This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Zero-Shot Domain Generalisation via Prompt-Driven Feature Refinement

Tingrui Qiao
University of Auckland
tqia361@aucklanduni.ac.nz

Di Zhao
University of Auckland
dzha866@auckland.ac.nz

Caroline Walker
University of Auckland
caroline.walker@auckland.ac.nz

Chris Cunningham
Massey University
C.W.Cunningham@massey.ac.nz

Yun Sing Koh
University of Auckland
y.koh@auckland.ac.nz

Abstract

Domain generalisation aims to develop models that generalise from source domains to unseen target domains. However, most existing methods assume access to source domain data and require additional training, which may not always be practical. We focus on a more flexible and broadly applicable setting, zero-shot domain generalisation, where models generalise without access to source data, target data, or any additional training. In this work, we propose ***Prefer*** *(**pr**ompt-driv**e**n **fe**ature **r**efinement), a simple and effective approach that enhances the zero-shot domain generalisation ability of vision-language foundation models. Prefer generates a diverse set of textual prompts for each class by imagining domain-specific variations (e.g., "a painting of a cat under a golden sunset with thick brush strokes"), and uses them to probe the model. We evaluate how reliably each feature channel represents a class across domains by measuring two quantities: (1) how strongly the channel aligns with the original class prompt (e.g., "a photo of a cat") across the generated domain-specific prompts, and (2) how stable the channel remains across those prompts, quantified by its variance. Channels that exhibit both high alignment and low variability are selected at inference time to improve class prediction under domain shift. Without any model updates or external data, Prefer achieves consistent improvements across domain generalisation benchmarks, outperforming existing state-of-the-art methods.*

1. Introduction

While modern deep learning models have achieved remarkable success across a wide range of applications [23–25], they typically rely on the assumption that training and test data are drawn from the same underlying distribution. In practice, this assumption is frequently violated due to distribution shifts arising from changes in style, context, viewpoint, background, or imaging conditions, leading to significant performance degradation at test time [40]. To address this challenge, the task of Domain Adaptation (DA) has been introduced, which aims to train models on one or more source domains along with limited target domain data to enable generalisation to the target domain [10]. Domain Generalisation (DG) further relaxes this assumption by removing the need for target domain data, seeking to learn models that generalise to unseen domains using only source domain data [37, 38]. A rich body of work has emerged around this problem, exploring approaches such as invariant representation learning [2], data augmentation [39], meta-learning [15], and distributionally robust optimisation [27].

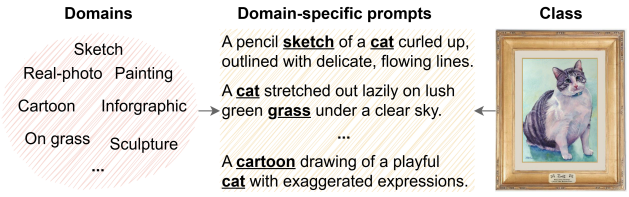

Figure 1. We use a language model to generate domain-specific prompts for each class, guiding robust feature selection for zero-shot inference on unseen domains without source data or training.

Despite the progress in domain generalisation, most existing methods still rely on access to source domain data and require training or fine-tuning [31]. In practice, however, access to source data may be restricted due to privacy regulations (e.g., in healthcare) [20], proprietary constraints (e.g., commercial systems) [21], or the high cost and time required to collect and curate domain-specific datasets. Moreover, while retraining or fine-tuning is feasible in controlled settings, it can be impractical in large-scale or rapidly changing environments. For instance, in conservation AI, researchers may need to deploy wildlife detection models in remote habitats, where collecting images of the animals is infeasible due to time, cost, and the need to min-

Setting	Source-free	Target-free	Training-free
DA	✗	✗	✗
DG	✗	✓	✗
SFDG	✓	✓	✗
ZSDG	✓	✓	✓

Table 1. Comparison of domain adaptation and generalisation settings in terms of data and training assumptions. **DA** (Domain Adaptation) requires access to source and target domain data and additional training. **DG** (Domain Generalisation) removes the need for target data but still relies on source data and training. **SFDG** (Source-Free Domain Generalisation) does not use source or target data but requires training on synthetic or proxy representations. **ZSDG** eliminates the need for any data or training, leveraging only pretrained models.

imise human interference [16]. These challenges motivate a more flexible setting called Zero-Shot Domain Generalisation (ZSDG), where models generalise to unseen domains without source data, target data, or any additional training. As illustrated in Table 1, ZSDG represents the most flexible and data-agnostic formulation within the broader domain generalisation landscape.

To address the ZSDG setting, we propose a method called **Prefer** (**pr**ompt-driv**e**n **fe**ature **r**efinement), which leverages the representational power of pretrained vision-language models without any additional training or access to external data. As shown in Figure 1, our method begins by prompting a large language model to generate a diverse set of visual domains, such as sketch, infographic, or on grass. For each target class, we then generate a range of domain-specific prompts that simulate how the object might appear across varied domains. These prompts are embedded using the pretrained model to obtain domain-specific text features for each class. To identify reliable feature channels, we compute two channel-wise scores. The similarity score measures how strongly each channel aligns with the class identity, computed by comparing the original class prompt (e.g., "a photo of a [class]") to the domain-specific features. The variance score quantifies how much each channel fluctuates across domains, indicating sensitivity to domain shift. These two scores are combined into a final channel score that favours class-consistent and domain-invariant features. The top-ranked channels are selected and used at inference time, allowing the model to compute class similarities using only stable and class-consistent features. This approach enables robust generalisation under domain shift without model updates or access to external data. Our main contributions are as follows :

- We introduce ZSDG, where models generalise to unseen domains without access to source data, target data, or additional training.
- We propose **Prefer** (**pr**ompt-driv**e**n **fe**ature **r**efinement), a simple and effective method that leverages LLM-generated domain-specific prompts and pretrained vision-language models to perform feature refinement without requiring any model updates or external data.
- We demonstrate that Prefer consistently improves zero-shot accuracy under domain shift across standard benchmarks, outperforming existing state-of-the-art domain generalisation methods.

