Domain Generalization via Text-Anchored Information Bottleneck

Eunyi Lyou, Yunjeong Choi, Junho Lee, and Joonseok Lee*

Seoul National University, Seoul, Republic of Korea
{onlyou0416, racheal0, joon2003, joonseok}@snu.ac.kr

Abstract. Visual recognition models often fail when deployed in new environments. Domain Generalization (DG) addresses this by learning representations that remain invariant to environment-specific variations. Recent approaches increasingly rely on large vision-language models, assuming that preserving their expressive visual representations improves robustness. However, we show that such visual expressiveness can instead propagate spurious cues that tie representations to the training environments, hindering invariant learning. We therefore discard visual guidance and instead treat the language embedding space as the primary source of domain invariance, naturally acting as an information bottleneck that preserves core semantics while suppressing domain-specific variations. Extensive experiments across diverse backbones exhibit state-of-the-art performance and further analyze what makes guidance effective for robust generalization. These findings shift the focus of DG from improving representations to designing supervision that enforces invariance.

Keywords: Domain generalization · Vision-Language Models · Information bottleneck

1 Introduction

Despite progress in computer vision, visual recognition models often fail to maintain performance under environmental changes at deployment, when test samples deviate from training distribution [9, 42]. Unlike domain adaptation [8, 9], which has access to target distribution during training, domain generalization (DG) [36] addresses the stricter scenario of generalizing to unseen environments without any prior exposure. Under this constraint, DG aims to learn domain-invariant representations that preserve only essential semantics of the input.

To pursue such invariance, standard DG frameworks train across multiple source domains, under the assumption that reducing distributional discrepancies among diverse sources will expose an underlying invariant subspace. For instance, aligning a simple sketch with a real photograph of an apple is expected to reveal the shared semantic "apple-ness" beneath domain-specific variations. Accordingly, traditional DG methods primarily align feature distributions across

* Corresponding author

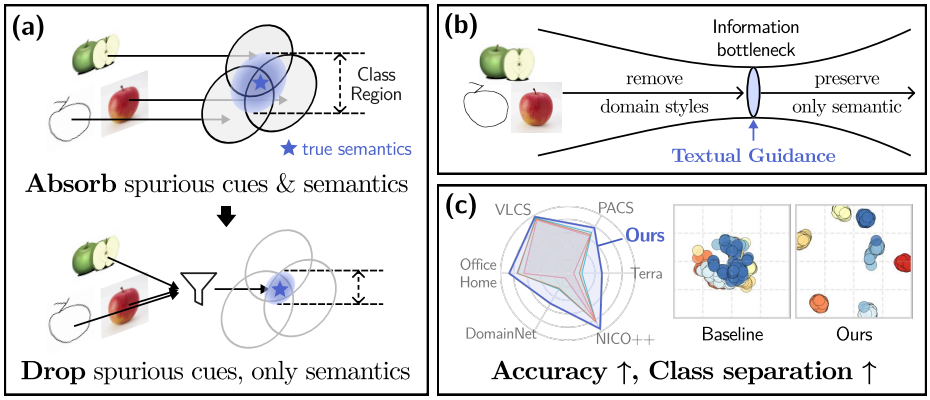

Fig. 1: (a) Visual encoders inevitably absorb spurious domain cues alongside domain-invariant semantics. This inflates class regions and blurs boundaries, hindering robustness. Ideally, models should drop domain-specific variations while preserving only core semantics. (b) Redefining invariance via a text-anchored Information Bottleneck (IB). Textual guidance acts as a semantic filter, preserving information shared between text and image as core semantics while dropping non-shared domain styles. (c) Our method achieves state-of-the-art performance across DG benchmarks, producing an embedding space with improved class separation.

domains [2, 4, 15, 18, 21, 29, 36, 46, 58] or employ robust training strategies to mitigate domain-specific correlations [10, 11, 51].

Recently, the field has turned to large-scale vision–language models (VLMs) such as CLIP [43], motivated by their strong zero-shot generalization ability. Because standard fine-tuning can compromise this robustness under distribution shift [40], current strategies aim to preserve the original zero-shot representations while adapting them to the DG task [55]—often freezing the text encoder while adapting the visual branch—via knowledge distillation [1, 23, 54], prompt tuning [13, 27, 33, 35, 54, 60, 62, 63], or weight ensembling [28, 44, 55].

These approaches share a common premise that preserving the pre-trained visual knowledge of VLMs (*e.g.*, CLIP) benefits generalization through zero-shot robustness or semantic expressiveness. *But does preserving the visual features truly benefit domain generalization?* Consider the 'apple' examples in Fig. 1(a). Highly expressive visual encoders entangle the core semantics (★) with visual styles, such as sketch strokes, photographic textures, or artistic abstractions. Without an explicit criterion to distinguish true semantics from these variations, models preserving such visual representations tend to retain any feature that is predictive of labels in the training domains. Consequently, domain-specific cues are absorbed alongside true semantics, expanding the range of a class beyond its core and blurring its boundary under domain shifts. In contrast, the ideal representation retains only core semantics by dropping domain spurious cues.

To avoid inheriting domain entanglement from visual guidance, we propose to discard it altogether. Instead, we elevate the text space to serve as the primary

source of domain invariance, drawing on both the stability of its anchors and their semantic structure (Sec. 3). Concretely, as shown in Fig. 1(b), features from diverse domains are constrained to align with text-defined anchors through a class-conditional Information Bottleneck (IB). These anchors remove domain-induced styles not shared across modalities while preserving information shared between input images and text.

Extensive experiments demonstrate that our approach consistently achieves the state-of-the-art performance across representative DG benchmarks with clearer separation in the learned embedding space, as illustrated in Fig. 1(c). Importantly, our evaluation spans diverse backbone architectures, further demonstrating the generality of the proposed framework.

Our contributions are summarized as follows:

1. Revisiting the visual guidance in DG, we reveal that highly expressive visual encoders can propagate domain-specific cues and hinder domain invariance under distribution shift.
2. We propose a purely text-guided approach based on IB theory, suppressing domain-specific variations while preserving shared semantics.
3. Through extensive experiments across diverse DG benchmarks and backbone architectures, we demonstrate consistent state-of-the-art performance and improved reliability under domain shift.
4. We further analyze guidance signals in DG and highlight supervision design as a key factor for learning invariant representations.

2 Related Work

Domain Generalization (DG). Visual recognition models often suffer from real-world distribution shift. To evaluate these failures, established benchmarks span diverse shifts: style variation (*e.g.*, sketch *vs.* photo) [30,39,50], differences in dataset acquisition process [48] or camera location [6], and background shifts [61]. Together, these benchmarks expose the inherent brittleness of models under unseen environments.

