Learning Informative Invariant Representations via Hierarchical Latent Decomposition

Hyungkwon Lee[*]
KAIST
hyungkwon@kaist.ac.kr

Hyewon Ryu[*]
KAIST
hw.ryu@kaist.ac.kr

Jong-Seok Lee[†]
KAIST
jongseok.lee@kaist.ac.kr

Abstract

Domain generalization aims to learn models that generalize to unseen target domains using labeled data from multiple source domains. Many existing approaches focus on learning domain-invariant representations, but enforcing invariance alone may fail to preserve sufficient task-relevant information under domain shift. We propose BLENDER, a reconstruction-aware framework for domain generalization that employs a hierarchical latent decomposition to separate domain-invariant and domain-specific factors. This structure prioritizes informativeness in the domain-invariant representation and captures residual domain-specific variation through conditional modeling. We further provide a theoretical analysis establishing a bound on the reconstruction risk for unseen target domains, revealing how domain invariance and latent disentanglement contribute to generalization beyond the observed domains. Experiments on standard domain generalization datasets demonstrate strong performance under substantial domain shifts, and qualitative analyses show that BLENDER promotes a structured allocation of information between domain-invariant and domain-specific latent representations.

1 Introduction

In many real-world scenarios, models are deployed under domain shift, where the data distribution at test time differs from that observed during training due to dataset bias, acquisition conditions, environmental variation, or temporal non-stationarity [1–6]. Domain generalization addresses this challenge by training a model on multiple labeled source domains and enabling generalization to unseen target domains, without access to target domain samples during training [7].

A dominant paradigm in domain generalization is invariant feature learning, which aims to learn representations that capture factors consistent across domains while suppressing domain-specific variation [8–10]. However, invariance alone does not ensure that the learned representation remains sufficiently informative, even though it can reduce spurious correlations associated with specific domains [11]. In particular, existing invariance-based approaches often achieve domain invariance by discarding substantial portions of the input, inadvertently removing task-relevant information needed for prediction across multiple domains.

This limitation is closely related to information loss arising from non-invertible representation mappings. When the representation mapping is not approximately invertible, it may fail to preserve task-relevant information under domain shift, even if domain invariance is achieved [12]. From this perspective, recent work has emphasized the importance of information preservation in learned representations for domain generalization and provided reconstruction as one practical mechanism to encourage such preservation [13]. Meanwhile, prior work has explored reconstruction-based latent

[*]Equal contribution.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

decomposition objectives that separate representations into domain-invariant and domain-specific factors [14, 15]. However, despite their empirical success, these approaches offer limited theoretical insight into when the domain-invariant component remains informative under unseen domain shift. This motivates us to learn an informative invariant representation by disentangling domain-invariant and domain-specific factors and provide theoretical evidence for its informativeness.

In this setting, we argue that architectural structure plays a critical role in constraining how retained information is distributed across latent representations. A widely used architecture is to map domain-invariant and domain-specific factors into separate latent representations via parallel encoders,

$$z_{\text{inv}} = f_{\text{inv}}(x), \quad z_{\text{spec}} = f_{\text{spec}}(x), \tag{1}$$

where f_{inv} and f_{spec} are encoders that produce the latent representations z_{inv} and z_{spec} for input x, respectively. However, such a parallel structure does not impose any constraint on information redundancy. To formalize this, we employ mutual information $I(\cdot;\cdot)$ to measure statistical dependence between two random variables and entropy $H(\cdot)$ to measure the information content of a random variable. In the extreme case where both encoders are bijective, each latent variable may encode the entire input, yielding

$$I(X; Z_{\text{inv}}) = H(X), \quad I(X; Z_{\text{spec}}) = H(X), \tag{2}$$

and thus

$$I(X; Z_{\text{inv}}) + I(X; Z_{\text{spec}}) \geq H(X), \tag{3}$$

indicating substantial duplication of information. As a result, parallel encoders do not provide a structural constraint for a meaningful partition of information between domain-invariant and domain-specific components.

To mitigate this issue, we consider the chain rule decomposition of the mutual information between the input and latent representations, which can be written as

$$I(X; Z_{\text{inv}}, Z_{\text{spec}}) = I(X; Z_{\text{inv}}) + I(X; Z_{\text{spec}}|Z_{\text{inv}}) \leq H(X). \tag{4}$$

This suggests interpreting the domain-specific representation as capturing information beyond that explained by the invariant component. Motivated by this observation, we design a hierarchical encoding structure that encourages a separation between the two latents. We prioritize informativeness in z_{inv} and then define z_{spec} conditioned on z_{inv}, such that

$$z_{\text{inv}} = f_{\text{inv}}(x), \quad z_{\text{spec}} = f_{\text{spec}}(x, z_{\text{inv}}). \tag{5}$$

This structure introduces an ordered dependency between latent components, encouraging the domain-specific representation to capture complementary information relative to the invariant features. It thus complements reconstruction-based objectives and regularization encouraging independence between the latents, and supports a more structured decomposition of factors within the total information content of X.

In this work, we propose a reconstruction-aware domain generalization approach that introduces a hierarchical latent decomposition to balance invariance and information preservation. To support its generalization to unseen domains, we derive an upper bound on unseen target domain reconstruction risk that explicitly accounts for domain shift. Our analysis clarifies when reconstruction objectives learned from source domains can generalize beyond the source distributions.

2 Literature Review

Invariant Feature Learning Many studies in domain generalization have aimed to obtain representations that remain invariant across domains and generalize to unseen targets [7]. Toward this goal, prior studies have developed a range of approaches that suppress domain-dependent variation, including explicit distribution alignment [8], adversarial learning for representation alignment [16, 17], and objectives that encourage a domain-invariant predictor across environments [10, 18, 19].

This paradigm is also supported by theoretical analyses that bound unseen target risk using source domain risks and domain discrepancy measures or related divergences [8, 20–22]. While these results offer useful insight into predictor-level generalization under domain shift, they provide limited understanding of whether invariant representations retain sufficient task-relevant information under potential information loss induced by invariance constraints [12, 13]. Recent theoretical studies have

further shown that predictor-level invariance alone may fail to recover the optimal invariant predictor or even the intended invariance across environments [23, 24]. These limitations have motivated direct representation-level approaches, such as invariant-feature subspace recovery [25]. Building on this perspective, we argue that explicitly structuring representations through decomposition provides a principled way to control how information is distributed.

Domain Separation Architectures A complementary direction in invariant feature learning is to decompose representations into domain-invariant and domain-specific parts, thereby encouraging a clearer partition of information. Domain Separation Networks introduce a parallel encoder architecture that employs distinct encoders to learn separate latent components [14]. This architecture has since inspired a range of disentanglement-based extensions that incorporate adversarial alignment, variational factorization, or information-theoretic regularization, and has been applied to diverse tasks, including visual recognition, time series forecasting, and anomaly detection [15, 26–31].

While empirically effective and providing a structured decomposition of latent representations, existing approaches offer limited theoretical analysis. In particular, it remains unclear how information is distributed across latent components and whether each component remains sufficiently informative. To address this gap, we formalize reconstruction-level risk in the context of domain-invariant and domain-specific representation learning. Under this perspective, we clarify how latent decomposition supports domain generalization by analyzing representation informativeness. This analysis also motivates a new architecture based on reconstruction risk, designed to better structure domain-invariant and domain-specific latent components.

3 Preliminaries

Domain Generalization We consider the supervised domain generalization setting. Let $\mathcal{D}$ denote the set of all possible domains, where each domain $d \in \mathcal{D}$ is associated with a marginal distribution P_d over the input space $\mathcal{X}$, and $\mathcal{Y}$ denote the label space. We assume that the conditional distribution $P(Y|X)$ is shared across domains, so that the joint distribution on domain d factorizes as

$$P_d(X, Y) = P_d(X)P(Y|X), \tag{6}$$

with $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. We denote by X and Y the random variables, and by $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ their realizations. During training, labeled samples are observed from a finite set of S source domains $\mathcal{D}_s \coloneqq \{d_1, \ldots, d_S\} \subset \mathcal{D}$. At test time, the model is evaluated on a target domain $d_t \in \mathcal{D} \setminus \mathcal{D}_s$, whose distribution P_{d_t} is not observed during training.

Following the standard representation learning paradigm in domain generalization, the predictor is parameterized via an intermediate latent representation. Specifically, an encoder $f\colon \mathcal{X} \to \mathcal{Z}$ maps the input to a latent space $\mathcal{Z}$, and predictions are made by a prediction head $h\colon \mathcal{Z} \to \mathcal{Y}$. Accordingly, the prediction risk of a domain-agnostic predictor on domain d is given by

$$\mathcal{R}_{\text{pred}}(d) \coloneqq \mathbb{E}_{X\sim P_d}[\mathbb{E}_{Y\sim P(Y|X)}\ell(h(f(X)), Y)], \tag{7}$$

where $\ell(\cdot, \cdot)$ denotes a task-relevant loss function.

Latent Decomposition Building on the general latent representation $z \in \mathcal{Z}$ introduced above, we now introduce a structured decomposition on the latent space. Specifically, we assume that the latent space factorizes as $\mathcal{Z} = \mathcal{Z}_{\text{inv}} \times \mathcal{Z}_{\text{spec}}$, such that each latent representation can be written as $z = (z_{\text{inv}}, z_{\text{spec}})$, where $z_{\text{inv}} \in \mathcal{Z}_{\text{inv}}$ captures domain-invariant factors and $z_{\text{spec}} \in \mathcal{Z}_{\text{spec}}$ captures domain-specific factors.

To realize this decomposition, we further specify a parameterized encoder f_ϕ with a hierarchical structure. In particular, the encoder consists of an invariant mapping $f_{\text{inv}}\colon \mathcal{X} \to \mathcal{Z}_{\text{inv}}$ and a conditional specific mapping $f_{\text{spec}}\colon \mathcal{X} \times \mathcal{Z}_{\text{inv}} \to \mathcal{Z}_{\text{spec}}$, as in (5). The prediction is given by $\hat{y} = h_\psi(z_{\text{inv}})$, where $h_\psi\colon \mathcal{Z}_{\text{inv}} \to \mathcal{Y}$ is a predictor that operates only on the invariant component. This enforces that task-relevant information required for prediction is encoded in an invariant representation. In addition, to encourage information preservation across the decomposed latent components, we employ a reconstruction-based objective implemented through a decoder $g_\theta\colon \mathcal{Z}_{\text{inv}} \times \mathcal{Z}_{\text{spec}} \to \mathcal{X}$, which reconstructs the input as $\hat{x} = g_\theta(z_{\text{inv}}, z_{\text{spec}})$.

4 Proposed Method

4.1 Domain-General Reconstruction Bound

We first present a reconstruction-level generalization bound under the assumption that the information in the input can be decomposed into domain-invariant and domain-specific components. For clarity, we provide a simplified instruction that highlights the essential structure of the bound. Preliminaries are provided in Appendix A. Detailed formulations, including Assumption 4.1, and the full proofs of Lemma 4.1 and Theorem 4.1 are deferred to Appendix B.

Definition 4.1 (Reconstruction risk)**.** *The reconstruction risk on domain d is defined as*

$$\mathcal{R}(d) \coloneqq \mathbb{E}_{X\sim P_d}[\ell(X,\hat{X})], \quad \hat{X} \coloneqq g_\theta(f_\phi(X)), \tag{8}$$

where $\ell(\cdot,\cdot)$ is a nonnegative loss that quantifies the discrepancy between the input random variable X and its reconstruction $\hat{X}$.

Lemma 4.1 (Reconstruction risk decomposition via source convex hull)**.** *Let Λ denote the convex hull of the source domain distributions, and let $\boldsymbol{\alpha}^*$ be a selected vector of source mixture coefficients that defines the auxiliary mixture distribution $P_{d_a} = \sum_{j=1}^{S} \alpha_j^* P_{d_j} \in \Lambda$. Let $\gamma_a^{\mathrm{inv}} \in \Pi(P_{d_t}, P_{d_a})$ denote the input space coupling obtained by lifting a Wasserstein-optimal coupling in the invariant latent space. Suppose that the reconstruction loss $\ell(x,\hat{x})$ is 1-Lipschitz continuous in each argument. Then the target reconstruction risk satisfies*

$$\mathcal{R}(d_t) \le \sum_{j=1}^{S} \alpha_j^* \mathcal{R}(d_j) + \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|g(z^{(d_t)}) - g(z^{(d_a)})\| + \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\|, \tag{9}$$

where $z^{(d_t)} \coloneqq f(x^{(d_t)})$ and $z^{(d_a)} \coloneqq f(x^{(d_a)})$. Here, d_a denotes the auxiliary domain associated with the mixture distribution P_{d_a}.

Lemma 4.1 leverages the convex hull of source domain distributions to convert the discrepancy between the target domain and individual source domains into that between the target domain and a source mixture. Under this reformulation, the target reconstruction risk decomposes into a mixture of source risks, a latent-induced discrepancy, and an input-space mismatch, where the latent-induced discrepancy can be further decomposed under the latent decomposition assumptions.

Assumption 4.1 (Regularity conditions)**.** *We assume the following regularity conditions.*

(i) *The invariant representation satisfies* $\mathrm{HSIC}(d, z_{\mathrm{inv}}) \le \varepsilon$*, where* $d \sim \mathrm{Unif}(\mathcal{D}_s)$.

(ii) *The reconstruction loss $\ell(x,\hat{x})$ is 1-Lipschitz continuous in each argument, and the decoder g is Lipschitz continuous in z_{inv} and z_{spec}, with constants L_{inv} and L_{spec}.*

(iii) *The specific encoder $f_{\mathrm{spec}}(x, z_{\mathrm{inv}})$ is $L_{s,x}$-Lipschitz in x and $L_{s,z}$-Lipschitz in z_{inv}.*

(iv) *We use the delta kernel for source domain labels and assume that the invariant latent space satisfies the compact-support, Matérn-kernel, and Wasserstein-MMD conditions in Assumption B.6 of Appendix B.*

Theorem 4.1 (Domain-general reconstruction bound)**.** *Under Assumption 4.1, the target reconstruction risk is bounded by*

$$\begin{aligned}\mathcal{R}(d_t) \le \underbrace{\sum_{j=1}^{S} \alpha_j^* \mathcal{R}(d_j)}_{\text{(A) Source Risk}} + \underbrace{\tilde{L} C_{\mathrm{ipm}} \|\mu_{d_t} - \mu\|_{\mathcal{H}_z}^{\delta}}_{\text{(B) Invariant Latent Gap}} + \underbrace{\tilde{L} C_{\mathrm{ipm}} \cdot S^{\delta}\, \mathrm{HSIC}(d, z_{\mathrm{inv}})^{\delta/2}}_{\text{(C) Invariant Leakage}} \\ + \underbrace{(1 + L_{\mathrm{spec}} L_{s,x}) \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\|}_{\text{(D) Input-Space Mismatch}},\end{aligned} \tag{10}$$

where $\tilde{L} \coloneqq L_{\mathrm{inv}} + L_{\mathrm{spec}} L_{s,z}$, $C_{\mathrm{ipm}} > 0$, and $\delta \in (0,1]$. We denote by μ_d the kernel mean embedding of the invariant latent distribution on domain d, with $\mu \coloneqq S^{-1} \sum_{j=1}^{S} \mu_{d_j}$ denoting the global source mean embedding in the RKHS $\mathcal{H}_z$.

