Superclass-Guided Representation Disentanglement for Spurious Correlation Mitigation

Chenruo Liu[1*] Hongjun Liu[1*] Zeyu Lai[3] Yiqiu Shen[1,2] Chen Zhao[1] Qi Lei[1]

[1]New York University [2]NYU Grossman School of Medicine [3]Zhejiang University

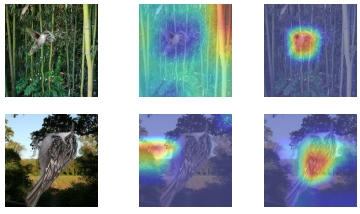
To enhance group robustness to spurious correlations, prior work often relies on auxiliary group annotations and assumes identical sets of groups across training and test domains. To overcome these limitations, we propose to leverage superclasses—categories that lie higher in the semantic hierarchy than the task's actual labels—as a more intrinsic signal than group labels for discerning spurious correlations. Our model incorporates superclass guidance from a pretrained vision-language model via gradient-based attention alignment, and then integrates feature disentanglement with a theoretically supported minimax-optimal feature-usage strategy. As a result, our approach attains robustness to more complex group structures and spurious correlations, without the need to annotate any training samples. Experiments across diverse domain generalization tasks show that our method significantly outperforms strong baselines and goes well beyond the vision-language model's guidance, with clear improvements in both quantitative metrics and qualitative visualizations.

1. Introduction

When the underlying group composition of the training and test distributions differs, certain input features may exhibit strong correlations with the target label during training, yet these correlations often fail to remain stable when evaluated on test data. When training machine learning models, such spurious correlations often lead to significantly degraded domain generalization performance [1, 2].

To improve model robustness across different groups under spurious correlations, many existing methods leverage group information to capture core features for prediction, including upweighting minority groups [3], downsampling majority groups [4], group distributionally robust optimization [3], and progressive data expansion [4]. When group annotations are unknown, another line of works aims to infer latent groups or identify biased samples during training [5–8].

However, these methods for mitigating spurious correlation typically fail or become substantially less effective when (1) group labels are unavailable because obtaining them is costly or even infeasible, or (2) certain test-time groups are absent from the training data, situations in which spurious correlations can become unidentifiable and significantly more severe. Moreover, the sensitivity of these methods to the specification of group information and to changes in the set of groups is further exacerbated by the unreliability of group labels in practice: although commonly used spurious correlation benchmarks deliberately define groups as combinations of $(\text{label}, \text{spurious feature})$ tuples [3, 9, 10], in real-world settings such clean partitions are rare, and the available grouping often fails to faithfully capture the underlying sources of spurious correlation. Consequently, this paper seeks to address the following question:



(a) Original (b) ERM (c) SupER

Figure 1: GradCAM maps of ERM baseline and our SupER approach on Waterbirds dataset. (a) original bird images, (b) ERM's GradCAM maps, and (c) SupER's GradCAM maps. Our approach focuses on core features for classification, while ERM tends to rely on spurious features.

*Equal contribution.

Third Conference on Parsimony and Learning (CPAL 2026).

What serves as a more intrinsic and reliable signal than group information to provide a fundamental criterion for discerning spurious correlations?

We propose that the answer lies in the superclass label, i.e., a label higher in the semantic hierarchy than the task's original class labels [11].

Consider a thought experiment in waterbirds ($Y = 0$) and landbirds ($Y = 1$) image classification, where during training we only observe two groups of bird–background combinations, represented as the tuples $(\text{waterbird}, \text{water background})$ and $(\text{landbird}, \text{land background})$. As shown in experiments with the Waterbirds dataset [3, 6], a model trained using Empirical Risk Minimization (ERM) will tend to make predictions based on backgrounds, which leads to spurious correlations (see Figure 1(b)). However, if we redefine the task as a background classification ($Y = 0$ for water background, $Y = 1$ for land background), all numeric labels remain the same and the group set is also unchanged, yet the background features we focus on are now non-spurious. This indicates that the meaning of the label Y goes far beyond the numeric values $\{0, 1\}$. In particular, unlike group annotations, the superclass label ("bird" versus "background" for these two tasks) captures the essential semantic content of Y by specifying what is being classified, and therefore serve as a key factor that can be exploited by appropriate representation and classifier design to mitigate spurious correlations.

Our Contributions. We propose a novel approach, **Sup**erclass-guided **E**mbedding **R**epresentation (SupER), which uses superclass semantic as a more intrinsic and reliable signal than group information for mitigating spurious correlations. Our contributions are summarized as follows:

- SupER combines feature disentanglement with superclass guidance from a pretrained vision–language model using gradient-based attention alignment, together with principled usage of different superclass-relevant and superclass-irrelevant features. To our knowledge, this is the first group-label-free framework to provide a formal extension and systematic evaluation for addressing group robustness against spurious correlations in settings where certain test-time groups are absent from the training data.
- We provide a theoretical analysis under a reasonable simplified setting, showing that our feature-usage strategy is minimax-optimal under superclass supervision.
- Extensive experiments indicate that SupER significantly outperforms baselines across multiple domain generalization benchmarks, especially when new groups appear at test time or spurious correlations become more severe.
- We further conduct visualizations and quantitative measurements to examine the behavior of SupER, showing that it effectively mitigates biases in the guiding model and achieves substantial improvements over it.

