Any-Shift Prompting for Generalization over Distributions

Zehao Xiao[1] Jiayi Shen[1] Mohammad Mahdi Derakhshani[1]
Shengcai Liao[2] Cees G. M. Snoek[1]
[1]University of Amsterdam [2]Core42

Abstract

Image-language models with prompt learning have shown remarkable advances in numerous downstream vision tasks. Nevertheless, conventional prompt learning methods overfit their training distribution and lose the generalization ability on test distributions. To improve generalization across various distribution shifts, we propose any-shift prompting: a general probabilistic inference framework that considers the relationship between training and test distributions during prompt learning. We explicitly connect training and test distributions in the latent space by constructing training and test prompts in a hierarchical architecture. Within this framework, the test prompt exploits the distribution relationships to guide the generalization of the CLIP image-language model from training to any test distribution. To effectively encode the distribution information and their relationships, we further introduce a transformer inference network with a pseudo-shift training mechanism. The network generates the tailored test prompt with both training and test information in a feedforward pass, avoiding extra training costs at test time. Extensive experiments on twenty-three datasets demonstrate the effectiveness of any-shift prompting on the generalization over various distribution shifts.

1. Introduction

Recent image-language foundation models like CLIP [52] show remarkable advances in various computer vision tasks. Benefiting from large image-text pairing datasets for pretraining, these models perform well when adapting to downstream tasks by manual prompts [37, 48, 53, 56] and prompt learning [82, 83]. However, it is difficult for conventional prompt learning approaches to handle distribution shifts in downstream tasks [8, 61]. The learned prompts usually overfit their training data, leading to performance degradation on unseen test distributions.

To improve generalization of prompt learning, recent methods introduce uncertainty into the learnable prompt [8] or fine-tune the prompt on each test sample with extra unsupervised optimizations [57, 61]. Nevertheless, these

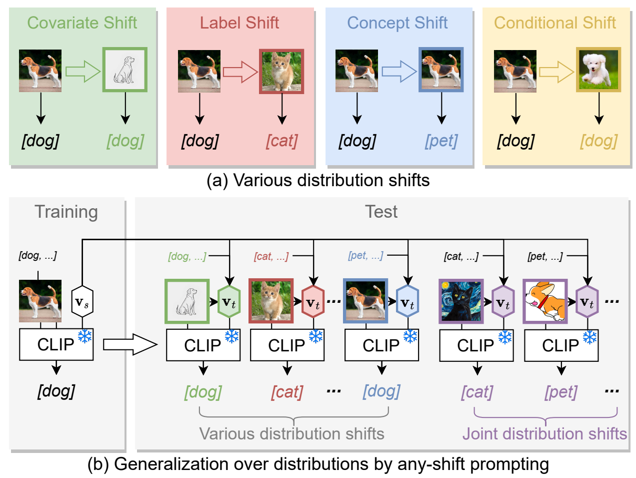

Figure 1. **Any-shift prompting.** (a) Various distribution shifts in real-world applications. (b) We propose any-shift prompting that aggregates training and test information for jointly handling individual distribution shifts and their combinations.

methods do not explicitly consider the relationships between training and test distributions of the downstream tasks. However, in real-world applications, the distribution shifts are usually complex and unpredictable, where models may encounter different distribution shifts (Figure 1 (a)), and even their combinations. Hence, we deem it crucial to explore the relationships between training and test distributions for the generalization of prompting across different distribution shifts. To this end, we make three contributions in this paper.

First, we propose any-shift prompting, a general probabilistic inference framework that can explore distribution relationships in prompt learning. Specifically, we introduce probabilistic training and test prompts in a hierarchical architecture to explicitly connect the training and test distributions. Within this framework, the test prompt encodes the test information and the relationships of the training and test distributions, thereby improving the generalization ability on various test distributions (Figure 1 (b)).

Second, we propose a pseudo-shift training mechanism, where the hierarchical probabilistic model learns the ability to encode distribution relationships by simulating distribution shifts. Consequently, at test time, our method generalizes to any specific distribution by generating a tailored prompt on the fly in just one feedforward process, without the need for re-learning or fine-tuning.

Third, to effectively and comprehensively encode the distribution information and their relationships, we design a transformer inference network for prompt generation. The transformer takes test information of both image and label space features, as well as the training prompts, as inputs. It then aggregates the training and test information and their relationships into the test-specific prompt. The test prompt is utilized to guide both the feature extraction and classification processes to generate test-specific features and classifiers, which bolsters robust predictions across distribution shifts.

We validate our method through extensive experiments on twenty-three benchmarks with various distribution shifts, including covariate shift, label shift, conditional shift, concept shift, and even joint shift. The results demonstrate the effectiveness of the proposed method on generalization across various distribution shifts.

2. Preliminary

We propose any-shift prompting based on CLIP [52] to handle various distribution shifts in a general way. Here we provide the technical background on CLIP as well as definitions of distribution shifts considered.

CLIP model. Contrastive Language-Image Pre-training (CLIP) [52] consists of an image encoder $f_{\Phi_I}(\mathbf{x})$ and a text encoder $f_{\Phi_T}(\mathbf{l})$, which are trained by a contrastive loss on a large dataset of image-language $(\mathbf{x}, \mathbf{l})$ pairs. For a downstream classification task with an input image $\mathbf{x}$ and a set of class names $\mathcal{Y}=\{c_i\}_{i=1}^{C}$, the image feature is extracted by $\mathbf{z}=f_{\Phi_I}(\mathbf{x})$ and the classifiers are composed of a set of text features $\{\mathbf{t}_i\}_{i=1}^{C}$, where $\mathbf{t}_i=f_{\Phi_T}(\mathbf{l}_i)$. Here, $\mathbf{l}_i$ is a manually crafted prompt to describe the corresponding class name c_i, e.g., "*an image of a [class]*." Thus, the prediction function of the CLIP model for downstream tasks without fine-tuning is formulated as:

$$p(\mathbf{y}|\mathbf{x}, \mathcal{Y})=\mathrm{softmax}(\mathbf{z}^\top \mathbf{t}). \tag{1}$$

This enables the pre-trained CLIP model to handle zero-shot learning classification in various downstream tasks.