DG seeks representations that remain invariant across such domain discrepancies without access to target data. Classical approaches primarily extract invariance by aligning source-domain feature distributions via statistical matching [31, 46], adversarial learning [18, 19, 31], or feature disentanglement [12, 21, 25, 26, 34, 36, 41, 58]. Others instead regularize optimization dynamics [2, 4], gradient constraints [28, 51], ensembling [5, 10, 24, 32], and consistency guidance [11]. These approaches define invariance based on patterns shared across the training domains, yet such shared structure often mix true semantics with spurious correlations, limiting generalization beyond source-like distributions.

Recent approaches increasingly build upon CLIP [43], adapting its visual encoder while attempting to preserve zero-shot robustness. Representative strategies include prompt optimization [13, 27, 33, 35, 54, 60, 62, 63], robust fine-tuning and weight ensembling [28,37,44,55], and knowledge distillation [1,23,54]. Across

these methods, the visual encoder is typically the component being updated during adaptation. Meanwhile, its pretrained knowledge is preserved and reused through mechanisms such as regularization, ensembling, or distillation.

In contrast, the text encoder is usually kept frozen and serves as a relatively stable semantic reference. Depending on the method, text embeddings are used (i) as auxiliary alignment targets alongside visual supervision [1, 23, 44, 54], (ii) as regularizers to constrain visual drift from the original zero-shot space [37], or (iii) as diversified prompts to steer visual representations via multimodal alignment [13, 33, 35]. We instead revisit the functional roles of CLIP's visual and text encoders in DG, minimizing reliance on visual guidance and treating textual semantics as the primary basis for defining invariance.

Information Bottleneck in DG. Empirical Risk Minimization (ERM) [49] often struggles under distribution shift due to spurious correlations. While Invariant Risk Minimization [4] enforces a shared classifier across environments, it does not fully eliminate such biases. To further constrain representations, several classical DG methods adopt the IB principle. As the IB objective is intractable, these approaches rely on variational formulations, introducing KL-based regularization toward simple, typically non-semantic priors [2, 29], with meta-learning [15], or coupling it with feature disentanglement objectives [58]. In contrast, we anchor the bottleneck to fixed text embeddings as an explicit semantic prior. Existing IB priors are either uninformative (*e.g.*, $\mathcal{N}(0, \mathbf{I})$), giving no signal to separate semantic from spurious factors, or learned from source images, inheriting domain bias. In contrast, our approach is class-conditional yet label-derived, signaling what to preserve without injecting domain bias.

3 Motivation

To assess whether visual and textual modalities provide structurally reliable basis for domain-invariant guidance, we investigate the following questions: 1) Do their embedding spaces remain stable across domains? 2) How much domain-invariant and domain-specific information do the modality-specific encoders encode? 3) How do visual and textual guidance signals affect learning dynamics?

3.1 Empirical Examination

Cross-domain Reliability of Embedding Space. We examine the geometry of multimodal embedding spaces across domains by extracting visual and textual CLIP features from image–caption pairs in the four PACS domains [30] (art painting, photo, cartoon, sketch). To obtain textual embeddings as rich as their visual counterparts, we generate captions with semantic and stylistic details (Sec. A of the suppl.), instead of using predefined prompts like `"a [domain] of [class]"`, and feed them into a frozen CLIP text encoder. This provides rich instance-level descriptions and enables a fair comparison between modalities.

As shown in Fig. 2(a), the visual space (top) exhibits pronounced domain-dependent dispersion. Even for the same class (*e.g.*, person, highlighted with

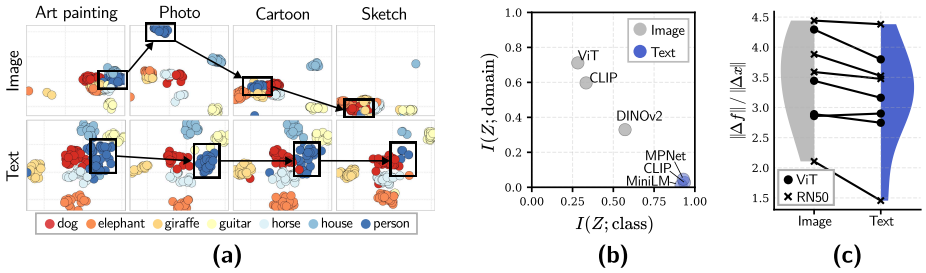

Fig. 2: Motivational experiments. (a) Textual embedding space (bottom) shows superior stability across domains than visual counterpart (top). (b) Visual encoders (gray) tend to possess both domain-specific and core-class information, while textual ones (blue) contain only the latter. (c) Text-guided models tend to yield lower Lipschitz constants, indicating smoother and more stable representations.

black boxes), the corresponding clusters significantly shift across domains. In contrast, text embeddings (bottom) remain relatively stable, largely insensitive to stylistic variations present in the captions. This contrast reveals a notable gap: while visual encoders capture fine-grained details including domain-specific cues, text embeddings remain semantically structured.

Encoded Information in Embedding Spaces. The previous experiment implies that the textual embedding space likely encode more domain-invariant information compared to the visual counterpart. Also, one might wonder whether this behavior is specific to CLIP. To answer these questions, we report in Fig. 2(b) the normalized mutual information between the learned embeddings and both class labels (x-axis) and domain labels (y-axis) across multiple encoders.

Across visual backbones (ViT [14], CLIP-Image [43], and DINOv2 [38]), their embeddings consistently encode substantial domain information $I(Z;\text{domain})$ alongside class information $I(Z;\text{class})$. In contrast, language models (MPNet [45], MiniLM [52], and CLIP-Text [43]) exhibit near-zero domain information, while primarily encoding class semantics. This result strongly implies that domain entanglement is indeed inherent to expressive visual representations. Preserving fine-grained variations, visual encoders inadvertently keep some domain-specific factors as well. In contrast, textual representations exhibit much less stylistic variation, leading them to align more consistently with semantic structure. Considering that domain generalization strictly requires isolating domain-invariant semantics from such variations, a natural hypothesis from this experiment is that textual space would be more appropriate to guide DG models.

Learning Dynamics under Domain-Dependent Guidance. A natural next question is if the guidance signals contain domain-dependent variations, how does this actually affect learning dynamics? To investigate this, we estimate the local Lipschitz constant, approximated by the norm of the gradient of learned features with respect to unseen target inputs [56]. This metric measures how sensitively the representation reacts to small perturbations under domain shift.