2. Related Work

Domain Generalisation (DG). DG seeks to train models on source domains that perform well on unseen target domains, without accessing any target data during training [31]. Traditional approaches focus on learning domain-invariant representations through techniques such as domain alignment [11], causal inference [17] and meta-learning [14]. Many methods also employ data augmentation [39] and learning strategies [37] to enhance robustness. Recent work has focused on domain generalisation of pretrained vision-language foundation models such as CLIP [26]. For instance, StyLIP learns disentangled style and content prompts to generalise across unseen domains [4]. CLIPCEIL refines feature channels using lightweight adapters to improve image–text alignment under domain shift [32]. Other approaches, such as disentangled prompt representations [5] and multi-modal prompt learning frameworks [1], also improve domain generalisation via prompt tuning. In contrast, source-free DG methods eliminate the need for source-domain images by adapting pretrained vision-language models through training on synthetic data via prompt tuning [6, 19, 29] or lightweight adapters [34]. While prior work has used the term "zero-shot DG" [18], they still rely on training on source domain data. We define and address zero-shot DG as a truly source-free, target-free, and training-free setting, where models must generalise to unseen domains using only pretrained models, without any additional training or access to external data.

Vision–Language Foundation Models. Contrastive Language–Image Pre-training (CLIP) [26] introduced a powerful paradigm by jointly learning image and text representations through large-scale contrastive learning on image–caption pairs. The resulting vision–language embedding space enables zero-shot classification using natural language prompts. Building upon CLIP, several parameter-efficient fine-tuning (PEFT) methods have been proposed to adapt its capabilities to downstream tasks without retraining the entire model. For example, CoOp [43] learns continuous soft prompts that replace manual textual templates, improving few-shot performance with minimal tuning. CLIP-Adapter [9] introduces lightweight adapter modules into one branch of CLIP, requiring only a small fraction of trainable parameters. MaPLe (Multi-modal Prompt

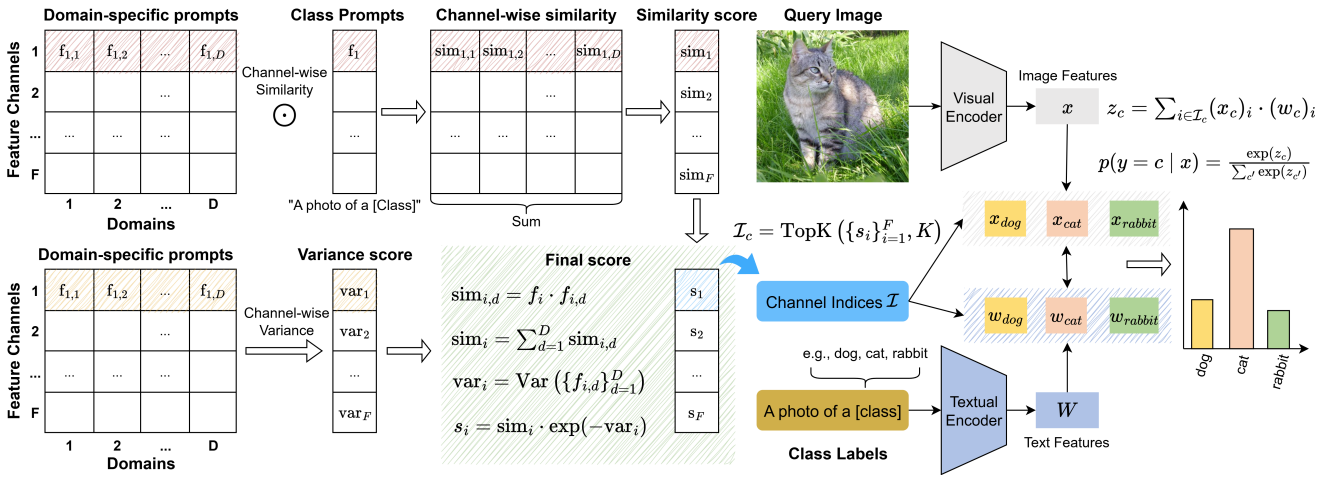

Figure 2. **Overview of Prefer.** We first generate domain-specific prompts and extract their features using a vision-language model. We compute channel-wise similarity between the class prompt and each domain-specific prompt, and aggregate the results to obtain a similarity score. Simultaneously, we compute channel-wise variance across domain-specific features to assess domain stability. The final score is defined as similarity weighted by inverse variance. For each class, we select the top-K feature channels with the highest scores. At inference, given an input image and prompts, we compute the logits using only the selected channels for each class.

Learning) [12] extends this direction by jointly learning prompts in both the vision and language branches of CLIP, improving cross-domain transfer through coupled, stage-wise prompting. These PEFT techniques strike a balance between efficiency and performance, and form the foundation of many CLIP DG methods such as StyLIP [4] and CLIPCEIL [32].