Theorem 4.1 shows that the bound on the target reconstruction risk decomposes into two groups of terms. The first group consists of {(B), (D)}, which capture discrepancies between the target and source domains in the invariant latent and input spaces, respectively. These terms reflect cross-domain variation and limited source coverage, and cannot be directly evaluated or minimized without access to the target domain. The second group consists of {(A), (C)}, all of which are directly related to the learned invariant representation and can therefore be reduced through training. In particular, achieving low source reconstruction risk in (A) while reducing the dependence between z_{inv} and d, as reflected in (C), encourages input information preservation while limiting domain leakage into the invariant representation. Consequently, the bound on the target reconstruction risk becomes tighter.

Remark 1 (Implication for prediction with invariant representations)**.** *Similarly, the prediction risk of $h_\psi(z_{\text{inv}})$ can be bounded. Assume that h_ψ is L_h-Lipschitz and $\ell_{\text{task}}(\cdot, y)$ is L_ℓ-Lipschitz and bounded by B. Under Assumption 4.1(i, iv),*

$$\begin{aligned}\mathcal{R}_{\text{pred}}(d_t) \leq \sum_{j=1}^{S} \alpha_j^* \mathcal{R}_{\text{pred}}(d_j) + L_\ell L_h C_{\text{ipm}} \|\mu_{d_t} - \mu\|_{\mathcal{H}_z}^{\delta} \\ + L_\ell L_h C_{\text{ipm}} S^\delta \,\text{HSIC}(d, z_{\text{inv}})^{\delta/2} + B\Lambda_a,\end{aligned} \tag{11}$$

where Λ_a is the expected total variation distance between $P(\cdot|x^{(d_t)})$ and $P(\cdot|x^{(d_a)})$ under γ_a^{inv}. Both bounds motivate minimizing their respective source risks while limiting domain dependence in z_{inv}. The full statement and proof are provided in Appendix B.6.

While the prediction analysis concerns task performance using z_{inv}, the reconstruction analysis examines information preservation across both latent components and explicitly accounts for their conditional structure under domain shift. We therefore focus our main analysis on reconstruction risk to study disentanglement as a mechanism for controlling information allocation, rather than merely enforcing invariance. Together, these analyses motivate the model architecture and training objective introduced in the following section.

4.2 Model Architecture and Training Objectives

We now introduce our proposed hierarchical architecture for latent decomposition, referred to as **B**ifurcated **L**atent **EN**coder for **D**omain **E**mbedding and **R**econstruction (BLENDER), as illustrated in Figure 1. Given an input x, the hierarchical encoder f_ϕ first obtains an embedded representation from a feature extractor and then processes it through stacked domain-invariant and domain-specific blocks. Let z^0 denote the embedded representation. At each hierarchical level $i = 1, \ldots, L$, the invariant block updates the invariant pathway, initialized with $z_{\text{inv}}^0 = z^0$, according to

$$z_{\text{inv}}^i = f_{\text{inv}}^i(z_{\text{inv}}^{i-1}), \tag{12}$$

where f_{inv}^i denotes the transformation associated with the i-th invariant block. This structure ensures that the invariant representation is updated independently of any domain-specific information.

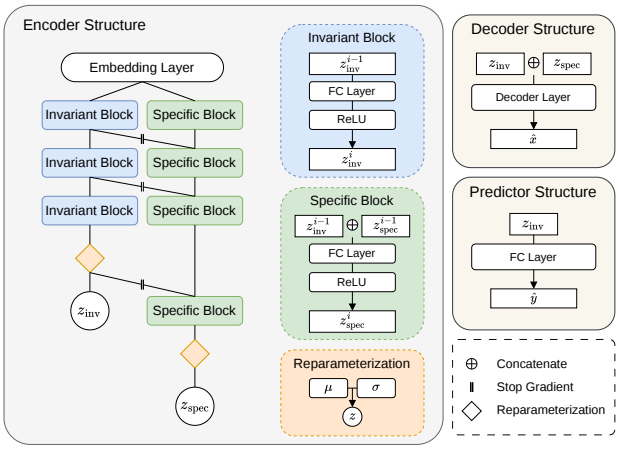

Figure 1: Overall architecture of BLENDER.

The domain-specific pathway is constructed with conditional dependence on the invariant representation. At the first level, the domain-specific representation is initialized as

$$z_{\text{spec}}^1 = f_{\text{spec}}^1(z^0), \tag{13}$$

where f_{spec}^1 takes only the shared embedding z^0 as input. For subsequent levels $i \geq 2$, the domain-specific representation is refined conditional on the invariant pathway using the previous domain-specific representation. Concretely,

$$z_{\text{spec}}^i = f_{\text{spec}}^i(\text{sg}(z_{\text{inv}}^{i-1}), z_{\text{spec}}^{i-1}), \tag{14}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator that blocks gradients flowing into the invariant pathway. This prevents the domain-specific branch from updating the invariant representation. Overall, the hierarchical encoder f_ϕ produces an invariant latent z_{inv} and a specific latent z_{spec} through the two pathways.

Under a reconstruction-aware paradigm, the decoder g_θ reconstructs the input as $\hat{x} = g_\theta(z_{\mathrm{inv}}, z_{\mathrm{spec}})$, where the inputs are concatenated. We minimize the reconstruction loss over all source domains by minimizing the negative evidence lower bound (ELBO). The resulting objective takes the form of a β-VAE [32], in which the reconstruction term is balanced against regularization terms on the domain-invariant and domain-specific latents. We first define the reconstruction term as

$$\mathcal{L}_{\mathrm{recon}} = -\mathbb{E}_{q(z_{\mathrm{inv}}|x)}\left[\mathbb{E}_{q(z_{\mathrm{spec}}|z_{\mathrm{inv}},x)}[\log p(x|z_{\mathrm{inv}}, z_{\mathrm{spec}})]\right], \tag{15}$$

which corresponds to the negative expected log-likelihood. The regularization terms encourage the approximate posteriors to remain close to their respective priors and are given by

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(q(z_{\mathrm{inv}}|x)\|p(z_{\mathrm{inv}})) + \mathbb{E}_{z_{\mathrm{inv}}\sim q(z_{\mathrm{inv}}|x)}[D_{\mathrm{KL}}(q(z_{\mathrm{spec}}|z_{\mathrm{inv}}, x)\|p(z_{\mathrm{spec}}|d))]. \tag{16}$$

The invariant latent is regularized toward a domain-agnostic prior, while the specific latent is regularized toward a domain-conditioned prior. A detailed derivation of the ELBO is provided in Appendix C.

To further enforce domain invariance and disentangle domain-invariant and domain-specific representations, we introduce HSIC-based regularization terms. Specifically, we penalize statistical dependence between the invariant latent and the domain label, as well as between the domain-invariant and domain-specific latents. Formally, we define the HSIC regularization loss as

$$\mathcal{L}_{\mathrm{HSIC}} = \mathrm{HSIC}(d, z_{\mathrm{inv}}) + \mathrm{HSIC}(z_{\mathrm{inv}}, z_{\mathrm{spec}}). \tag{17}$$

The first term encourages the invariant representation to be independent of the domain, directly promoting domain invariance. The second term discourages statistical dependence between the domain-invariant and domain-specific latents, which reduces information leakage and redundancy across the two components. This regularization complements the hierarchical encoder design by enforcing representation separation at the distributional level, thereby reinforcing the assumptions underlying the reconstruction-based generalization bound in Theorem 4.1.

In addition to these invariance objectives, we train a predictor h_ψ that operates exclusively on the invariant representation z_{inv} to produce a prediction $\hat{y} = h_\psi(z_{\mathrm{inv}})$ for the downstream task. By restricting the predictor to z_{inv}, we encourage discriminative information required for prediction to be encoded in a domain-stable representation. Accordingly, we define the task loss as the discrepancy between predicted and ground-truth labels,

$$\mathcal{L}_{\mathrm{task}} = \mathbb{E}_{z_{\mathrm{inv}}\sim q(z_{\mathrm{inv}}|x)}[\ell(h_\psi(z_{\mathrm{inv}}), y)], \tag{18}$$

where $\ell(\cdot,\cdot)$ denotes the task-specific loss function. This loss implicitly pressures the invariant encoder to preserve predictive information while remaining independent of domain-specific variation.

Combining the reconstruction objective, the HSIC-based regularization, and the task objective, the full training objective is given by

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \mathcal{L}_{\mathrm{recon}} + \beta\mathcal{L}_{\mathrm{KL}} + \gamma\mathcal{L}_{\mathrm{HSIC}}, \tag{19}$$

where β and γ are nonnegative hyperparameters that balance information preservation and domain invariance, thereby facilitating improved predictive performance under domain shift.

5 Experimental Results and Analyses

5.1 Experimental Setup

Datasets We conduct quantitative evaluation on widely used domain generalization datasets, including VLCS [33], PACS [34], OfficeHome [35], Terra Incognita [5], and DomainNet [36], which consist of natural images collected from multiple distinct domains. For qualitative evaluation, we additionally employ RotatedMNIST [37], a variant of MNIST in which domain shifts are synthetically induced by rotating digit images at different angles, thereby enabling clearer qualitative analysis of learned representations. As in standard domain generalization evaluation, we adopt a leave-one-domain-out protocol, where one domain is held out as the unseen target domain while the remaining domains are used for training.

Table 1: Domain generalization accuracies (%) under the leave-one-domain-out protocol. For each dataset, we report the mean accuracy averaged over target domains (Avg), and the standard error computed across random seeds. The worst-domain accuracy (Worst) is also reported. The total average summarizes performance across datasets. **Bold** and underlined numbers indicate the best and second-best results in each column, respectively.

Algorithm	VLCS		PACS		OfficeHome		Terra Incognita		DomainNet		Total Average	
	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
ERM [40]	77.5 ± 0.4	64.3 ± 0.9	85.5 ± 0.2	79.3 ± 1.0	66.5 ± 0.3	52.4 ± 0.3	46.1 ± 1.8	35.7 ± 3.9	40.9 ± 0.1	12.2 ± 0.4	63.3	48.8
IRM [10]	78.5 ± 0.5	64.9 ± 0.9	83.5 ± 0.8	76.1 ± 1.0	64.3 ± 2.2	52.2 ± 1.6	47.6 ± 0.8	39.6 ± 0.8	33.9 ± 2.8	10.9 ± 0.5	61.6	48.7
GroupDRO [18]	76.7 ± 0.6	63.4 ± 0.9	84.4 ± 0.8	78.3 ± 2.0	66.0 ± 0.7	52.7 ± 1.0	43.2 ± 1.1	36.4 ± 2.1	33.3 ± 0.2	9.3 ± 0.3	60.7	48.0
Mixup [41]	77.4 ± 0.6	64.8 ± 1.0	84.6 ± 0.6	75.8 ± 1.8	68.1 ± 0.3	54.8 ± 0.6	47.9 ± 0.8	33.9 ± 1.4	39.2 ± 0.1	12.5 ± 0.4	63.4	48.4
MLDG [42]	77.2 ± 0.4	65.2 ± 0.7	84.9 ± 1.0	76.6 ± 1.1	66.8 ± 0.6	53.2 ± 0.6	47.7 ± 0.9	36.9 ± 2.2	41.2 ± 0.1	13.4 ± 0.3	63.6	49.1
CORAL [43]	78.8 ± 0.6	$\mathbf{66.1 \pm 1.2}$	86.2 ± 0.3	78.8 ± 1.3	68.7 ± 0.3	54.4 ± 0.5	47.6 ± 1.0	39.8 ± 2.9	41.5 ± 0.1	13.4 ± 0.4	64.6	**50.5**
MMD [44]	77.5 ± 0.9	64.0 ± 1.1	84.6 ± 0.5	76.5 ± 0.5	66.3 ± 0.1	53.3 ± 0.3	42.2 ± 1.6	35.2 ± 1.8	23.4 ± 9.5	8.7 ± 2.1	58.8	47.5
DANN [16]	78.6 ± 0.4	65.1 ± 1.4	83.6 ± 0.4	73.5 ± 2.3	65.9 ± 0.6	53.0 ± 0.3	46.7 ± 0.5	37.7 ± 1.8	38.3 ± 0.1	11.8 ± 0.1	62.6	48.2
CDANN [17]	77.5 ± 0.1	65.1 ± 1.2	82.6 ± 0.9	73.5 ± 0.6	65.8 ± 1.3	50.4 ± 2.4	45.8 ± 1.6	39.8 ± 2.3	38.3 ± 0.3	12.1 ± 0.7	62.0	48.2
MTL [45]	77.2 ± 0.4	64.3 ± 0.3	84.6 ± 0.5	77.3 ± 1.8	66.4 ± 0.5	52.4 ± 0.6	45.6 ± 1.2	37.8 ± 0.8	40.6 ± 0.1	12.5 ± 0.1	62.9	48.9
SagNet [46]	77.8 ± 0.5	64.5 ± 0.5	$\mathbf{86.3 \pm 0.2}$	80.0 ± 0.4	68.1 ± 0.1	54.8 ± 0.4	48.6 ± 1.0	40.4 ± 1.3	40.3 ± 0.1	12.7 ± 0.5	64.2	**50.5**
ARM [47]	77.6 ± 0.3	63.6 ± 0.7	85.1 ± 0.4	79.3 ± 1.2	64.8 ± 0.3	51.0 ± 0.5	45.5 ± 0.3	38.7 ± 1.3	35.5 ± 0.2	9.4 ± 0.1	61.7	48.4
VREx [19]	78.3 ± 0.2	64.4 ± 1.4	84.9 ± 0.6	77.7 ± 1.7	66.4 ± 0.6	53.0 ± 0.9	46.4 ± 0.6	38.7 ± 3.1	33.6 ± 2.9	10.9 ± 0.3	61.9	48.9
RSC [48]	77.1 ± 0.5	62.5 ± 0.7	85.2 ± 0.9	78.2 ± 1.2	65.5 ± 0.9	51.4 ± 0.3	46.6 ± 1.0	40.8 ± 0.6	38.9 ± 0.5	12.2 ± 0.2	62.7	49.0
DIVA [15]	77.5 ± 0.1	61.3 ± 0.9	68.8 ± 0.5	47.1 ± 0.9	71.3 ± 0.1	52.2 ± 0.2	36.4 ± 0.7	34.9 ± 0.9	35.0 ± 0.0	4.1 ± 0.1	57.8	39.9
DIRT [49]	74.5 ± 0.6	64.5 ± 0.4	82.1 ± 0.6	75.3 ± 1.0	63.2 ± 0.7	41.9 ± 2.5	36.0 ± 1.5	28.9 ± 2.9	45.6 ± 0.1	13.7 ± 0.1	60.3	44.9
LFME [50]	74.8 ± 0.2	64.4 ± 0.5	79.2 ± 0.7	74.2 ± 1.3	61.2 ± 0.4	48.6 ± 1.0	43.5 ± 0.8	32.0 ± 1.5	41.5 ± 0.1	$\mathbf{14.2 \pm 0.2}$	60.0	46.7
Arith [51]	76.1 ± 0.4	60.5 ± 1.2	84.8 ± 0.1	$\mathbf{81.6 \pm 0.4}$	66.4 ± 0.3	52.2 ± 0.6	43.2 ± 1.1	37.8 ± 0.6	42.8 ± 0.1	11.7 ± 0.4	62.6	48.8
BLENDER	$\mathbf{79.4 \pm 0.3}$	64.7 ± 0.7	85.5 ± 0.5	76.9 ± 1.2	$\mathbf{71.6 \pm 0.1}$	$\mathbf{55.1 \pm 0.5}$	$\mathbf{49.8 \pm 0.6}$	$\mathbf{41.5 \pm 0.5}$	$\mathbf{47.5 \pm 0.1}$	13.5 ± 0.2	**66.8**	50.3