As shown in Fig. 2(c), student models (ViT [14] and ResNet-50 [22]) distilled with visual signals exhibit consistently higher Lipschitz values than those trained with textual guidance, indicating more input-sensitive mappings. This result verifies our earlier hypothesis that visual guidance would expose the student to cross-domain conflicting cues (Fig. 2(a)). Fitting to such inconsistent signals makes the representation more sensitive to small input changes and compromises stability. In contrast, textual guidance mitigates cross-domain conflict, yielding smoother and more stable representations under domain shift.

3.2 Information-theoretic Perspective

Beyond our empirical diagnostics, we further provide an information-theoretic perspective on our two design choices: removing domain-contaminated visual guidance and introducing an explicit information bottleneck. We interpret this through the lens of finite model capacity (formal assumptions and derivations are provided in Suppl. Sec. B).

An input image X can be decomposed into domain-invariant and domain-specific (spurious) components: $X = (X_{\text{inv}}, X_{\text{sp}})$. Denoting the output of the student model ϕ as $S = \phi(X)$ with finite capacity C, we have $I(S; X_{\text{inv}}) + I(S; X_{\text{sp}}) \leq C$. This constraint implies a trade-off between invariant semantics and spurious domain styles. Because visual guidance depends on the input image, it carries domain-specific variation X_{sp} into the supervision signal. Learning from such guidance increases $I(S; X_{\text{sp}})$, thereby reducing the capacity available for X_{inv}. Since labels depend only on X_{inv}, this limits the attainable predictive information and degrades generalization to unseen domains. This provides additional theoretical support for removing visual guidance.

However, the removal alone does not fully resolve the capacity allocation problem. Although domain signals X_{sp} are no longer injected through supervision, they are still present in the input and may be encoded by the model during training. In practice, a high-capacity model can learn separate domain-specific feature pathways that reach the same semantic target, leaving the representation entangled with X_{sp}. To explicitly restrict this spurious capacity allocation, we introduce a Text-Anchored Information Bottleneck in the next section.

4 Method

Our previous diagnostics verify that domain invariance is largely determined by the structure of supervision. Instead of relying on visually entangled guidance, which inevitably propagates spurious domain variations, we propose to anchor the training entirely to fixed, domain-invariant text embeddings, namely, Text-Anchored Information Bottleneck, illustrated in Fig. 3. We first formalize the domain generalization problem and introduce the Information Bottleneck formulations (Sec. 4.1), then present our text-anchored framework with two mechanisms: preserving semantic structure in the text space and suppressing spurious visual cues (Sec. 4.2).

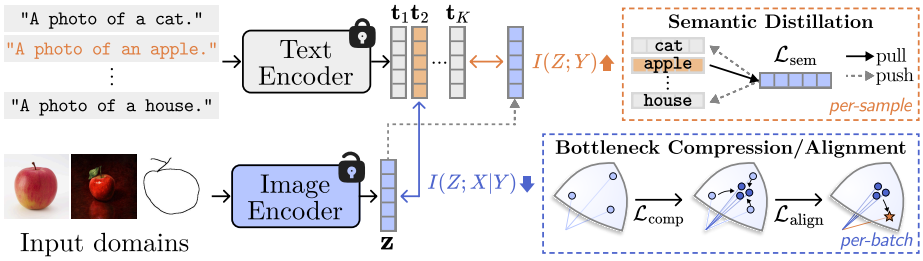

Fig. 3: Overview of Text-Anchored Information Bottleneck. Text guidance is the primary source of domain-invariance under our Conditional Entropy Bottleneck (CEB) formulation, composed of two parts: i) Semantic distillation ($\mathcal{L}_{\text{sem}}$) maximizes $I(Z;Y)$ by pulling image representations toward text anchors, and ii) Bottleneck compression and alignment minimizes $I(Z;X|Y)$ to suppress domain-specific variations, achieved by encouraging intra-class concentration via $\mathcal{L}_{\text{comp}}$ and aligning class-wise mean feature with text anchors via $\mathcal{L}_{\text{align}}$.

4.1 Preliminary

We first formulate the DG problem and introduce the Information Bottleneck and its conditional variant, which governs our semantic purification strategy.

Problem Formulation. In domain generalization, training data from K source domains $\mathcal{D}_S = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$ are given, where $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}$ from each domain k consists of N_k samples and follows a probability distribution $P_k(X,Y)$. The core challenge is distribution shift, where $P(Y|X)$ remains invariant while the marginals differ across domains ($P_i(X) \neq P_j(X)$). The task aims to learn a feature encoder $Z = f_\theta(X)$ that generalizes to an unseen target domain $\mathcal{D}_T$, where $\mathcal{D}_T \cap \mathcal{D}_S = \emptyset$.

Information Bottleneck. To learn robust representations Z, we adopt the Information Bottleneck (IB) [47], which seeks features that are maximally predictive of labels Y while compressing the input X:

$$\mathcal{L}_{\text{IB}} = -I(Z;Y) + \beta I(Z;X), \tag{1}$$

where $I(\cdot;\cdot)$ denotes mutual information and $\beta > 0$ controls the trade-off between prediction and compression. Assuming the Markov chain $Y \leftrightarrow X \leftrightarrow Z$, IB encourages Z to retain label-relevant information while discarding irrelevant variations in X.

Minimizing $I(Z;X)$ alone, however, does not explicitly distinguish label-relevant semantics from domain-specific factors, since domain cues may still be predictive of Y within the training domains. Therefore, we minimize the Conditional Information Bottleneck (CEB) [16]:

$$\mathcal{L}_{\text{CEB}} = -I(Z;Y) + \beta I(Z;X|Y), \tag{2}$$

which replaces $I(Z;X)$ with $I(Z;X|Y)$. By conditioning on the label Y, CEB removes input information that is unnecessary given Y. This provides a direct mechanism for suppressing spurious domain-specific variations, as labels serve as semantic anchors that isolate task-relevant information.

4.2 Text-Anchored Information Bottleneck

Fig. 3 presents our architecture-agnostic training framework for optimizing the CEB under fixed textual semantics. With K classes, we obtain semantic anchors $T = [\mathbf{t}_1, \ldots, \mathbf{t}_K]^\top \in \mathbb{R}^{K \times d}$, where each $\mathbf{t}_k \in \mathbb{R}^d$ is the frozen CLIP text embedding of the prompt `'a photo of a [class]'`. Although we have used rich captions in Sec. 3 to fairly match image richness, here we adopt the simplest prompt, which strips instance-level noise while keeping class identity. A trainable visual encoder f maps an input image $\boldsymbol{x}$ to $\boldsymbol{z} = f_\theta(\boldsymbol{x}) \in \mathbb{R}^d$. Motivated by Sec. 3, we treat these fixed text embeddings as the primary source of domain invariance, explicitly anchoring the representation space to domain-stable semantics rather than relying on invariance to emerge implicitly or using text as auxiliary guidance. We derive concrete forms of the two complementary objectives, maximizing predictive sufficiency $I(Z;Y)$ and minimizing anchor-inconsistent variation $I(Z;X|Y)$, comprising our CEB formulation in Eq. (2). This yields three terms: a semantic distillation loss $\mathcal{L}_{\text{sem}}$ that maximizes $I(Z;Y)$, and compression and alignment losses $\mathcal{L}_{\text{comp}}$, $\mathcal{L}_{\text{align}}$ that minimize $I(Z;X|Y)$, together forming our final objective (Eq. (12)). We further detail each below.