3. Preliminaries

Consider $\mathcal{X}$ as the input space with dimension d, and $\mathcal{Y}$ as the target label space. A domain consists of data points sampled from a distribution $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, and n is the number of examples. The data are drawn from the joint distribution $P(X, Y)$, where X and Y are the corresponding random variables [31, 41].

Domain Generalisation (DG). The DG problem assumes access to N source domains $\mathcal{S} = \{\mathcal{D}^{(i)}\}_{i=1}^{N}$, where each domain $\mathcal{D}^{(i)} = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$ has a distinct joint distribution $P^{(i)}(X, Y)$. The distributions differ across domains: $P^{(j)}(X, Y) \neq P^{(k)}(X, Y)$ for $j \neq k$. The goal is to learn a predictive function $f : \mathcal{X} \to \mathcal{Y}$ that generalises well to an unseen target domain $\mathcal{D}_T$, where $P^{(T)}(X, Y) \neq P^{(i)}(X, Y)$ for all $i \in \{1, \ldots, N\}$.

Vision-Language Models (VLMs). A vision-language foundation model such as CLIP [26] consists of an image encoder $\mathcal{E}_I$ and a text encoder $\mathcal{E}_T$. For each class $c_m \in \mathcal{C}$, a textual prompt p_m is constructed, e.g., "a photo of a $\{c_m\}$", where $m \in \{1, \ldots, M\}$ and M is the total number of classes. The text encoder produces a text embedding $T_m = \mathcal{E}_T(p_m)$ for each prompt, while the image encoder generates an embedding $I_x = \mathcal{E}_I(x)$ for an input image x. Similarity between image and text embeddings is typically computed using cosine similarity: $\cos(\theta) = \langle I_x, T_m \rangle$, which aligns image and text pairs in a shared embedding space, commonly optimised via contrastive learning.

Zero-Shot Domain Generalisation (ZSDG). We define ZSDG as the task of learning a predictive function $f : \mathcal{X} \to \mathcal{Y}$ that generalises to unseen target domains $\mathcal{D}_T$, under the most minimal assumptions: no access to source domain data $\mathcal{S} = \{\mathcal{D}^{(i)}\}$, no access to target domain data $\mathcal{D}_T$, and no additional training. The model must generalise to target domains where $P^{(T)}(X, Y) \neq P^{(i)}(X, Y)$ for all $i \in \{1, \ldots, N\}$, using only a pretrained model and inference-time mechanisms. This setting differs from source-free DG, which typically involves fine-tuning adapters or prompts using auxiliary data or training steps, and from prior zero-shot DG formulations that still assume access to source domain data. In ZSDG, the model operates entirely without access to external data or retraining.

4. Prompt-Driven Feature Refinement

In this section, we present **Prefer** (prompt-driven feature refinement), our approach for achieving ZSDG using only pretrained vision–language models (VLMs) and inference-time feature refinement. The key idea is to identify feature channels that consistently represent each class across a diverse range of imagined domain shifts, without access to source or target data or model retraining. Prefer consists of three main components: (1) generating domain-specific prompts for each class to simulate unseen domain variations; (2) computing similarity and variance scores

Algorithm 1 Feature Channel Refinement via Similarity and Variance Scoring

Input: Class prompt p_c; domain-specific prompts $\{p_{c,d}\}_{d=1}^{D}$; text encoder $\mathcal{E}_T$; elbow-point finder $\mathcal{K}$
Output: Selected feature channel indices $\mathcal{I}_c$ for class c
Compute reference embedding: $T_c = \mathcal{E}_T(p_c)$
Compute domain-specific embeddings: $T_{c,d} = \mathcal{E}_T(p_{c,d})$ for $d = 1, \ldots, D$
for $i = 1$ to F **do**
  for $d = 1$ to D **do**
    Compute per-channel cosine similarity:

$$\text{sim}_{i,d} = (T_c)_i \cdot (T_{c,d})_i$$

  end for
  Aggregate similarity: $\text{sim}_i = \sum_{d=1}^{D} \text{sim}_{i,d}$
  Compute channel variance:

$$\text{var}_i = \text{Var}\left(\{(T_{c,d})_i\}_{d=1}^{D}\right)$$

  Compute final score: $s_i = \text{sim}_i \cdot \exp(-\text{var}_i)$
end for
Determine $K = \mathcal{K}(\{s_i\}_{i=1}^{F})$ using the elbow-point heuristic
Select top-K channels: $\mathcal{I}_c = \text{TopK}(\{s_i\}_{i=1}^{F}, K)$
Return $\mathcal{I}_c$

across feature channels to select the most robust and class-consistent channels for each class; (3) performing inference using only the selected feature channels for both image and text embeddings to improve robustness under domain shift. The Prefer framework is illustrated in Figure 2. We detail each component in the following sections.

4.1. Generation of Domain-Specific Prompts

To simulate a wide range of potential domain shifts, we first generate a diverse set of domain-specific textual prompts for each class. The goal is to model how objects might appear across different visual styles, contexts, or environments, without requiring access to any source or target data. We begin by prompting a large language model (LLM) to produce a list of visually distinct domains. The exact instruction is:

"List [D] visually distinct domains in which objects might appear. Domains should reflect a wide range of variation, such as artistic or media styles (e.g., painting, sketch, infographic), contextual settings or environments (e.g., on grass, in water, indoors, or at night). Do not include object names; focus only on the type of domain or visual context that affects how the object appears."

