribution Moving ρ .

	ρ	Zeroshot	Linear Prob	Fine-Tune
CC3M	0.05	17.28	54.13	80.08
	0.20	18.79	55.11	80.61
	0.50	20.33	56.14	81.12
	0.70	19.82	55.44	80.09
	1.00	16.68	52.92	79.14
CC12M	0.05	36.44	69.27	80.28
	0.20	38.37	71.17	83.11
	0.50	39.34	72.12	84.67
	0.70	37.34	69.12	82.89
	1.00	35.19	68.94	81.74

the ImageNet 1K dataset. We adopt the implementation and learning hyper-parameter setting from [6] for fair comparisons. The majority of our experiments are evaluated on the ImageNet 1K dataset. To further illustrate the generability of our model, we also perform the zero-shot evaluation on six different zero-shot benchmarks STL-10, Country-211, Caltech-101, Flowers, Pets, and SUN-397.

5.2 Ablation Studies

Effectiveness of Distribution Moving ρ The results in Table 1 illustrate the effectiveness of the distance between distribution ρ . When the value of ρ is small, i.e., $\rho = 0.05$, the CLIP gains a little improvement due to the small distribution shift. Then, the performance is gradually improved when the value of ρ increases from 0.05 to 0.5. When the value of ρ is increased, the CLIP model can improve its generalizability to unknown distributions. However, if we keep increasing the value of ρ , the performance tends to drop. This is because if we shift the new data distribution

in the latent space far away from the original data distribution ($\mathcal{N}(\mathbf{0}, \mathbf{I})$), the quality of synthetic images generated by the latent diffusion model will dramatically drop in both realism and content information. Our best performance of CLIP is achieved at ρ of 0.5.

Effectiveness of Number of Augmented Images As shown in Table 2, the performance of our domain generalization approach evaluated on ImageNet1K is gradually increased when the number of adversarial images is increased. When we use only 3 adversarial images, the CLIP model gains a minor performance. Meanwhile, when we use the 10 adversarial images during training, the zero-shot classification performance of CLIP trained on CC3M and CC12M archives up to 20.33% and 39.34%. The performance of linear probing and fine-tuning is also significantly improved when the number of adversarial images is increased. However, if we keep increasing the number of images, we have

Evaluation Setup In our experiments, we consider three different evaluation metrics, i.e., Zero-shot Classification Accuracy, Linear Probing Accuracy, and Fine-tuning Accuracy. For zero-shot classification, we adopt the template of prompts and evaluation protocol as described in CLIP [60]. For linear probing, following the common practices [60, 39, 48], we use our frozen pre-trained image encoder to extract features followed by training a linear classifier. For a fair comparison, we adopt the hyper-parameter setting from [60, 6]. For fine-tuning evaluation, we fine-tune the end-to-end image encoder with a linear classifier on

Table 2: The Effectiveness of Number of Generated Samples.

	M	Zeroshot	Linear Prob	Fine-Tune
CC3M	0	17.10	53.50	79.50
	3	18.36	54.05	79.90
	5	19.05	55.17	80.82
	10	20.33	56.14	81.12
	15	20.40	57.26	81.17
	20	20.28	56.18	81.11
CC12M	0	36.50	69.00	82.10
	3	37.34	70.25	82.84
	5	38.10	71.44	83.58
	10	39.34	72.12	84.67
	15	39.21	72.18	84.65
	20	39.49	72.15	84.68

observed that the performance of the CLIP model is becoming stable. Therefore, generating 10 adversarial images for each real image is a good trade-off between performance and time efficiency.

Table 3: The Effectiveness of Transport Transformation.

	Zeroshot	Linear Prob	Fine-Tune
CC3M	CLIP	17.10	53.50
	Random	15.34	50.10
	T	20.33	56.14
CC12M	CLIP	36.50	69.00
	Random	34.90	67.35
	T	39.34	72.12
			84.67

prompt \mathbf{p}_s . As shown in Table 3, our transport transformation T significantly outperforms the random transformation. Indeed, using the random transformation even downgrades the performance of the CLIP model since the generation of the diffusion model by using random transformation is uncontrolled. Meanwhile, by controlling the latent variable \mathbf{z}_t^* via \mathbf{z}_s and ρ , as defined in Eqn. (9), the generation of adversarial samples is oriented and significantly improves the CLIP’s performance.

