ExpertDiff: Head-less Model Reprogramming with Diffusion Classifiers for Out-of-Distribution Generalization

Jee Seok Yoon[1] , **Junghyo Sohn**[2] , **Wootaek Jeong**[2] and **Heung-Il Suk**[2,*]

[1]Deptartment of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea
[2]Deptartment of Artificial Intelligence, Korea University, Seoul, Republic of Korea
{wltjr1007, jhsohn0633, wtjeong, hisuk}@korea.ac.kr

Abstract

Vision-language models have achieved remarkable performance across various tasks by leveraging large-scale multimodal training data. However, their ability to generalize to out-of-distribution (OOD) domains requiring expert-level knowledge remains an open challenge. To address this, we investigate cross-domain transfer learning approaches for efficiently adapting diffusion classifiers to new target domains demanding expert-level domain knowledge. Specifically, we propose ExpertDiff, a head-less model reprogramming technique that optimizes the instruction-following abilities of text-to-image diffusion models via learnable prompts, while leveraging the diffusion classifier objective as a modular plug-and-play adaptor. Our approach eliminates the need for conventional output mapping layers (*e.g.*, linear probes), enabling seamless integration with off-the-shelf diffusion frameworks like Stable Diffusion. We demonstrate the effectiveness of ExpertDiff on the various OOD datasets (*i.e.*, medical and satellite imagery). Furthermore, we qualitatively showcase ExpertDiff's ability to faithfully reconstruct input images, highlighting its potential for both downstream discriminative and upstream generative tasks. Our work paves the way for effectively repurposing powerful foundation models for novel OOD applications requiring domain expertise.

1 Introduction

Vision-language models trained on internet-scale datasets have shown ground-breaking performance improvements across various downstream tasks. Many of these large-scale models have surpassed human capabilities in in-distribution (ID) generalization tasks, with some even achieving comparable out-of-distribution (OOD) generalization performances [Jaini *et al.*, 2024]. The use of large-scale datasets has significantly enhanced vision-language models' capabilities (*e.g.*, emergent abilities [Wei *et al.*, 2022; Zhou *et al.*, 2024]) by effectively bridging textual and visual information.

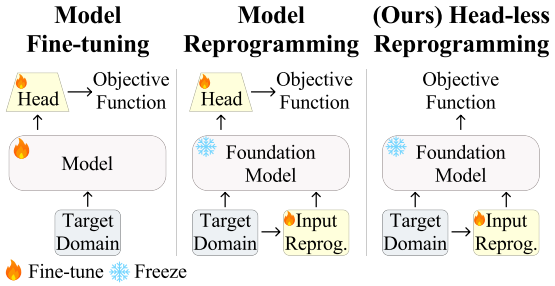

Figure 1: Conceptual comparison of different transfer learning approaches. (Left) Traditional fine-tuning adjusts the entire model or specific layers. (Middle) Model reprogramming focuses on modifying input transformations and output mappings while keeping the foundation model fixed. (Right) Our proposed head-less reprogramming method eliminates the need for an output mapping layer, directly leveraging the foundation model for the target domain task.

This integration of textual data is crucial as it infuses semantic depth into visual embeddings, which enriches the model's ability to generate more contextually accurate and semantically coherent images.

Building on these advancements, recent studies have investigated the robustness of off-the-shelf diffusion models (*e.g.*, Stable Diffusion [Rombach *et al.*, 2022]) through *diffusion classifiers* [Li *et al.*, 2023; Clark and Jaini, 2023], which utilize the approximated likelihood to calculate the posterior probabilities via Bayes' theorem (see Section 3.2). These diffusion classifiers have shown an ability to generalize shapes in a manner similar to human perception [Jaini *et al.*, 2024], enabling them to develop a detailed understanding of objects and generate even spuriously correlated images under language drift [Gandelsman *et al.*, 2024; Ruiz *et al.*, 2023; Huang *et al.*, 2024].

However, whether suitability of off-the-shelf diffusion models as a foundation model [Bommasani *et al.*, 2022] for OOD tasks requiring *expert-level domain knowledge* remains a topic of debate. For example, performance of diffusion classifiers on complex tasks like tumor diagnosis with histological images remain near chance levels [Radford *et al.*, 2021; Clark and Jaini, 2023; Fini *et al.*, 2023]. In a similar manner, previous studies have found that vision-language models generalize very well with known concepts and data types [Udandarao *et al.*, 2024a; Yang *et al.*, 2023b], but they require an

*Correspondng author.

exponential amount of data for linear performance gains on new concepts [Udandarao *et al.*, 2024b]. Hence, it is evident that while large diffusion models excel in certain generalization tasks, their ability to adapt to new and unfamiliar domains is still largely underexplored. This gap in their capability forms a compelling ground for further exploration.

Model reprogramming [Chen, 2024] has emerged as an efficient transfer learning technique which only require training the input transformation and output mapping projections, while keeping the pre-trained backbone model fixed. First, the input transformation for diffusion models conventionally utilize prompt learning methods, which transforms the initial prompt embedding for target downstream task. Intuitively, prompt learning shifts the focus from directly adjusting model weights to crafting prompts that "instruct" the model to better perform the target downstream tasks. Then, the output mapping projections (*e.g.*, linear probing heads) pools the discrimintative features from the pretrained backbone model.

Drawing from these insights, we introduce *head-less model reprogramming* leveraging diffusion classifiers, a novel approach that eliminates the need for output mapping projections while enabling efficient adaptation of text-to-image diffusion models to OOD tasks (see Figure 1). Our key contributions include:

- We propose a fully modular plug-and-play model reprogramming technique that is applicable to both upstream and downstream tasks, paving the way for effectively repurposing powerful foundational text-to-image diffusion models for novel OOD applications requiring expert-level domain knowledge.
- Our approach is highly efficient, as it only requires training a single parameter for the input prompt.
- The proposed method achieves state-of-the-art performance in zero-shot, few-shot domain generalization, as well as fully supervised learning scenarios.