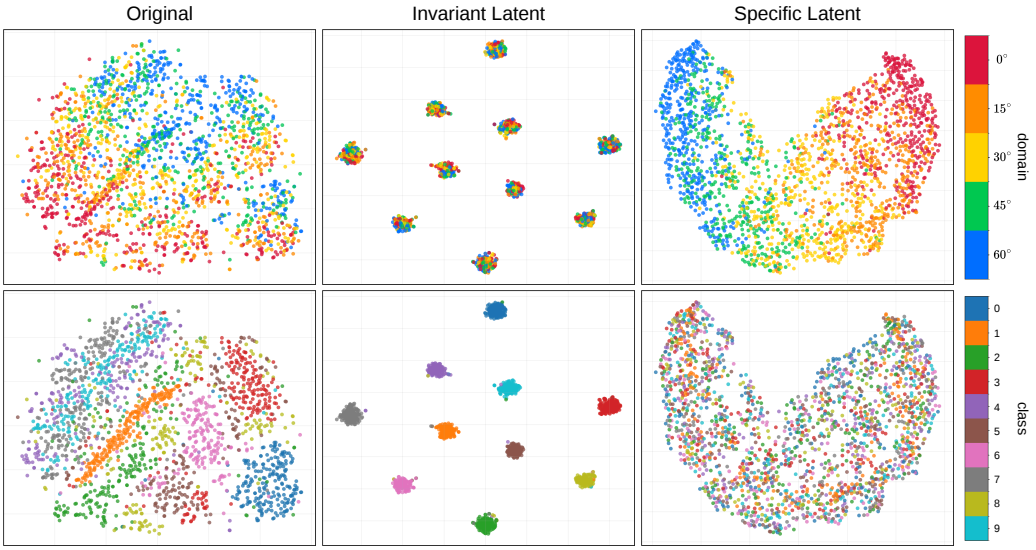

Figure 2: Latent space visualization of BLENDER on RotatedMNIST for the target domain 75°. The original feature space, invariant latent space, and specific latent space are projected to two dimensions using t-SNE. The top row is colored by domain labels and the bottom row is colored by class labels.

Implementation Details We follow the DomainBed evaluation protocol and adopt the training-domain validation set model selection criterion [38]. Model selection is performed exclusively on the training domains by reserving a portion of the aggregated training data for validation. The checkpoint that achieves the highest validation accuracy is selected and subsequently evaluated on the unseen target domain. Our experiments are repeated five times with different random seeds to account for stochastic variability, and we report the mean and standard error of performance across runs. We train all models using Adam [39]. Detailed architectural specifications and hyperparameter configurations are provided in Appendix D.

5.2 Results

Performance on Unseen Target Domains Table 1 reports domain generalization performance under the leave-one-domain-out protocol. BLENDER achieves the highest total average accuracy among the compared methods. In particular, BLENDER achieves higher average accuracy than the

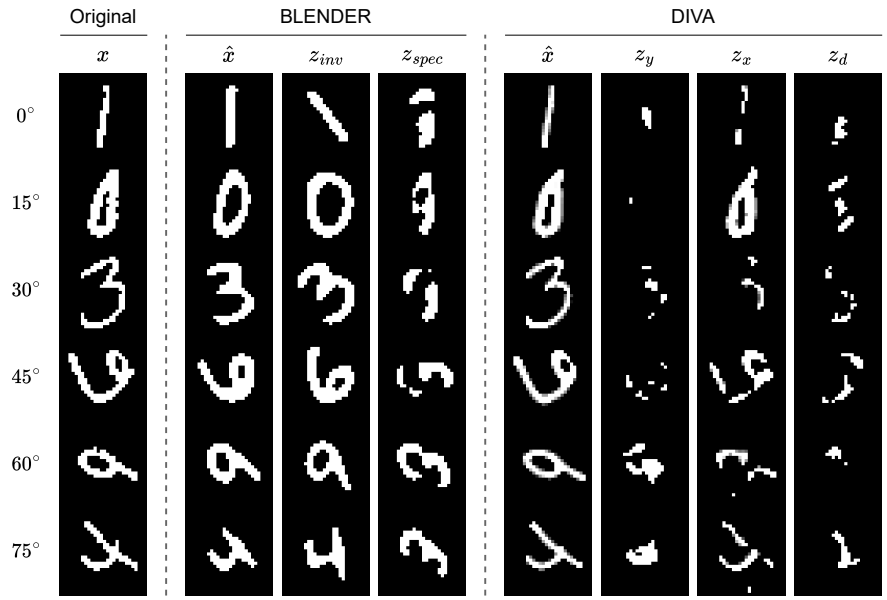

Figure 3: Reconstruction comparison between BLENDER (hierarchical encoder) and DIVA (parallel encoder) on RotatedMNIST. For each rotation angle (rows), the original input x and the full reconstruction $\hat{x}$ are displayed. To analyze the semantic contribution of each latent component, reconstructions from individual latent variables are also shown. BLENDER uses z_{inv} and z_{spec}, whereas DIVA uses z_y, z_x, and z_d. In each case, non-selected latent variables are set to zero vectors.

other methods on OfficeHome, Terra Incognita, and DomainNet, where overall performance tends to be lower. BLENDER also remains competitive in terms of worst-domain accuracy, achieving the highest scores on OfficeHome and Terra Incognita. These results suggest that explicitly modeling domain-specific variation can be beneficial for generalization to unseen domains. Detailed results for individual datasets are provided in Appendix E.

Structured Latent Decomposition Figure 2 illustrates the organized decomposition of the latent space learned by BLENDER on RotatedMNIST. The invariant latent representation concentrates task-relevant information, forming compact clusters aligned with class labels while exhibiting reduced sensitivity to domain variation. In this space, samples sharing the same class tend to group together regardless of their domain origin, indicating that class-discriminative information is preserved independently of domain-specific variation.

In contrast, the specific latent representation primarily encodes variation that is specific to each domain. Therefore, the latent geometry is organized by domain rather than class, indicating that domain shifts are encoded in the domain-specific component. This complementary behavior demonstrates that the model achieves effective disentanglement between domain-invariant and domain-specific factors, consistent with the intended separation enforced during training. Additional qualitative analyses supporting this interpretation are provided in Appendix F.1.

Information Allocation Across Latent Components To evaluate whether each latent captures its intended semantic content, we analyze partial reconstructions obtained by activating a single latent component while fixing all remaining latents to zero. Figure 3 illustrates the reconstruction behavior under a leave-one-domain-out protocol on RotatedMNIST. For each rotation angle designated as the target domain, a model is trained using all remaining domains. Accordingly, each row corresponds to a distinct model trained without access to the shown target domain, thereby reflecting the domain generalization scenario.

For BLENDER, the learned decomposition faithfully reflects the intended semantic structure of the dataset. Reconstructions from z_{inv} consistently preserve digit identity while exhibiting minimal sensitivity to rotational variation, indicating that class-discriminative information is encoded in a

domain-invariant manner. In contrast, reconstructions from z_{spec} primarily reflect rotation-dependent structure while suppressing class semantics, confirming that domain-specific transformations are isolated within the domain-specific latent component. The full reconstruction $\hat{x}$ demonstrates that the invariant representation provided by z_{inv} is systematically transformed according to the rotation information encoded in z_{spec}.

By comparison, DIVA decomposes the latent space into three components where z_y, z_d and z_x are intended to capture label-related, domain-related, and residual variation, respectively. Despite this design, it yields a noticeably less structured information allocation. Partial reconstructions indicate that label-related and domain-related information are not clearly isolated across components. Notably, z_x captures both label and domain cues in a redundant manner, while neither z_y nor z_d alone yields a semantically coherent reconstruction. This suggests that information is distributed in an entangled and overlapping manner across latent variables.

Overall, these observations indicate that the hierarchical encoder combined with HSIC-based disentanglement in BLENDER reduces information redundancy between the two latent representations and promotes a more principled allocation of information. By explicitly encouraging independence between domain-invariant and domain-specific components while preserving task-relevant structure in z_{inv}, the model achieves a more interpretable and structurally coherent factorization of the latent space. An additional experiment on domain-conditioned generation is presented in Appendix F.2.

Relationship between HSIC and Performance Figure 4 illustrates the training trajectory in the HSIC plane for the target domain 75° in RotatedMNIST. The top panel shows the trajectory measured on the training domains, while the bottom panel reports the corresponding validation trajectory. Each point represents a training epoch and is colored by classification accuracy, with the first and last epochs explicitly marked.

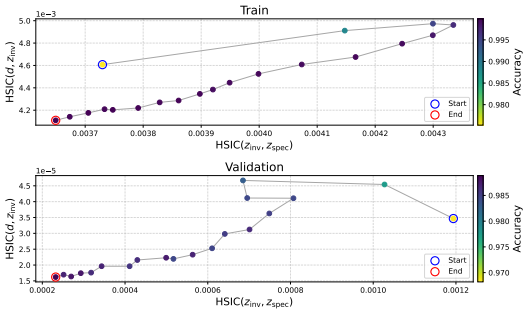

Figure 4: Training trajectory of HSIC on RotatedMNIST for the target domain 75°.

In both panels, the horizontal axis corresponds to $\text{HSIC}(z_{\text{inv}}, z_{\text{spec}})$, measuring dependence between invariant and specific latents, and the vertical axis corresponds to $\text{HSIC}(d, z_{\text{inv}})$, quantifying domain dependence of the invariant representation. As training progresses, the trajectories move toward lower $\text{HSIC}(d, z_{\text{inv}})$, reflecting increasing domain invariance, while the concurrent reduction in $\text{HSIC}(z_{\text{inv}}, z_{\text{spec}})$ indicates that the two representations become progressively disentangled.

On the training set, this reduction is relatively smooth, whereas the validation trajectory exhibits larger fluctuations but still maintains an overall decreasing tendency. Notably, lower values of $\text{HSIC}(d, z_{\text{inv}})$ are associated with higher classification accuracy, supporting our claim that reducing domain dependence in the domain-invariant latent improves cross-domain generalization.

Ablation Study We compare BLENDER with fully connected and parallel encoder variants to assess the benefit of hierarchically organizing domain-invariant and domain-specific representations. We also remove the reconstruction loss, HSIC regularization, and stop-gradient one at a time. Table 2 reports average accuracy across datasets, with five random seeds per configuration.

Table 2: Comparison of model variants.

Variants	Method	Accuracy
Architecture Variants	Fully Connected Encoder	66.5 ± 0.2
	Parallel Encoder	65.9 ± 0.2
Component Ablation	w/o Recon Loss	66.5 ± 0.1
	w/o HSIC Loss	66.7 ± 0.2
	w/o Stop-Gradient	66.5 ± 0.1
	BLENDER (Ours)	66.8 ± 0.1

BLENDER achieves higher average accuracy than the fully connected and parallel encoder variants, with a more pronounced advantage over the parallel encoder. This supports the benefit of hierarchically organizing domain-invariant and domain-specific representations. The component ablations favor the full BLENDER model over variants with individual components removed. This pattern is consistent with a positive contribution from the reconstruction loss, HSIC regularization, and stop-gradient to the complete model.

6 Conclusion

In this work, we introduced BLENDER, a reconstruction-aware framework for domain generalization that employs a hierarchical latent decomposition to explicitly separate domain-invariant and domain-specific factors. We provided a theoretical analysis showing that the reconstruction risk on an unseen target domain can be bounded by a combination of source domain reconstruction risk, latent distribution mismatch, and residual domain dependence in the invariant representation. This result clarifies how representation-level properties, such as domain invariance and latent disentanglement, influence generalization beyond the observed domains.

Empirically, the proposed method demonstrates strong performance across standard domain generalization datasets, particularly on datasets with substantial domain shifts. Qualitative analyses further confirm that the hierarchical encoder combined with HSIC-based regularization promotes a structured allocation of information between domain-invariant and domain-specific latent components. Overall, our results highlight the importance of combining architectural constraints with objectives that preserve information for learning robust representations under domain shift.

Acknowledgments and Disclosure of Funding

This research was supported in part by the National Research Foundation of Korea grant funded by the Korean government (MSIT) (Grant numbers RS-2022-NR068758 and RS-2024-00361377), in part by the Korea Evaluation Institute of Industrial Technology grant funded by the Korean government (MOTIR) (Grant number RS-2025-25458052), and in part by the National IT Industry Promotion Agency grant funded by the Korean government (MSIT) (Grant number RS-2026-25621650).

References

[1] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

[2] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 2021.

[3] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision – ECCV 2010*, pages 213–226, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15561-1.

[4] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021.

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[6] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019. doi: 10.1109/TKDE.2018.2876857.

[7] 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, 2023. doi: 10.1109/TPAMI.2022.3195549.

[8] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 10–18, Atlanta, Georgia, USA, 2013. PMLR.

[9] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

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

[11] Robin Winter, Marco Bertolini, Tuan Le, Frank Noe, and Djork-Arné Clevert. Unsupervised learning of group invariant and equivariant representations. In *Advances in Neural Information Processing Systems*, volume 35, pages 31942–31956. Curran Associates, Inc., 2022.

[12] Fredrik D. Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 527–536. PMLR, 2019.

[13] Thuan Nguyen, Boyang Lyu, Prakash Ishwar, Matthias Scheutz, and Shuchin Aeron. Trade-off between reconstruction loss and feature alignment for domain generalization. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 794–801, 2022. doi: 10.1109/ICMLA55696.2022.00132.

[14] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[15] Maximilian Ilse, Jakub M. Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 322–348. PMLR, 2020.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[17] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[18] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[19] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 2021.

[20] 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(1):151–175, 2010.

[21] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

[22] Yu Ding, Lei Wang, Bin Liang, Shuming Liang, Yang Wang, and Fang Chen. Domain generalization by learning and removing domain-specific features. In *Advances in Neural Information Processing Systems*, volume 35, pages 24226–24239. Curran Associates, Inc., 2022.

[23] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

[24] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 4069–4077. PMLR, 2021.

[25] Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23018–23033. PMLR, 2022.

[26] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 2060–2066. AAAI Press, 2019. ISBN 9780999241141.

[27] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *Computer Vision – ECCV 2020*, pages 205–221, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58542-6.

[28] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 21189–21201. Curran Associates, Inc., 2021.