Effectiveness of Pre-trained and Re-trained Diffusion Model

We compared the pre-trained LDM with a re-trained latent diffusion model on CC3M and CC12M. We only re-train the second stage of the LDM while we adopt the pre-trained VQ-VAE of LDM [62] for the first stage. As shown in Table 4, the experimental results show that using adversarial samples generated via our transport transformation has significantly improved the performance in both cases of using re-trained and pre-trained LDM. However, practically, the performance of using the pre-trained diffusion model outperforms re-training the diffusion model on the corresponding dataset. This is because the pre-trained latent diffusion model was trained on the large-scale dataset and is able to model the data distribution better than the re-trained latent diffusion on the specific datasets. Therefore, using the pre-trained latent diffusion model is beneficial in terms of not only time efficiency but also performance improvement.

Effectiveness of Our Domain Generalization on Different Datasets and CLIP-based Models

Table 5 illustrates the results of our proposed approach on three datasets at different scales and CLIP-based models, i.e., CLIP [60], LaCLIP [12], and SLIP [48]. The zero-shot classification results have illustrated the generalizability of our proposed approach on different dataset scales. In particular, our proposed approach improves the zero-shot results of CLIP by +3.23%, +2.48%, and 3.11% on CC3M, CC12M, and LAION400M, respectively. Further fine-tuning the model via linear probing or end-to-end fine-tuning significantly improves the performance of the CLIP model. The results of fine-tuned models on ImageNet achieved 81.12%, 84.67%, and 86.98% on CC3M, CC12M, and LAION400M, respectively. Our proposed approach is effective not only on different datasets but also with different CLIP-based approaches. By further using better CLIP-based training approaches, i.e., LaCLIP or SLIP, the performance of zero-shot results is significantly improved, up to 75.58% trained LAION-400M using SLIP. By further fine-tuning the SLIP model, our proposed approach achieved state-of-the-art performance on ImageNet1K, i.e., 87.49%. The results in Table 5 have confirmed the scalability and generalizability of our approach across the training datasets and CLIP-based models.

5.3 Comparisons With State-of-the-Art Approaches

In this section, we present the results of our approach compared with other augmentation and domain generalization approaches, i.e., ADA [75], AdvStyle [88], and Masking Augmentation (FLIP) [39].

Table 5: The Effectiveness of Our Proposed Approach on Different Datasets and Different Language-Image Pretraining Models.

	CC3M			CC12M			LAION400M		
	Zeroshot	Linear Prob	Fine-Tune	Zeroshot	Linear Prob	Fine-Tune	Zeroshot	Linear Prob	Fine-Tune
CLIP	17.10	53.50	79.50	36.50	69.00	82.10	67.00	78.60	84.70
Ours + CLIP	20.33	56.14	81.12	39.34	72.12	84.67	70.11	80.74	86.98
Δ	+3.23	+2.64	+1.62	+2.84	+3.12	+2.57	+3.11	+2.14	+2.28
LaCLIP	21.50	56.50	81.15	48.40	72.30	82.53	—	—	—
Ours + LaCLIP	24.12	58.03	83.11	51.16	74.34	84.68	—	—	—
Δ	+2.62	+1.53	+1.95	+2.76	+2.04	+2.15	—	—	—
SLIP	23.00	65.40	81.40	40.70	73.70	83.10	70.21	80.34	85.83
SLIP+ Our	26.97	67.60	83.18	43.13	75.58	84.95	72.53	83.21	87.49
Δ	+3.97	+2.20	+1.78	+2.43	1.88	+1.85	+2.33	+2.87	+1.67

Table 6: The Comparison With Other Augmentation and Generalization Approaches.