Maximizing $I(Z;Y)$. The mutual information term $I(Z;Y)$ in Eq. (2) quantifies the dependency between the representation Z and the target Y, defined as

$$I(Z;Y) = \mathbb{E}_{Z,Y}\left[\log \frac{p(Z,Y)}{p(Z)p(Y)}\right] = \mathbb{E}_{Z,Y}\Big[\log p(Y|Z)\Big] + H(Y), \tag{3}$$

where $H(Y)$ is constant. Thus, maximizing $I(Z;Y)$ is equivalent to maximizing $\mathbb{E}_{Z,Y}[\log p(Y|Z)]$, *i.e.*, minimizing the cross-entropy loss [3]. We parameterize $p(Y|Z)$ using fixed text embeddings as class prototypes. For a sample with label k, we define:

$$\mathcal{L}_{\text{sem}} = -\log \frac{\exp(\cos(\boldsymbol{z}, \mathbf{t}_k)/\tau)}{\sum_{k' \in K} \exp(\cos(\boldsymbol{z}, \mathbf{t}_{k'})/\tau)}, \tag{4}$$

where $\mathbf{t}_k$ is the text embedding of class k and τ is a temperature. This objective pulls representations toward their class anchors and pushes them away from others, thereby distilling the semantic structure of the textual space into the learned representation (Fig. 3, upper right).

Minimizing $I(Z;X|Y)$. We minimize the second term $I(Z;X|Y)$ in Eq. (2) to suppress spurious domain variations. We first derive its variational upper bound, and then convert it into a tractable training objective.

Starting from the definition of conditional mutual information,

$$I(Z;X|Y) \coloneqq \mathbb{E}\left[\log \frac{p(X,Z|Y)}{p(X|Y)p(Z|Y)}\right] = \mathbb{E}\left[\log \frac{p(Z|X,Y)}{p(Z|Y)}\right] = \mathbb{E}\left[\log \frac{p(Z|X)}{p(Z|Y)}\right], \tag{5}$$

where the last equality follows from the Markov chain $Y \leftrightarrow X \leftrightarrow Z$ assumption, implying $p(Z|X,Y) = p(Z|X)$. Since the true marginal $p(Z|Y)$ is intractable, we introduce a variational approximation $q(Z|Y)$. By non-negativity of KL divergence,

$$I(Z;X|Y) \le \mathbb{E}[\log p(Z|X)] - \mathbb{E}[\log q(Z|Y)] = \mathbb{E}\big[\mathrm{KL}(p(Z|X) \| q(Z|Y))\big]. \tag{6}$$

We now derive a stable training objective from this bound. For samples belonging to class k, Eq. (6) becomes

$$\mathbb{E}_{X|y=k}\big[\mathrm{KL}(p(Z|X)\|q(Z|y=k))\big] = \mathbb{E}_{X|y=k}\big[\mathbb{E}_Z\big[\log p(Z|X) - \log q(Z|y=k)\big]\big], \tag{7}$$

Since CLIP embeddings are l_2-normalized and lie on the unit hypersphere, we model both p and q as von Mises–Fisher (vMF) distributions [17], parameterized by a mean direction $\boldsymbol{\mu}$ and concentration κ. A vMF can be viewed as the spherical analogue of a Gaussian, concentrating probability mass around a direction on the unit sphere. Specifically, $q(Z|y=k)$ is defined as a vMF centered at the fixed text embedding $\mathbf{t}_k$ obtained from the frozen text encoder, with a fixed concentration κ_q, while $p(Z|x)$ is centered at the image feature $\boldsymbol{z} = f_\theta(x)$, with concentration κ_p. With κ_p fixed, the first term in Eq. (7), $\mathbb{E}_Z[\log p(Z|X)]$, becomes constant with respect to θ. Thus, the optimization reduces to minimizing the second term, $-\mathbb{E}_{X|y=k}\mathbb{E}_Z[\log q(Z|y=k)]$.

Approximating $Z \sim p(Z|x)$ by its mean feature $\boldsymbol{z} = f_\theta(x)$ (with a large κ_p), and using the vMF density $q(\boldsymbol{z}|y=k) = C_d(\kappa_q)\exp(\kappa_q \boldsymbol{t}_k^\top \boldsymbol{z})$, the log-density can be substituted into Eq. (7). Applying the empirical expectation over a minibatch $\mathcal{B}_k$ yields

$$-\frac{1}{|\mathcal{B}_k|}\sum_{i\in\mathcal{B}_k}[\log C_d(\kappa_q) + \kappa_q \mathbf{t}_k^\top \boldsymbol{z}_i] = \text{const} - \kappa_q\, \mathbf{t}_k^\top \left(\frac{1}{|\mathcal{B}_k|}\sum_{i\in\mathcal{B}_k}\boldsymbol{z}_i\right), \tag{8}$$

where both κ_q and $C_d(\kappa_q)$ are constants. Let $\bar{\boldsymbol{z}}_k = \sum_{i\in\mathcal{B}_k}\boldsymbol{z}_i/|\mathcal{B}_k|$ denote the mean feature of the class k, the objective reduces to

$$-\boldsymbol{t}_k^\top \bar{\boldsymbol{z}}_k = -\|\boldsymbol{t}_k\|\|\bar{\boldsymbol{z}}_k\|\cos(\boldsymbol{t}_k, \bar{\boldsymbol{z}}_k) = -\|\bar{\boldsymbol{z}}_k\|\cos(\boldsymbol{t}_k, \bar{\boldsymbol{z}}_k), \tag{9}$$

since $\|\boldsymbol{t}_k\| = 1$. Empirically, however, directly optimizing this multiplicative form couples the magnitude and the cosine terms, which can lead to poorly conditioned gradients during training (*e.g.*, the cosine term may quickly saturate, weakening gradients acting on the $\|\bar{\boldsymbol{z}}\|$ term). We therefore adopt a separable surrogate objective that independently encourages (i) a large intra-class resultant length $\|\bar{\boldsymbol{z}}_k\|$, and (ii) strong directional alignment $\cos(\mathbf{t}_k, \bar{\boldsymbol{z}}_k)$ with the text anchor $\mathbf{t}_k$, to stabilize optimization. From the magnitude term, we define:

$$\mathcal{L}_{\text{comp}} = -\frac{1}{|K_\mathcal{B}|}\sum_{k\in K_\mathcal{B}}\|\bar{\boldsymbol{z}}_k\|, \tag{10}$$

where $K_\mathcal{B}$ denotes the set of classes present in a mini-batch $\mathcal{B}$. This encourages features of the same class to concentrate along a coherent direction (Fig. 3, lower right). From the cosine term, we define

$$\mathcal{L}_{\text{align}} = -\frac{1}{|K_\mathcal{B}|}\sum_{k\in K_\mathcal{B}}\cos\left(\mathbf{t}_k, \frac{\bar{\boldsymbol{z}}_k}{\|\bar{\boldsymbol{z}}_k\|}\right) \tag{11}$$

which aligns each class mean with its text anchor. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{sem}} + \beta_1\mathcal{L}_{\text{align}} + \beta_2\mathcal{L}_{\text{comp}}, \tag{12}$$

where β_1 and β_2 balance the two effects.

Unlike prior IB-based DG methods [15, 29, 58] that learn explicit Gaussian parameters (*e.g.*, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$) from each image to match a simple unconditional prior (*e.g.*, $\mathcal{N}(0, \mathbf{I})$), irrespective of class semantics, we impose a class-conditional prior anchored by fixed text embeddings, enabling more reliable compression of the representations.

5 Experiments

5.1 Experimental Setting

Datasets. We evaluate on six standard DG benchmarks: TerraIncognita [6], OfficeHome [50], VLCS [48], PACS [30], DomainNet [39], and NICO^{++} [61]. The first four contain four domains each with 10, 65, 5, and 7 classes, respectively. For large-scale evaluation, we use the latter two, DomainNet (6 domains, 345 classes) and NICO^{++} (6 domains, 60 classes).

Baselines. Across diverse backbones, representative DG methods vary by architecture. For consistent comparison, we focus on a set of key baselines. Linear probing (LP) evaluates frozen features, while MIRO [11] represents classical DG without external guidance. RISE [23] and VL2V [1] perform cross-modal distillation, explicitly using a CLIP image encoder as the primary teacher with simple alignment objectives for text supervision. In contrast, CLIP-dedicated methods such as CLIPood [44] do not employ an external teacher but implicitly retain original CLIP visual features through zero-shot preservation. When applied to non-CLIP backbones, however, this preservation operates on backbone-specific weights, so CLIP visual features are no longer involved in adaptation and guidance is provided only by CLIP text embeddings.

Implementation Details. We follow the DomainBed [20] protocol. All datasets use the standard leave-one-domain-out setting, where one domain is held out for testing while training on the remaining domains. For NICO^{++}, we adopt a leave-one-group-out protocol, where multiple domains are grouped and one entire group is reserved for evaluation. Model selection is based on validation accuracy using a 20% split of training data. We use $\beta_1 = 0.1$ and $\beta_2 = 1.0$ based on ResNet-50 on PACS, and apply them across all datasets and backbones. Results are averaged over three runs. See Sec. C–E of the suppl. for more details.

5.2 Comparison with Baselines

Main Results. Tab. 1 shows that our method achieves the state-of-the-art performance across various standard backbones. Existing approaches exhibit architectural bias: KD methods (*e.g.*, RISE, VL2V) favor conventional encoders, while CLIP-specialized methods favor CLIP backbones, and both degrade when evaluated beyond their original setup. In contrast, ours consistently improves the performance without backbone-specific design, demonstrating its universal robustness.

Table 1: DG performance comparison across representative backbones. The 'G' column denotes CLIP visual (V) or textual (T) guidance. The **best** and second-best results are highlighted. * denotes our implementation.

Method	G	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Average
ResNet-50 pretrained on ImageNet-1k							
LP	-	78.1	86.2	68.4	46.3	41.2	64.0
MIRO [11]	-	79.0	85.4	70.5	50.4	44.3	65.9
SAGM [51]	-	80.0	86.6	70.1	48.8	45.0	66.1
GESTUR [28]	-	80.1	88.0	71.1	51.3	46.3	67.4
INSURE [58]	-	-	89.3	72.0	53.1	48.0	-
CLIPood* [44]	T	76.7	88.8	70.3	44.7	-	-
RISE [23]	V,T	**81.7**	89.4	71.6	52.3	46.5	68.3
VL2V [1]	V,T	79.2	86.7	74.4	53.5	47.7	68.3
Ours	T	**81.7**	**96.9**	**79.0**	**59.9**	**58.3**	**75.4**
ViT-B/16 pretrained on ImageNet-1k							
LP	-	79.5	81.5	82.8	42.2	50.5	67.3
MIRO* [11]	-	80.4	81.5	74.9	44.5	-	-
CLIPood* [44]	T	80.4	87.1	80.9	44.3	-	-
RISE* [23]	V,T	84.2	91.0	80.3	44.6	56.6	71.3
VL2V [1]	V,T	81.9	**94.9**	85.7	55.4	59.4	75.5
Ours	T	**86.2**	94.1	**86.4**	**62.2**	**68.8**	**79.5**
CLIP-ViT-B/16 pretrained on private dataset (400M)							
LP	-	83.4	97.2	82.3	57.3	58.2	75.7
CLIP-ZS	-	82.4	96.1	82.3	34.4	49.7	69.0
MIRO [11]	-	82.2	95.6	82.5	54.3	54.0	73.7
GESTUR [28]	-	82.8	96.0	84.2	55.7	58.9	75.5
CAR-FT [35]	V,T	85.5	96.8	85.7	61.9	62.5	78.5
CLIPood [44]	V,T	85.0	97.3	87.0	60.4	63.5	78.6
CLIPCEIL [59]	V,T	85.2	97.2	87.7	62.0	63.6	79.1
CLIP-DPR [13]	V,T	86.4	97.5	86.1	57.1	62.1	77.8
CLIP-DTP [54]	V,T	84.8	97.0	87.7	63.3	63.1	79.2
RISE [23]	V,T	80.6	93.3	78.4	49.6	55.4	71.5
VL2V [1]	V,T	83.3	96.7	87.4	58.5	62.8	77.7
Ours	T	**89.0**	**98.5**	**93.2**	**75.1**	**75.8**	**86.3**

Robustness to Background Shift. We further compare the performance of competing DG models in Tab. 2 on NICO^{++}, which focuses on background-driven shift. Notably, the margin becomes more pronounced; ResNet-50 reaches 95.7% accuracy, nearly closing the gap to CLIP (97.5%). This suggests that this setting is particularly better-aligned with our text-anchored bottleneck, since background variations are more readily separable from foreground semantics than the style shifts in Tab. 1, where object appearance itself can vary. In such cases, varying background is more likely to be filtered as spurious style. We further discuss in the supplementary (Sec. G) that this purification improves overall robustness, though it may fail in rare cases where contextual cues are genuinely required for identification.