[29] Yufei Wang, Haoliang Li, Hao Cheng, Bihan Wen, Lap-Pui Chau, and Alex Kot. Variational disentanglement for domain generalization. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[30] Yunchuan Shi, Wei Li, and Albert Y. Zomaya. Domain generalization for time-series forecasting via extended domain-invariant representations. In *2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT)*, pages 110–116, 2024. doi: 10.1109/AIoT63253.2024.00031.

[31] Vincent Jacob and Yanlei Diao. Unsupervised anomaly detection in multivariate time series across heterogeneous domains. *Proc. VLDB Endow.*, 18(6):1691–1704, 2025. ISSN 2150-8097. doi: 10.14778/3725688.3725699.

[32] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[33] 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 International Conference on Computer Vision (ICCV)*, 2013.

[34] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[35] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[36] 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 (ICCV)*, 2019.

[37] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

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

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[40] V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5): 988–999, 1999. doi: 10.1109/72.788640.

[41] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[42] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11596.

[43] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

[44] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021.

[46] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8690–8699, 2021.

[47] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 23664–23678. Curran Associates, Inc., 2021.

[48] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision – ECCV 2020*, pages 124–140, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5.

[49] A. Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. In *Advances in Neural Information Processing Systems*, volume 34, pages 5264–5275. Curran Associates, Inc., 2021.

[50] Liang Chen, Yong Zhang, Yibing Song, Zhiqiang Shen, and Lingqiao Liu. Lfme: A simple framework for learning from multiple experts in domain generalization. In *Advances in Neural Information Processing Systems*, volume 37, pages 102919–102947. Curran Associates, Inc., 2024. doi: 10.52202/079017-3269.

[51] Xiran Wang, Jian Zhang, Lei Qi, and Yinghuan Shi. Balanced direction from multifarious choices: Arithmetic meta-learning for domain generalization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30577–30587, 2025. doi: 10.1109/CVPR52734.2025.02847.

[52] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68: 337–404, 1950. ISSN 0002-9947,1088-6850. doi: 10.2307/1990404.

[53] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2011. ISBN 9781441990969.

[54] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31696-1.

[55] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[56] Titouan Vayer and Rémi Gribonval. Controlling wasserstein distances by kernel norms with application to compressive statistical learning. *Journal of Machine Learning Research*, 24(149):1–51, 2023.

[57] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

A Preliminaries

Reproducing Kernel Hilbert Space (RKHS) Given a symmetric positive definite kernel $\kappa\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a unique RKHS $\mathcal{H}_\kappa$ of functions on $\mathcal{X}$ such that $\kappa(x, \cdot) \in \mathcal{H}_\kappa$ and the reproducing property holds, namely

$$f(x) = \langle f, \kappa(x, \cdot)\rangle_{\mathcal{H}_\kappa}, \quad \forall f \in \mathcal{H}_\kappa, \forall x \in \mathcal{X}. \tag{20}$$

One may use a feature map $\varphi\colon \mathcal{X} \to \mathcal{H}_\kappa$ satisfying $\kappa(x, x') = \langle \varphi(x), \varphi(x')\rangle_{\mathcal{H}_\kappa}$ [52, 53].

Hilbert-Schmidt Independence Criterion (HSIC) Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random variables with joint distribution P_{XY}. Given positive definite kernels κ on $\mathcal{X}$ and l on $\mathcal{Y}$, with RKHSs $\mathcal{H}_\kappa$ and $\mathcal{H}_l$ and feature maps Φ, Ψ, define mean elements $\mu_X = \mathbb{E}[\Phi(X)]$ and $\mu_Y = \mathbb{E}[\Psi(Y)]$. The cross-covariance operator $C_{XY}\colon \mathcal{H}_l \to \mathcal{H}_\kappa$ is

$$C_{XY} \coloneqq \mathbb{E}[(\Phi(X) - \mu_X) \otimes (\Psi(Y) - \mu_Y)]. \tag{21}$$

The Hilbert-Schmidt Independence Criterion is defined as

$$\mathrm{HSIC}(X, Y) \coloneqq \|C_{XY}\|^2_{\mathrm{HS}}, \tag{22}$$

which equals zero if and only if X and Y are independent for characteristic kernels [54, 55]. Given samples $\{(x_i, y_i)\}_{i=1}^n$ with Gram matrices $K_{ij} = \kappa(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$, a common biased estimator is

$$\widehat{\mathrm{HSIC}} = \frac{1}{n^2}\mathrm{tr}(KHLH), \tag{23}$$

where $H = \mathrm{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix [54].

B Proofs

B.1 Definitions

We define the latent representations for each domain $d \in \mathcal{D}$ as

$$z^{(d)}_{\mathrm{inv}} \coloneqq f_{\mathrm{inv}}(x^{(d)}), \quad z^{(d)}_{\mathrm{spec}} \coloneqq f_{\mathrm{spec}}(x^{(d)}, z^{(d)}_{\mathrm{inv}}), \quad z^{(d)} \coloneqq (z^{(d)}_{\mathrm{inv}}, z^{(d)}_{\mathrm{spec}}) = f_\phi(x^{(d)}), \tag{24}$$

where $x^{(d)} \sim P_d$. Throughout the theoretical analysis, the encoder and decoder are treated as fixed deterministic mappings. For notational simplicity, we omit the parameters of f_ϕ and g_θ.

Kernels, RKHSs, and Feature Maps We use two kernels: a domain kernel $\kappa_d\colon \mathcal{D}_s \times \mathcal{D}_s \to \mathbb{R}$ and an invariant latent kernel $\kappa_z\colon \mathcal{Z}_{\mathrm{inv}} \times \mathcal{Z}_{\mathrm{inv}} \to \mathbb{R}$. Let $\mathcal{H}_d \coloneqq \mathcal{H}_{\kappa_d}$ and $\mathcal{H}_z \coloneqq \mathcal{H}_{\kappa_z}$ denote their associated RKHSs. We write $\Psi\colon \mathcal{D}_s \to \mathcal{H}_d$ and $\Phi\colon \mathcal{Z}_{\mathrm{inv}} \to \mathcal{H}_z$ for the corresponding feature maps, so that $\kappa_d(d, d') = \langle \Psi(d), \Psi(d')\rangle_{\mathcal{H}_d}$, $\kappa_z(z, z') = \langle \Phi(z), \Phi(z')\rangle_{\mathcal{H}_z}$. In particular, we take κ_d to be the delta kernel on $\mathcal{D}_s$, $\kappa_d(d, d') = \mathbf{1}\{d = d'\}$, and choose κ_z to be a Matérn kernel on $\mathcal{Z}_{\mathrm{inv}}$.

Definition B.1 (Restatement of Definition 4.1)**.** *The reconstruction risk on domain d is defined as*

$$\mathcal{R}(d) \coloneqq \mathbb{E}_{X \sim P_d}[\ell(X, \hat{X})], \quad \hat{X} \coloneqq g_\theta(f_\phi(X)), \tag{25}$$

where $\ell(\cdot, \cdot)$ is a nonnegative loss that quantifies the discrepancy between the input random variable X and its reconstruction $\hat{X}$.

Definition B.2 (Convex hull of source domain distributions)**.** *Let $\{P_{d_j}\}_{j=1}^S$ denote the input distributions of the source domains. The convex hull of the source domain distributions is defined as*

$$\Lambda \coloneqq \left\{ \sum_{j=1}^S \alpha_j P_{d_j} \;\middle|\; \alpha_j \geq 0,\ \sum_{j=1}^S \alpha_j = 1 \right\}, \tag{26}$$

where the nonnegativity condition applies to all $j = 1, \ldots, S$. For a selected coefficient vector $\boldsymbol{\alpha}^*$, we define the corresponding mixture distribution for an auxiliary domain d_a as

$$P_{d_a} \coloneqq \sum_{j=1}^S \alpha^*_j P_{d_j}. \tag{27}$$

Definition B.3 (Induced latent distributions)**.** *Given the invariant encoder f_{inv}, the induced latent distribution for domain d is defined as*

$$Q_d^{\mathrm{inv}} \coloneqq (f_{\mathrm{inv}})_{\#} P_d, \tag{28}$$

where $(f)_{\#}$ denotes the pushforward operator. For the mixture distribution P_{d_a}, we have

$$Q_{d_a}^{\mathrm{inv}} = (f_{\mathrm{inv}})_{\#} P_{d_a} = \sum_{j=1}^{S} \alpha_j^* Q_{d_j}^{\mathrm{inv}}. \tag{29}$$

Definition B.4 (Invariant-optimal lifted coupling)**.** *For a selected coefficient vector $\boldsymbol{\alpha}^*$, let*

$$\pi_a^{\mathrm{inv}} \in \arg\min_{\pi} \mathbb{E}_{(u,v)\sim\pi} \|u - v\|, \tag{30}$$

where $\pi \in \Pi(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}})$ is a coupling between $Q_{d_t}^{\mathrm{inv}}$ and $Q_{d_a}^{\mathrm{inv}}$. Thus,

$$\mathbb{E}_{(u,v)\sim\pi_a^{\mathrm{inv}}} \|u - v\| = W_1(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}). \tag{31}$$

Let $\eta_{d_t}(dx|u)$ and $\eta_{d_a}(dx'|v)$ denote regular conditional distributions of $X \sim P_{d_t}$ given $f_{\mathrm{inv}}(X) = u$, and of $X' \sim P_{d_a}$ given $f_{\mathrm{inv}}(X') = v$, respectively. We define the lifted input space coupling $\gamma_a^{\mathrm{inv}} \in \Pi(P_{d_t}, P_{d_a})$ by

$$\gamma_a^{\mathrm{inv}}(dx, dx') \coloneqq \int_{\mathcal{Z}_{\mathrm{inv}} \times \mathcal{Z}_{\mathrm{inv}}} \eta_{d_t}(dx|u)\eta_{d_a}(dx'|v)\, \pi_a^{\mathrm{inv}}(du, dv). \tag{32}$$

By construction, if $(x, x') \sim \gamma_a^{\mathrm{inv}}$, then

$$(f_{\mathrm{inv}}(x), f_{\mathrm{inv}}(x')) \sim \pi_a^{\mathrm{inv}}. \tag{33}$$

Accordingly, we use the shorthand notation, $\mathbb{E}_{\gamma_a^{\mathrm{inv}}}[\cdot] \equiv \mathbb{E}_{(x^{(d_t)}, x^{(d_a)})\sim\gamma_a^{\mathrm{inv}}}[\cdot]$.

B.2 Assumptions

The following assumptions provide a more explicit and technical restatement of Assumption 4.1, detailed for the subsequent theoretical analysis.

Assumption B.1 (Domain invariance)**.** *In the ideal case, the domain-invariant latent is independent of the domain,*

$$z_{\mathrm{inv}} \perp d. \tag{34}$$

In practice, we relax this via Assumption B.2.

Assumption B.2 (Controlled domain leakage via HSIC)**.** *The dependence between the domain-invariant latent variable and the source domain index is bounded in terms of HSIC. Specifically, we assume*

$$\mathrm{HSIC}(d, z_{\mathrm{inv}}) \le \varepsilon, \quad d \sim \mathrm{Unif}(\mathcal{D}_s), \quad z_{\mathrm{inv}}|d \sim Q_d^{\mathrm{inv}}. \tag{35}$$

Assumption B.3 (Loss regularity)**.** *The reconstruction loss $\ell(x, \hat{x})$ is 1-Lipschitz continuous in each argument, meaning that*

$$|\ell(x, \hat{x}) - \ell(x', \hat{x})| \le \|x - x'\|, \tag{36}$$

$$|\ell(x, \hat{x}) - \ell(x, \hat{x}')| \le \|\hat{x} - \hat{x}'\|. \tag{37}$$

Assumption B.4 (Decoder regularity)**.** *The decoder g is L_{inv}-Lipschitz in z_{inv} and L_{spec}-Lipschitz in z_{spec}, meaning that*

$$\|g(z_{\mathrm{inv}}, z_{\mathrm{spec}}) - g(z'_{\mathrm{inv}}, z_{\mathrm{spec}})\| \le L_{\mathrm{inv}} \|z_{\mathrm{inv}} - z'_{\mathrm{inv}}\|, \tag{38}$$

$$\|g(z_{\mathrm{inv}}, z_{\mathrm{spec}}) - g(z_{\mathrm{inv}}, z'_{\mathrm{spec}})\| \le L_{\mathrm{spec}} \|z_{\mathrm{spec}} - z'_{\mathrm{spec}}\|. \tag{39}$$

Assumption B.5 (Conditional specific encoder regularity)**.** *The conditional specific encoder $f_{\mathrm{spec}}(x, z_{\mathrm{inv}})$ is $L_{s,x}$-Lipschitz in its input and $L_{s,z}$-Lipschitz in its invariant latent argument as*

$$\|f_{\mathrm{spec}}(x, z) - f_{\mathrm{spec}}(x', z)\| \le L_{s,x} \|x - x'\|, \tag{40}$$

$$\|f_{\mathrm{spec}}(x, z) - f_{\mathrm{spec}}(x, z')\| \le L_{s,z} \|z - z'\|. \tag{41}$$

Assumption B.6 (Latent space kernel and Wasserstein-MMD control)**.** *The domain-invariant latent variable z_{inv} takes values in a compact set $\mathcal{Z}_{\mathrm{inv}} \subset \mathbb{R}^m$. We consider a Matérn kernel κ_z on $\mathcal{Z}_{\mathrm{inv}}$ with smoothness parameter $\nu > m/2$ and lengthscale $l > 0$, and denote by $\mathcal{H}_{\kappa_z}$ the corresponding RKHS. We assume that the associated feature map $\Phi\colon \mathcal{Z}_{\mathrm{inv}} \to \mathcal{H}_{\kappa_z}$ is bounded,*

$$\sup_{z \in \mathcal{Z}_{\mathrm{inv}}} \|\Phi(z)\|_{\mathcal{H}_{\kappa_z}} \leq M < \infty. \tag{42}$$

Under these conditions, we assume that there exist constants $C_{\mathrm{ipm}} > 0$ and $\delta \in (0, 1]$, depending only on $(\mathcal{Z}_{\mathrm{inv}}, \kappa_z)$, such that for any two probability measures P and Q supported on $\mathcal{Z}_{\mathrm{inv}}$,

$$W_1(P, Q) \leq C_{\mathrm{ipm}} \mathrm{MMD}_{\kappa_z}(P, Q)^{\delta}. \tag{43}$$

Remark 2. *For translation-invariant kernels, the control of the Wasserstein-1 distance by the Maximum Mean Discrepancy (MMD) depends in part on the spectral decay of the kernel relative to the dimension of the space [56]. In particular, for Matérn kernels with smoothness parameter $\nu > m/2$, the associated RKHS admits a Sobolev space characterization, ensuring the Wasserstein-MMD control stated in Assumption B.6. The Matérn family includes the Laplacian kernel ($\nu = 1/2$) and approaches the Gaussian kernel ($\nu \to \infty$). Therefore, minimizing the MMD provides a principled and computationally efficient surrogate for controlling the Wasserstein distance in latent space [57].*

B.3 Key Lemmas

Lemma B.1 (Restatement of Lemma 4.1)**.** *Let Λ denote the convex hull of the source domain distributions, and let $\boldsymbol{\alpha}^*$ be a selected vector of source mixture coefficients that defines the auxiliary mixture distribution $P_{d_a} = \sum_{j=1}^{S} \alpha_j^* P_{d_j} \in \Lambda$. Let $\gamma_a^{\mathrm{inv}} \in \Pi(P_{d_t}, P_{d_a})$ denote the input space coupling obtained by lifting a Wasserstein-optimal coupling in the invariant latent space. Suppose that the reconstruction loss $\ell(x, \hat{x})$ is 1-Lipschitz continuous in each argument. Then the target reconstruction risk satisfies*

$$\mathcal{R}(d_t) \leq \sum_{j=1}^{S} \alpha_j^* \mathcal{R}(d_j) + \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|g(z^{(d_t)}) - g(z^{(d_a)})\| + \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\|, \tag{44}$$

where $z^{(d_t)} \coloneqq f(x^{(d_t)})$ and $z^{(d_a)} \coloneqq f(x^{(d_a)})$. Here, d_a denotes the auxiliary domain associated with the mixture distribution P_{d_a}.