This produces a set of domains: $\mathcal{D} = \{d_1, d_2, \ldots, d_D\}$, where D is the number of imagined domains. Next, for each class $c \in \mathcal{C}$, we generate natural language descriptions depicting how the object might appear in each domain. The following prompt is used:

"Given an object class and a list of visual domains, generate a short and vivid natural language description for each domain that depicts how the object might appear within that domain. Each description should include the object and reflect the unique visual characteristics, style, medium, or environment implied by the domain. The descriptions should be semantically plausible and visually grounded, capturing domain-specific attributes without changing the identity of the object. For example, if the object is "bicycle" and the domain is "sketch," the description could be: "a pencil sketch of a bicycle with fine lines and light shading." Now generate descriptions for the class "[CLASS]" across the following domains: [domain1, domain2, domain3, ...]."

This yields a set of domain-specific prompts for each class $\mathcal{P}_c = \{p_{c,d} \mid d \in \mathcal{D}\}$. We then encode each prompt using the pretrained text encoder $\mathcal{E}_T$ to obtain domain-specific text embeddings $T_{c,d} = \mathcal{E}_T(p_{c,d})$. These embeddings serve as domain-augmented representations for each class, which will be used to probe feature channels for their stability and class-consistency across diverse domain shifts. By leveraging the compositional capabilities of large language models, this approach generates rich and plausible cross-domain descriptions that reflect how object appearance may vary across different styles and contexts. Encoding these prompts through a VLM encourages the alignment of class concepts with diverse visual conditions. This allows us to identify and rely on robust, domain-invariant feature channels, thereby improving the generalisation ability of the foundation model without requiring access to any visual domain data.

4.2. Computing Similarity and Variance Scores

Given the domain-specific embeddings $T_{c,d}$ generated in the previous step, we aim to identify feature channels that consistently represent each class across diverse domain shifts. These channels should be both *class-consistent*, meaning they remain aligned with the original class concept, and *domain-invariant*, meaning they exhibit low variability across different domains.

For each class c, we first encode its original class prompt p_c using the text encoder $\mathcal{E}_T$, obtaining the reference embedding $T_c = \mathcal{E}_T(p_c)$. Each embedding T_c or $T_{c,d}$ is an F-dimensional feature vector:

$$T_c = [f_1, f_2, \ldots, f_F], \quad T_{c,d} = [f_{1,d}, f_{2,d}, \ldots, f_{F,d}]$$

with unit norm $\|T_c\| = 1$ and $\|T_{c,d}\| = 1$, as is typical in vision–language contrastive learning. The cosine similarity between T_c and $T_{c,d}$ is:

$$\cos(\theta) = \langle T_c, T_{c,d} \rangle = \sum_{i=1}^{F} f_i \cdot f_{i,d}$$

which decomposes into channel-wise contributions $\text{sim}_{i,d} = f_i \cdot f_{i,d}$. Thus, cosine similarity naturally motivates defining a per-channel similarity as the product of corresponding channels. We then aggregate these similarities across domains $\text{sim}_i = \sum_{d=1}^{D} \text{sim}_{i,d}$. To quantify the stability of each channel across domains, we compute the variance of the raw feature activations $\text{var}_i = \text{Var}\left(\{f_{i,d}\}_{d=1}^{D}\right)$.

Finally, we compute a combined score for each feature channel $s_i = \text{sim}_i \cdot \exp(-\text{var}_i)$. This formulation balances *class-consistency* and *domain-invariance*: channels that are highly aligned with the class (sim_i) and stable across domains (low var_i) receive higher scores, while channels with unstable behaviour are exponentially down-weighted. This allows us to select features that are both discriminative and robust to domain shift. We select the top-K scoring channels for each class c, $\mathcal{I}_c = \text{TopK}\left(\{s_i\}_{i=1}^{F}, K\right)$. The value of K is automatically determined for each class using an *elbow-point heuristic*, which identifies the point of diminishing returns in the sorted score curve $\{s_i\}$. Specifically, we sort all channel scores s_i in descending order and fit a curve. The elbow point is then estimated as the index K that maximises the distance to the straight line connecting the highest and lowest scores [28]. This allows Prefer to adaptively select a different number of robust channels per class without requiring manual tuning or access to validation data. The selected channels $\mathcal{I}_c$ are then used at inference time to improve robustness under domain shift, as described in the next section.

4.3. Theoretical Discussion

Our method is grounded in the principle that effective ZSDG requires selecting feature channels that preserve class-specific semantic information while being robust to domain-specific variations. We formalise this intuition through an information-theoretic perspective, acknowledging the heuristic nature of certain approximations, yet providing clear theoretical motivations.

Definition 1 (Semantic-Domain Decomposition)**.** *For each class c, we define $\mathcal{S}_c$ as invariant visual attributes that do not change across domains. We denote the domain-specific variations as latent factors $\mathcal{D}$. Domain-specific prompts $\{p_{c,d}\}_{d=1}^{D}$ act as controlled textual interventions that simulate variations due to domain shifts. We assume each channel-level embedding f_i can be approximately decomposed additively as $f_i = \phi_i(\mathcal{S}_c) + \psi_i(\mathcal{D}) + \epsilon_i$, where $\phi_i(\mathcal{S}_c)$ captures class-specific semantic information, $\psi_i(\mathcal{D})$ represents variations induced by domain-specific factors, and ϵ_i denotes unstructured noise independent of both semantic and domain attributes. A robust feature channel i is thus characterised by a minimal contribution from $\psi_i(\mathcal{D})$.*

Core Assumptions:

- **A1** *Per-Channel Semantic Alignment*: The per-channel cosine similarity sim_i correlates positively with mutual information between semantic content and the feature channel $\text{sim}_i \propto I(\mathcal{S}_c; f_i)$. This assumption leverages principles of mutual information maximisation typically employed in contrastive VLM training [26].
- **A2** *Variance as Domain Sensitivity*: The per-channel variance across domain-specific prompt embeddings, $\nu_i = \text{Var}_d[(T_{c,d})_i]$, serves as a proxy for conditional mutual information between domain attributes and features, given semantic content $\nu_i \propto I(\mathcal{D}; f_i \mid \mathcal{S}_c)$. This assumption relies on the standard Gaussian channel approximation [7], in which higher variance indicates greater sensitivity to domain interventions.
- **A3** *Cross-Modal Consistency*: The semantic information captured by the visual encoder is consistently related to the semantic information captured by the textual encoder, i.e.: $I^{\text{vis}}(\mathcal{S}_c; f_i) \geq \gamma \cdot I^{\text{txt}}(\mathcal{S}_c; f_i)$,, with $\gamma > 0$. This assumption reflects empirical cross-modal alignment properties of CLIP-style models [12].

Proposition 1 (Scoring Function as Information-Theoretic Proxy)**.** *Under assumptions A1 and A2, the scoring function is defined as $s_i = sim_i \cdot \exp(-\nu_i)$, serves as a heuristic proxy correlating with an information-theoretic objective of balancing semantic relevance and domain invariance:*

$$s_i \sim I(\mathcal{S}_c; f_i) \cdot \exp\big(-I(\mathcal{D}; f_i \mid \mathcal{S}_c)\big).$$

Channels achieving higher s_i scores effectively prioritise semantic relevance while penalising sensitivity to domain-specific variations.

The similarity term sim_i empirically captures channel-level semantic information, approximating $I(\mathcal{S}_c; f_i)$. The variance term ν_i penalises domain sensitivity and approximates $I(\mathcal{D}; f_i \mid \mathcal{S}_c)$ under Gaussian channel assumptions, as variance corresponds to conditional entropy. The exponential form emerges naturally from entropy-based regularisation, as lower entropy (variance) indicates greater certainty and stability across domains.

Proposition 2 (Generalisation Error Heuristic Bound)**.** *Consider $\mathcal{I}_c$ as the selected robust channel subset for class c. The generalisation error on a novel domain d, denoted as ϵ_d, is heuristically bounded by:*

$$\epsilon_d \leq C_1 \sum_{i \notin \mathcal{I}_c} (1 - s_i) + C_2 \cdot \mathbb{E}_d \left[\frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} Var^{vis}(f_i) \right] + \epsilon_0,$$

where $C_1, C_2 > 0$ are constants controlling the relative importance of semantic information loss and residual domain variance; ϵ_0 denotes irreducible error due to noise or model limitations; $Var^{vis}(f_i)$ denotes empirical variance computed in visual feature embeddings across domains.

The first term quantifies information loss due to discarding channels that carry meaningful semantic information but have high domain sensitivity. The second term penalises any residual domain variance present in selected channels, measuring the empirical stability of visual features under domain interventions. Together, this heuristic bound aligns closely with principles of domain adaptation and generalisation theory [3], adapted specifically to zero-shot settings where no explicit training on source domains occurs.

Proposition 3 (Optimality of Elbow-Point Heuristic)**.** *Selecting the number of robust channels K via the elbow-point heuristic approximately minimises the following regularised objective:*

$$\mathcal{L}(K) = \sum_{i=K+1}^{F} (1 - s_{(i)}) + \eta \cdot K,$$

where:

$$\eta = -\frac{\partial^2}{\partial K^2} \left(\sum_{k=1}^{K} s_{(k)} \right)$$

is automatically determined by identifying the point of maximal curvature (elbow-point).

The sorted channel scores $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(F)}$ form a concave and submodular cumulative gain curve. The elbow-point corresponds to the knee region of this curve, where the marginal information gain from adding further channels sharply declines relative to increased model complexity. This channel selection method aligns with standard practices for heuristic model complexity control [28].

Corollary 1 (Cross-Modal Invariance Transfer)**.** *Under assumption A3, selecting robust channels based on textual embeddings effectively constrains visual-domain variance:*

$$\mathbb{E}_d \left[Var^{vis}(f_i) \right] \leq \frac{1}{\gamma} \mathbb{E}_d \left[Var^{txt}(f_i) \right], \quad \forall i \in \mathcal{I}_c.$$

This ensures semantic robustness identified in text embeddings generalises effectively to the visual domain.

Exact mutual information computations are generally intractable for high-dimensional embeddings used by VLMs. Therefore, our theoretical framework is intentionally heuristic and relies on plausible proxies to operationalise information-theoretic intuition.

4.4. Inference

At inference time, given a test image x, we encode it using the pretrained image encoder:

$$f^{(I)} = \mathcal{E}_I(x) = [f_1^{(I)}, f_2^{(I)}, \ldots, f_F^{(I)}].$$

For each class c, we also compute the text embedding of its original class prompt:

$$f_c^{(T)} = \mathcal{E}_T(p_c) = [f_{c,1}^{(T)}, f_{c,2}^{(T)}, \ldots, f_{c,F}^{(T)}].$$

We use only the subset of feature channels $\mathcal{I}_c$ selected in the previous step. The class-specific representations after channel selection are:

$$x_c = \{f_i^{(I)} \mid i \in \mathcal{I}_c\}, \quad w_c = \{f_{c,i}^{(T)} \mid i \in \mathcal{I}_c\}.$$

The similarity score between image x and class c is computed as:

$$z_c = \sum_i (x_c)_i \cdot (w_c)_i.$$

Finally, a softmax over all classes gives the predicted probability:

$$P(y = c \mid x) = \frac{\exp(z_c)}{\sum_{c'} \exp(z_{c'})}.$$

This inference strategy ensures that the prediction relies only on the most robust and class-consistent feature channels, improving generalisation under domain shift, without requiring any model updates or adaptation.