Distribution shifts. A data distribution is generally denoted as $p(\mathbf{x}, \mathbf{y})$, which is a joint distribution of the input data $\mathbf{x}$ and the label $\mathbf{y}$. The models are usually trained on a training distribution $p(\mathbf{x}_s, \mathbf{y}_s)$ and then deployed on test distributions $p(\mathbf{x}_t, \mathbf{y}_t)$. In real-world applications, differences between the training and test distributions are known as the joint distribution shift:

$$p(\mathbf{x}_s, \mathbf{y}_s) \neq p(\mathbf{x}_t, \mathbf{y}_t). \tag{2}$$

Common distribution shifts in the literature. Due to the joint distribution shift, the performance of the trained model degrades on the test data [36, 67], sometimes significantly so. Since the joint distribution shift is complex, previous methods limit the scope of the problem and simplify the joint distribution shift to different partial distribution shifts. From a Bayesian perspective, the joint distribution is decomposed into $p(\mathbf{x}, \mathbf{y})=p(\mathbf{x})p(\mathbf{y}|\mathbf{x})=p(\mathbf{y})p(\mathbf{x}|\mathbf{y})$. According to the different components in the decomposition, we summarize the partial distribution shifts into four different definitions in Table 1 and detail them one by one.

Joint distribution shift	$p(\mathbf{x}_s, \mathbf{y}_s) \neq p(\mathbf{x}_t, \mathbf{y}_t)$	
Partial distribution shifts		
Covariate shift	$p(\mathbf{x}_s) \neq p(\mathbf{x}_t)$	$p(\mathbf{y}_s|\mathbf{x}_s) = p(\mathbf{y}_t|\mathbf{x}_t)$
Label shift	$p(\mathbf{y}_s) \neq p(\mathbf{y}_t)$	$p(\mathbf{x}_s|\mathbf{y}_s) = p(\mathbf{x}_t|\mathbf{y}_t)$
Concept shift	$p(\mathbf{x}_s) = p(\mathbf{x}_t)$	$p(\mathbf{y}_s|\mathbf{x}_s) \neq p(\mathbf{y}_t|\mathbf{x}_t)$
Conditional shift	$p(\mathbf{y}_s) = p(\mathbf{y}_t)$	$p(\mathbf{x}_s|\mathbf{y}_s) \neq p(\mathbf{x}_t|\mathbf{y}_t)$

Table 1. **Common distribution shifts.** The joint distribution shift is usually decomposed into four partial shifts, which are investigated individually in the literature. By contrast, we focus in this paper on various shifts and even consider their combinations.

Covariate shift [31, 59, 64] assumes the distribution shifts occur only in the input space $p(\mathbf{x})$ while the labels given the input features $p(\mathbf{y}|\mathbf{x})$ remain the same, e.g., by image corruptions [21] or changing image styles [31, 51]. Covariate shift is widely investigated by domain generalization [31, 73, 81] and domain adaptation methods [36, 67]. *Label shift* focuses on the opposite problem, where the label distributions $p(\mathbf{y})$ are different, but the label-conditional distributions $p(\mathbf{x}|\mathbf{y})$ are the same [55, 65]. Previous methods generate datasets with uniform distribution $p(\mathbf{y})$ during training and different distributions at test time [2, 19, 70]. The classification of unknown classes can be treated as a specific and worse case of the label shift [38, 60, 82], where $p(\mathbf{y})=0$ for the unknown classes. *Concept shift* treats the distribution of input $p(\mathbf{x})$ the same while the conditional distributions $p(\mathbf{y}|\mathbf{x})$ are different, indicating different annotation methods for the same data distribution [39]. *Conditional shift* assumes the label distribution is the same while the conditional distribution $p(\mathbf{x}|\mathbf{y})$ are different [16, 38, 77], where different classes can have their own shift protocols on the input data, e.g., sub-population problems [29, 58].

Distribution shifts in this paper. Conventional prompting methods [82, 83] learn the prompt on the training distribution of the downstream task, which is easy to overfit and vulnerable to the above shifts [8, 61]. Moreover, in real-world scenarios, all distribution shifts may happen unpredictably, and even simultaneously. Hence, we propose to encode test information and the training-test relationships for generalization over distributions. Our method is not designed for specific partial distribution shifts. Instead, it is proposed to handle various shifts, even when they occur simultaneously.

3. Any-Shift Prompting

3.1. Prompt modeling

We propose any-shift prompting, a general probabilistic inference framework to explore distribution relationships. Specifically, we introduce training and test prompts as latent variables in a hierarchical architecture. The graphical model of our method is provided in Figure 2.

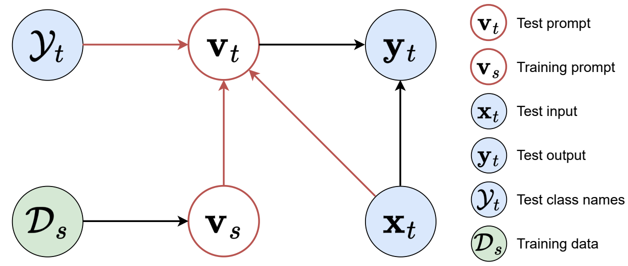

Figure 2. **Graphical model for any-shift prompting.** We introduce probabilistic training and test prompts in a hierarchical inference framework to explore distribution relationships.

Training prompt. The intuitive idea of adapting the CLIP model is to inject the downstream training data $\mathcal{D}_s$ in a training prompt for prediction (eq. 1). $\mathcal{D}_s$ consists of training input-output pairs sampled from the distribution $p(\mathbf{x}_s, \mathbf{y}_s)$. The predictive function of CLIP for the test distribution $p(\mathbf{x}_t, \mathbf{y}_t)$ is then formulated as:

$$p_\Phi(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_t, \mathcal{D}_s) \propto p_\Phi(\mathbf{y}_t|\mathbf{x}_t, \mathbf{v}_s, \mathcal{Y}_t) p(\mathbf{v}_s|\mathcal{D}_s), \quad (3)$$

where Φ denotes the frozen parameters of the image and text encoders of the CLIP model. Here $\mathbf{v}_s$ is the training prompt that encodes the training downstream task information, which improves the performance of the CLIP model on the training distribution. However, the prompt $\mathbf{v}_s$ usually overfits the training data, which may not benefit and even harm the prediction on the unseen test distribution due to the distribution shifts at test time.

Probabilistic test prompt. To generalize across distribution shifts in downstream tasks at test time, we further introduce a probabilistic test prompt within a hierarchical Bayes framework to encode the information of test distributions. Specifically, the test prompt $\mathbf{v}_t$ is inferred from the training prompt $\mathbf{v}_s$ and the accessible test information, i.e., a test image $\mathbf{x}_t$ and the class names $\mathcal{Y}_t$. To build the connections between the training and test prompts, we take the training prompt $\mathbf{v}_s$ as a condition for the generation of the test prompt. This enables the method to generate the test prompt across different shifts by considering the relationships between training and test distributions and exploring relevant training information. By introducing $\mathbf{v}_t$, the CLIP prediction function is formulated as:

$$\begin{aligned} &p_{\Phi,\theta}(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_t, \mathcal{D}_s) \\ &= \int\int p_\Phi(\mathbf{y}_t|\mathbf{x}_t, \mathbf{v}_t, \mathcal{Y}_t) p_{\boldsymbol{\theta}}(\mathbf{v}_t|\mathbf{v}_s, \mathbf{x}_t, \mathcal{Y}_t) p(\mathbf{v}_s|\mathcal{D}_s) d\mathbf{v}_t d\mathbf{v}_s, \end{aligned} \quad (4)$$

where $\boldsymbol{\theta}$ denotes the learnable inference network for the test prompt. With the probabilistic test prompt, we provide a general way to incorporate the training and test information, as well as their relationships, into the prediction of the CLIP model, enabling it to generalize on any test distribution.