	CC3M			CC12M			LAION400M		
	Zeroshot	Linear Prob	Fine-Tune	Zeroshot	Linear Prob	Fine-Tune	Zeroshot	Linear Prob	Fine-Tune
CLIP	17.10	53.50	79.50	36.50	69.00	82.10	67.00	78.60	84.70
CLIP + Masking	17.69	54.13	80.08	37.34	70.56	82.28	68.06	78.95	85.03
CLIP + ADA	18.36	55.75	80.43	38.10	70.95	82.93	68.59	79.54	85.23
CLIP + AdvStyle	19.01	55.55	80.40	38.77	71.22	81.21	69.47	79.90	85.57
CLIP + Ours	20.33	56.14	81.12	39.34	72.12	84.67	70.11	80.74	86.98
SLIP	23.00	65.40	81.40	40.70	73.70	83.10	70.21	80.34	85.83
SLIP + Masking	24.13	65.98	81.91	41.01	73.97	83.29	70.47	80.92	86.05
SLIP + ADA	24.89	66.26	82.09	41.64	74.02	83.56	70.95	81.49	86.23
SLIP + AdvStyle	25.50	66.55	82.59	42.30	74.46	84.01	71.37	81.74	86.58
SLIP + Ours	26.97	67.60	83.18	43.13	75.58	84.95	72.53	83.21	87.49

Zero-shot Classification Table 6 compares our approach with other augmentation and domain generalization methods. Our proposed approach consistently improves the performance of zero-shot classification. While the masking augmentation generates masked augmented samples, ADA [75] and AdvStyle [88] generate the adversarial samples via adversarial training. However, the distribution shift in these methods remains limited compared to our diffusion-based approach. As a result, our proposed approach significantly outperforms other augmentation and domain generalization approaches. In particular, by pre-training on the large-scale LAION400M dataset, our model achieves the state-of-the-art zero-shot classification performance, i.e., 70.11% and 72.53% by using CLIP and SLIP training. The results have shown our advantages in improving the generalizability of vision-language models against unknown data distributions.

Linear-Probing and End-to-end Fine-tuning Classification Table 6 illustrates the results of our linear probing and end-to-end fine-tuning experiments. Similar to the zero-shot classification results, our linear probing and end-to-end fine-tuning results consistently improve the performance of CLIP [60] and SLIP [48] and outperform other augmentation approaches. By pre-training on LAION-400M and further fine-tuning on ImageNet-1K, our training approach achieved state-of-the-art performance, with the accuracy of CLIP and SLIP improved to 86.98% and 87.49%. These results have further confirmed the effectiveness of our approach across evaluation settings and pre-training datasets.

Other Zero-shot Classification Benchmarks

Table 7 illustrates the results of our proposed approach (pretrained on LAION400M) on six different zero-shot benchmarks. Our approach consistently improves the performance of CLIP and SLIP on all zero-shot classification benchmarks which have illustrated the generalizability of our approach to unseen domains. Thanks to our generalization approach, the vision-language foundation model is able to learn better visual representation against the data distribution shift. Therefore, the vision-language model can be later well generalized to various downstream tasks.

Table 7: Zero-shot Classification Results on Six Benchmarks, i.e., STL-10, Country-211, Caltech-101, Flowers, Pets, and SUN-397.

	STL-10	Coun-211	Cal-101	Flowers	Pets	SUN-397
CLIP	97.30	17.80	91.20	63.90	90.10	66.80
CLIP + Our	97.58	18.34	93.14	77.12	91.74	68.85
SLIP	97.50	19.90	92.10	75.62	91.00	67.40
SLIP + Our	98.87	21.73	94.63	81.35	94.67	70.41

6 Conclusions, Limitations, and Broader Impact

Conclusions: This paper has introduced a novel, simple yet efficient diffusion-based domain generalization approach to the vision-language foundation model. Under our theoretical analysis, we introduced a new efficient sampling to generate new diffusion-based adversarial samples based on our proposed transport transformation to improve the generalizability of the vision-language foundation model. Our experimental results on various benchmarks have illustrated the effectiveness of our generalization approach to the vision-language foundation model.

Limitations: Our paper has chosen specific network configurations and learning hyperparameters to support our hypothesis. However, the other aspects of learning have not been fully investigated due to hardware constraints, e.g., the larger network sizes, the different diffusion models, the larger scale of pre-training datasets, etc. Additionally, the larger pre-training dataset may require larger computational time to generate diffusion-based adversarial samples.

Broader Impacts: Our paper studies the problem of domain generalization, which is a step forward in improving the generalizability of the vision-language foundation model. Our contributions have emphasized the relationship between diffusion and domain generalization, which can later be used to improve the performance of vision-language models. Our approach helps to increase the robustness of foundation models across various zero-shot downstream tasks.