Generalization across Backbones. While Tab. 1 evaluates standard architectures, Fig. 4 extends the comparison to a broader range of backbones, including lightweight models, CNNs, and diverse transformers. Across all backbones, ours is the only method that consistently maintains a clear margin over LP.

Interestingly, when the pretrained backbone is already strong (*i.e.*, high LP performance), other state-of-the-art methods often shrink to near-zero or

Table 2: Performance on NICO^{++}. Results for ResNet-50 and CLIP backbones evaluated via a leave-one-group-out protocol on predefined target pairs: Autumn & Rock, Dim & Grass, and Outdoor & Water.

	ResNet-50 pretrained on ImageNet-1k							*CLIP-ViT-B/16*						
Method	A	R	D	G	O	W	Avg	A	R	D	G	O	W	Avg
LP	85.3	85.6	78.6	86.5	82.0	76.7	82.5	91.4	92.2	89.8	92.7	90.4	84.6	90.2
MIRO* [11]	82.3	81.2	73.5	81.5	77.9	72.0	78.1	92.1	91.4	86.7	92.3	89.3	83.8	89.3
VL2V* [1]	85.2	84.3	77.1	87.1	82.7	78.9	82.6	92.7	92.8	90.1	94.7	92.3	88.5	91.8
SRE [53]	-	-	-	-	-	-	-	91.4	92.3	90.4	93.2	90.8	86.4	90.8
CLIPood* [44]	-	-	-	-	-	-	-	93.0	94.3	89.6	93.7	92.0	86.8	91.5
Ours	**95.7**	**97.3**	**94.4**	**97.7**	**96.4**	**92.9**	**95.7**	**97.9**	**98.5**	**97.7**	**98.7**	**98.3**	**94.0**	**97.5**

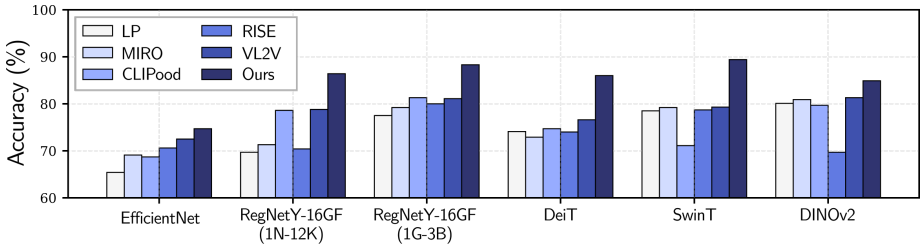

Fig. 4: DG performance across diverse backbones. Average accuracy is reported. We evaluate across CNNs (EfficientNet, RegNetY) and Transformers with different architectures and pretraining schemes (SwinT, DeiT, DINOv2).

even negative margins (*e.g.*, MIRO on SwinT and RISE on DINOv2). This implies that KD approaches (RISE, VL2V) risk distorting well-structured semantic spaces, while regularization-based methods (MIRO, CLIPood) largely preserve them without actively removing domain-specific factors, yielding only marginal gains. In contrast, our text-anchored compression explicitly filters spurious variations, sustaining positive improvements regardless of backbone strength.

Recent work [57] notes that some DG benchmarks may be affected by data leakage, as supervised backbones could have encountered similar domains during large-scale pretraining. To partially control for this effect, we additionally evaluate on DINOv2, a self-supervised backbone less likely to exhibit such leakage. While most methods remain close to the LP baseline in this setting, our method maintains a clear margin, suggesting that the gains are not solely attributable to inherited exposure, but instead stem from the intended domain-invariance mechanism.

5.3 Further Analysis

Ablation on Loss Components. Tab. 3 ablates the three terms across three backbones on OfficeHome, PACS, and DomainNet. Adding $\mathcal{L}_{\text{comp}}$ to $\mathcal{L}_{\text{sem}}$ drives the largest gains, while $\mathcal{L}_{\text{align}}$ provides a consistent further improvement; their combination performs the best across all backbones and datasets, confirming that

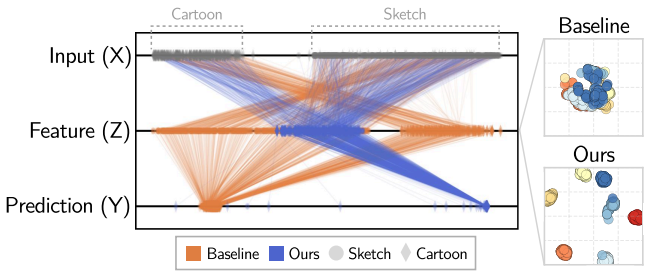

Fig. 5: Information flow ($X \to Z \to Y$) visualization using t-SNE on PACS. *Left*: For the *house* class, our method (blue) collapses inputs from different domains into a single cluster, while RISE (orange) retains domain-dependent separation, showing reduced cross-domain variation. *Right*: Within a single domain, our embeddings form more compact, well-separated class clusters, showing sharper class margins.

Table 3: Ablation of loss components

$\mathcal{L}_{\text{sem}}$	$\mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{comp}}$	OfficeHome			PACS			DomainNet		
			RN50	ViT	CLIP	RN50	ViT	CLIP	RN50	ViT	CLIP
✓			73.4	82.4	85.2	92.6	92.3	96.8	35.5	52.8	60.3
✓	✓		76.0	81.9	85.9	94.2	93.6	98.2	37.3	67.9	65.9
✓		✓	78.9	85.1	90.8	95.8	94.0	98.1	57.6	68.7	75.2
✓	✓	✓	**79.0**	**86.4**	**93.2**	**96.9**	**94.1**	**98.5**	**58.3**	**68.8**	**75.8**

compression and alignment are complementary. Sec. H of the suppl. extends this ablation to other backbones.