Variational test prompt. To optimize the model for generating the probabilistic test prompt in eq. (4), we use variational inference to approximate the true posterior $p(\mathbf{v}_t, \mathbf{v}_s|\mathcal{D}_t, \mathcal{Y}_t, \mathcal{D}_s)$, which is factorized as:

$$q_{\boldsymbol{\theta}}(\mathbf{v}_t, \mathbf{v}_s|\mathcal{D}_t, \mathcal{Y}_t, \mathcal{D}_s) = q_{\boldsymbol{\theta}}(\mathbf{v}_t|\mathbf{v}_s, \mathcal{D}_t, \mathcal{Y}_t) p(\mathbf{v}_s|\mathcal{D}_s), \quad (5)$$

where $\mathcal{D}_t$ consists of test input-output pairs sampled from the test distribution $p(\mathbf{x}_t, \mathbf{y}_t)$. The variational posterior of the test prompt shares the same inference model $\boldsymbol{\theta}$ with its prior. By integrating eq. (5) into eq. (4), we derive the evidence lower bound (ELBO) of the predictive function as:

$$\begin{aligned} \log p_{\Phi,\boldsymbol{\theta}}(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_t, \mathcal{D}_s) \geq \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{v}_t, \mathbf{v}_s)}\big[\log p_\Phi(\mathbf{y}_t|\mathbf{x}_t, \mathbf{v}_t, \mathcal{Y}_t)\big] \\ - \mathbb{D}_{\mathrm{KL}}\big[q_{\boldsymbol{\theta}}(\mathbf{v}_t|\mathbf{v}_s, \mathcal{D}_t, \mathcal{Y}_t) || p_{\boldsymbol{\theta}}(\mathbf{v}_t|\mathbf{v}_s, \mathbf{x}_t, \mathcal{Y}_t)\big]. \end{aligned} \quad (6)$$

The variational posterior of the test prompt $q_{\boldsymbol{\theta}}(\mathbf{v}_t)$ encodes more input-output information of the test distribution and their relationships, yielding a more representative test prompt for better generalization on the test distributions. We provide the step-by-step derivations in the supplemental material.

Notably, the variational posteriors and ELBO are intractable since large numbers of test samples and their ground truth labels in $\mathcal{D}_t$ are usually unavailable at test time. Thus, in the next section, we propose a pseudo-shift training setup to approximate the ELBO for any-shift prompting.

3.2. Training and inference

Pseudo-shift training mechanism. To approximate the intractable ELBO in eq. (6), we develop a pseudo-shift training mechanism. Specifically, the mini-batch data in the current iteration is treated as the pseudo-test data $\mathcal{D}_{t'}$ from the pseudo-test distribution $p(\mathbf{x}_{t'}, \mathbf{y}_{t'})$. Likewise, the mini-batches in previous iterations are treated as the pseudo-training data $\mathcal{D}_{s'}$ from the pseudo-training distribution $p(\mathbf{x}_{s'}, \mathbf{y}_{s'})$. In this case, the ground truth labels of the pseudo-test data are available during training. We then approximate the ELBO and obtain the optimization function for any-shift prompting as:

$$\begin{aligned} \mathcal{L} = &-\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{v}_{t'}, \mathbf{v}_{s'})}\big[\log p_\Phi(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathbf{v}_{t'}, \mathcal{Y}_{t'})\big] \\ &+ \mathbb{D}_{\mathrm{KL}}\big[q_{\boldsymbol{\theta}}(\mathbf{v}_{t'}|\mathbf{v}_{s'}, \mathcal{D}_{t'}, \mathcal{Y}_{t'}) || p_{\boldsymbol{\theta}}(\mathbf{v}_{t'}|\mathbf{v}_{s'}, \mathbf{x}_{t'}, \mathcal{Y}_{t'})\big], \end{aligned} \quad (7)$$

where $\mathbf{v}_{t'}$ and $\mathbf{v}_{s'}$ denote the pseudo-test and pseudo-training prompts, respectively. In practice, we assume the prompts follow the standard Gaussian distributions. The negative log-likelihood in eq. (7) is implemented by a cross-entropy loss. The mini-batch training mechanism mimics the distribution shifts and trains the any-shift prompting to handle the distribution shifts during training, where the model never

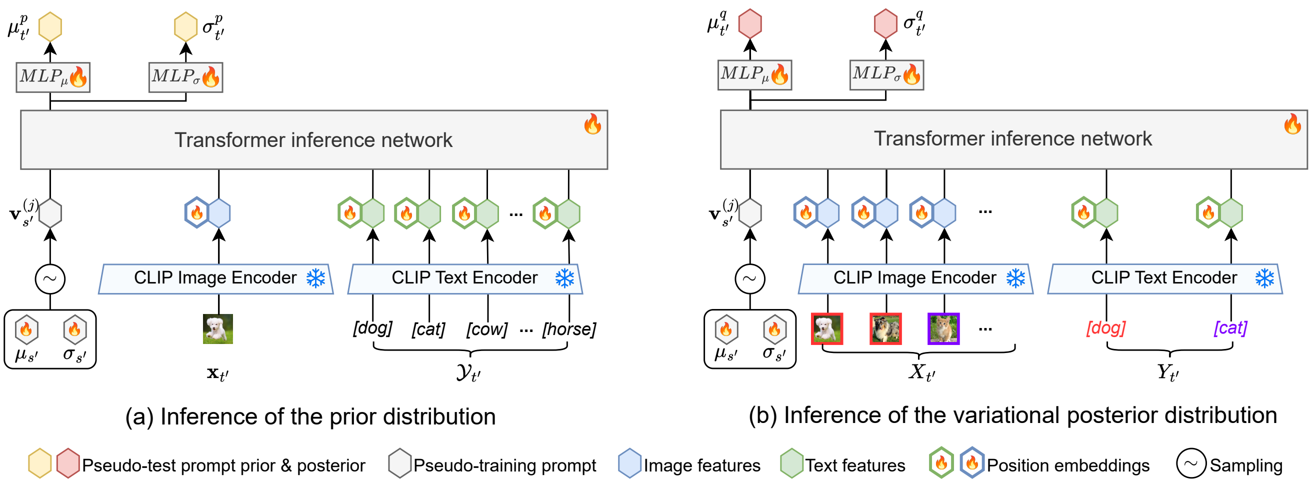

Figure 3. **Transformer inference network of the pseudo-test prompt.** The prior (a) of the pseudo-test prompt is inferred by aggregating the pseudo-training prompt, a single image, and all class names of the pseudo-test distribution. The posterior (b) is inferred from the shared pseudo-training prompt, a batch of pseudo-test images, and corresponding class names. Therefore, the posterior incorporates more pseudo-test information and relationships and guides the prior to learn the same knowledge by KL divergence. The image and text encoders of CLIP are frozen. Only the shared transformer, pseudo-training prompt distribution, and MLP networks are trainable, saving training costs.