Lemma B.2 (Conditional specific contribution)**.** *Under Assumptions B.4 and B.5, the domain-specific contribution satisfies*

$$\begin{aligned} \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|g(z_{\mathrm{inv}}^{(d_t)}, z_{\mathrm{spec}}^{(d_t)}) - g(z_{\mathrm{inv}}^{(d_t)}, z_{\mathrm{spec}}^{(d_a)})\| &\leq L_{\mathrm{spec}} L_{s,x} \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\| \\ &\quad + L_{\mathrm{spec}} L_{s,z} W_1(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}). \end{aligned} \tag{45}$$

Lemma B.3 (Relation between HSIC and geometry of latent means)**.** *Let $p(d) = 1/S$ denote the uniform distribution over the S source domains. Under the introduced kernel and feature-map definitions, for each $d \in \mathcal{D}_s$, define the conditional mean embedding $\mu_d \coloneqq \mathbb{E}_{z_{\mathrm{inv}} | d \sim Q_d^{\mathrm{inv}}}[\Phi(z_{\mathrm{inv}})]$, and the global mean embedding $\mu \coloneqq \mathbb{E}_{d \sim p(d)}[\mu_d] = \frac{1}{S} \sum_{d \in \mathcal{D}_s} \mu_d$. Then, HSIC between the domain index d and the domain-invariant latent variable z_{inv} admits the closed-form expression*

$$\mathrm{HSIC}(d, z_{\mathrm{inv}}) = \frac{1}{S^2} \sum_{d \in \mathcal{D}_s} \|\mu_d - \mu\|_{\mathcal{H}_z}^2. \tag{46}$$

Lemma B.4 (Invariant contribution bound via HSIC)**.** *Under Assumption B.6, there exists a constant $C_{\mathrm{ipm}} > 0$ such that*

$$W_1(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}) \leq C_{\mathrm{ipm}} \Big(\|\mu_{d_t} - \mu\|_{\mathcal{H}_z} + S \sqrt{\mathrm{HSIC}(d, z_{\mathrm{inv}})} \Big)^{\delta}, \tag{47}$$

where $0 < \delta \leq 1$.

B.4 Proof of Theorem 4.1

Theorem B.1 (Restatement of Theorem 4.1)**.** *Under Assumptions B.2–B.6, the target reconstruction risk is bounded by*

$$\begin{aligned}\mathcal{R}(d_t) \le \underbrace{\sum_{j=1}^{S} \alpha_j^* \mathcal{R}(d_j)}_{\text{(A) Source Risk}} + \underbrace{\tilde{L} C_{\mathrm{ipm}} \|\mu_{d_t} - \mu\|_{\mathcal{H}_z}^{\delta}}_{\text{(B) Invariant Latent Gap}} + \underbrace{\tilde{L} C_{\mathrm{ipm}} S^{\delta}\, \mathrm{HSIC}(d, z_{\mathrm{inv}})^{\delta/2}}_{\text{(C) Invariant Leakage}} \\ + \underbrace{(1 + L_{\mathrm{spec}} L_{s,x}) \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\|}_{\text{(D) Input-Space Mismatch}},\end{aligned} \tag{48}$$

where $\tilde{L} \coloneqq L_{\mathrm{inv}} + L_{\mathrm{spec}} L_{s,z}$*,* $C_{\mathrm{ipm}} > 0$*, and* $\delta \in (0, 1]$*.*

Proof. The proof combines Lemmas B.1–B.4.

Step 1. Risk decomposition By Lemma B.1,

$$\mathcal{R}(d_t) \le \sum_{j=1}^{S} \alpha_j^* \mathcal{R}(d_j) + \underbrace{\mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|g(z^{(d_t)}) - g(z^{(d_a)})\|}_{\text{Latent-Space Mismatch}} + \underbrace{\mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\|}_{\text{Input-Space Mismatch}}. \tag{49}$$

Step 2. Latent variable decomposition Using $z = (z_{\mathrm{inv}}, z_{\mathrm{spec}})$ and the triangle inequality,

$$\begin{aligned}\text{Latent-Space Mismatch} \le \underbrace{\mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|g(z_{\mathrm{inv}}^{(d_t)}, z_{\mathrm{spec}}^{(d_t)}) - g(z_{\mathrm{inv}}^{(d_t)}, z_{\mathrm{spec}}^{(d_a)})\|}_{\text{Specific Part}} \\ + \underbrace{\mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|g(z_{\mathrm{inv}}^{(d_t)}, z_{\mathrm{spec}}^{(d_a)}) - g(z_{\mathrm{inv}}^{(d_a)}, z_{\mathrm{spec}}^{(d_a)})\|}_{\text{Invariant Part}}.\end{aligned} \tag{50}$$

Step 3. Bounding the specific part By Lemma B.2,

$$\text{Specific Part} \le L_{\mathrm{spec}} L_{s,x} \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\| + L_{\mathrm{spec}} L_{s,z} W_1(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}). \tag{51}$$

Step 4. Bounding the invariant contribution By Assumption B.4,

$$\begin{aligned}\text{Invariant Part} &\le L_{\mathrm{inv}} \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|z_{\mathrm{inv}}^{(d_t)} - z_{\mathrm{inv}}^{(d_a)}\| &(52)\\ &= L_{\mathrm{inv}} W_1(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}), &(53)\end{aligned}$$

where the equality follows from Definition B.4.

Combining this result with (51) gives

$$\text{Latent-Space Mismatch} \le L_{\mathrm{spec}} L_{s,x} \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\| + \tilde{L} W_1(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}), \tag{54}$$

where $\tilde{L} = L_{\mathrm{inv}} + L_{\mathrm{spec}} L_{s,z}$.

Applying Lemma B.4,

$$\begin{aligned}\text{Latent-Space Mismatch} \le L_{\mathrm{spec}} L_{s,x} \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\| \\ + \tilde{L} C_{\mathrm{ipm}} \Big(\|\mu_{d_t} - \mu\|_{\mathcal{H}_z} + S \sqrt{\mathrm{HSIC}(d, z_{\mathrm{inv}})} \Big)^{\delta}.\end{aligned} \tag{55}$$

Since $0 < \delta \le 1$, $r \mapsto r^{\delta}$ is subadditive on $\mathbb{R}_+$, and hence

$$\begin{aligned}\text{Latent-Space Mismatch} \le L_{\mathrm{spec}} L_{s,x} \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\| \\ + \tilde{L} C_{\mathrm{ipm}} \|\mu_{d_t} - \mu\|_{\mathcal{H}_z}^{\delta} + \tilde{L} C_{\mathrm{ipm}} S^{\delta}\, \mathrm{HSIC}(d, z_{\mathrm{inv}})^{\delta/2}.\end{aligned} \tag{56}$$

Step 5. Final synthesis Substituting (56) into (49) yields

$$\begin{aligned}\mathcal{R}(d_t) \le \sum_{j=1}^{S} \alpha_j^* \mathcal{R}(d_j) + \tilde{L} C_{\mathrm{ipm}} \|\mu_{d_t} - \mu\|_{\mathcal{H}_z}^{\delta} + \tilde{L} C_{\mathrm{ipm}} S^{\delta} \,\mathrm{HSIC}(d, z_{\mathrm{inv}})^{\delta/2} \\ + (1 + L_{\mathrm{spec}} L_{s,x}) \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\|,\end{aligned} \tag{57}$$

which proves the result. □

B.5 Proofs of Lemmas

B.5.1 Lemma B.1

Proof. We aim to bound the target reconstruction risk

$$\mathcal{R}(d_t) = \mathbb{E}_{x^{(d_t)} \sim P_{d_t}} [\ell(x^{(d_t)}, g(z^{(d_t)}))], \quad z^{(d_t)} \coloneqq f(x^{(d_t)}). \tag{58}$$

Let $P_{d_a} = \sum_{j=1}^{S} \alpha_j^* P_{d_j}$ be an auxiliary mixture distribution. Since $\gamma_a^{\mathrm{inv}} \in \Pi(P_{d_t}, P_{d_a})$, its first marginal is P_{d_t} and its second marginal is P_{d_a}. Therefore,

$$\mathcal{R}(d_t) = \mathbb{E}_{\gamma_a^{\mathrm{inv}}} [\ell(x^{(d_t)}, g(z^{(d_t)}))], \tag{59}$$

and similarly,

$$\mathcal{R}(d_a) = \mathbb{E}_{\gamma_a^{\mathrm{inv}}} [\ell(x^{(d_a)}, g(z^{(d_a)}))]. \tag{60}$$

Applying the triangle inequality yields

$$\mathcal{R}(d_t) \le \mathcal{R}(d_a) + |\mathcal{R}(d_t) - \mathcal{R}(d_a)|. \tag{61}$$

By linearity of expectation, the first term satisfies

$$\mathcal{R}(d_a) = \sum_{j=1}^{S} \alpha_j^* \mathcal{R}(d_j), \tag{62}$$

and by using the common coupling γ_a^{inv}, the second term can be written as

$$\begin{aligned}|\mathcal{R}(d_t) - \mathcal{R}(d_a)| &= \left|\mathbb{E}_{\gamma_a^{\mathrm{inv}}} \left[\ell(x^{(d_t)}, g(z^{(d_t)})) - \ell(x^{(d_a)}, g(z^{(d_a)}))\right]\right| \\ &\le \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \left[\left|\ell(x^{(d_t)}, g(z^{(d_t)})) - \ell(x^{(d_a)}, g(z^{(d_a)}))\right|\right].\end{aligned} \tag{63}$$

For each coupled pair $(x^{(d_t)}, x^{(d_a)})$, we insert and subtract the intermediate term as

$$\begin{aligned}&\left|\ell(x^{(d_t)}, g(z^{(d_t)})) - \ell(x^{(d_a)}, g(z^{(d_a)}))\right| \\ &= \left|\ell(x^{(d_t)}, g(z^{(d_t)})) - \ell(x^{(d_t)}, g(z^{(d_a)})) + \ell(x^{(d_t)}, g(z^{(d_a)})) - \ell(x^{(d_a)}, g(z^{(d_a)}))\right| \\ &\le \left|\ell(x^{(d_t)}, g(z^{(d_t)})) - \ell(x^{(d_t)}, g(z^{(d_a)}))\right| + \left|\ell(x^{(d_t)}, g(z^{(d_a)})) - \ell(x^{(d_a)}, g(z^{(d_a)}))\right|.\end{aligned} \tag{64}$$

Using Assumption B.3,

$$\left|\ell(x^{(d_t)}, g(z^{(d_t)})) - \ell(x^{(d_t)}, g(z^{(d_a)}))\right| \le \|g(z^{(d_t)}) - g(z^{(d_a)})\|, \tag{65}$$

and similarly,

$$\left|\ell(x^{(d_t)}, g(z^{(d_a)})) - \ell(x^{(d_a)}, g(z^{(d_a)}))\right| \le \|x^{(d_t)} - x^{(d_a)}\|. \tag{66}$$

Combining the above bounds, we obtain

$$\mathcal{R}(d_t) \le \sum_{j=1}^{S} \alpha_j^* \mathcal{R}(d_j) + \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|g(z^{(d_t)}) - g(z^{(d_a)})\| + \mathbb{E}_{\gamma_a^{\mathrm{inv}}} \|x^{(d_t)} - x^{(d_a)}\|. \tag{67}$$

□

B.5.2 Lemma B.2

Proof. By Assumption B.4,

$$\|g(z_{\text{inv}}^{(d_t)}, z_{\text{spec}}^{(d_t)}) - g(z_{\text{inv}}^{(d_t)}, z_{\text{spec}}^{(d_a)})\| \leq L_{\text{spec}} \|z_{\text{spec}}^{(d_t)} - z_{\text{spec}}^{(d_a)}\|. \tag{68}$$

Using the hierarchical definition $z_{\text{spec}} = f_{\text{spec}}(x, z_{\text{inv}})$, we insert $f_{\text{spec}}(x^{(d_a)}, z_{\text{inv}}^{(d_t)})$ and then apply the triangle inequality.

$$\begin{aligned} \|z_{\text{spec}}^{(d_t)} - z_{\text{spec}}^{(d_a)}\| \leq \|f_{\text{spec}}(x^{(d_t)}, z_{\text{inv}}^{(d_t)}) - f_{\text{spec}}(x^{(d_a)}, z_{\text{inv}}^{(d_t)})\| \\ + \|f_{\text{spec}}(x^{(d_a)}, z_{\text{inv}}^{(d_t)}) - f_{\text{spec}}(x^{(d_a)}, z_{\text{inv}}^{(d_a)})\|. \end{aligned} \tag{69}$$

By Assumption B.5,

$$\|z_{\text{spec}}^{(d_t)} - z_{\text{spec}}^{(d_a)}\| \leq L_{s,x} \|x^{(d_t)} - x^{(d_a)}\| + L_{s,z} \|z_{\text{inv}}^{(d_t)} - z_{\text{inv}}^{(d_a)}\|. \tag{70}$$

Taking expectation under γ_a^{inv} and using Definition B.4,

$$\mathbb{E}_{\gamma_a^{\text{inv}}} \|z_{\text{spec}}^{(d_t)} - z_{\text{spec}}^{(d_a)}\| \leq L_{s,x} \mathbb{E}_{\gamma_a^{\text{inv}}} \|x^{(d_t)} - x^{(d_a)}\| + L_{s,z} W_1(Q_{d_t}^{\text{inv}}, Q_{d_a}^{\text{inv}}). \tag{71}$$