Does the bottleneck indeed suppress domain-specific information? To verify this, we visualize the information flow for a specific class (*house*) using t-SNE in Fig. 5. Across the cartoon and sketch domains in PACS, the inputs (top) are clearly domain-separated. Our features (blue in the middle, Z) concentrate into a unified cluster successfully, discarding domain information. On the other hand, the baseline (RISE) features (orange in the middle, Z) largely retain the domain-dependent separation. Within each single domain, our embedding space (right) forms more compact class clusters, indicating that the text-anchored bottleneck removes cross-domain variation, enhancing class separability with larger margins.

To examine the underlying mechanism, we track the CEB term $I(Z;X|Y)$ during training. Fig. 6(a) estimates this quantity using MINE [7]. In practice, $I(Z;X|Y)$ is approximated by the alignment and compression terms ($\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{comp}}$ in Eq. (12)), which decrease together during training, indicating that the objective suppresses dependence on input-specific variations. This mechanism also facilitates semantic learning. Fig. 6(b) compares the full model with a variant trained only with $\mathcal{L}_{\text{sem}}$. By filtering domain-specific variations, the model allows $\mathcal{L}_{\text{sem}}$ to converge to a lower value, resulting in higher accuracy.

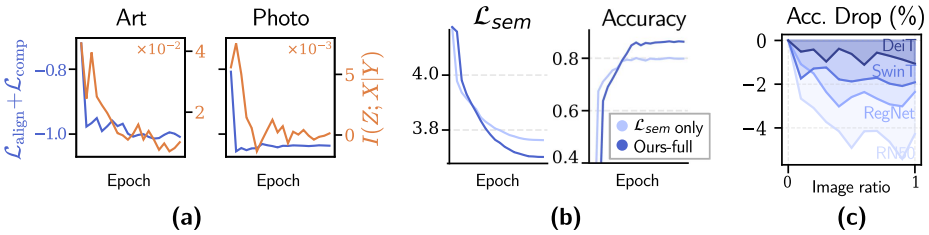

Fig. 6: (a) $I(Z;X|Y)$ decreases with the regularizers $\mathcal{L}_{\text{align}}, \mathcal{L}_{\text{comp}}$ during training. (b) Training with the regularizers achieves lower semantic loss and higher accuracy. (c) Accuracy decreases as the image guidance ratio increases.

Does visual guidance reintroduce domain entanglement? In Fig. 6(c), we test this by injecting CLIP image supervision into our objective in Eq. (12) with varied guidance ratios. As the ratio increases, the accuracy consistently drops, with CNNs (RegNet, RN50) degrading more than Transformers (DeiT, SwinT), likely because such guidance propagates domain-specific statistics that CNNs are more sensitive to. This indicates that aligning with expressive visual features disrupts the text-driven abstraction. Interestingly, this trend contrasts with prior KD approaches [1,23], where a small amount of textual guidance often improves performance, suggesting different roles of text under the two settings.

Why do text embeddings serve as effective anchors? Since our bottleneck relies on textual anchors, we examine whether it depends on specific CLIP properties or reflects a more general characteristic of language spaces. Tab. 4(a) compares anchors from different language encoders (MiniLM [52], MPNet [45]) and prompt templates on OfficeHome. All variants perform similarly, indicating that the framework is largely agnostic to particular language model and not tied to CLIP-specific representations (see Suppl. Sec. I–J for richer prompt variants and class-similarity analysis).

Even random anchors achieve competitive performance, surpassing the previous state-of-the-art (76.6 *vs.* 74.4), suggesting that stable external anchors are the key to the framework's effectiveness, while pretrained text embeddings further benefit from their learned semantic structure.

Crucially, this semantic advantage grows as class discrimination becomes harder; on DomainNet (345 classes), the gap between CLIP and random anchors widens sharply (58.3 *vs.* 36.4). It is most evident in open-set generalization, where unseen test classes make random anchors inapplicable, yet text anchors still generalize well (78.8 *vs.* CLIPood 78.2; Tab. 5).

How should anchors be integrated into training? Tab. 4(b) varies three factors on OfficeHome and PACS: the anchor source (random *vs.* CLIP), whether anchors are kept fixed (✓), and the alignment objective. Freezing inconsistently helps for contrastive and L2 while consistently under ours, as learnable anchors otherwise absorb domain-specific statistics and drift from invariant references. Likewise, our CEB-based objective is the only one that turns semantic CLIP

Table 4: Analysis of anchor design: (a) what to use and (b) how to integrate it

(a) Effect of anchor encoder and template

Encoder	Template	Acc.
CLIP (VL)	'a photo of [cls]'	79.0
	ImageNet prompts [43]	**77.9**
MiniLM (L)	'a photo of [cls]'	79.1
	ImageNet prompts [43]	77.7
MPNet (L)	'a photo of [cls]'	78.7
	ImageNet prompts [43]	78.9
random	-	76.6

(b) Effect of anchor source and objective

	Contrastive		L2		Ours	
Fixed?	✗	✓	✗	✓	✗	✓
OfficeHome						
random	71.4	73.4	13.4	11.9	75.2	76.6
CLIP	74.6	73.4	68.3	68.4	77.2	**79.0**
PACS						
random	93.7	94.8	92.3	93.3	83.4	95.5
CLIP	94.6	94.9	91.7	88.0	95.2	**96.9**

Table 5: Comparison on Open-Set DG.

	Known classes	Unseen classes
CLIP	86.1	77.6
CLIPood	89.4	78.2
Ours	**90.6**	**78.8**

anchors into reliable gains, via compression and directional constraints. Overall, domain invariance emerges only when stable external anchors are paired with an objective that explicitly enforces such invariance.

6 Conclusion

Expressive visual features can propagate spurious cues, challenging the common belief that greater capacity ensures robustness. We introduce a Text-Anchored Information Bottleneck that explicitly grounds supervision in external anchors to directly address the source of domain bias. Rather than distilling invariance from domain-shifting visual data, we utilize stable external signals to prove that the structural source of supervision is more critical than model capacity. While fixed anchors lack contextual cues and benchmark data leakage [57] currently obscures true generalization gains, it will be an interesting future direction to explore adaptive anchors and alternative signals to resolve these limitations and better isolate core semantics.

Acknowledgments

This work was supported by Samsung Electronics, Youlchon Foundation, National Research Foundation of Korea (NRF) grants (RS-2021-NR05515, RS-2024-00336576, RS-2023-0022663), and the Institute for Information & Communication Technology Planning & Evaluation (IITP) grants (RS-2022-II220264, RS-2024-00353131) funded by the Korean government.