accesses any test data. Minimizing the KL terms encourages the prior to implicitly learn more comprehensive pseudo-test information from the variational posterior, which aggregates more data information together with the ground truth labels.

Transformer inference network. The pseudo-test prompt in eq. (7) is inferred from: the pseudo-training information in $\mathbf{v}_{s'}$, the pseudo-test image $\mathbf{x}_{t'}$, and the class names $\mathcal{Y}_{t'}$. To better aggregate the different information sources and consider their relationships, we introduce a transformer inference network to generate the pseudo-test prompt.

In our model, the prior $p_{\boldsymbol{\theta}}(\mathbf{v}_{t'}|\mathbf{v}_{s'}, \mathbf{x}_{t'}, \mathcal{Y}_{t'})$ and variational posterior $q_{\boldsymbol{\theta}}(\mathbf{v}_{t'}|\mathbf{v}_{s'}, \mathcal{D}_{t'}, \mathcal{Y}_{t'})$ of the pseudo-test prompt share the same inference network to encode the different conditions. Compared with the prior, the variational posterior has access to one batch of pseudo-test images with the corresponding ground-truth labels. Figure 3 illustrates the deployment of the shared transformer inference network. In the following, we provide the detailed inference of the prior and variational posterior.

As shown in Figure 3 (a), the prior of the pseudo-test prompt is generated by the pseudo-training prompt $\mathbf{v}_{s'}$, the pseudo-test image $\mathbf{x}'_t$, and class names $\mathcal{Y}'_t$. Specifically, we sample a pseudo-training prompt $\mathbf{v}_{s'}^{(j)}$ from a Gaussian distribution $\mathcal{N}(\mathbf{v}_{s'}; \mu_{s'}, \sigma_{s'})$ by the reparameterization trick [27]. The mean $\mu_{s'}$ and variance $\sigma_{s'}$ are two sets of parameters trained with the pseudo-training data $\mathcal{D}_{s'}$ in the previous iterations. The pseudo-test image is fed into the fixed CLIP image encoder to get the image feature $f_{\Phi_I}(\mathbf{x}_{t'})$. The class names of the pseudo-test distribution are processed by the fixed text encoder to extract the textual features $f_{\Phi_T}(\mathcal{Y}_{t'})$. After the pre-processing, we take the sampled pseudo-training prompt, pseudo-test image feature, and textual features as input tokens of our transformer inference network to generate the prior of the pseudo-test prompt:

$$[\widetilde{\mathbf{v}}_{t'}^{p}; \cdot; \cdot] = \mathtt{Trans}([\mathbf{v}_{s'}^{(j)}; f_{\Phi_I}(\mathbf{x}_{t'}); f_{\Phi_T}(\mathcal{Y}_{t'})]), \tag{8}$$

$$\mu_{t'}^{p} = \mathrm{MLP}_{\mu}(\widetilde{\mathbf{v}}_{t'}^{p}), \quad \sigma_{t'}^{p} = \mathrm{MLP}_{\sigma}(\widetilde{\mathbf{v}}_{t'}^{p}), \tag{9}$$

$$p_{\boldsymbol{\theta}}(\mathbf{v}_{t'}|\mathbf{v}_{s'}, \mathbf{x}_{t'}; \mathcal{Y}_{t'}) = \mathcal{N}(\mathbf{v}_{t'}; \mu_{t'}^{p}, \sigma_{t'}^{p}). \tag{10}$$

The prior of the pseudo-test prompt follows the Gaussian distribution in eq. (10), whose mean and variance are obtained by two MLP networks on the output of the transformer $\widetilde{\mathbf{v}}_{t'}^{p}$.

In Figure 3 (b), with the pseudo-test data $\mathcal{D}_{t'}$, the variational posterior learns more distribution information as well as the relations between inputs and outputs. To be clearer, we rewrite the variational posterior $q_{\boldsymbol{\theta}}(\mathbf{v}_{t'}|\mathbf{v}_{s'}, \mathcal{D}_{t'}, \mathcal{Y}_{t'})$ as $q_{\boldsymbol{\theta}}(\mathbf{v}_{t'}|\mathbf{v}_{s'}, X_{t'}, Y_{t'})$, where $X_{t'}$ contains a batch of pseudo-test images in $\mathcal{D}_{t'}$ and $Y_{t'}$ consists of the ground truth class names of $X_{t'}$ in $\mathcal{Y}_{t'}$. Hence, the shared transformer takes all image features and their corresponding label features as input tokens to infer the variational posterior:

$$[\widetilde{\mathbf{v}}_{t'}^{q}; \cdot; \cdot] = \mathtt{Trans}([\mathbf{v}_{s'}^{(j)}; f_{\Phi_I}(X_{t'}); f_{\Phi_T}(Y_{t'})]), \tag{11}$$

$$\mu_{t'}^{q} = \mathrm{MLP}_{\mu}(\widetilde{\mathbf{v}}_{t'}^{q}), \quad \sigma_{t'}^{q} = \mathrm{MLP}_{\sigma}(\widetilde{\mathbf{v}}_{t'}^{q}), \tag{12}$$

$$q_{\boldsymbol{\theta}}(\mathbf{v}_{t'}|\mathbf{v}_{s'}, \mathcal{D}_{t'}, \mathcal{Y}_{t'}) = \mathcal{N}(\mathbf{v}_{t'}; \mu_{t'}^{q}, \sigma_{t'}^{q}). \tag{13}$$

With the inferred pseudo-test prompt, we take its samples from the variational posterior as the input tokens for both image and text encoders of CLIP to make predictions during training. Thus, although the encoders are fixed, the image and textual features are generalized by utilizing the distribution information in the prompts during the feature extraction and classification procedure, enabling the method to handle different distribution shifts.