Multiplying by L_{spec} gives the claimed bound. □

B.5.3 Lemma B.3

Proof. We give an explicit expansion of $\text{HSIC}(d, z_{\text{inv}})$ under the delta kernel on the domain space and the uniform prior over source domains. Let $p(d) = 1/S$ be the uniform distribution over $\mathcal{D}_s$ and recall that $z_{\text{inv}}|d \sim Q_d^{\text{inv}}$. Let $\Psi \colon \mathcal{D}_s \to \mathcal{H}_d$ and $\Phi \colon \mathcal{Z}_{\text{inv}} \to \mathcal{H}_z$ be the feature maps of κ_d and κ_z, respectively. For each domain $d \in \mathcal{D}_s$, define the conditional mean embedding

$$\mu_d \coloneqq \mathbb{E}_{z_{\text{inv}}|d \sim Q_d^{\text{inv}}}[\Phi(z_{\text{inv}})] \in \mathcal{H}_z, \tag{72}$$

and define the global mean embeddings

$$\mu_\Psi \coloneqq \mathbb{E}_{d \sim p(d)}[\Psi(d)] \in \mathcal{H}_d, \quad \mu \coloneqq \mathbb{E}_{d \sim p(d)}[\mu_d] = \frac{1}{S} \sum_{d \in \mathcal{D}_s} \mu_d \in \mathcal{H}_z. \tag{73}$$

The cross-covariance operator is defined as

$$C_{dz} \coloneqq \mathbb{E}_{d, z_{\text{inv}}}[(\Psi(d) - \mu_\Psi) \otimes (\Phi(z_{\text{inv}}) - \mu)], \tag{74}$$

and

$$\text{HSIC}(d, z_{\text{inv}}) = \|C_{dz}\|_{\text{HS}}^2. \tag{75}$$

Conditioning on d, we obtain

$$\begin{aligned} C_{dz} &= \mathbb{E}_d\big[\mathbb{E}_{z_{\text{inv}}|d}[(\Psi(d) - \mu_\Psi) \otimes (\Phi(z_{\text{inv}}) - \mu)]\big] && (76) \\ &= \mathbb{E}_d\big[(\Psi(d) - \mu_\Psi) \otimes \mathbb{E}_{z_{\text{inv}}|d}[\Phi(z_{\text{inv}}) - \mu]\big] && (77) \\ &= \mathbb{E}_d[(\Psi(d) - \mu_\Psi) \otimes (\mu_d - \mu)] && (78) \\ &= \sum_{d \in \mathcal{D}_s} p(d)(\Psi(d) - \mu_\Psi) \otimes (\mu_d - \mu). && (79) \end{aligned}$$

Expanding the right-hand side of (79) yields

$$C_{dz} = \sum_d p(d)\Psi(d) \otimes (\mu_d - \mu) - \sum_d p(d)\mu_\Psi \otimes (\mu_d - \mu). \tag{80}$$

The second term vanishes since

$$\sum_d p(d)(\mu_d - \mu) = \Big(\sum_d p(d)\mu_d\Big) - \mu = \mu - \mu = 0. \tag{81}$$

Hence,

$$C_{dz} = \sum_{d \in \mathcal{D}_s} p(d)\Psi(d) \otimes (\mu_d - \mu) = \frac{1}{S} \sum_{d \in \mathcal{D}_s} \Psi(d) \otimes (\mu_d - \mu). \tag{82}$$

Using (82), we have

$$\mathrm{HSIC}(d, z_{\mathrm{inv}}) = \|C_{dz}\|_{\mathrm{HS}}^2 \tag{83}$$

$$= \Big\| \frac{1}{S} \sum_{d \in \mathcal{D}_s} \Psi(d) \otimes (\mu_d - \mu) \Big\|_{\mathrm{HS}}^2 \tag{84}$$

$$= \frac{1}{S^2} \sum_{i \in \mathcal{D}_s} \sum_{j \in \mathcal{D}_s} \langle \Psi(i) \otimes (\mu_i - \mu), \Psi(j) \otimes (\mu_j - \mu) \rangle_{\mathrm{HS}}. \tag{85}$$

For rank-one operators, the Hilbert-Schmidt inner product factorizes as

$$\langle a \otimes b, c \otimes d \rangle_{\mathrm{HS}} = \langle a, c \rangle_{\mathcal{H}_d} \langle b, d \rangle_{\mathcal{H}_z}. \tag{86}$$

Therefore,

$$\mathrm{HSIC}(d, z_{\mathrm{inv}}) = \frac{1}{S^2} \sum_{i \in \mathcal{D}_s} \sum_{j \in \mathcal{D}_s} \langle \Psi(i), \Psi(j) \rangle_{\mathcal{H}_d} \langle \mu_i - \mu, \mu_j - \mu \rangle_{\mathcal{H}_z}. \tag{87}$$

Since κ_d is the delta kernel, the feature maps $\Psi(d)$ are orthogonal in $\mathcal{H}_d$, which implies that

$$\langle \Psi(i), \Psi(j) \rangle_{\mathcal{H}_d} = \kappa_d(i, j) = \mathbf{1}\{i = j\}. \tag{88}$$

Consequently all cross terms with $i \neq j$ vanish. Hence,

$$\mathrm{HSIC}(d, z_{\mathrm{inv}}) = \frac{1}{S^2} \sum_{d \in \mathcal{D}_s} \|\mu_d - \mu\|_{\mathcal{H}_z}^2. \tag{89}$$

□

B.5.4 Lemma B.4

Proof. By Assumption B.6,

$$W_1(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}) \leq C_{\mathrm{ipm}} \mathrm{MMD}_{\kappa_z}(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}})^{\delta}. \tag{90}$$

Let $\mu_Q \coloneqq \mathbb{E}_{z \sim Q}[\Phi(z)]$. By the definition of MMD,

$$\mathrm{MMD}_{\kappa_z}(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}) = \|\mu_{d_t} - \mu_{d_a}\|_{\mathcal{H}_z} \tag{91}$$

$$\leq \|\mu_{d_t} - \mu\|_{\mathcal{H}_z} + \|\mu_{d_a} - \mu\|_{\mathcal{H}_z}. \tag{92}$$

Since $\mu_{d_a} = \sum_{d \in \mathcal{D}_s} \alpha_d^* \mu_d$,

$$\|\mu_{d_a} - \mu\|_{\mathcal{H}_z} = \Big\| \sum_{d \in \mathcal{D}_s} \alpha_d^* (\mu_d - \mu) \Big\|_{\mathcal{H}_z} \tag{93}$$

$$\leq \sum_{d \in \mathcal{D}_s} \alpha_d^* \|\mu_d - \mu\|_{\mathcal{H}_z} \tag{94}$$

$$\leq \sqrt{\sum_{d \in \mathcal{D}_s} (\alpha_d^*)^2} \sqrt{\sum_{d \in \mathcal{D}_s} \|\mu_d - \mu\|_{\mathcal{H}_z}^2} \tag{95}$$

$$\leq \sqrt{\sum_{d \in \mathcal{D}_s} \|\mu_d - \mu\|_{\mathcal{H}_z}^2}. \tag{96}$$

By Lemma B.3,

$$\|\mu_{d_a} - \mu\|_{\mathcal{H}_z} \leq S \sqrt{\mathrm{HSIC}(d, z_{\mathrm{inv}})}. \tag{97}$$

Therefore,

$$\mathrm{MMD}_{\kappa_z}(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}) \leq \|\mu_{d_t} - \mu\|_{\mathcal{H}_z} + S \sqrt{\mathrm{HSIC}(d, z_{\mathrm{inv}})}, \tag{98}$$

and substitution into the Wasserstein-MMD control gives

$$W_1(Q_{d_t}^{\mathrm{inv}}, Q_{d_a}^{\mathrm{inv}}) \leq C_{\mathrm{ipm}} \Big(\|\mu_{d_t} - \mu\|_{\mathcal{H}_z} + S \sqrt{\mathrm{HSIC}(d, z_{\mathrm{inv}})} \Big)^{\delta}. \tag{99}$$

□

B.6 Prediction-Risk Bound

This section provides the proof of the prediction-risk bound stated in Remark 1. For the invariant predictor $h_\psi(z_{\mathrm{inv}})$, define

$$\mathcal{R}_{\mathrm{pred}}(d) \coloneqq \mathbb{E}_{X\sim P_d}[\mathbb{E}_{Y\sim P(\cdot|X)}\ell_{\mathrm{task}}(h_\psi(f_{\mathrm{inv}}(X)),Y)]. \tag{100}$$

For the lifted coupling γ_a^{inv}, define

$$\Lambda_a \coloneqq \mathbb{E}_{\gamma_a^{\mathrm{inv}}}[\mathrm{TV}(P(\cdot|x^{(d_t)}),P(\cdot|x^{(d_a)}))]. \tag{101}$$

Thus, Λ_a measures the expected discrepancy between the conditional label distributions of paired target and auxiliary-source inputs.

Proposition B.1 (Prediction-risk bound)**.** *Suppose h_ψ is L_h-Lipschitz, $\ell_{\mathrm{task}}(\cdot,y)$ is L_ℓ-Lipschitz for every y, and $0\le\ell_{\mathrm{task}}\le B$. Under Assumptions B.2 and B.6,*

$$\begin{aligned}\mathcal{R}_{\mathrm{pred}}(d_t) \le \sum_{j=1}^{S}\alpha_j^*\mathcal{R}_{\mathrm{pred}}(d_j) + L_\ell L_h C_{\mathrm{ipm}}\|\mu_{d_t}-\mu\|_{\mathcal{H}_z}^{\delta} \\ + L_\ell L_h C_{\mathrm{ipm}} S^{\delta}\,\mathrm{HSIC}(d,z_{\mathrm{inv}})^{\delta/2} + B\Lambda_a.\end{aligned} \tag{102}$$

Proof. For notational convenience, define

$$r(x,z) \coloneqq \mathbb{E}_{Y\sim P(\cdot|x)}[\ell_{\mathrm{task}}(h_\psi(z),Y)]. \tag{103}$$

Step 1. Risk decomposition Since $P_{d_a}=\sum_{j=1}^{S}\alpha_j^*P_{d_j}$, linearity of expectation gives

$$\mathcal{R}_{\mathrm{pred}}(d_a) = \sum_{j=1}^{S}\alpha_j^*\mathcal{R}_{\mathrm{pred}}(d_j). \tag{104}$$

Moreover, because γ_a^{inv} has marginals P_{d_t} and P_{d_a},

$$\mathcal{R}_{\mathrm{pred}}(d_t) \le \mathcal{R}_{\mathrm{pred}}(d_a) + \mathbb{E}_{\gamma_a^{\mathrm{inv}}}|r(x^{(d_t)},z_{\mathrm{inv}}^{(d_t)}) - r(x^{(d_a)},z_{\mathrm{inv}}^{(d_a)})|. \tag{105}$$

Step 2. Separating representation and label discrepancies For each coupled pair, we insert $r(x^{(d_t)},z_{\mathrm{inv}}^{(d_a)})$ and then apply the triangle inequality.

$$\begin{aligned}&|r(x^{(d_t)},z_{\mathrm{inv}}^{(d_t)}) - r(x^{(d_a)},z_{\mathrm{inv}}^{(d_a)})| \\ &\le |r(x^{(d_t)},z_{\mathrm{inv}}^{(d_t)}) - r(x^{(d_t)},z_{\mathrm{inv}}^{(d_a)})| + |r(x^{(d_t)},z_{\mathrm{inv}}^{(d_a)}) - r(x^{(d_a)},z_{\mathrm{inv}}^{(d_a)})|.\end{aligned} \tag{106}$$

The first term changes only the invariant representation while keeping the conditional label distribution fixed. Since h_ψ is L_h-Lipschitz and $\ell_{\mathrm{task}}(\cdot,y)$ is L_ℓ-Lipschitz,

$$|r(x^{(d_t)},z_{\mathrm{inv}}^{(d_t)}) - r(x^{(d_t)},z_{\mathrm{inv}}^{(d_a)})| \le L_\ell L_h\|z_{\mathrm{inv}}^{(d_t)} - z_{\mathrm{inv}}^{(d_a)}\|. \tag{107}$$

The second term keeps the prediction fixed at $h_\psi(z_{\mathrm{inv}}^{(d_a)})$ and changes only the conditional label distribution. Since $0\le\ell_{\mathrm{task}}\le B$,

$$|r(x^{(d_t)},z_{\mathrm{inv}}^{(d_a)}) - r(x^{(d_a)},z_{\mathrm{inv}}^{(d_a)})| \le B\ \mathrm{TV}(P(\cdot|x^{(d_t)}),P(\cdot|x^{(d_a)})). \tag{108}$$

Combining these two bounds yields

$$\begin{aligned}|r(x^{(d_t)},z_{\mathrm{inv}}^{(d_t)}) - r(x^{(d_a)},z_{\mathrm{inv}}^{(d_a)})| \le L_\ell L_h\|z_{\mathrm{inv}}^{(d_t)} - z_{\mathrm{inv}}^{(d_a)}\| \\ + B\ \mathrm{TV}(P(\cdot|x^{(d_t)}),P(\cdot|x^{(d_a)})).\end{aligned} \tag{109}$$

Step 3. Taking expectation under the lifted coupling Taking expectation with respect to γ_a^{inv}, the first discrepancy becomes

$$\mathbb{E}_{\gamma_a^{\mathrm{inv}}}\|z_{\mathrm{inv}}^{(d_t)} - z_{\mathrm{inv}}^{(d_a)}\| = W_1(Q_{d_t}^{\mathrm{inv}},Q_{d_a}^{\mathrm{inv}}) \tag{110}$$

by Definition B.4, while the second equals Λ_a by definition. Therefore,

$$\mathcal{R}_{\mathrm{pred}}(d_t) \le \sum_{j=1}^{S}\alpha_j^*\mathcal{R}_{\mathrm{pred}}(d_j) + L_\ell L_h W_1(Q_{d_t}^{\mathrm{inv}},Q_{d_a}^{\mathrm{inv}}) + B\Lambda_a. \tag{111}$$

Step 4. Controlling the invariant latent discrepancy By Lemma B.4,

$$W_1(Q_{d_t}^{\text{inv}}, Q_{d_a}^{\text{inv}}) \le C_{\text{ipm}}\Big(\|\mu_{d_t} - \mu\|_{\mathcal{H}_z} + S\sqrt{\text{HSIC}(d, z_{\text{inv}})}\Big)^{\delta}. \tag{112}$$

Since $0 < \delta \le 1$, the map $u \mapsto u^{\delta}$ is subadditive on $\mathbb{R}_+$. Hence,

$$W_1(Q_{d_t}^{\text{inv}}, Q_{d_a}^{\text{inv}}) \le C_{\text{ipm}}\|\mu_{d_t} - \mu\|_{\mathcal{H}_z}^{\delta} + C_{\text{ipm}}S^{\delta}\,\text{HSIC}(d, z_{\text{inv}})^{\delta/2}. \tag{113}$$

Substituting this result into the preceding risk bound gives

$$\begin{aligned}\mathcal{R}_{\text{pred}}(d_t) \le \sum_{j=1}^{S} \alpha_j^* \mathcal{R}_{\text{pred}}(d_j) + L_\ell L_h C_{\text{ipm}}\|\mu_{d_t} - \mu\|_{\mathcal{H}_z}^{\delta} \\ + L_\ell L_h C_{\text{ipm}} S^{\delta}\,\text{HSIC}(d, z_{\text{inv}})^{\delta/2} + B\Lambda_a,\end{aligned} \tag{114}$$

which proves the result. □

C Derivation of the Evidence Lower Bound

We derive the evidence lower bound (ELBO) by introducing a variational posterior $q(z_{\text{inv}}, z_{\text{spec}}|x)$ and applying Jensen's inequality to the marginal log-likelihood $\log p(x|d)$. Using the hierarchical factorization $q(z_{\text{inv}}, z_{\text{spec}}|x) = q(z_{\text{inv}}|x)q(z_{\text{spec}}|z_{\text{inv}}, x)$, the Kullback-Leibler (KL) divergence decomposes into separate regularization terms for the domain-invariant and domain-specific latents, yielding the objective used in Section 4.2.