References

- [1] D. Arpit, H. Wang, Y. Zhou, and C. Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35:8265–8277, 2022.
- [2] M.-H. Bui, T. Tran, A. Tran, and D. Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021.
- [3] T. Chang, X. Yang, T. Zhang, and M. Wang. Domain generalized stereo matching via hierarchical visual transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9559–9568, June 2023.
- [4] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
- [5] J. Chen, Z. Gao, X. Wu, and J. Luo. Meta-causal learning for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7683–7692, June 2023.
- [6] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.
- [7] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.
- [8] S. Dehdashtian, L. Wang, and V. N. Boddeti. Fairerclip: Debiasing clip’s zero-shot predictions using functions in rkhss. In *The Twelfth International Conference on Learning Representations*, 2024.
- [9] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [10] J. Ding, N. Xue, G.-S. Xia, B. Schiele, and D. Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15413–15423, 2023.
- [11] T. Dockhorn, A. Vahdat, and K. Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021.
- [12] L. Fan, D. Krishnan, P. Isola, D. Katabi, and Y. Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023.
- [13] X. Fan, Q. Wang, J. Ke, F. Yang, B. Gong, and M. Zhou. Adversarially adaptive normalization for single domain generalization. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8208–8217, 2021.
- [14] H. Fang, B. Han, S. Zhang, S. Zhou, C. Hu, and W.-M. Ye. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1257–1266, 2024.
- [15] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.
- [16] C.-M. Feng, K. Yu, Y. Liu, S. Khan, and W. Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023.
- [17] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.
- [18] Y. Fu, C. Chen, Y. Qiao, and Y. Yu. Dreamda: Generative data augmentation with diffusion models. *arXiv preprint arXiv:2403.12803*, 2024.
- [19] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015.

- [20] J. Guo, L. Qi, and Y. Shi. Domaindrop: Suppressing domain-sensitive channels for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19114–19124, 2023.
- [21] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [22] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.
- [23] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [24] T. Hoang Nguyen and A. Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. *arXiv e-prints*, pages arXiv–2312, 2023.
- [25] L. Hoyer, D. Dai, and L. Van Gool. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [26] L. Hu, M. Kan, S. Shan, and X. Chen. Dandelionnet: Domain composition with instance adaptive classification for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19050–19059, 2023.
- [27] J. Huang, D. Guan, A. Xiao, and S. Lu. Fsdrl: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021.
- [28] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren. Prodif: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605, 2022.
- [29] W. Huang, C. Chen, Y. Li, J. Li, C. Li, F. Song, Y. Yan, and Z. Xiong. Style projected clustering for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3071, June 2023.
- [30] K. Islam, M. Z. Zaheer, A. Mahmood, and K. Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *CVPR*, 2024.
- [31] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [32] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.
- [33] Y. Kim, B. Na, M. Park, J. Jang, D. Kim, W. Kang, and I.-C. Moon. Training unbiased diffusion models from biased dataset. *arXiv preprint arXiv:2403.01189*, 2024.
- [34] S. Lee, J. Bae, and H. Y. Kim. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11776–11785, 2023.
- [35] S. Lee, H. Seong, S. Lee, and E. Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9936–9946, 2022.
- [36] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018.
- [37] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [38] L. Li, K. Gao, J. Cao, Z. Huang, Y. Weng, X. Mi, Z. Yu, X. Li, and B. Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021.