2 Related Works

2.1 Generative Classifiers

Generative classifiers have long been explored as an alternative to discriminative approaches, with early works demonstrating their potential benefits [Ng and Jordan, 2001]. Recent advances in diffusion models have led to a resurgence of interest in generative classification, particularly in the form of diffusion classifiers [Jaini *et al.*, 2024; Clark and Jaini, 2023; Li *et al.*, 2023; Prabhudesai *et al.*, 2023; He *et al.*, 2023; Chen *et al.*, 2024; Chen *et al.*, 2023; Bhattacharya and Prasanna, 2024; Vilouras *et al.*, 2024; Krojer *et al.*, 2023]. These approaches leverage the approximate likelihood of conditional diffusion models to estimate the class probabilities, showing impressive zero-shot generalization capabilities. A key advantage of diffusion-based classifiers is their focus on geometric and shape-based features of data, in contrast to traditional models that may overly rely on texture [Jaini *et al.*, 2024]. This property aligns well with human perception and contributes to their strong OOD performance. Additionally, text-to-image diffusion models integrate the generative capabilities of diffusion models with the semantic understanding of vision-language models (VLMs) like CLIP [Radford *et al.*, 2021], thereby further improving their zero-shot classification performance [Clark and Jaini, 2023]. However, challenges remain in adapting these models to domains with characteristics that are difficult to express in layman's terms or fall outside their training distribution. Our work addresses this gap by proposing efficient learning techniques for improving classification performance on such domains.

2.2 Prompt Learning and Model Reprogramming

Prompt learning has emerged as an effective technique for adapting large pre-trained VLMs like CLIP to downstream tasks without full fine-tuning. Early works like CoOp [Zhou *et al.*, 2022b] and CoCoOp [Zhou *et al.*, 2022a] focused on learning continuous prompt vectors for the text encoder of CLIP. More recent approaches like Visual Prompt Tuning [Jia *et al.*, 2022] and MaPLe [Khattak *et al.*, 2023] have explored prompt learning in both visual and textual modalities. However, most of these works focus on CLIP-based classification. To the best of our knowledge, we are the first to propose a prompt learning technique for diffusion-based classifiers.

In parallel, model reprogramming [Chen, 2024] has emerged as an efficient transfer learning technique that adapts pre-trained models to new tasks by modifying only the input transformation and output mapping, while keeping the pre-trained backbone fixed. In the context of VLMs, input transformation often take the form of prompt learning, while output mapping typically involve linear probing or other lightweight classification heads. However, recent observations suggest that using projected features from output mapping layers can result in lower classification performance compared to using backbone features directly [Bordes *et al.*, 2022]. This has led to some VLMs discarding output mapping projections during inference. Our work builds on these insights by proposing a head-less model reprogramming approach that leverages diffusion classifiers, offering a fully modular, plug-and-play adaptor for off-the-shelf diffusion models without the need for output mapping layers.

3 Preliminary

In this section, we first briefly derive the approximate likelihood in discrete-time conditional diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020], and use Bayes' theorem to calculate the class conditional density estimates for diffusion classifiers [Li *et al.*, 2023; Clark and Jaini, 2023].

3.1 Diffusion Models

Diffusion models are a class of probabilistic generative models that are based on principle of reversing a diffusion process. The *forward* diffusion process incrementally adds Gaussian noise to the original input $\mathbf{x} := \mathbf{x}_0 \in \mathbb{R}^D$ into a sequence of noisy latent variables $\mathbf{x}_{1:T} := \{\mathbf{x}_1, ..., \mathbf{x}_T\}$ with a fixed schedule defined by:

$$\begin{aligned} q(\mathbf{x}_{1:T}|\mathbf{x}_0) &:= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) &:= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \end{aligned} \tag{1}$$

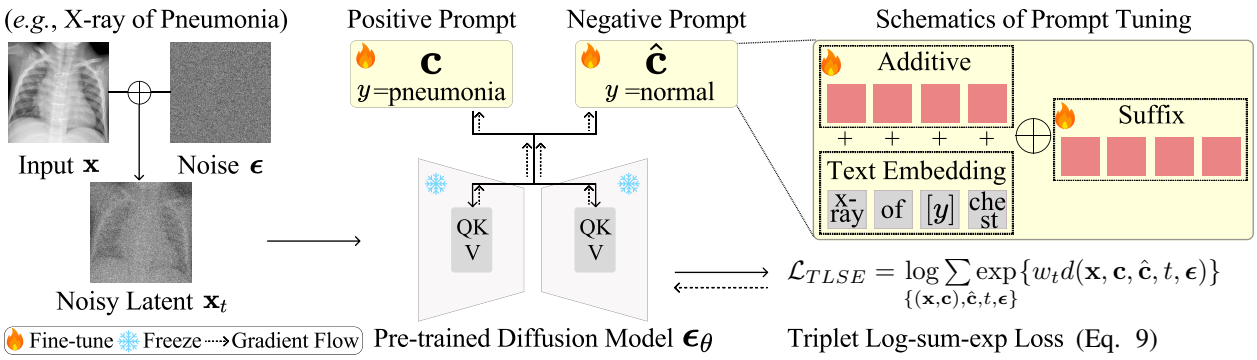

Figure 2: Schematic overview of our proposed ExpertDiff head-less model reprogramming approach. The method combines a pre-trained diffusion model with prompt tuning techniques, utilizing learnable additive and suffix prompts intialized with text prompt embedding (*e.g.*, embedding of "x-ray of pneumonia chest"). The model is optimized using a triplet log-sum-exp loss, with fine-tuning applied only to the prompt parameters while keeping everything else fixed. $+$ denotes element-wise addition, and $\oplus$ denotes concatenation.