5. Experiments

We conduct comprehensive experiments to assess the effectiveness of **Prefer** in the ZSDG setting. Our goal is to evaluate whether Prefer can enhance the generalisation capabilities of vision-language models across unseen domains, without relying on any source data or training. We compare **Prefer** against a broad range of recent domain generalisation approaches, spanning both multi-source and source-free methods, across benchmarks and backbones.

5.1. Experimental Settings

Baselines. We evaluate **Prefer** against a wide range of DG baselines, including multi-source DG methods [4, 9, 12, 33, 36, 42, 43], which are based on parameter-efficient fine-tuning on sourced domain data with adapter or prompt learning; source-free DG methods [6, 19, 29, 34], which fine-tunes CLIP on synthetic data; and the zero-shot CLIP baseline [26]. These comparisons allow us to contextualise the performance of Prefer both within the ZSDG setting and relative to methods that rely on stronger assumptions, such as access to source data or additional training.

Datasets. Following previous work on source-free domain generalisation [6, 19, 29, 34], we conduct experiments on four standard domain generalisation benchmarks: *PACS*[13], *OfficeHome*[30], *VLCS*[8], and *DomainNet*[22]. For all datasets, we adopt the leave-one-domain-out evaluation protocol, reporting the average accuracy across held-out domains. We repeat each experiment ten times with different random seeds and report mean accuracies with standard errors. Statistical significance is assessed using the

Methods	ResNet50				ViT-B/16			
	PACS	OfficeHome	VLCS	DomainNet	PACS	OfficeHome	VLCS	DomainNet
Multi-source DG								
CLIP-Adapter [9]	92.3±.2	73.4±.2	80.1±.1	50.1±.1	97.2±.1	83.0±.2	84.4±.1	58.8±.2
CoOp [43]	92.8±.1	75.8±.0	81.2±.1	49.5±.1	96.4±.1	81.7±.2	83.8±.1	58.9±.1
CoCoOp [42]	92.9±.0	76.5±.2	82.4±.1	49.8±.1	97.4±.1	83.3±.2	85.3±.1	59.7±.2
MaPLE [12]	93.0±.0	**77.0±.1**	80.9±.0	49.6±.2	97.6±.1	83.4±.2	84.6±.1	61.3±.2
DPL [35]	92.8±.1	75.2±.2	82.1±.1	49.9±.1	97.8±.1	83.6±.2	84.8±.1	60.3±.2
StyLIP [4]	**93.8±.1**	74.8±.2	83.2±.1	51.3±.1	**98.1±.1**	84.6±.2	86.9±.1	**62.3±.2**
CLIPCEIL [32]	92.8±.1	76.6±.1	**83.5±.1**	**51.6±.1**	97.6±.1	**85.4±.2**	**87.4±.2**	62.0±.0
Source-free DG								
DUPRG [19]	93.1±.1	73.0±.1	81.2±.1	48.4±.1	96.9±.1	82.2±.2	82.9±.1	57.2±.2
PromptStyler [6]	93.1±.1	72.9±.1	82.3±.1	48.5±.1	96.8±.2	81.8±.1	83.7±.2	56.7±.2
DPStyler [29]	93.2±.2	72.5±.2	82.5±.2	48.0±.1	97.1±.1	82.8±.1	84.0±.1	58.4±.2
PromptTA [34]	93.1±.1	73.1±.1	82.8±.1	48.6±.1	97.2±.1	82.6±.2	84.2±.2	58.3±.1
Zero-shot DG								
CLIP-ZS [26]	91.8±.1	71.5±.1	76.0±.1	47.5±.1	96.5±.1	79.6±.1	76.7±.1	57.4±.1
Prefer (Ours)	**93.6±.2**	**75.8±.2**	**83.5±.2**	**50.3±.2**	**98.0±.2**	**84.2±.1**	**86.1±.2**	**60.6±.1**

Table 2. Leave-one-domain-out classification accuracy on four DG benchmarks, comparing ResNet50 and ViT-B/16 backbones. Ours achieves state-of-the-art performance among source-free methods, and remains competitive with multi-source DG methods.

Methods	ResNet50				ViT-B/16			
	PACS	OfficeHome	VLCS	DomainNet	PACS	OfficeHome	VLCS	DomainNet
CLIP-ZS [26]	91.8±.1	71.5±.1	76.0±.1	47.5±.1	96.5±.1	79.6±.1	76.7±.1	57.4±.1
+ Similarity only	92.5±.2	72.8±.1	78.4±.2	48.1±.1	97.1±.2	80.7±.1	81.8±.1	58.2±.1
+ Variance only	92.3±.1	73.0±.2	78.6±.1	48.0±.2	97.0±.1	81.1±.2	82.0±.2	58.0±.2
+ Combination	**93.6±.2**	**75.8±.2**	**83.5±.2**	**50.3±.2**	**98.0±.2**	**84.2±.1**	**86.1±.2**	**60.6±.1**

Table 3. Ablation study on similarity and variance scores for robust feature selection. Using similarity or variance alone gives marginal improvement over CLIP-ZS, but their combination significantly outperforms all variants across all benchmarks.

Wilcoxon signed-rank test, and the best-performing results with statistical significance are highlighted in bold.