Prediction. At test time, we make predictions on each test image $\mathbf{x}_t$ with the test prompt generated by the transformer inference network. Since the test data and labels in $\mathcal{D}_t$ are unavailable, the variational posterior becomes intractable. Thus, we sample the test prompt $\mathbf{v}_t^{(i)}$ from the prior distribution $p_{\boldsymbol{\theta}}(\mathbf{v}_t|\mathbf{v}_s^{(j)}, \mathbf{x}_t, \mathcal{Y}_t)$, where $\mathbf{v}_s^{(j)}$ is a sample of the training prompt following $p(\mathbf{v}_s|\mathcal{D}_s)$. $\mathbf{v}_t^{(i)}$ is then introduced into both the image and text encoders of the CLIP model for generalization and prediction as:

$$
\begin{aligned}
p_\Phi(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_t, \mathcal{D}_s) &= \frac{1}{N_t}\frac{1}{N_s}\sum_{i=1}^{N_t}\sum_{j=1}^{N_s} p_\Phi(\mathbf{y}_t|\mathbf{x}_t, \mathbf{v}_t^{(i)}, \mathcal{Y}_t), \\
\mathbf{v}_t^{(i)} \sim p_{\boldsymbol{\theta}}(\mathbf{v}_t|\mathbf{v}_s^{(j)}, \mathbf{x}_t, \mathcal{Y}_t), &\quad \mathbf{v}_s^{(j)} \sim p(\mathbf{v}_s|\mathcal{D}_s).
\end{aligned} \tag{14}
$$

Although the test data and their labels are not available at test time, the information in each test sample and all class names in the vocabulary of the test task are available to infer the prior of the test prompt. The ability to encode test information from a single test image and the class vocabulary is learned during training by minimizing the KL divergence between the prior and posterior. Note the CLIP image encoder and text encoder are always frozen. Only the test prompt changes for different test distributions by aggregating the training and test information in each test sample $\mathbf{x}_t$ and the class names $\mathcal{Y}_t$. In this case, we utilize the original generalization ability of CLIP to generate the test prompt for generalization on downstream tasks across various distribution shifts.

4. Related Work

Prompt learning. Image-language foundation models such as CLIP [52] and ALIGN [25] achieve significant advances in various downstream tasks. To adapt the foundation models to downstream tasks, adapter [14] and prompt learning methods [30, 34, 83] are proposed. Zhou *et al*. [83] propose a learnable prompt as the input of the language model in CLIP. To avoid forgetting the original knowledge in the CLIP model, Zhu *et al*. [84] and Yao *et al*. [75] guide prompt learning with hand-crafted prompts. Instead of generating prompts for the language model, Bahng *et al*. [3] introduce prompting of the image model. Khattak *et al*. [26] learn a joint prompt for both image and language encoders. Zhou *et al*. [82] introduce the imaging conditions into the language prompt to enhance the generalization ability of zero-shot performance. To further improve the generalization ability, Derakhshani *et al*. [8] propose Bayesian prompt learning, which considers the uncertainty in the learned prompts for zero-shot generalization. Shu *et al*. [61] and Hassan *et al*. [57] fine-tune the prompt at test time to a specific distribution. We also improve the generalization of prompt learning. Different from previous methods that consider uncertainty or fine-tune the prompt for specific distributions, we propose any-shift prompting that explicitly explores distribution information and relationships within a hierarchical probabilistic framework. The method generates the test-specific prompt on the fly for any test distribution.

Distribution shift generalization. Domain generalization [18, 33, 44, 81] and domain adaptation [35, 41, 69, 76] are the most widely investigated methods for handling distribution shifts. Some domain generalization methods train invariant models on the training distributions [1, 43, 73], which are assumed to be invariant on the test distributions also. To further improve the generalization ability, some methods [4, 10, 32] introduce meta-learning in domain generalization to mimic domain shifts during training. In this paper, we also simulate the distribution shift by a pseudo-shift training mechanism, which uses different mini-batches as distributions. To better utilize the test information for generalization without accessing the test data during training, Sun *et al*. [64] and Wang *et al*. [67] propose test-time adaptation, which fine-tunes the trained model on test data with self-supervised losses. The method is followed by many methods [17, 40, 46, 47, 78] due to its good generalization ability on covariate shift. In addition, test-time adaptation is also investigated with other methods like normalization statistics re-estimation [36, 59], or classifier adjustment [24, 74, 80]. Most of these methods focus on covariate shift [11, 17, 67, 74], such as changes of the image styles [31, 51] and corruptions [21]. Some other methods work on the conditional shift [15, 16, 38, 77] or label shift [15, 49, 65, 77]. We also utilize the test information for generalization, but without any test-time optimization. Different from the previous methods, we explicitly bridge the training and test information and explore their relationships to address various distribution shifts in a general way.

5. Experiments

Twenty-three datasets. To demonstrate the generalization ability of any-shift prompting, we evaluate the method on datasets with different distribution shifts. For covariate shift, we conduct experiments on the common domain generalization datasets, `PACS` [31], `Office-Home` [66], `VLCS` [12], and `DomainNet` [51], which contain images from different domains such as image styles. We also evaluate the model on covariate shifts of ImageNet [7] following Zhou *et al*. [82], where the model is trained on ImageNet with 16-shot images and evaluated on other variants `ImageNet-V2` [54], `ImageNet-(S)ketch` [68], `ImageNet-A` [23], and `ImageNet-R` [22]. For label shift, we follow the base-to-new class generalization from Zhou *et al*. [83], with 11 datasets that cover various tasks, `ImageNet` [7], `Caltech101` [13], `OxfordPets` [50], `StanfordCars` [28], `Flowers102` [45], `Food101` [5], `FGVCAircraft` [42], `SUN397` [72], `DTD` [6], `EuroSAT` [20], and `UCF101` [63]. For concept shift, we build a