where β is a hyperparameter for variance schedule. The *conditional reverse* diffusion process approximates the reverse step conditionally with $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, \mathbf{c}, t))$ by gradually denoising from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$:

$$p_\theta(\mathbf{x} \mid \mathbf{c}) = \int_{\mathbf{x}_{1:T}} p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) d\mathbf{x}_{1:T}. \quad (2)$$

The parameter θ can be trained by optimizing the variational lower bound on the log likelihood:

$$\begin{aligned} \log p_\theta(\mathbf{x} \mid \mathbf{c}) &\geq -\mathbb{E}_{t,\boldsymbol{\epsilon}} \left[w_t \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t) - \boldsymbol{\epsilon}\|^2 \right] + C \\ p_\theta(\mathbf{x} \mid \mathbf{c}) &= \exp\left\{ -\mathbb{E}_{t,\boldsymbol{\epsilon}} \left[w_t \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t) - \boldsymbol{\epsilon}\|^2 \right] \right\}, \end{aligned} \quad (3)$$

where w_t is the timestep weighting and C is a small constant that is negligible. Stable diffusion [Rombach *et al.*, 2022] and Imagen [Saharia *et al.*, 2022] use $w_t = \mathrm{SNR}(t)$, *i.e.*, signal-to-noise ratio, whereas $w_t = \exp(-7t)$ have empirically resulted in better performance for diffusion classifiers across different backbone models and tasks [Clark and Jaini, 2023].

3.2 Diffusion Classifiers

Bayes' theorem states $p_\theta(\mathbf{c}|\mathbf{x}) = \frac{p(\mathbf{c})p_\theta(\mathbf{x}|\mathbf{c})}{\sum_{\hat{\mathbf{c}}} p(\hat{\mathbf{c}})p_\theta(\mathbf{x}|\hat{\mathbf{c}})}$, and assuming class probabilities $p(\mathbf{c})$ are uniform,

$$p_\theta(\mathbf{c} \mid \mathbf{x}) = \frac{p_\theta(\mathbf{x} \mid \mathbf{c})}{\sum_{\hat{\mathbf{c}} \in \mathcal{C}} [p_\theta(\mathbf{x} \mid \hat{\mathbf{c}})]}, \quad (4)$$

where $\hat{\mathbf{c}} \in \mathcal{C}$ is set of all possible classes. Inserting the approximate likelihood of conditional diffusion model in Eq. 3 into Eq. 4, the predictive posterior probability is derived as

$$p_\theta(\mathbf{c} \mid \mathbf{x}) = \frac{\exp\left\{ -\mathbb{E}_{t,\boldsymbol{\epsilon}} \left[w_t \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t) - \boldsymbol{\epsilon}\|^2 \right] \right\}}{\sum_{\hat{\mathbf{c}} \in \mathcal{C}} \exp\left\{ -\mathbb{E}_{t,\boldsymbol{\epsilon}} \left[w_t \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \hat{\mathbf{c}}, t) - \boldsymbol{\epsilon}\|^2 \right] \right\}}, \quad (5)$$

where the expectation term can be approximated using the Monte Carlo sampling [Li *et al.*, 2023]. Finally, diffusion classifier typically makes its decision based on the most probable class assignments, *i.e.*,

$$\begin{aligned} y^* &= \arg\max_i p(\mathbf{c} = \mathbf{c}_i \mid \mathbf{x}) \\ &= \arg\min_i \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[w_t \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_i, t) - \boldsymbol{\epsilon}\|^2 \right], \end{aligned} \quad (6)$$

where $\mathbf{c}_i \in \mathbb{R}^{|\text{token}| \times \text{dim}}$ is the prompt embedding for i-th class label y_i. These prompts are typically generated using a task template [Zhang *et al.*, 2023] which transforms a class name into a task-specific prompt, for example, class name 'car' combined with the task template 'a photo of $[y]$' can be transformed into 'a photo of car'. Refer to Appendix C for templates used in this paper.

4 ExpertDiff: Head-less Model Reprogramming with Diffusion Classifiers

In this section, we present **ExpertDiff** for head-less reprogramming of text-to-image diffusion model for cross-domain classification tasks. Specifically, ExpertDiff builds upon diffusion classifier [Li *et al.*, 2023; Clark and Jaini, 2023] by optimizing the prompt embeddings for better instructions-following capabilities on the target downstream task. By reprogramming the diffusion model in a head-less manner and aligning it with the diffusion classifier's objective, ExpertDiff eliminates the need for a traditional output mapping layer. Consequently, this approach enables the diffusion model to effectively handle cross-domain data for both downstream classification and upstream generation tasks. Our model integrates seamlessly with off-the-shelf diffusion frameworks (*e.g.*, Stable Diffusion [Rombach *et al.*, 2022]), making it a fully modular plug-and-play adaptor for a wide range of OOD classification tasks.

The proposed ExpertDiff converts the prompts $\mathbf{c}$ used as conditions in text-to-image diffusion models into learnable parameters, while keeping the weights of the rest of the model. The design of this parameter is detailed in Section 4.3. We then replace the traditional output mapping layer (e.g., MLP classifier head) with the diffusion classifier's objective as described in Eq. 5. Consequently, the negative log-

likelihood of diffusion classifier is

$$\begin{aligned}\mathcal{L} &= -\sum_{(\mathbf{x},\mathbf{c})} \log p_\theta(\mathbf{c} \mid \mathbf{x}) \\ &= -\sum_{(\mathbf{x},\mathbf{c})} \log \frac{\exp\left\{-\mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w_t \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\mathbf{c},t)-\boldsymbol{\epsilon}\|^2\right]\right\}}{\sum_{\hat{\mathbf{c}}\in\mathcal{C}} \exp\left\{-\mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w_t \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\hat{\mathbf{c}},t)-\boldsymbol{\epsilon}\|^2\right]\right\}},\end{aligned} \tag{7}$$

where $(\mathbf{x},\mathbf{c}) \in (\mathcal{X},\mathcal{C})$ represents all possible pairs of image and prompt. However, calculating the likelihood $p_\theta(\mathbf{c}|\mathbf{x})$ requires calculating the expectation over t and ϵ for each class. This makes the training exponentially inefficient for datasets with many classes. In the following sections, we propose an efficient form of the likelihood (Section 4.1) and explain how to search for the optimal timestep t^* to replace the expectation $\mathbb{E}_t$ (Section 4.2).