Implementation Details. We use the OpenAI CLIP model [26] as our vision-language model, employing publicly available pretrained weights. We use ResNet50 and ViT-B/16 backbones for image encoding, with corresponding Transformer-based text encoders. The dimensionality of image and text embeddings is $F = 1024$ for ResNet50 and $F = 512$ for ViT-B/16. Our approach requires no additional training; feature selection is computed entirely at inference time using LLM-generated prompts. We use GPT-4o to generate domain names and domain-specific prompts for each class, following the procedure described in Section 4. We set $D = 30$ domains per class. To automatically determine the optimal number of feature channels per class, we apply the elbow heuristic using the `kneed` Python package [28]. The method identifies the point of maximum curvature on the sorted feature scores, avoiding manual selection of K and ensuring a consistent, data-free cutoff across datasets and backbones. For training-required methods, we adopt a batch size of 32 and a stochastic gradient descent optimiser with a cosine annealing scheduler. We set the learning rate at 0.002 and trained for 200 epochs. All experiments are conducted on an NVIDIA Tesla A100 GPU.

5.2. Main Results

Table 2 presents leave-one-domain-out classification accuracy on four standard DG benchmarks, comparing both ResNet50 and ViT-B/16 backbones. Prefer consistently outperforms existing source-free DG methods across all datasets and architectures, establishing new state-of-the-art results in this setting. Compared to the zero-shot CLIP baseline (CLIP-ZS), Prefer yields substantial gains on both small and large backbones (e.g., +1.9% on PACS and +2.8% on DomainNet with ResNet50), demonstrating that prompt-driven feature refinement significantly enhances zero-shot domain generalisation. While multi-source DG methods still achieve higher overall accuracy, benefiting from access to source-domain data and supervised training, Prefer narrows this gap. On ViT-B/16, Prefer approaches the performance of strong trained methods such as CLIP-CEIL and StyLIP, despite requiring no source data or train-

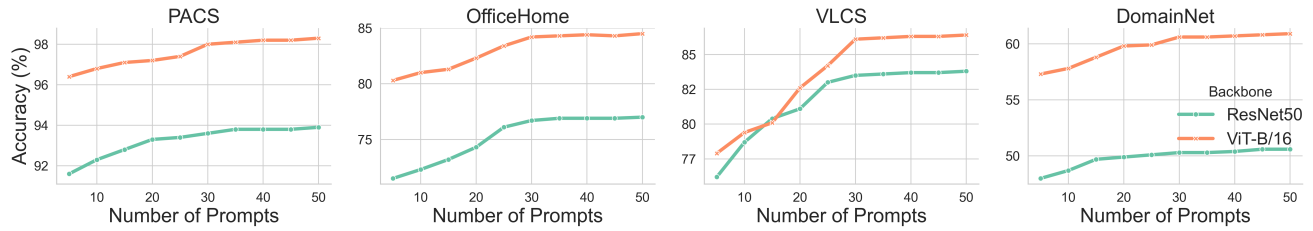

Figure 3. We evaluate the performance of Prefer as the number of generated prompts increases from 5 to 50. Results are shown across four benchmarks using ResNet50 and ViT-B/16 backbones.

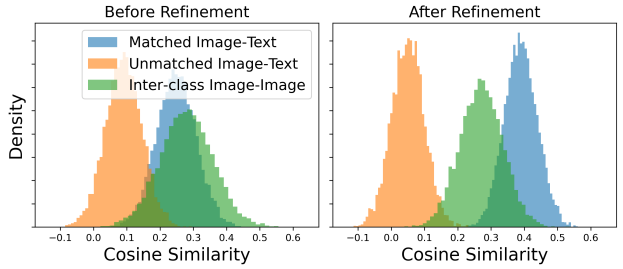

Figure 4. Comparison of cosine similarity distributions between matched image–text pairs, unmatched image–text pairs, and inter-class image–image pairs, before and after applying Prefer.

ing. These results highlight the flexibility and effectiveness of Prefer for scenarios where source data is unavailable or training is impractical.

5.3. Further Analysis

Combination of Similarity and Variance Scores. To validate the effectiveness of our channel selection strategy, we conduct an ablation study comparing different variants of the scoring function. Starting from the zero-shot baseline (CLIP-ZS), we first consider using only the channel-wise similarity to the original class embedding ("Similarity only") and only the variance-based stability across domains ("Variance only"). As shown in Table 3, both variants yield modest improvements over CLIP-ZS, suggesting that either metric alone captures some useful signal for identifying robust features. However, our full method ("Combination") achieves the highest accuracy on all benchmarks and backbones, consistently outperforming the individual variants. This demonstrates that the combination of class-consistency and domain-invariance provides complementary benefits, and that our multiplicative scoring formulation effectively balances these objectives to select more reliable feature channels.

Effectiveness of Feature Channel Refinement. To assess whether our channel selection method improves cross-modal alignment, we visualise cosine similarity distributions on the OfficeHome dataset in Figure 4. We compare similarity between matched image–text pairs, unmatched image–text pairs, and inter-class image–image pairs, both before and after refinement. After applying Prefer, the similarity of matched image–text pairs shifts notably higher. This indicates that our feature refinement effectively enhances discriminability and reduces spurious cross-modal correlations, leading to more robust generalisation across domains.