- [39] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.
- [40] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.
- [41] Y. Li, D. Zhang, M. Keuper, and A. Khoreva. Intra-source style augmentation for improved domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 509–519, 2023.
- [42] Z. Li, K. Ren, X. JIANG, Y. Shen, H. Zhang, and D. Li. SIMPLE: Specialized model-sample matching for domain generalization. In *The Eleventh International Conference on Learning Representations*, 2023.
- [43] C. Lin, Z. Yuan, S. Zhao, P. Sun, C. Wang, and J. Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8771–8780, 2021.
- [44] S. Lin, Z. Zhang, Z. Huang, Y. Lu, C. Lan, P. Chu, Q. You, J. Wang, Z. Liu, A. Parulkar, et al. Deep frequency filtering for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11797–11807, 2023.
- [45] S. Long, Q. Zhou, C. Ying, L. Ma, and Y. Luo. Rethinking domain generalization: Discriminability and generalizability. *arXiv preprint arXiv:2309.16483*, 2023.
- [46] Z. Luo, P. Zhao, C. Xu, X. Geng, T. Shen, C. Tao, J. Ma, Q. Lin, and D. Jiang. Lexlip: Lexicon-bottlenecked language-image pre-training for large-scale image-text sparse retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11206–11217, 2023.
- [47] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.
- [48] N. Mu, A. Kirillov, D. Wagner, and S. Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022.
- [49] T. Nguyen, Y. Li, U. Ojha, and Y. J. Lee. Visual instruction inversion: Image editing via visual prompting. *arXiv preprint arXiv:2307.14331*, 2023.
- [50] X.-B. Nguyen, C. N. Duong, X. Li, S. Gauch, H.-S. Seo, and K. Luu. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1492, 2023.
- [51] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.
- [52] W. Nie, A. Vahdat, and A. Anandkumar. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021.
- [53] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*, volume 2, pages 1447–1454. IEEE, 2006.
- [54] X. Pan, P. Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pages 464–479, 2018.
- [55] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.
- [56] H. Pham, Z. Dai, G. Ghiasi, K. Kawaguchi, H. Liu, A. W. Yu, J. Yu, Y.-T. Chen, M.-T. Luong, Y. Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.
- [57] H. Phung, Q. Dao, and A. Tran. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10199–10208, 2023.

- [58] F. Qiao and X. Peng. Uncertainty-guided model generalization to unseen domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6790–6800, 2021.
- [59] F. Qiao, L. Zhao, and X. Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565, 2020.
- [60] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [61] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- [62] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [63] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.
- [64] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [65] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Goncalves Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [66] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [67] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [68] S. Shen, Z. Zhu, L. Fan, H. Zhang, and X. Wu. Diffclip: Leveraging stable diffusion for language grounded 3d classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3596–3605, 2024.
- [69] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- [70] Z. Tang, Y. Gao, Y. Zhu, Z. Zhang, M. Li, and D. N. Metaxas. Selfnorm and crossnorm for out-of-distribution robustness. *arXiv*, 2020.
- [71] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [72] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.
- [73] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- [74] V. Vudit, M. Engilberge, and M. Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3219–3229, 2023.
- [75] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.
- [76] J. Wang, R. Du, D. Chang, K. Liang, and Z. Ma. Domain generalization via frequency-domain-based feature disentanglement and interaction. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4821–4829, 2022.

- [77] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
- [78] P. Wang, Z. Zhang, Z. Lei, and L. Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3769–3778, June 2023.
- [79] T. Wang, K. Lin, L. Li, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang. Equivariant similarity for vision-language foundation models. *arXiv preprint arXiv:2303.14465*, 2023.
- [80] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.
- [81] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou. Diffusion-GAN: Training GANs with diffusion. In *The Eleventh International Conference on Learning Representations*, 2023.
- [82] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119:3–22, 2016.
- [83] X. Yao, Y. Bai, X. Zhang, Y. Zhang, Q. Sun, R. Chen, R. Li, and B. Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7097–7107, 2022.
- [84] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [85] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.
- [86] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.
- [87] Y. Zhang, X. Wang, K. Jin, K. Yuan, Z. Zhang, L. Wang, R. Jin, and T. Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International Conference on Machine Learning*, pages 41647–41676. PMLR, 2023.
- [88] Z. Zhong, Y. Zhao, G. H. Lee, and N. Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in Neural Information Processing Systems*, 35:338–350, 2022.

Appendix

1 Proof of Proposition 1

Proposition 1: Given $\mathbf{z}_s \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\alpha (-\rho \leq \alpha \leq \rho)$, the condition of distance between distributions $\mathcal{D}(p_s(\mathbf{z}_s), p_t(\mathbf{z}_t^*)) \leq \rho$ holds if the transport transformation \mathcal{T} is defined as $\mathcal{T}(\mathbf{z}_s, \rho) = \frac{\mathbf{z}_s + \mathcal{N}(\alpha \sqrt{2}\mathbf{I}, \mathbf{I})}{\sqrt{2}}$.