4.1 Efficient Likelihood Calculation

In this section, we derive an upper-bound of negative log-likelihood for diffusion classifiers, which we use for a more efficient learning paradigm:

Theorem 4.1. *(Proof in Appendix A.3)*
The upper-bound of expected negative log-likelihood in diffusion classifier is given by

$$\begin{aligned}\mathcal{L} &= -\mathbb{E}_{(\mathbf{x},\mathbf{c})} \log p_\theta(\mathbf{c} \mid \mathbf{x}) \\ &\le \log \mathbb{E}_{(\mathbf{x},\mathbf{c}),\hat{\mathbf{c}},t,\boldsymbol{\epsilon}} \exp\left\{w_t d(\mathbf{x},\mathbf{c},\hat{\mathbf{c}},t,\boldsymbol{\epsilon})\right\},\end{aligned} \tag{8}$$

where $(\mathbf{x},\mathbf{c}) \in (\mathcal{X},\mathcal{C})$ represent pairs of image and prompt, $\hat{\mathbf{c}} \in \mathcal{C}$ represent all possible prompts, and $d(\mathbf{x},\mathbf{c},\hat{\mathbf{c}},t,\boldsymbol{\epsilon}) := \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\mathbf{c},t)-\boldsymbol{\epsilon}\|^2 - \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\hat{\mathbf{c}},t)-\boldsymbol{\epsilon}\|^2$.

To this end, we use Monte-carlo estimate of the expectation $\mathbb{E}_{(\mathbf{x},\mathbf{c}),\hat{\mathbf{c}},t,\boldsymbol{\epsilon}}$ by sampling tuples of $\{(\mathbf{x},\mathbf{c}),\hat{\mathbf{c}},t,\boldsymbol{\epsilon}\}$. This transforms the negative log-likelihood loss into a log-sum-exp form of the triple loss, where $\boldsymbol{\epsilon}$ is the anchor, $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\mathbf{c},t)$ is the positive sample, and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\hat{\mathbf{c}},t)$ is the negative sample:

$$\begin{aligned}\mathcal{L}_{TLSE} := \log \sum_{\{(\mathbf{x},\mathbf{c}),\hat{\mathbf{c}},t,\boldsymbol{\epsilon}\}} \exp\{w_t(\| \underbrace{\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\mathbf{c},t)}_{\text{pos. sample}} - \overbrace{\boldsymbol{\epsilon}}^{\text{anc.}} \|^2 \\ - \| \underbrace{\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\hat{\mathbf{c}},t)}_{\text{neg. sample}} - \overbrace{\boldsymbol{\epsilon}}^{\text{anc.}} \|^2)\}.\end{aligned} \tag{9}$$

This triplet Log-Sum-Exp (TLSE) loss exhibits several desirable properties that make it well-suited for reprogramming diffusion models (see Section 4.4).

4.2 Optimal Timestep Search

Previous studies on diffusion classifiers have employed a monotonically decreasing timestep weighting in Eq. 6, such as $w_t = \text{SNR}(t)$ [Li *et al.*, 2023] and $w_t = \exp(-7t)$ [Clark and Jaini, 2023], based on the assumption that scores from less noisy latents are more discriminative. However, recent findings have challenged this assumption [Mukhopadhyay *et al.*, 2023; Yue *et al.*, 2024]. In accordance to these recent findings, we have also observed that less noisy latents can sometimes be less discriminative, and that scores from most timesteps contribute minimally to the predictive posterior, with some even consistently degrading performance (see Figure 5). These findings suggest a more nuanced relationship between timestep, noise level, and discriminative power than previously assumed. Moreover, we observe that a single, carefully chosen timestep can yield performance comparable to using all timesteps, suggesting the potential for a more efficient approach that eliminates the need for timestep iteration during both training and inference.

Algorithm 1 Training ExpertDiff with Optimal Timestep

Require: Paired dataset $\mathcal{D} = \{(\mathbf{x}, y)\}$, learnable prompt embeddings $\mathcal{C} = \{(\mathbf{c})_i\}$, pre-trained diffusion model $\boldsymbol{\epsilon}_\theta$, learning rate η, timestep search iterations N

for $i = 1$ to # of class **do** ▷ *Initialize Prompt Emb.*
  $\mathbf{c}_i \leftarrow$ task-template(y_i)
scores $\leftarrow$ zeros(T)
for $i = 1$ to N **do** ▷ *Optimal Timestep Search*
  Sample $(\mathbf{x},\mathbf{c}) \sim \mathcal{D}, \hat{\mathbf{c}} \sim \mathcal{C}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
  for $t = 1$ to T **do**
    score $\leftarrow \frac{1}{2}\left[1 - \text{erf}\left(\frac{\sqrt{\bar{\alpha}_t} d(\mathbf{x},\mathbf{c},\hat{\mathbf{c}},t,\boldsymbol{\epsilon})}{2\sqrt{2(1-\bar{\alpha}_t)}}\right)\right]$
    scores$[t] \leftarrow$ scores$[t]$ + score
$t^* \leftarrow \arg\max_t$ scores$[t]$
while not converged **do** ▷ *Prompt Optimization*
  Sample $(\mathbf{x},\mathbf{c}) \sim \mathcal{D}, \hat{\mathbf{c}} \sim \mathcal{C}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
  $\mathcal{L} \leftarrow \log \sum_{\{(\mathbf{x},\mathbf{c}),\hat{\mathbf{c}},\boldsymbol{\epsilon}\}} \exp\{d(\mathbf{x},\mathbf{c},\hat{\mathbf{c}},t^*,\boldsymbol{\epsilon})\}$
  Update $\mathbf{c} \leftarrow \mathbf{c} - \eta\nabla_{\mathbf{c}}\mathcal{L}$
return Prompt embeddings $\mathbf{c}$, timestep t^*