Number of Domain-specific Prompts. To understand the effect of prompt diversity, we vary the number of domain-specific prompts per class from 5 to 50 and evaluate the resulting performance on four domain generalisation benchmarks using both ResNet50 and ViT-B/16 backbones (Figure 3). When using only a small number of prompts (e.g., 5), PREFER performs similarly to zero-shot CLIP, likely due to insufficient domain variation failing to reliably identify stable feature channels. As the number of prompts increases, accuracy improves consistently, indicating that more diverse prompts help better simulate the possible distributional shifts in unseen domains. The performance gains become saturated around 25–35 prompts, beyond which additional prompts yield only marginal improvements. This suggests that a moderate number of domain-specific prompts is sufficient to capture transferable semantics while avoiding excessive noise or redundancy.

6. Conclusion

We presented **Prefer**, a plug-and-play feature refinement method that enhances vision-language foundation models for domain generalisation without requiring access to source or target data. By generating domain-specific prompts and scoring feature channels based on their class-consistency and domain-invariance, Prefer selects a subset of robust channels tailored to each class. Extensive experiments across four standard benchmarks and multiple backbones demonstrate significant over state-of-the-art baselines. Our results highlight the potential of language-driven feature analysis as a scalable and effective strategy for enhancing model robustness under distribution shift.

References

[1] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23922–23932, 2024. 2

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

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 6

[4] Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5542–5552, 2024. 2, 3, 6, 7

[5] De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23595–23604, 2024. 2

[6] Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15702–15712, 2023. 2, 6, 7

[7] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 5

[8] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1657–1664, 2013. 6

[9] 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*, 132(2): 581–595, 2024. 2, 6, 7

[10] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 1

[11] Sobhan Hemati, Guojun Zhang, Amir Estiri, and Xi Chen. Understanding hessian alignment for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19004–19014, 2023. 2

[12] 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*, pages 19113–19122, 2023. 3, 5, 6, 7

[13] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5542–5550, 2017. 6

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[15] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision*, pages 624–639, 2018. 1

[16] Yuzhuo Li, Di Zhao, Tingrui Qiao, Yihao Wu, Bo Pang, and Yun Sing Koh. Metawild: A multimodal dataset for animal re-identification with environmental metadata. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13009–13015, 2025. 2

[17] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pages 7313–7324. PMLR, 2021. 2

[18] Udit Maniyar, K. J. Joseph, Aniket Anand Deshmukh, Urun Dogan, and Vineeth N. Balasubramanian. Zero–shot domain generalization. In *BMVC*, 2020. 2

[19] Hongjing Niu, Hanting Li, Feng Zhao, and Bin Li. Domain-unified prompt representations for source-free domain generalization. *arXiv preprint arXiv:2209.14926*, 2022. 2, 6, 7

[20] Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Cabin: Debiasing vision-language models using backdoor adjustments. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 484–492, 2025. 1

[21] Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Libra: Measuring bias of large language model from a local context. In *European Conference on Information Retrieval*, pages 1–16. Springer, 2025. 1

[22] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 6

[23] Tingrui Qiao, Caroline Walker, Chris Cunningham, Adam Jang-Jones, Susan Morton, Kane Meissel, and Yun Sing Koh. Thematic bottleneck models for multimodal analysis of school attendance. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5971–5979, 2025. 1

[24] Tingrui Qiao, Caroline Walker, Chris Cunningham, Adam Jang-Jones, Susan Morton, Kane Meissel, and Yun Sing Koh. Longitudinal surveys are texts: Llm-enhanced analysis of school attendance in new zealand. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 310–327. Springer, 2025.

[25] Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Thematic-lm: a llm-based multi-agent system for large-scale thematic analysis. In *Proceedings of the ACM on Web Conference 2025*, pages 649–658, 2025. 1

[26] 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. 2, 3, 5, 6, 7

[27] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1

[28] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011. 5, 6, 7

[29] Yunlong Tang, Yuxuan Wan, Lei Qi, and Xin Geng. Dpstyler: dynamic promptstyler for source-free domain generalization. *IEEE Transactions on Multimedia*, 2025. 2, 6, 7

[30] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 6

[31] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2022. 1, 2, 3

[32] Xi Yu, Shinjae Yoo, and Yuewei Lin. Clipceil: Domain generalization through clip via channel refinement and image-text alignment. *Advances in Neural Information Processing Systems*, 37:4267–4294, 2024. 2, 3, 7

[33] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019. 6

[34] Haoran Zhang, Shuanghao Bai, Wanqi Zhou, Jingwen Fu, and Badong Chen. Promptta: Prompt-driven text adapter for source-free domain generalization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2, 6, 7

[35] Xin Zhang, Shixiang Shane Gu, Yutaka Matsuo, and Yusuke Iwasawa. Domain prompt learning for efficiently adapting clip to unseen domains. *Transactions of the Japanese Society for Artificial Intelligence*, 38(6):B–MC2_1, 2023. 7

[36] Yi-Fan Zhang, Jindong Wang, Jian Liang, Zhang Zhang, Baosheng Yu, Liang Wang, Dacheng Tao, and Xing Xie. Domain-specific risk minimization for domain generalization. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3409–3421, 2023. 6

[37] Di Zhao, Yun Sing Koh, Gillian Dobbie, Hongsheng Hu, and Philippe Fournier-Viger. Symmetric self-paced learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16961–16969, 2024. 1, 2

[38] Di Zhao, Jingfeng zhang, Hongsheng Hu, Philippe Fournier-Viger, Gillian Dobbie, and Yun Sing Koh. Balancing invariant and specific knowledge for domain generalization with online knowledge distillation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 2440–2448, 2025. 1

[39] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13025–13032, 2020. 1, 2

[40] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021. 1

[41] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022. 3

[42] 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. 6, 7

[43] 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. 2, 6, 7