References

1. Addepalli, S., Asokan, A.R., Sharma, L., Babu, R.V.: Leveraging vision-language models for improving domain generalization in image classification. In: CVPR (2024)
2. Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I., Rish, I.: Invariance principle meets information bottleneck for out-of-distribution generalization. In: NeurIPS (2021)
3. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: ICLR (2017)
4. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv:1907.02893 (2020)
5. Arpit, D., Wang, H., Zhou, Y., Xiong, C.: Ensemble of averages: Improving model selection and boosting performance in domain generalization. In: NeurIPS (2022)
6. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: ECCV (2018)
7. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: ICML (2018)
8. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.: A theory of learning from different domains. Machine Learning **79** (2010)
9. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: NIPS (2006)
10. Cha, J., Chun, S., Lee, K., Cho, H.C., Park, S., Lee, Y., Park, S.: SWAD: Domain generalization by seeking flat minima. In: NeurIPS (2021)
11. Cha, J., Lee, K., Park, S., Chun, S.: Domain generalization by mutual-information regularization with pre-trained models. In: ECCV (2022)
12. Chattopadhyay, P., Balaji, Y., Hoffman, J.: Learning to balance specificity and invariance for in and out of domain generalization. In: ECCV (2020)
13. Cheng, D., Xu, Z., Jiang, X., Wang, N., Li, D., Gao, X.: Disentangled prompt representation for domain generalization. CVPR (2024)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
15. Du, Y., Xu, J., Xiong, H., Qiu, Q., Zhen, X., Snoek, C.G.M., Shao, L.: Learning to learn with variational information bottleneck for domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV (2020)
16. Fischer, I.: The conditional entropy bottleneck. Entropy **22**(9), 999 (Sep 2020)
17. Fisher, R.A.: Dispersion on a sphere. Proceedings of the royal society of London. Series A. Mathematical and physical sciences **217**(1130), 295–305 (1953)

18. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. pp. 1180–1189 (2015)
19. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR (2016)
20. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: ICLR (2021)
21. Guo, J., Qi, L., Shi, Y.: DomainDrop: Suppressing domain-sensitive channels for domain generalization. In: ICCV (2023)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
23. Huang, Z., Zhou, A., Lin, Z., Cai, M., Wang, H., Lee, Y.J.: A sentence speaks a thousand images: Domain generalization through distilling CLIP with language guidance. ICCV (2023)
24. Jain, S., Addepalli, S., Sahu, P.K., Dey, P., Babu, R.V.: DART: Diversify-aggregate-repeat training improves generalization of neural networks. CVPR (2023)
25. Jeon, M., Kang, M., Lee, J.: A unified framework for robustness on diverse sampling errors. In: ICCV (2023)
26. Jeon, M., Kim, D., Lee, W., Kang, M., Lee, J.: A conservative approach for unbiased learning on unknown biases. In: CVPR (2022)
27. khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: MaPLe: Multi-modal prompt learning. arXiv:2210.03117 (2022)
28. Lew, B., Son, D., Chang, B.: Gradient estimation for unseen domain risk minimization with pre-trained models. ICCVW (2023)
29. Li, B., Shen, Y., Wang, Y., Zhu, W., Reed, C., Zhang, J., Li, D., Keutzer, K., Zhao, H.: Invariant information bottleneck for domain generalization. In: AAAI (2021)
30. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: ICCV (2017)
31. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR (2018)
32. Li, Z., Ren, K., Jiang, X., Li, B., Zhang, H., Li, D.: Domain generalization using pretrained models without fine-tuning. arXiv:2203.04600 (2022)
33. Liu, G.M., Wang, Y.: TDG: Text-guided domain generalization. arXiv:2308.09931 (2023)
34. Lv, F., Liang, J., Li, S., Zang, B., Liu, C.H., Wang, Z., Liu, D.: Causality inspired representation learning for domain generalization. In: CVPR (2022)
35. Mao, X., Chen, Y., Jia, X., Zhang, R., Xue, H., Li, Z.: Context-aware robust fine-tuning. IJCV (Dec 2023)
36. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: ICML (2013)
37. Nam, G.C., Heo, B., Lee, J.: Lipsum-FT: Robust fine-tuning of zero-shot models using random text guidance. ICLR (2024)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision (2023)
39. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019)

40. Pham, H., Dai, Z., Ghiasi, G., Liu, H., Yu, A.W., Luong, M.T., Tan, M., Le, Q.V.: Combined scaling for zero-shot transfer learning. Neurocomputing (2021)
41. Piratla, V., Netrapalli, P., Sarawagi, S.: Efficient domain generalization via common-specific low-rank decomposition. In: ICML (2020)
42. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset Shift in Machine Learning. The MIT Press (2009)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
44. Shu, Y., Guo, X., Wu, J., Wang, X., Wang, J., Long, M.: CLIPood: Generalizing CLIP to out-of-distributions. In: ICML (2023)
45. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: MPNet: masked and permuted pre-training for language understanding. In: NIPS (2020)
46. Sun, B., Saenko, K.: Deep CORAL: Correlation alignment for deep domain adaptation. In: ECCV (2016)
47. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. In: Proc. of the Annual Allerton Conference on Communication, Control and Computing (1999)
48. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
49. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
50. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: CVPR (2017)
51. Wang, P., Zhang, Z., Lei, Z., Zhang, L.: Sharpness-aware gradient matching for domain generalization. CVPR (2023)
52. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: NeurIPS (2020)
53. Wang, Z., Gao, Z., Chen, J., Zhao, Q., Wu, X., Luo, J.: Simulate, refocus and ensemble: An attention-refocusing scheme for domain generalization. arXiv:2507.12851 (2025)
54. Wen, C., Peng, Z., Huang, Y., Yang, X., Shen, W.: Domain generalization in CLIP via learning with diverse text prompts. In: CVPR (2025)
55. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. In: CVPR (2022)
56. Yang, D., Lee, J., Kim, Y.: TAROT: Towards essentially domain-invariant robustness with theoretical justification. In: CVPR (2025)
57. Yu, H., Zhang, X., Xu, R., Liu, J., He, Y., Cui, P.: Rethinking the evaluation protocol of domain generalization. CVPR (2023)
58. Yu, X., Tseng, H.H., Yoo, S., Ling, H., Lin, Y.: INSURE: An information theory inspired disentanglement and purification model for domain generalization. IEEE TIP (2023)
59. Yu, X., Yoo, S., Lin, Y.: CLIPCEIL: Domain generalization through CLIP via channel refinement and image-text alignment. In: NeurIPS (2024)
60. Zhang, X., Iwasawa, Y., Matsuo, Y., Gu, S.S.: Domain prompt learning for efficiently adapting CLIP to unseen domains. Transactions of the Japanese Society for Artificial Intelligence (2021)
61. Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., Liu, H.: NICO++: Towards better benchmarking for domain generalization. CVPR (2022)
62. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022)

63. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (Jul 2022)