To this end, we propose a modified version of the class attribute loss introduced by [Yue *et al.*, 2024] to search for the optimal timestep t^* that maximizes the discriminative power of the latents:

$$t^* := \arg\max_t \frac{1}{2}\left[1 - \text{erf}\left(\frac{\sqrt{\bar{\alpha}_t} d(\mathbf{x},\mathbf{c},\hat{\mathbf{c}},t,\boldsymbol{\epsilon})}{2\sqrt{2(1-\bar{\alpha}_t)}}\right)\right], \tag{10}$$

where $\bar{\alpha}_t$ is the variance schedule, $d(\cdot)$ is the triplet distance in Eq. 8, and $\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2} dt$ denotes the error function which normalizes the triplet distances. This equation calculates the pixel-level difference between images conditioned on different classes, serving as a measure of their distinguishability at a given timestep t. However, for tasks where diffusion classifier (*i.e.*, Eq. 6) perform at chance levels, the class attribute loss is ineffective in identifying the optimal timepoint prior to training. In such cases, we can instead select a timepoint that has empirically demonstrated consistent effectiveness across diverse datasets (for Stable Diffusion, we found $t^* = 200$ to be a reliable choice across domains).

Finally, we reformulate the TLSE training loss in Eq. 9 and the decision rule in Eq. 6 with this optimal timestep (see Algorithm 1):

$$\mathcal{L}^*_{TLSE} := \log \sum_{\{(\mathbf{x},\mathbf{c}),\hat{\mathbf{c}},\boldsymbol{\epsilon}\}} \exp\{d(\mathbf{x},\mathbf{c},\hat{\mathbf{c}},t^*,\boldsymbol{\epsilon})\}, \tag{11}$$

$$y^* = \arg\min_i \mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_{t^*},\mathbf{c}_i,t^*) - \boldsymbol{\epsilon}\|^2\right]. \tag{12}$$

Method (%)	Zero-shot					Fully Supervised				
	CLIP	MedCLIP	DiffCLS	DiffTTA	**Ours**	CoOp	CoCoOp	MaPLe	DiffTTA	**Ours**
Breast	69.11	**81.23**	52.13	56.69	<u>73.98</u>	76.19	77.04	**81.20**	59.23	<u>79.44</u>
Chest	62.52	**87.52**	63.12	57.24	<u>65.23</u>	85.88	86.06	<u>87.11</u>	65.32	**88.62**
Camelyon	58.07	**65.40**	51.31	53.00	<u>64.91</u>	61.35	59.29	<u>67.29</u>	59.29	**93.70**
EuroSAT	58.13	36.24	12.40	**72.34**	<u>62.33</u>	83.47	75.74	<u>91.67</u>	75.39	**93.22**

Table 1: Comparison across different methods on various datasets in zero-shot and fully supervised settings (measured in accuracy (%), highest value is **bolded**, second highest value is <u>underlined</u>). ExpertDiff demonstrates competitive performance in both zero-shot and fully supervised settings, with the exception of MedCLIP, which excel on medical data but struggle with non-medical domains like EuroSAT.

4.3 Prompt Design

We design the prompt $\mathbf{c}$ as a learnable parameter in the diffusion model. This design allows us to optimize the prompt specifically for the downstream task, enhancing the model's ability to follow task-specific instructions. Our prompt design incorporates two visual prompting techniques commonly used in transformer architectures [Li and Liang, 2021]: (1) an additive prompt, and (2) a suffix prompt. The additive prompt is added to the input embeddings, while the suffix prompt is concatenated with the input embeddings. We initialize the additive prompt using the transformed task template (e.g., 'a photo of airplane'), while the suffix prompts are initialized as zero vectors. The suffix prompt is then concatenated with the additive prompt embedding, resulting in a learnable prompt embedding $\mathbf{c}_i \in \mathbb{R}^{2|\text{token}| \times \text{dim}}$. In off-the-shelf Stable Diffusion models, the number of tokens (*i.e.*, $|\text{token}|$) is typically 77, and the embedding dimension (*i.e.*, dim) is 1024. This configuration allows for flexible and learnable prompt representations within the established Stable Diffusion framework.

4.4 Interesting Properties of the TLSE Loss

Proposition 1 (Proof in Appendix A.1)**.** *The TLSE loss is robust to outlier negative samples, assigning them lower gradients compared to boundary negative samples.*

Proposition 2 (Proof in Appendix A.2)**.** *The TLSE loss is smooth with respect to the model parameters θ, allowing for stable optimization.*

The TLSE loss's robustness to outlier negatives prevents overfitting to difficult samples, which is particularly beneficial when reprogramming diffusion models for new domains where some class distinctions may be subtle. Additionally, the smoothness of TLSE with respect to model parameters allows for stable optimization, enabling effective fine-tuning of the prompt embeddings without destabilizing the pre-trained diffusion model weights.