$$\begin{aligned}\log p(x|d) &= \log \iint p(x, z_{\text{inv}}, z_{\text{spec}}|d)\, dz_{\text{inv}}\, dz_{\text{spec}} \\ &= \log \iint p(z_{\text{inv}})p(z_{\text{spec}}|d)p(x|z_{\text{inv}}, z_{\text{spec}})\, dz_{\text{inv}}\, dz_{\text{spec}} \\ &\ge \iint q(z_{\text{inv}}, z_{\text{spec}}|x) \log\left[\frac{p(z_{\text{inv}})p(z_{\text{spec}}|d)p(x|z_{\text{inv}}, z_{\text{spec}})}{q(z_{\text{inv}}, z_{\text{spec}}|x)}\right] dz_{\text{inv}}\, dz_{\text{spec}} \\ &= \iint q(z_{\text{inv}}, z_{\text{spec}}|x) \log p(x|z_{\text{inv}}, z_{\text{spec}})\, dz_{\text{inv}}\, dz_{\text{spec}} \\ &\quad - \iint q(z_{\text{inv}}, z_{\text{spec}}|x) \log\left[\frac{q(z_{\text{inv}}, z_{\text{spec}}|x)}{p(z_{\text{inv}})p(z_{\text{spec}}|d)}\right] dz_{\text{inv}}\, dz_{\text{spec}} \\ &= \mathbb{E}_{q(z_{\text{inv}}, z_{\text{spec}}|x)}[\log p(x|z_{\text{inv}}, z_{\text{spec}})] - D_{\text{KL}}(q(z_{\text{inv}}, z_{\text{spec}}|x)\|p(z_{\text{inv}})p(z_{\text{spec}}|d))\end{aligned} \tag{115}$$

Therefore, maximizing the ELBO is equivalent to minimizing the negative ELBO, which consists of a reconstruction term and KL regularization terms. In particular, the negative expected log-likelihood corresponds to

$$\mathcal{L}_{\text{recon}} = -\mathbb{E}_{q(z_{\text{inv}}|x)}\big[\mathbb{E}_{q(z_{\text{spec}}|z_{\text{inv}}, x)}[\log p(x|z_{\text{inv}}, z_{\text{spec}})]\big], \tag{116}$$

while the KL regularization corresponds to

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q(z_{\text{inv}}|x)\|p(z_{\text{inv}})) + \mathbb{E}_{z_{\text{inv}}\sim q(z_{\text{inv}}|x)}[D_{\text{KL}}(q(z_{\text{spec}}|z_{\text{inv}}, x)\|p(z_{\text{spec}}|d))]. \tag{117}$$

D Additional Implementation Details

Table 3: BLENDER architecture with CNN and ResNet-50 backbone. MLP* configurations are used for VLCS and PACS, while the other ResNet-50 MLP configurations are used for OfficeHome, Terra Incognita, and DomainNet.

Module	**CNN**	**ResNet-50**
Embedding Layer	CNN: $1 \to 32 \to 64$	ResNet-50 (pretrained, fully fine-tuned)
Invariant Block	MLP: $E \to 256 \to 128 \to 128 \to 64$	MLP: $E \to 1024 \to 512 \to 256 \to 128$ MLP*: $E \to 512$
Specific Block	MLP: $E \to 256 \to 128 \to 128 \to 64$	MLP: $E \to 1024 \to 512 \to 256 \to 128$ MLP*: $E \to 512$
Latent Dim	$z_{\text{inv}} = 64, z_{\text{spec}} = 64$	$z_{\text{inv}} = 128, z_{\text{spec}} = 128$
Predict $(\mu_{\text{inv}}, \log \sigma^2_{\text{inv}})$	Individual Linear Head	Individual Linear Head
Predict $(\mu_{\text{spec}}, \log \sigma^2_{\text{spec}})$	Individual Linear Head	Individual Linear Head
Decoder	MLP: $[z_{\text{inv}}; z_{\text{spec}}] \to 128 \to 256 \to 512 \to E$ Reverse CNN: $64 \to 32 \to 1$	MLP: $[z_{\text{inv}}; z_{\text{spec}}] \to 256 \to 512 \to 1024 \to E$ MLP*: $[z_{\text{inv}}; z_{\text{spec}}] \to 512 \to E$
Reconstruction Space	Image space	Embedding space
Classifier	Linear: $z_{\text{inv}} \to y$	Linear: $z_{\text{inv}} \to y$
Predict Domain-Specific Distribution	MLP / Linear Head: $d \to (\mu_{\text{spec}}, \log \sigma^2_{\text{spec}})$	MLP / Linear Head: $d \to (\mu_{\text{spec}}, \log \sigma^2_{\text{spec}})$

We adopt a unified training configuration across datasets. Model selection follows the training-domain validation protocol, where a fixed 8:2 validation split of the aggregated source domain data is used. All models are trained for up to 50 epochs with early stopping based on validation performance (patience = 10). Optimization is performed using Adam with a learning rate of 10^{-4}. The batch size is set to 128 for RotatedMNIST and DomainNet, and 32 for VLCS, PACS, OfficeHome, and Terra Incognita to balance computational efficiency and stability. All experiments are conducted on NVIDIA GeForce RTX 5070 Ti, except for DomainNet, which is trained on an NVIDIA RTX A6000. For quantitative evaluation, each of our experiments is repeated five times with random seeds 41 to 45, and we report the mean and standard error across runs.

Architectural choices vary across datasets in the backbone, MLP configuration, and latent dimensionality. RotatedMNIST uses a CNN backbone with 64-dimensional domain-invariant and domain-specific latents, while all natural image datasets adopt a pretrained ResNet-50 [58] with 128-dimensional latents. In Table 3, E denotes the embedding dimension. The label y and domain indicator d are encoded as one-hot vectors.

The objective function combines reconstruction, classification, and HSIC-based independence regularization. For RotatedMNIST, reconstruction is optimized using Binary Cross-Entropy on grayscale pixels, whereas natural image datasets use Mean Squared Error in the embedding space. Classification is trained with Cross-Entropy loss. For HSIC computation, a delta kernel is used for the discrete domain label d, and Gaussian kernels are applied to z_{inv} and z_{spec}, with a fixed bandwidth $\sigma = 1.0$. The regularization coefficients β and γ are linearly increased from 0 to 0.05 during training.

E Domain generalization accuracies per algorithm, dataset, and domain

This section presents detailed domain generalization results for each of the five datasets. For every dataset, we report accuracy on each individual target domain under the leave-one-domain-out protocol. **Bold** and underlined numbers indicate the best and second-best results, respectively. For baseline comparisons, results of most methods are taken from the DomainBed benchmark. Since DIVA, DIRT, LFME, and Arith are not included in DomainBed, we reproduce their results under the same experimental protocol for fair comparison. For all methods, worst-domain accuracy is evaluated on L (VLCS), S (PACS), C (OfficeHome), L46 (Terra Incognita), and quick (DomainNet).

Code is available at `https://github.com/lhk6565/BLENDER`

E.1 VLCS

Table 4: Domain generalization accuracies for each target domain on the VLCS dataset.

Algorithm	C	L	S	V	Avg
ERM	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5
IRM	98.6 ± 0.1	64.9 ± 0.9	73.4 ± 0.6	77.3 ± 0.9	78.5
GroupDRO	97.3 ± 0.3	63.4 ± 0.9	69.5 ± 0.8	76.7 ± 0.7	76.7
Mixup	98.3 ± 0.6	64.8 ± 1.0	72.1 ± 0.5	74.3 ± 0.8	77.4
MLDG	97.4 ± 0.2	$\underline{65.2 \pm 0.7}$	71.0 ± 1.4	75.3 ± 1.0	77.2
CORAL	98.3 ± 0.1	$\mathbf{66.1 \pm 1.2}$	73.4 ± 0.3	77.5 ± 1.2	$\underline{78.8}$
MMD	97.7 ± 0.1	64.0 ± 1.1	72.8 ± 0.2	75.3 ± 3.3	77.5
DANN	$\mathbf{99.0 \pm 0.3}$	65.1 ± 1.4	73.1 ± 0.3	77.2 ± 0.6	78.6
CDANN	97.1 ± 0.3	65.1 ± 1.2	70.7 ± 0.8	77.1 ± 1.5	77.5
MTL	97.8 ± 0.4	64.3 ± 0.3	71.5 ± 0.7	75.3 ± 1.7	77.2
SagNet	97.9 ± 0.4	64.5 ± 0.5	71.4 ± 1.3	77.5 ± 0.5	77.8
ARM	$\underline{98.7 \pm 0.2}$	63.6 ± 0.7	71.3 ± 1.2	76.7 ± 0.6	77.6
VREx	98.4 ± 0.3	64.4 ± 1.4	$\underline{74.1 \pm 0.4}$	76.2 ± 1.3	78.3
RSC	97.9 ± 0.1	62.5 ± 0.7	72.3 ± 1.2	75.6 ± 0.8	77.1
DIVA	98.5 ± 0.3	61.3 ± 0.9	72.4 ± 0.2	$\underline{77.8 \pm 0.6}$	77.5
DIRT	94.3 ± 0.7	64.5 ± 0.4	68.6 ± 2.0	70.8 ± 0.9	74.5
LFME	94.3 ± 0.2	64.4 ± 0.5	70.3 ± 0.8	70.2 ± 0.3	74.8
Arith	98.2 ± 0.2	60.5 ± 1.2	72.8 ± 0.6	73.0 ± 0.5	76.1
BLENDER	97.9 ± 0.3	64.7 ± 0.7	$\mathbf{74.3 \pm 0.1}$	$\mathbf{80.7 \pm 0.6}$	**79.4**

E.2 PACS

Table 5: Domain generalization accuracies for each target domain on the PACS dataset.

Algorithm	A	C	P	S	Avg
ERM	84.7 ± 0.4	$\underline{80.8 \pm 0.6}$	97.2 ± 0.3	79.3 ± 1.0	85.5
IRM	84.8 ± 1.3	76.4 ± 1.1	96.7 ± 0.6	76.1 ± 1.0	83.5
GroupDRO	83.5 ± 0.9	79.1 ± 0.6	96.7 ± 0.3	78.3 ± 2.0	84.4
Mixup	86.1 ± 0.5	78.9 ± 0.8	$\underline{97.6 \pm 0.1}$	75.8 ± 1.8	84.6
MLDG	85.5 ± 1.4	80.1 ± 1.7	97.4 ± 0.3	76.6 ± 1.1	84.9
CORAL	$\mathbf{88.3 \pm 0.2}$	80.0 ± 0.5	97.5 ± 0.3	78.8 ± 1.3	$\underline{86.2}$
MMD	86.1 ± 1.4	79.4 ± 0.9	96.6 ± 0.2	76.5 ± 0.5	84.6
DANN	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6
CDANN	84.6 ± 1.8	75.5 ± 0.9	96.8 ± 0.3	73.5 ± 0.6	82.6
MTL	$\underline{87.5 \pm 0.8}$	77.1 ± 0.5	96.4 ± 0.8	77.3 ± 1.8	84.6
SagNet	87.4 ± 1.0	80.7 ± 0.6	97.1 ± 0.1	$\underline{80.0 \pm 0.4}$	**86.3**
ARM	86.8 ± 0.6	76.8 ± 0.5	97.4 ± 0.3	79.3 ± 1.2	85.1
VREx	86.0 ± 1.6	79.1 ± 0.6	96.9 ± 0.5	77.7 ± 1.7	84.9
RSC	85.4 ± 0.8	79.7 ± 1.8	$\underline{97.6 \pm 0.3}$	78.2 ± 1.2	85.2
DIVA	77.3 ± 0.6	53.3 ± 1.2	97.5 ± 0.2	47.1 ± 0.9	68.8
DIRT	79.8 ± 1.0	78.4 ± 1.2	94.9 ± 0.5	75.3 ± 1.0	82.1
LFME	76.4 ± 1.2	76.2 ± 1.6	90.1 ± 0.9	74.2 ± 1.3	79.2
Arith	80.8 ± 0.4	$\mathbf{82.3 \pm 0.3}$	94.4 ± 0.3	$\mathbf{81.6 \pm 0.4}$	84.8
BLENDER	87.2 ± 0.8	79.9 ± 1.2	$\mathbf{98.1 \pm 0.1}$	76.9 ± 1.2	85.5

E.3 OfficeHome

Table 6: Domain generalization accuracies for each target domain on the OfficeHome dataset.

Algorithm	A	C	P	R	Avg
ERM	61.3 ± 0.7	52.4 ± 0.3	75.8 ± 0.1	76.6 ± 0.3	66.5
IRM	58.9 ± 2.3	52.2 ± 1.6	72.1 ± 2.9	74.0 ± 2.5	64.3
GroupDRO	60.4 ± 0.7	52.7 ± 1.0	75.0 ± 0.7	76.0 ± 0.7	66.0
Mixup	62.4 ± 0.8	$\underline{54.8 \pm 0.6}$	76.9 ± 0.3	78.3 ± 0.2	68.1
MLDG	61.5 ± 0.9	53.2 ± 0.6	75.0 ± 1.2	77.5 ± 0.4	66.8
CORAL	65.3 ± 0.4	54.4 ± 0.5	76.5 ± 0.1	78.4 ± 0.5	68.7
MMD	60.4 ± 0.2	53.3 ± 0.3	74.3 ± 0.1	77.4 ± 0.6	66.3
DANN	59.9 ± 1.3	53.0 ± 0.3	73.6 ± 0.7	76.9 ± 0.5	65.9
CDANN	61.5 ± 1.4	50.4 ± 2.4	74.4 ± 0.9	76.6 ± 0.8	65.8
MTL	61.5 ± 0.7	52.4 ± 0.6	74.9 ± 0.4	76.8 ± 0.4	66.4
SagNet	63.4 ± 0.2	$\underline{54.8 \pm 0.4}$	75.8 ± 0.4	78.3 ± 0.3	68.1
ARM	58.9 ± 0.8	51.0 ± 0.5	74.1 ± 0.1	75.2 ± 0.3	64.8
VREx	60.7 ± 0.9	53.0 ± 0.9	75.3 ± 0.1	76.6 ± 0.5	66.4
RSC	60.7 ± 1.4	51.4 ± 0.3	74.8 ± 1.1	75.1 ± 1.3	65.5
DIVA	$\underline{68.4 \pm 0.3}$	52.2 ± 0.2	$\mathbf{82.1 \pm 0.1}$	$\mathbf{82.3 \pm 0.3}$	$\underline{71.3}$
DIRT	60.5 ± 1.0	41.9 ± 2.5	73.8 ± 0.3	76.6 ± 0.3	63.2
LFME	55.0 ± 0.5	48.6 ± 1.0	69.2 ± 0.4	71.9 ± 0.3	61.2
Arith	60.6 ± 0.6	52.2 ± 0.6	76.1 ± 0.1	76.5 ± 0.4	66.4
BLENDER	$\mathbf{69.0 \pm 0.2}$	$\mathbf{55.1 \pm 0.5}$	$\underline{80.3 \pm 0.4}$	$\underline{82.0 \pm 0.2}$	**71.6**

E.4 Terra Incognita

Table 7: Domain generalization accuracies for each target domain on the Terra Incognita dataset.