Proof: The proposition can be sufficiently proven via the Wasserstein distance between two distributions. As the latent variable \mathbf{z}_s belong to the Normal distribution, i.e., $\mathbf{z}_s \in \mathcal{N}(\mathbf{0}, \mathbf{I})$, the transformed latent variable \mathbf{z}_t^* via the transformation \mathcal{T} should belong to $\mathbf{z}_t^* = \frac{\mathbf{z}_s + \mathcal{N}(\alpha \sqrt{2}\mathbf{I}, \mathbf{I})}{\sqrt{2}} \sim \mathcal{N}(\alpha, \mathbf{I})$. Then, the transportation cost between two distributions p_s and p_t measured via the Wasserstein distance can be defined as follows:

$$\mathcal{D}(\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s), \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)) = \|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\|_2^2 + \text{tr} \left[\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_t - 2(\boldsymbol{\Sigma}_s^{1/2} \boldsymbol{\Sigma}_t \boldsymbol{\Sigma}_t^{1/2})^{1/2} \right]. \quad (12)$$

Since $\boldsymbol{\mu}_s = \mathbf{0}$, $\boldsymbol{\mu}_t = \alpha$, the $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Sigma}_t$ of the two data distributions is \mathbf{I} , the distance between two data distribution defined in Eqn. (12) can be rewritten as:

$$\begin{aligned} \mathcal{D}(\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s), \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)) &= \|\mathbf{0} - \alpha\|_2^2 = \alpha \leq \rho \\ \Rightarrow \quad \mathcal{D}(p_s(\mathbf{z}_s), p_t(\mathbf{z}_t^*)) &\leq \rho \quad (\text{Q.E.D}) \end{aligned} \quad (13)$$

2 Additional Ablation Study

Effectiveness of Different Backbone The results in Table 8 illustrate the effectiveness of our approach in different backbones, i.e., ResNet-50 (RN50) and ViT-B/16. By using the better backbone, the performance of the CLIP model is major improved. In particular, on the zero-shot classification benchmarks, the performance of our approach trained on CC3M and CC12M is improved from 18.77% to 20.33% and from 37.05% to 39.34%, respectively. The performance is even further majorly improved in both backbones with further fine-tuning.

Table 8: The Effectiveness of Vision Network Backbone.

	Different Work	Zeroshot	Linear Prob	Fine-Tune
CC3M	RN50	16.13	50.02	78.12
	RN50 + Our	18.87	52.11	80.08
	ViT-B-16	17.10	53.50	79.30
CC12M	ViT-B-16 + Our	20.33	56.14	81.12
	RN50	35.02	67.42	80.81
	RN50 + Our	37.05	70.21	82.36
ViT-B-16	ViT-B-16	36.50	69.00	82.10
	ViT-B-16 + Our	39.34	72.12	84.67

3 Discussion of Limitations and Broader Impact

Limitations In our paper, we opt for a particular set of hyperparameters and learning methodologies to bolster our hypothesis. Although our proposed approach has shown its effectiveness in improving the generalizability of the vision-language foundation model, it could potentially consist of several limitations. First, the choice of different contrastive learning losses in the vision-language models should be exploited. Second, the different visual and textual encoders have not been fully investigated in our study. Besides, the larger number of DDIM steps in the data generation process of image generation should be studied in future work. Third, the data generation process via the diffusion model requires high computational resources and a large amount of time. Additionally, in our paper, we only consider the images conditioned on the text prompts. However, the different conditions, e.g., image or object layouts, semantic segmentation, etc., should be considered in future studies. These constraints may motivate new research studies to enhance the diffusion-based domain generalization approach to vision-language foundation models.

Other Potential Social Broader Impacts Our paper has introduced a novel diffusion-based approach to domain generalization in the vision-language foundation model. Our approach has improved the performance of the foundation model on various downstream tasks. However, we acknowledge that the large-scale diffusion model, i.e., LDM [62], trained on extreme-scale data could potentially produce inappropriate images and even hallucinations. Thus, the vision-language models could accidentally learn these pieces of information. In addition, since the data generation process requires high computational resources and a large amount of time, it could potentially produce a higher carbon footprint.

4 Adversarial Samples

Fig. 5 illustrates our diffusion-based adversarial samples generated via our proposed transport transformation with the latent diffusion model [62].