5 Experiments

In this section, we present a comprehensive evaluation of our proposed ExpertDiff method across various experimental settings. Our experiments are designed to assess ExpertDiff's performance and adaptability in three key OOD scenarios: zero-shot learning, few-shot domain generalization, and fully supervised learning. Through these experiments, we aim to answer couple of important questions: (1) How well does ExpertDiff perform in domains requiring expert knowledge, particularly in scenarios with limited labeled data? (2) Can ExpertDiff effectively generalize to unseen domains, especially in challenging medical imaging tasks? Additionally, we conduct qualitative evaluations to examine ExpertDiff's ability to reconstruct input images and perform image editing, demonstrating its potential for both downstream discriminative and upstream generative tasks. Finally, we present ablation studies to analyze the impact of various components of our method.

5.1 Experimental Setup

In this section, we briefly describe the experimental setup for evaluating the proposed method. Further details are provided in Appendix C.

Dataset

To test the effectiveness of our proposed method, we conduct experiments across 3 medical datasets (**Breast** ultrasound [Al-Dhabyani *et al.*, 2020], **Chest** X-ray [Kermany and others, 2018], and **Camelyon**17-WILDS breast cancer miscropy [Sagawa *et al.*, 2022]) and the **EuroSAT** satellite dataset [Helber *et al.*, 2018]. This diverse selection of datasets is designed to encompass a broad spectrum of domains and modalities, each demanding varying degrees of expert knowledge.

Implementation Details

For our proposed method, we utilize the publicly available pre-trained Stable Diffusion[1] as the backbone architecture. However, our method is not limited to this specific model and can be applied to most text-to-image diffusion models. The core components of our method, *i.e.*, the TLSE loss (Eq. 9, 11) and optimal timestep search (Eq. 10), are model-agnostic and only require access to the noise prediction network, which is a standard component in most text-to-image diffusion models. We trained our model for 100,000 iterations for fully supervised learning and 20,000 iterations for few-shot learning using a single Nvidia RTX 4090. Following DiffTTA's zero-shot setting [Prabhudesai *et al.*, 2023], we sample random Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$, pair it with each class's prompt embedding $\mathbf{c}$, and partially reverse-diffuse it via Eq. 2, producing synthetic triplets $(\mathbf{x}_{t^\star}, \mathbf{c}, \hat{\mathbf{c}})$. Then, we train the model for 1,000 iterations using these synthetic samples.

[1] https://hf.co/stabilityai/stable-diffusion-2-1

Method	1-Shot	10-Shot	100-Shot	Full
ContriMix	50.0±1.21	50.0±0.97	53.1±1.42	**94.6**±0.27
CoOp	56.2±3.48	57.3±1.32	59.3±1.76	61.4±0.62
CoCoOp	57.2±3.36	58.3±2.08	59.1±2.61	59.2±1.73
MaPLe	62.3±4.02	64.2±1.37	66.4±0.64	67.2±0.55
CLIP-LP	61.2±0.83	61.6±0.72	69.3±0.38	85.4±0.16
DiffTTA	52.5±1.25	54.4±1.08	55.0±0.97	59.2±0.79
Ours	**83.7**±3.05	**83.7**±1.91	**84.2**±0.68	93.7±0.41

Table 2: Performance comparison of different methods on the Camelyon17-WILDS dataset for few-shot and fully supervised domain generalization (measured in accuracy (%), best results in bold, standard deviation calculated across 25 random seeds). While ExpertDiff shows consistent improvement with increased data, other methods struggle to scale effectively, and the discriminative model excels only with full supervision but performs at change levels in few-shot scenarios.

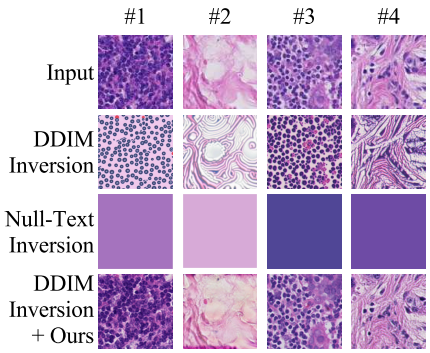

Figure 3: Image reconstruction results. DDIM inversion and Null-text inversion were performed using the prompt *"a histopathological image of lymph node containing metastatic tumor tissue."*

5.2 Quantitative Evaluation

To evaluate the effectiveness of our proposed ExpertDiff method, we conducted extensive experiments in zero-shot, fully supervised, and few-shot domain generalization settings (see Table 1 and 2).

Zero-shot and Fully Supervised Learning

We compared ExpertDiff against three categories of methods:

1. CLIP-based zero-shot learning: **CLIP** [Radford *et al.*, 2021] and **MedCLIP** [Wang *et al.*, 2022],
2. Diffusion classifiers: **DiffCLS** [Clark and Jaini, 2023], **DiffTTA** [Prabhudesai *et al.*, 2023],
3. Prompt learning: **CoOp** [Zhou *et al.*, 2022b], **CoCoOp** [Zhou *et al.*, 2022a], and **MaPLe** [Khattak *et al.*, 2023].

We use the ViT-H/14 CLIP model[2] for all of the methods and use Stable Diffusion V2.1[1] for diffusion classifiers. For fully supervised learning, all prompt learning methods were trained for 20 epochs, and ExpertDiff and DiffTTA was trained for 100,000 iterations. For a fair competition, we fixed the number of Monte-carlo sampling to 1,000 for diffusion classifiers (*i.e.*, 50 uniformly distributed timepoints $\times$ 20 noise samples per timepoint) and ExpertDiff (*i.e.*, 1 optimal timepoint $\times$ 1,000 noise samples). As shown in Table 1, ExpertDiff achieves competitive performance across all datasets in the zero-shot and fully supervised settings. While MedCLIP shows superior performance on medical datasets in the zero-shot scenario, it's important to note that MedCLIP is specifically pre-trained on medical data, resulting in a detrimental OOD performance on EuroSAT dataset.