Method	PACS	VLCS	Office-Home	DomainNet	ImageNet-V2	ImageNet-S	ImageNet-A	ImageNet-R
Prompting without test-time optimization								
CLIP [52]	96.13	81.43	80.35	54.08	60.83	46.15	47.77	73.96
CLIP-D [52]	96.65	80.70	81.51	56.24	-	-	-	-
CoOp [83]	96.45	82.51	82.12	58.82	64.20	47.99	49.71	75.21
CoCoOp [82]	97.00	83.89	82.77	59.43	64.07	48.75	50.63	76.18
DPL [79]	97.07	83.99	83.00	59.86	-	-	-	-
BPL [8]	-	-	-	-	64.23	49.20	**51.33**	77.00
This paper	**98.16** $\pm$ 0.4	**86.54** $\pm$ 0.4	**85.16** $\pm$ 0.6	**60.93** $\pm$ 0.6	**64.53** $\pm$ 0.2	**49.80** $\pm$ 0.5	**51.52** $\pm$ 0.6	**77.56** $\pm$ 0.4
Prompting with test-time optimization								
TPT [61]	97.25	84.33	83.45	59.90	63.45	47.94	54.77	77.06
CoOp + TPT [61]	97.85	85.06	84.32	60.65	**66.83**	49.29	57.95	77.27
CoCoOp + TPT [61]	97.95	85.55	84.54	60.44	64.85	48.47	**58.47**	78.65
This paper + TPT	**98.47** $\pm$ 0.4	**86.98** $\pm$ 0.4	**86.00** $\pm$ 0.8	**61.75** $\pm$ 0.8	**67.08** $\pm$ 0.6	**50.83** $\pm$ 0.6	**58.05** $\pm$ 0.5	**79.23** $\pm$ 0.5

Table 2. **Covariate shift comparison.** The experiments are conducted on eight domain generalization datasets, with average classification accuracy reported. Any-shift prompting achieves the best results compared with the original CLIP and other prompt learning methods, which demonstrates the generalization ability of our method on covariate shift. When combined with TPT's test-time optimization, promting methods in general, as well as our method improves further.

`ImageNet-Superclass` dataset, where we evaluate the ImageNet-trained model on super-classes in [58]. For conditional shift, we evaluate on the sub-population datasets `Living-17` and `Entity-30` [58], where the training and test distributions consist of the same classes with different subpopulations. To evaluate our method on the combination of different distribution shifts, we follow the open-domain generalization setting [62] on the Office-Home dataset, which contains four domains, Art, Clipart, Product, and Real-world. We refer to it as `Open-Office-Home`, which combines covariate shift and label shift. The detailed settings are provided in the supplemental materials.

Implementation details. Our model consists of the pretrained image and language encoders of CLIP [52], and the proposed transformer inference network to generate the test prompt. We use the ViT-B/16 [9] as the image encoder following [8, 82]. The pretrained image and language encoders of CLIP are frozen during training and inference. To generate the prior and variational posterior of the prompt, we use a 2-layer transformer in the inference network. As shown in Figure 3, the inputs of the transformer include the training prompt, the image features, and the class-name features. The distribution of the training prompts consists of two trainable vectors as the mean and variance respectively. The class-name tokens are generated by the hand-crafted tokens *"an image of a [class]"*. The transformer also contains two kinds of trainable position embeddings to indicate the image and language tokens. The introduced prompts are sampled from the corresponding distributions by the reparameterization trick [27]. More detailed implementations and hyperparameters are provided in the supplemental materials.

5.1. Results on various distribution shifts

Covariate shift. We conduct experiments on eight domain generalization datasets with covariate shift. The averaged results of classification accuracy for each dataset are provided in Table 2. We follow the leave-one-out protocol [31] for evaluation on the first four datasets, where the model evaluated on each test domain is trained on the other domains. The detailed results on each test domain are provided in the supplemental materials. For the last four datasets, we evaluate the same ImageNet-pretrained model on them individually. Our method outperforms the other prompt learning methods CoOp, CoCoOp, and DPL on all eight datasets. Note that the comparisons with the other prompt learning methods are fair since we generate the test prompt and make predictions in a single feedforward pass, without any optimization or backpropagation at test time. The proposed method also performs better on seven of the eight datasets compared with the test-time tuning method TPT, securing the second position on `ImageNet-A`. Moreover, since the proposed method learns the prompt and transformer network only during training, it can also be combined with test-time optimization. Then we obtain even better results, which are also competitive on `ImageNet-A`, indicating the effectiveness of any-shift prompting on covariate shift.

Label shift. We conduct the experiments on label shift following the base-to-new class generalization setting in Zhou *et al.* [82]. The results on eleven datasets and the averaged performance are provided in Table 3. Since our any-shift prompts encode both training and test information, as well as their relationships, it performs well in both base and new classes, therefore achieving the best overall Harmonic mean on the eleven datasets. Compared with the original CLIP model, the proposed method achieves better performance in the base classes, showing good adaptation to the downstream tasks with the training information. Compared with the other prompt learning methods CoOp [83], CoCoOp [82], BPL [8], and MaPLe [26], our method performs best in the new classes on seven of the eleven datasets and is competitive on the other four. This demonstrates the ability of the method to handle label shift by incorporating the distribution information and their relationships.

Concept shift. For concept shift, we conduct experiments on

(a) **Average over 11 datasets**.

	Base	New	H
CLIP	69.34	74.22	71.70
CoOp	**82.69**	63.22	71.66
CoCoOp	80.47	71.69	75.83
BPL	80.10	74.94	77.43
MaPLe	82.28	75.14	78.55
This paper	82.36	**76.30**	**79.21**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	70.43	73.10
BPL	-	70.93	-
MaPLe	**76.66**	70.54	73.47
This paper	76.63	**71.33**	**73.88**

(c) Caltech101.

	Base	New	H
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
CoCoOp	97.96	93.81	95.84
BPL	-	**94.93**	-
MaPLe	97.74	94.36	96.02
This paper	**98.28**	94.27	**96.23**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	97.69	96.43
BPL	-	**98.00**	-
MaPLe	95.43	97.76	96.58
This paper	**95.78**	97.80	**96.78**

(e) StanfordCars.

	Base	New	H
CLIP	63.37	74.89	68.65
CoOp	**78.12**	60.40	68.13
CoCoOp	70.49	73.59	72.01
BPL	-	73.23	-
MaPLe	72.94	74.00	73.47
This paper	73.05	**75.83**	**74.41**

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	**97.60**	59.67	74.06
CoCoOp	94.87	71.75	81.71
BPL	-	70.40	-
MaPLe	95.92	72.46	82.56
This paper	96.50	76.20	**85.16**

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	90.70	91.29	90.99
BPL	-	**92.13**	-
MaPLe	90.71	92.05	**91.38**
This paper	**90.87**	91.35	91.11

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	**36.29**	31.09
CoOp	**40.44**	22.30	28.75
CoCoOp	33.41	23.71	27.74
BPL	-	35.00	-
MaPLe	37.44	35.61	**36.50**
This paper	37.10	35.70	36.39

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
BPL	-	77.87	-
MaPLe	**80.82**	**78.70**	**79.75**
This paper	80.50	78.50	79.48

(j) DTD.

	Base	New	H
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
BPL	-	60.80	-
MaPLe	**80.36**	59.18	68.16
This paper	79.63	**61.98**	**69.71**

(k) EuroSAT.