Algorithm	L100	L38	L43	L46	Avg
ERM	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
IRM	54.6 ± 1.3	39.8 ± 1.9	56.2 ± 1.8	39.6 ± 0.8	47.6
GroupDRO	41.2 ± 0.7	38.6 ± 2.1	56.7 ± 0.9	36.4 ± 2.1	43.2
Mixup	$\mathbf{59.6 \pm 2.0}$	42.2 ± 1.4	55.9 ± 0.8	33.9 ± 1.4	47.9
MLDG	54.2 ± 3.0	44.3 ± 1.1	55.6 ± 0.3	36.9 ± 2.2	47.7
CORAL	51.6 ± 2.4	42.2 ± 1.0	57.0 ± 1.0	39.8 ± 2.9	47.6
MMD	41.9 ± 3.0	34.8 ± 1.0	57.0 ± 1.9	35.2 ± 1.8	42.2
DANN	51.1 ± 3.5	40.6 ± 0.6	57.4 ± 0.5	37.7 ± 1.8	46.7
CDANN	47.0 ± 1.9	41.3 ± 4.8	54.9 ± 1.7	39.8 ± 2.3	45.8
MTL	49.3 ± 1.2	39.6 ± 6.3	55.6 ± 1.1	37.8 ± 0.8	45.6
SagNet	53.0 ± 2.9	43.0 ± 2.5	$\mathbf{57.9 \pm 0.6}$	40.4 ± 1.3	48.6
ARM	49.3 ± 0.7	38.3 ± 2.4	55.8 ± 0.8	38.7 ± 1.3	45.5
VREx	48.2 ± 4.3	41.7 ± 1.3	56.8 ± 0.8	38.7 ± 3.1	46.4
RSC	50.2 ± 2.2	39.2 ± 1.4	56.3 ± 1.4	40.8 ± 0.6	46.6
DIVA	40.4 ± 0.5	27.8 ± 1.7	42.4 ± 0.6	34.9 ± 0.9	36.4
DIRT	43.2 ± 1.7	29.1 ± 5.1	42.9 ± 2.9	28.9 ± 2.9	36.0
LFME	47.8 ± 1.1	41.0 ± 2.3	53.1 ± 1.3	32.0 ± 1.5	43.5
Arith	50.1 ± 2.2	39.4 ± 1.6	45.3 ± 1.7	37.8 ± 0.6	43.2
BLENDER	52.9 ± 2.4	$\mathbf{47.2 \pm 0.9}$	57.6 ± 0.9	$\mathbf{41.5 \pm 0.5}$	**49.8**

E.5 DomainNet

Table 8: Domain generalization accuracies for each target domain on the DomainNet dataset.

Algorithm	clip	info	paint	quick	real	sketch	Avg
ERM	58.1 ± 0.3	18.8 ± 0.3	46.7 ± 0.3	12.2 ± 0.4	59.6 ± 0.1	49.8 ± 0.4	40.9
IRM	48.5 ± 2.8	15.0 ± 1.5	38.3 ± 4.3	10.9 ± 0.5	48.2 ± 5.2	42.3 ± 3.1	33.9
GroupDRO	47.2 ± 0.5	17.5 ± 0.4	33.8 ± 0.5	9.3 ± 0.3	51.6 ± 0.4	40.1 ± 0.6	33.3
Mixup	55.7 ± 0.3	18.5 ± 0.5	44.3 ± 0.5	12.5 ± 0.4	55.8 ± 0.3	48.2 ± 0.5	39.2
MLDG	59.1 ± 0.2	19.1 ± 0.3	45.8 ± 0.7	13.4 ± 0.3	59.6 ± 0.2	50.2 ± 0.4	41.2
CORAL	59.2 ± 0.1	19.7 ± 0.2	46.6 ± 0.3	13.4 ± 0.4	59.8 ± 0.2	50.1 ± 0.6	41.5
MMD	32.1 ± 13.3	11.0 ± 4.6	26.8 ± 11.3	8.7 ± 2.1	32.7 ± 13.8	28.9 ± 11.9	23.4
DANN	53.1 ± 0.2	18.3 ± 0.1	44.2 ± 0.7	11.8 ± 0.1	55.5 ± 0.4	46.8 ± 0.6	38.3
CDANN	54.6 ± 0.4	17.3 ± 0.1	43.7 ± 0.9	12.1 ± 0.7	56.2 ± 0.4	45.9 ± 0.5	38.3
MTL	57.9 ± 0.5	18.5 ± 0.4	46.0 ± 0.1	12.5 ± 0.1	59.5 ± 0.3	49.2 ± 0.1	40.6
SagNet	57.7 ± 0.3	19.0 ± 0.2	45.3 ± 0.3	12.7 ± 0.5	58.1 ± 0.5	48.8 ± 0.2	40.3
ARM	49.7 ± 0.3	16.3 ± 0.5	40.9 ± 1.1	9.4 ± 0.1	53.4 ± 0.4	43.5 ± 0.4	35.5
VREx	47.3 ± 3.5	16.0 ± 1.5	35.8 ± 4.6	10.9 ± 0.3	49.6 ± 4.9	42.0 ± 3.0	33.6
RSC	55.0 ± 1.2	18.3 ± 0.5	44.4 ± 0.6	12.2 ± 0.2	55.7 ± 0.7	47.8 ± 0.9	38.9
DIVA	46.9 ± 0.1	17.7 ± 0.1	43.0 ± 0.1	4.1 ± 0.1	63.1 ± 0.1	35.4 ± 0.1	35.0
DIRT	67.1 ± 0.1	22.0 ± 0.4	52.6 ± 0.4	13.7 ± 0.1	64.0 ± 0.3	54.3 ± 0.2	45.6
LFME	62.5 ± 0.1	19.7 ± 0.2	47.1 ± 0.2	$\mathbf{14.2 \pm 0.2}$	55.3 ± 0.2	50.4 ± 0.2	41.5
Arith	63.9 ± 0.1	19.5 ± 0.2	47.1 ± 0.3	11.7 ± 0.4	60.9 ± 0.1	53.4 ± 0.1	42.8
BLENDER	$\mathbf{68.6 \pm 0.1}$	$\mathbf{24.2 \pm 0.2}$	$\mathbf{55.5 \pm 0.1}$	13.5 ± 0.2	$\mathbf{67.3 \pm 0.1}$	$\mathbf{56.0 \pm 0.3}$	**47.5**

F Additional Analyses

To provide further insight into the behavior of the learned representations, we conduct additional analyses on RotatedMNIST. These analyses help verify whether the proposed decomposition successfully captures domain-invariant semantics and domain-specific variations as intended.

F.1 Domain-Specific Latent Space Geometry under Domain Exclusion

Figure 5 visualizes the geometry of the learned domain-specific latent representations when the model is trained under a leave-one-domain-out protocol. For each panel, the model is trained while excluding the designated target domain. A clear geometric pattern consistently emerges across all training configurations, where the latent space forms a structured manifold, reflecting the intended domain-specific variation of RotatedMNIST. Notably, when a domain is excluded during training, the corresponding region of the latent space remains largely unoccupied, creating a visible gap between neighboring domains. Despite the absence of training samples from that domain, the surrounding domains maintain their relative ordering along the rotation trajectory. This behavior suggests that the model successfully captures the underlying semantic structure of domain variation rather than merely memorizing discrete domain identities.

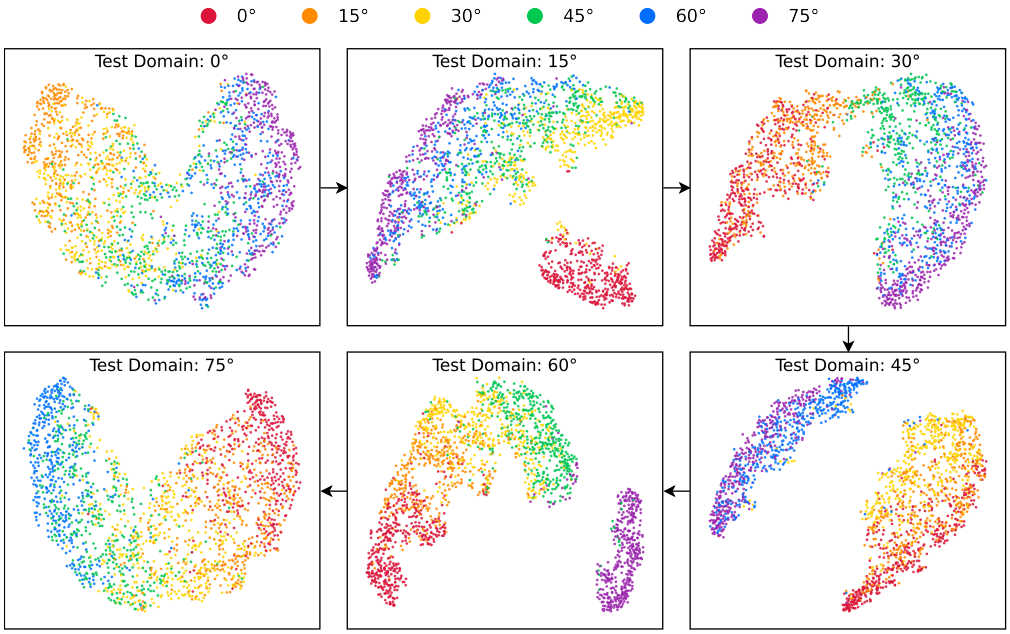

Figure 5: Domain-specific latent space geometry under leave-one-domain-out training. For each panel, the model is trained while excluding the designated target domain. The resulting domain-specific latent representations are projected to two dimensions using t-SNE. Each point corresponds to a sample and is colored according to its domain label.

F.2 Domain-Conditioned Generation from Fixed Invariant Representations

To further examine the functional role of the latent decomposition, we perform a domain-conditioned generation experiment in which the domain-invariant representation is held fixed while the domain-specific latent is varied, as shown in Figure 6. The resulting generations demonstrate that digit identity remains stable across all rows, while systematic changes in rotation angle emerge as the sampled domain-specific latent varies. This indicates that the domain-invariant component indeed preserves label-related information, whereas the domain-specific latent controls domain-dependent transformations.

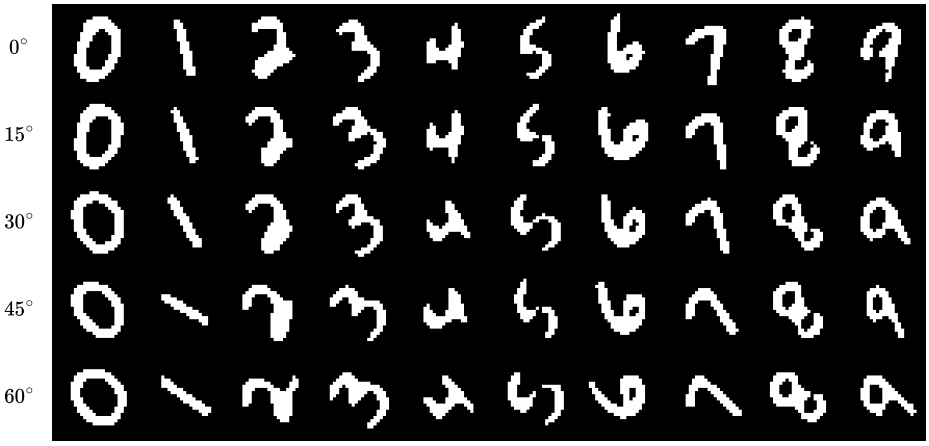

Figure 6: Domain-conditioned generation from a fixed invariant representation. A base image from the domain $0°$ is encoded to obtain its invariant latent representation z_{inv}. For each training domain, a domain-specific latent vector z_{spec} is sampled from the Gaussian distribution parameterized by the learned (μ_d, σ_d). Images are reconstructed from the combined latent pair $(z_{\text{inv}}, z_{\text{spec}})$.

F.3 Sensitivity Analysis

We examine the sensitivity of BLENDER to the maximum values of the scheduling coefficients β and γ. Both coefficients are initialized at zero and increased linearly during training to a shared maximum value. We jointly vary this maximum over $\{0.05, 0.1, 0.2\}$ and evaluate performance on VLCS and PACS datasets. As shown in Table 9, the accuracy averaged across the two datasets ranges from 81.8% to 82.5%, while the corresponding average worst-domain accuracy ranges from 69.7% to 70.9%. These results suggest that aggregate performance remains relatively consistent across the evaluated settings, although worst-domain accuracy exhibits somewhat greater variation.

Table 9: Sensitivity to the shared maximum value of the scheduling coefficients β and γ. Dataset accuracies (%) are reported as mean $\pm$ standard error over five random seeds. Total Accuracy denotes the average across VLCS and PACS datasets.

	VLCS		PACS		Total Accuracy	
$\beta_{\max} = \gamma_{\max}$	Avg	Worst	Avg	Worst	Avg	Worst
0.05	79.4 ± 0.3	64.7 ± 0.7	85.5 ± 0.5	76.9 ± 1.2	82.5	70.8
0.1	78.9 ± 0.2	64.5 ± 0.3	84.7 ± 0.2	74.8 ± 0.7	81.8	69.7
0.2	79.2 ± 0.2	65.6 ± 0.6	85.0 ± 0.4	76.2 ± 1.5	82.1	70.9

G Limitations

This work provides a reconstruction-aware framework for learning informative invariant representations, and its analysis relies on assumptions that the distribution shift across domains is bounded. Accordingly, the derived bound should be interpreted as highlighting key factors affecting target-domain reconstruction risk. In this context, we do not analyze how much the unseen target domain can deviate from the source domains, either empirically or theoretically, and this remains an important direction for future work. Such analysis is also not commonly addressed in existing domain generalization studies. Moreover, the derived error bound includes terms induced by domain-specific latent variation and input-space mismatch that are not explicitly controllable. If these factors could be controlled or characterized, the bound could be further tightened.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are consistent with the theoretical analysis in Section 4 and the experimental results in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All theoretical results are stated with explicit assumptions, and complete proofs are provided in Appendix B and properly referenced in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient details to enable reproduction of the main experimental results, with additional information in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code and instructions for reproducing the main experimental results and use publicly available datasets.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sufficient experimental details are provided, with additional information in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are reported with mean and standard error over multiple runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are provided in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper focuses on methodological contributions and does not explicitly address broader societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve models or data with high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets and code assets are properly cited, and their licenses and terms of use are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is provided with a detailed README.md for its use.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not involve LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.