Few-shot Domain Generalization

To evaluate the generalizability of our proposed ExpertDiff method, we conducted extensive experiments using the Camelyon17-WILDS dataset [Koh and others, 2021], where the task is to classify samples from unseen domains in domain generalization settings (see Table 2). Specifically, this dataset presents a challenging task of classifying breast cancer metastases in whole-slide histological images of lymph node sections, with images sourced from multiple hospitals representing different domains. We compared ExpertDiff against **ContriMix** [Nguyen *et al.*, 2024], which is the highest ranking discriminative model in the official leaderboard[3] at the time of writing, **CLIP-LP** [Radford *et al.*, 2021], which is the fine-tuned linear probe with CLIP embeddings, prompt learning methods, and DiffTTA.

Notably, ExpertDiff's performance shows consistent superiority across all data regimes, from zero-shot to fully supervised. This is in contrast to other methods like CoOp, CoCoOp, and DiffTTA, which show limited improvement or even decreased performance as more training data becomes available. The discriminative model ourperforms the proposed method in fully supervised settings, but tends to overfit or perform at near chance levels in few-shot scenarios. These results demonstrate that ExpertDiff can effectively adapt to challenging domain generalization scenarios, showing robust performance even with limited data.

5.3 Qualitative Evaluation

To assess our proposed method's modularity and its ability to preserve crucial visual information, we evaluated its performance on the original upstream generative task. Specifically, we examined its capacity to faithfully reconstruct input images, which is a critical feature, particularly in domains like medical imaging where precise detail is paramount. Figure 3 presents reconstructions using three methods: DDIM inversion, null-text inversion [Mokady *et al.*, 2023], and DDIM inversion with the fine-tuned prompt embedding from the proposed ExpertDiff. DDIM inversion employs a reverse application of the DDIM sampling algorithm to recover an input image's latent representation. Null-text inversion [Mokady *et al.*, 2023], on the other hand, optimizes the unconditional embedding used in classifier-free guidance while maintaining model weights and conditional embedding.

As shown in Figure 3, DDIM inversion produces an unrealistic reconstruction, while null-text inversion fails entirely in this context. This rather supurising result adds evidence that

[2]https://github.com/mlfoundations/open_clip

[3]https://wilds.stanford.edu/leaderboard/#camelyon17

Dataset	Add.	Suffix	Both	Linear
Breast	78.23	78.11	79.44	**80.15**
Chest	82.44	87.64	**88.62**	83.26
Camelyon	92.13	91.23	**93.70**	91.61
EuroSAT	92.33	92.22	**93.22**	92.63

Table 3: Comparison of classification accuracy (%) across different datasets using various prompt designs (additive, suffix, both) and a linear projector.

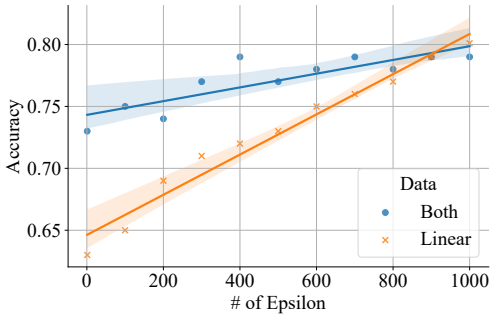

Figure 4: Classification accurary with regards to number of noise samples used for inference on Breast dataset. Model is trained and tested using all timepoints (*i.e.*, Eq. 9 and 6).

off-the-shelf diffusion models do not generalize well to new and unfamiliar concepts. In contrast, ExpertDiff's head-less reprogramming approach can effectively leverage the pre-trained diffusion model's capabilities for upstream generative tasks in novel domains. The enhanced reconstruction quality not only indicates ExpertDiff's successful adaptation to new classification tasks but also highlights its ability to retain the model's generative capabilities.

5.4 Ablation Studies

To evaluate the effectiveness of different components in our proposed ExpertDiff method, we conducted a series of ablation studies. These experiments focus on three key aspects: the impact of prompt design, the effect of timestep selection, and the influence of the number of noise samples used during inference.

Prompt Design

We first examined the impact of different prompt designs on classification accuracy across four datasets. Table 3 presents the results comparing additive prompts, suffix prompts, a combination of both, and a linear projection layer (*i.e.*, prompt embedding is fixed for this case). The results demonstrate that combining both additive and suffix prompts consistently outperforms using either type alone across most datasets. Interestingly, for the Breast dataset, the linear classifier baseline slightly outperforms our prompt-based methods, suggesting that simpler models may sometimes be sufficient for certain tasks.

Timestep Selection

We next investigated the impact of timestep selection on classification accuracy. Figure 5 shows the classification accuracy

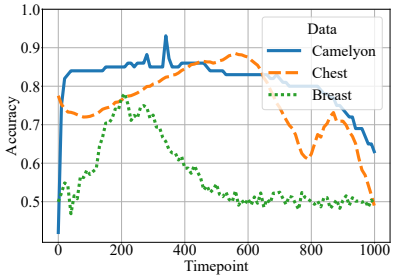

Figure 5: Classification accurary with regards to single timepoint. Model is trained using all timepoints (*i.e.*, Eq. 9), but used only a single timepoint during inference (*i.e.*, Eq. 12).

for different single timepoints across three datasets (Camelyon, Chest, and Breast). The results reveal that the optimal timepoint varies across datasets, with Camelyon showing peak performance around timepoint 200, while Chest and Breast datasets exhibit more complex patterns. This variability underscores the importance of careful timepoint selection for each specific task or dataset. Notably, the performance tends to degrade at very early and very late timepoints, suggesting that intermediate timepoints often provide the most discriminative features for classification.