	Base	New	H
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
CoCoOp	87.49	60.04	71.21
BPL	-	75.30	-
MaPLe	**94.07**	73.23	82.35
This paper	93.07	**77.63**	**84.65**

(l) UCF101.

	Base	New	H
CLIP	70.53	77.50	73.85
CoOp	**84.69**	56.05	67.46
CoCoOp	82.33	73.45	77.64
BPL	-	75.77	-
MaPLe	83.00	78.66	80.77
This paper	84.60	**78.70**	**81.54**

Table 3. **Label shift comparison**. The models are trained on the base classes with 16 shots and evaluated on both the base and new classes. We bold the **best** results and underline the runner-up. H denotes the Harmonic mean [71]. Our method performs well on both base and new classes, therefore achieving the best overall Harmonic mean, demonstrating the generalization ability across label shifts.

	Concept Shift	Conditional Shift	
Method	ImageNet-Superclass	Living-17	Entity-30
CLIP†	69.23	86.94	67.95
CoOp†	69.35	87.11	78.02
CoCoOp†	69.77	87.24	79.52
This paper	**71.12** ± 0.6	**88.41** ± 0.3	**81.74** ± 0.4

Table 4. **Concept shift and conditional shift comparison**. Results of the compared methods are based on the author-provided code.

Method	Art	Clipart	Product	Real	***Mean***
CLIP†	79.32	67.70	86.93	87.46	80.35
CLIP-D†	80.47	68.83	87.93	88.80	81.51
CoOp†	80.50	69.05	88.26	89.01	81.71
CoCoOp†	80.93	69.51	88.85	89.32	82.19
This paper	**83.40** ± 0.8	**72.53** ± 0.5	**91.24** ± 0.6	**90.84** ± 0.3	**84.50** ± 0.4

Table 5. **Multiple shifts comparison** on `Open-Office-Home`, including both covariate and label shifts.

the introduced `ImageNet-Superclass` dataset, where the same images are assigned with different annotations. To do so, we evaluate the ImageNet-trained model on the validation set with the superclass annotations. As shown in Table 4, the prompt learning methods achieve similar performance compared with the original CLIP. By contrast, our method improves the performance of CLIP by about 2%, indicating the ability to handle concept shift.

Conditional shift. We also conduct experiments on two datasets with conditional shift. The results are also reported in Table 4. The prompt learning methods perform similarly to CLIP while achieving more improvement on `Entity-30`. The reason can be that the class names of `Living-17` (e.g., wolf, fox) are more detailed than `Entity-30` (e.g., crustacean, carnivore, insect), revealing the importance of adapting the original CLIP model to downstream tasks in specific scenarios. Moreover, compared with the conventional prompt learning methods CoOp and CoCoOp, our method consistently improves the performance on both datasets and performs better, demonstrating the effectiveness of any-shift prompting for the conditional shift.

Joint distribution shift. In Table 5, we report the results on `Open-Office-Home` for the joint distribution shifts. Following Shu *et al.* [62], we assign data from different parts of classes in the training domains and evaluate the model on the test domain with both seen and unseen classes. Therefore, the model encounters covariate and label shifts jointly. As shown in Table 5, the CLIP-based zero-shot methods keep the same performance as the close-set generalization setting (Table 2) since they are kept frozen. The prompt learning methods perform slightly worse than the close-set setting. Our method outperforms the others on all test domains, showing the ability to handle joint distribution shifts.

Overall, our method achieves good performance on covariate, label, concept, conditional, and even joint shifts, demonstrating the effectiveness of handling various distribution shifts by considering the distribution information and their relationship with any-shift prompting.

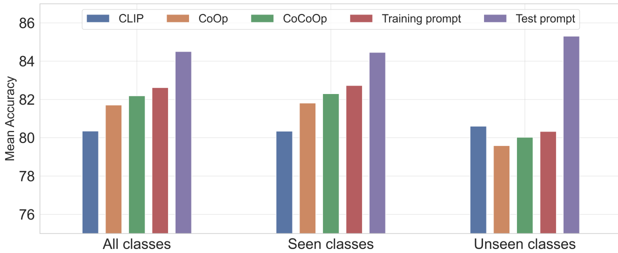

Figure 4. **Effectiveness of training and test prompts.** The test prompt in the proposed any-shift prompting achieves good generalization on both seen and unseen classes, indicating its ability to handle different shifts jointly.

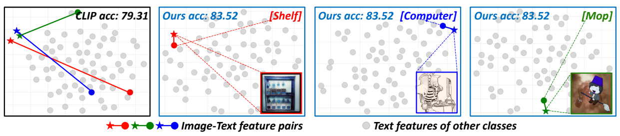

Figure 5. **Visualization of generalization effect** on the image and text features before and after generalization. Different colors denote different classes. The image and text features with the same categories get closer after generalization by our method, leading to more accurate predictions.

Training prompt $\mathbf{v}_s$	Test text feature of $\mathcal{Y}_t$	Test image feature of $\mathbf{x}_t$	Accuracy
✓			82.62
	✓		82.67
		✓	83.11
	✓	✓	83.63
✓	✓	✓	**84.50**

Table 6. **Benefits of training and test information in any-shift prompt.** The experiments are conducted across the joint shifts on `Open-Office-Home`. Both training and test information in the prompt benefit the method across joint shifts.

5.2. Ablation studies

Effectiveness of training and test prompts. To investigate the benefits of the training and test prompts of any-shift prompting, we evaluate our method with training and test prompts separately. The experiments are conducted on `Open-office-Home` with joint distribution shift. We compare the prompts with the original CLIP model as well as CoOp and CoCoOp in Figure 4, and provide the accuracy on all classes, seen classes, and unseen classes, respectively. CoOp and CoCoOp show better performance on seen classes across covariate shift but struggle in the unseen classes where both covariate shift and label shift exist. The training prompt in our method encounters the same problem since it encodes the training information with seen classes but also tends to overfit the training distribution. The performance is slightly better since it considers uncertainty in the prompt. By contrast, the test prompt in our method encodes the test information with the relationships between the training and test distribution. This enables the method to achieve good generalization across different shifts, leading to higher performance on both seen (covariate shift) and unseen classes (both covariate shift and label shift).

Visualization of generalization effect. To further show the benefits of generalization with our method, we visualize the image and text features before and after generalization by any-shift prompting. The experiments are conducted on the "Art" domain under `Open-Office-Home`. The image and text features before generalization are generated by the fixed CLIP image and language encoders respectively. As shown in Figure 5, after generalization by any-shift prompting, the image features get closer to the text features of the corresponding ground truth labels, which leads to more accurate predictions.