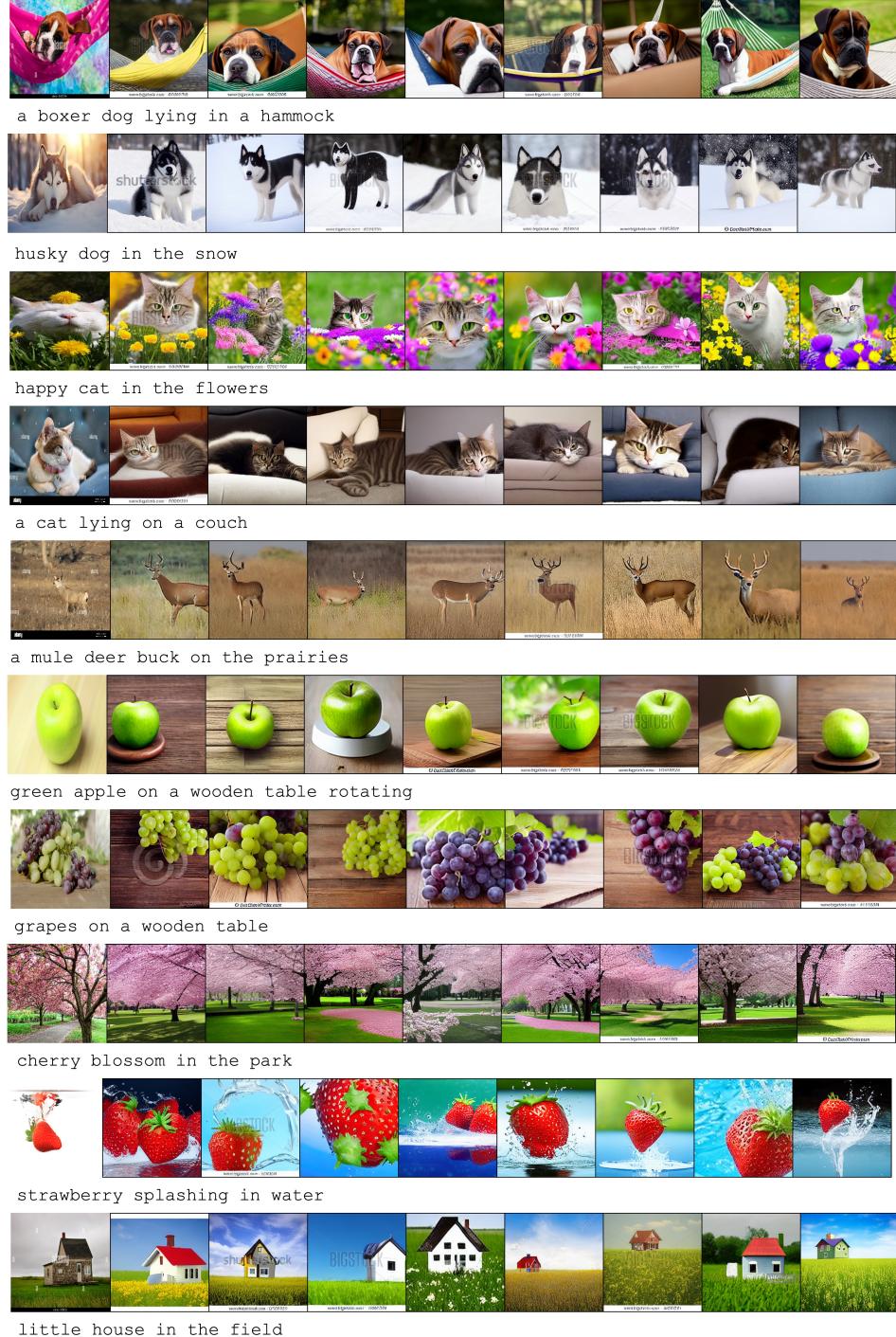


Figure 5: Our Diffusion-based Adversarial Samples. The first figure of each row is the original image.

5 Qualitative Results of Zeroshot Predictions

Fig. 6 visualizes our zero-shot predictions on ImageNet-1K compared to CLIP trained on LAION400M.



Figure 6: Our Zeroshot Predictions on ImageNet-1K.