Number of Noise Samples

Lastly, we examined how the number of noise samples used during inference affects classification accuracy. Interestingly, Figure 4 and Table 3 reveals that linear projection methods may indeed perform better in some cases, particularly when a large number of inference noise samples are used. However, their performance degrades significantly as the number of noise samples decreases, especially in the range of 0-100 samples. This observation highlights a key advantage of our ExpertDiff method: it maintains more robust performance across a wider range of noise sample counts, making it potentially more versatile and reliable in scenarios where computational resources may be limited or variable.

6 Conclusion

In conclusion, we presented ExpertDiff, a head-less model reprogramming technique that effectively adapts diffusion classifiers to out-of-distribution domains requiring expert knowledge. Our method demonstrates superior performance across zero-shot, few-shot, and fully supervised learning scenarios, while also preserving the model's generative capabilities. This work paves the way for efficiently repurposing powerful foundation diffusion models for novel OOD applications in domains like medical imaging. One **limitation and future research direction** of our proposed ExpertDiff is counterfactual image editing. As the goal of the ExpertDiff's loss function (*i.e.*, Eq 9) is to minimize the error for positive samples and maximize it for negative samples, conditional image generation with incorrect prompt results in samples filled with artifacts. Although this property is desirable for downstream discriminative tasks, it limits the applications in upstream generative tasks. We hope this work brings new insights into foundation model adaptation and efficient cross-domain transfer learning.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00457882, National AI Research Lab Project; No. RS-2022-II220959, (Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making)

References

[Al-Dhabyani *et al.*, 2020] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

[Bhattacharya and Prasanna, 2024] Moinak Bhattacharya and Prateek Prasanna. Gazediff: A radiologist visual attention guided diffusion model for zero-shot disease classification. In *MIDL*, 2024.

[Bommasani *et al.*, 2022] Rishi Bommasani, Drew A. Hudson, et al. On the opportunities and risks of foundation models, 2022.

[Bordes *et al.*, 2022] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *Transactions on Machine Learning Research*, 2022.

[Chen *et al.*, 2023] Huanran Chen, Yinpeng Dong, et al. Robust classification via a single diffusion model, 2023.

[Chen *et al.*, 2024] Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, et al. Your diffusion model is secretly a certifiably robust classifier, 2024.

[Chen, 2024] Pin-Yu Chen. Model reprogramming: resource-efficient cross-domain machine learning. In *AAAI Conference on Artificial Intelligence*, 2024.

[Clark and Jaini, 2023] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Advances in Neural Information Processing Systems*, 2023.

[Fini *et al.*, 2023] Enrico Fini, Pietro Astolfi, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. Improved baselines for vision-language pre-training. *Transactions on Machine Learning Research*, 2023. Featured Certification.

[Gandelsman *et al.*, 2024] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.

[He *et al.*, 2023] Xuehai He, Weixi Feng, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. Discffusion: Discriminative diffusion models as few-shot vision and language learners, 2023.

[Helber *et al.*, 2018] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 204–207, 2018.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[Huang *et al.*, 2024] Zhuo Huang, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, and Tongliang Liu. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. In *Forty-first International Conference on Machine Learning*, 2024.

[Jaini *et al.*, 2024] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *The Twelfth International Conference on Learning Representations*, 2024.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[Kermany and others, 2018] Daniel S. Kermany et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018.

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, June 2023.

[Koh and others, 2021] Pang Wei Koh et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

[Krojer *et al.*, 2023] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Chris Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 8385–8405. Curran Associates, Inc., 2023.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[Li *et al.*, 2023] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2206–2217, October 2023.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[Mukhopadhyay *et al.*, 2023] Soumik Mukhopadhyay, Matthew Gwilliam, et al. Do text-free diffusion models learn discriminative visual representations?, 2023.

[Ng and Jordan, 2001] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[Nguyen *et al.*, 2024] Tan H. Nguyen, Dinkar Juyal, et al. Contrimix: Scalable stain color augmentation for domain generalization without domain labels in digital pathology, 2024.

[Prabhudesai *et al.*, 2023] Mihir Prabhudesai, Tsung-Wei Ke, Alexander Cong Li, Deepak Pathak, and Katerina Fragkiadaki. Test-time adaptation of discriminative models via diffusion generative feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Radford *et al.*, 2021] Alec Radford, Kim Jong Wook, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[Ruiz *et al.*, 2023] Nataniel Ruiz, Yuanzhen Li, et al. 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 (CVPR)*, pages 22500–22510, June 2023.

[Sagawa *et al.*, 2022] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. In *International Conference on Learning Representations*, 2022.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[Udandarao *et al.*, 2024a] Vishaal Udandarao, Max F Burg, Samuel Albanie, and Matthias Bethge. Visual data-type understanding does not emerge from scaling vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[Udandarao *et al.*, 2024b] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H. S. Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance, 2024.

[Vilouras *et al.*, 2024] Konstantinos Vilouras, Pedro Sanchez, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Zero-shot medical phrase grounding with off-the-shelf diffusion models, 2024.

[Wang *et al.*, 2022] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

[Wei *et al.*, 2022] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.

[Yang *et al.*, 2023a] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[Yang *et al.*, 2023b] Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. Out-of-distribution generalization in natural language processing: Past, present, and future. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore, December 2023. Association for Computational Linguistics.

[Yue *et al.*, 2024] Zhongqi Yue, Pan Zhou, Richang Hong, Hanwang Zhang, and Qianru Sun. Few-shot learner parameterization by diffusion time-steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23263–23272, June 2024.

[Zhang *et al.*, 2023] Shengyu Zhang, Linfeng Dong, et al. Instruction tuning for large language models: A survey, 2023.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[Zhou *et al.*, 2024] Jingqiu Zhou, Aojun Zhou, and Hongsheng Li. Nodi: Out-of-distribution detection with noise from diffusion, 2024.