Benefits of training and test information in any-shift prompt. To show the benefits of considering different information in the test prompt, we conduct experiments on `Open-Office-Home`, which contains both covariate and label shifts. As shown in Table 6, using only the training prompt achieves better performance than CLIP (80.35) and we get similar results with only test text features or test image features. The information from the test images gains more improvement. The reason can be that test images include more unseen information in this setting. The test prompt generated by both image and text information further improves the generalization of test distributions, indicating the importance of considering test information for generalization. Moreover, including the training prompt provides the relationships and shift information between training and test distribution in the prompt, leading to the best performance.

6. Conclusion

We propose any-shift prompting to adapt the large image-language model (CLIP) to downstream tasks while enhancing the generalization ability across different distribution shifts at test time. The proposed method bridges the training and test distributions under a hierarchical probabilistic framework, which generates the specific prompt for each test sample by encoding the distribution information and relationships of the training and test distributions. Once trained, we generate the test-specific prompt across any distribution shift in a single feedforward pass without any fine-tuning or backpropagation. The test prompt generalizes both the image and language encoders of CLIP to the specific test distribution. Experiments on various distribution shifts, including covariate shift, label shift, conditional shift, concept shift, and joint shift, demonstrate the effectiveness of the proposed method on the generalization of any test distribution.

Acknowledgment

This work is financially supported by the Inception Institute of Artificial Intelligence, the University of Amsterdam, and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 5

[2] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019. 2

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 5

[4] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. MetaReg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, pages 998–1008, 2018. 5

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014. 5

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[8] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor G Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Bayesian prompt learning for image-language model generalization. In *IEEE International Conference on Computer Vision*, pages 15237–15246, 2023. 1, 2, 5, 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 6

[10] Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, 2019. 5

[11] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14340–14349, 2021. 5

[12] Yuming Fang, Weisi Lin, Zhenzhong Chen, Chia-Ming Tsai, and Chia-Wen Lin. A video saliency detection model in compressed domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(1):27–38, 2013. 5

[13] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE, 2004. 5

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 5

[15] Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. Rls-bench: Domain adaptation under relaxed label shift. In *International Conference on Machine Learning*, pages 10879–10928. PMLR, 2023. 5

[16] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848. PMLR, 2016. 2, 5

[17] Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and Zico Kolter. Test-time adaptation via conjugate pseudo-labels. In *Advances in Neural Information Processing Systems*, 2022. 5

[18] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020. 5

[19] Jiaxian Guo, Mingming Gong, Tongliang Liu, Kun Zhang, and Dacheng Tao. Ltf: A label transformation framework for correcting label shift. In *International Conference on Machine Learning*, pages 3843–3853. PMLR, 2020. 2

[20] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5

[21] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2, 5

[22] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE International Conference on Computer Vision*, pages 8340–8349, 2021. 5

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 5

[24] Yusuke Iwasawa et al. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, 2021. 5

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 5

[26] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 5, 6

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4, 6

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 5

[29] Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *International Conference on Learning Representations*, 2023. 2

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 5

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision*, pages 5542–5550, 2017. 2, 5, 6

[32] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, 2018. 5

[33] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 5

[34] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 5

[35] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 5

[36] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *International Conference on Learning Representations*, 2023. 2, 5

[37] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 1

[38] Xiaofeng Liu, Zhenhua Guo, Site Li, Fangxu Xing, Jane You, C-C Jay Kuo, Georges El Fakhri, and Jonghye Woo. Adversarial unsupervised domain adaptation with conditional and label shift: Infer, align and iterate. In *IEEE International Conference on Computer Vision*, pages 10367–10376, 2021. 2, 5

[39] Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, Jonghye Woo, et al. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022. 2

[40] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems*, 2021. 5

[41] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015. 5

[42] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[43] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *IEEE International Conference on Computer Vision*, pages 5715–5725, 2017. 5

[44] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 5

[45] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 5

[46] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, pages 16888–16905. PMLR, 2022. 5

[47] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. 5

[48] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 1

[49] Sunghyun Park, Seunghan Yang, Jaegul Choo, and Sungrack Yun. Label shift adapter for test-time adaptation under covariate and label shifts. In *IEEE International Conference on Computer Vision*, pages 16421–16431, 2023. 5

[50] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. 5

[51] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 2, 5

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5, 6

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[54] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to im-

agenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 5

[55] Manley Roberts, Pranav Mani, Saurabh Garg, and Zachary Lipton. Unsupervised learning under latent label shift. In *Advances in Neural Information Processing Systems*, pages 18763–18778, 2022. 2

[56] Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. *arXiv preprint arXiv:2306.07282*, 2023. 1

[57] Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Advances in Neural Information Processing Systems*, 2023. 1, 5

[58] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020. 2, 6

[59] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 11539–11551, 2020. 2, 5

[60] Jiayi Shen, Zehao Xiao, Xiantong Zhen, Cees Snoek, and Marcel Worring. Association graph learning for multi-task classification with category shifts. In *Advances in Neural Information Processing Systems*, pages 4503–4516, 2022. 2

[61] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, pages 14274–14289, 2022. 1, 2, 5, 6

[62] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9624–9633, 2021. 6, 7

[63] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[64] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020. 2, 5

[65] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In *Advances in Neural Information Processing Systems*, pages 19276–19289, 2020. 2, 5

[66] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 5

[67] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 2, 5

[68] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019. 5

[69] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 5

[70] Ruihan Wu, Chuan Guo, Yi Su, and Kilian Q Weinberger. Online adaptation to label distribution shift. In *Advances in Neural Information Processing Systems*, pages 11340–11351, 2021. 2

[71] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2018. 7

[72] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 5

[73] Zehao Xiao, Jiayi Shen, Xiantong Zhen, Ling Shao, and Cees G M Snoek. A bit more bayesian: Domain-invariant learning with uncertainty. In *International Conference on Machine Learning*. PMLR, 2021. 2, 5

[74] Zehao Xiao, Xiantong Zhen, Ling Shao, and Cees G M Snoek. Learning to generalize across domains on single test samples. In *International Conference on Learning Representations*, 2022. 5

[75] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 5

[76] Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *International Conference on Learning Representations*, 2023. 5

[77] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013. 2, 5

[78] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, pages 38629–38642, 2022. 5

[79] Xin Zhang, Shixiang Shane Gu, Yutaka Matsuo, and Yusuke Iwasawa. Domain prompt learning for efficiently adapting clip to unseen domains. *arXiv e-prints*, pages arXiv–2111, 2021. 6

[80] Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International Conference on Machine Learning*, pages 41647–41676. PMLR, 2023. 5

[81] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 2022. 2, 5

[82] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2, 5, 6

[83] 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. 1, 2, 5, 6

[84] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *IEEE International Conference on Computer Vision*, pages 15659–15669, 2023. 5