2. Related Work

Group robustness to spurious correlation.[1] To mitigate spurious correlations caused by group imbalance, when group labels are accessible, methods employ strategies such as upweighting minority groups [3], downsampling majority groups [4], group distributionally robust optimization [3], and progressive data expansion [4], with the shared goal of balancing performance across groups. When group information is not available, another line of work attempts to infer group labels or identify biased samples [5–8, 12], or leverage auxiliary information such as knowledge of spurious attributes [13–15]. However, these methods become ineffective when the sets of groups across training and test domains differ, as spurious correlations can not be reliably identified.

Feature learning through disentangled representation. Disentangled representation learning aim to separate independent generative factors of data variation [16]. Building on this principle, various approaches have sought to disentangle representations of X into core and spurious features, and then use only core features for prediction [17–19]. Additionally, sparsity-based methods [20, 21]

[1] Due to space constraints, we defer discussions on additional related work to Appendix C.

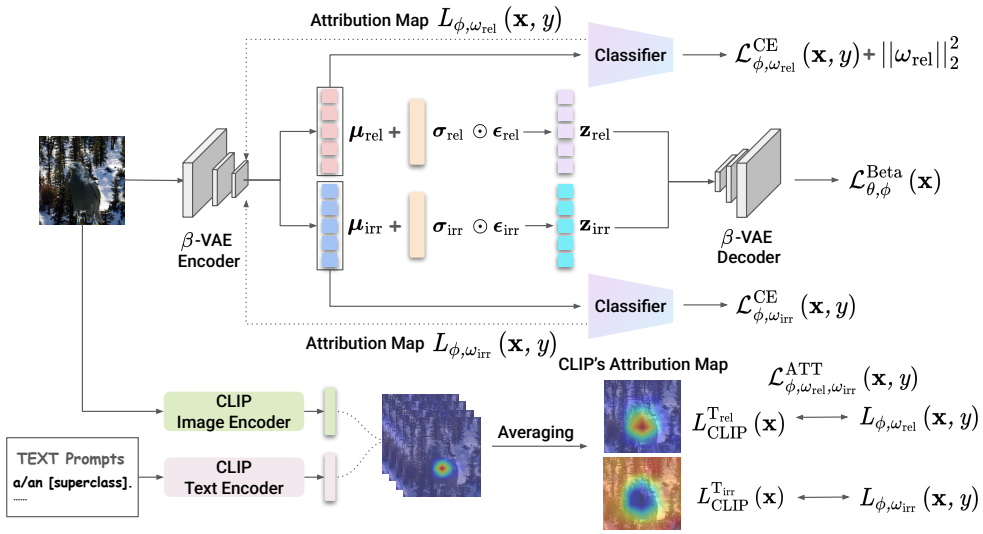


Figure 2: Overview of SupER architecture. The model processes each input image $(\mathbf{x}, y)$ through four key components: (1) A β-VAE architecture that optimizes $\mathcal{L}^{\text{Beta}}_{\theta,\phi}(\mathbf{x})$ to disentangle the input into latent features $\mathbf{z} = [\mathbf{z}_{\text{rel}}; \mathbf{z}_{\text{irr}}]$. (2) Two classifiers, ω_{rel} and ω_{irr}, are trained separately to predict the label y from $\boldsymbol{\mu}_{\text{rel}}$ (mean of $\mathbf{z}_{\text{rel}}$) and $\boldsymbol{\mu}_{\text{irr}}$ (mean of $\mathbf{z}_{\text{irr}}$), by optimizing $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{rel}}}(\mathbf{x}, y)$ and $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{irr}}}(\mathbf{x}, y)$, respectively. (3) A CLIP-guided mechanism that generates attribution maps through text-image alignment, which guides $\mathbf{z}_{\text{rel}}$ and $\mathbf{z}_{\text{irr}}$ to capture superclass-relevant and irrelevant features via $\mathcal{L}^{\text{ATT}}_{\phi,\omega_{\text{rel}},\omega_{\text{irr}}}(\mathbf{x})$. (4) An L_2 regularization term $\|\omega_{\text{rel}}\|_2^2$ that encourages the utilization of all superclass features during classification.

and diverse classifier training [22, 23] have shown effectiveness in feature disentanglement and enhancing generalization. However, these approaches still rely on group or environment annotations, or become less effective when the test domain contains groups that do not appear during training.

3. Method

3.1. Problem Setup

We study a prediction task with inputs $X \in \mathcal{X}$ and labels $Y \in \mathcal{Y}$. The training dataset $\mathcal{D}_s$ (drawn from P_s) and test dataset $\mathcal{D}_t$ (drawn from P_t) consist of groups collected in the sets $\mathcal{G}_s$ and $\mathcal{G}_t$, with each group specified by a label $y \in \mathcal{Y}$ and an attribute $z \in \mathcal{Z}$. When the mixture weights of these groups differ, $P_s \neq P_t$, and z may correlate spuriously with y. We assume that the label set $\mathcal{Y}$ is shared between $\mathcal{D}_s$ and $\mathcal{D}_t$. Our goal is to leverage the superclass label y^{super}, defined as a label higher in the semantic hierarchy than the task's original class labels $\mathcal{Y}$ [11], to learn a predictor on $\mathcal{D}_s$ that maximizes worst group accuracy on $\mathcal{D}_t$. Unlike prior work that often assumes $\mathcal{G}_s = \mathcal{G}_t$, we consider a more general setting: (1) $\mathcal{G}_s$ and $\mathcal{G}_t$ may differ, allowing unseen groups at test time (or equivalently, missing groups during training). In this case, the spurious correlation between z and y in $\mathcal{D}_s$ can become more severe, or z may fail to faithfully capture the underlying sources of spurious correlation. (2) no group information is available during training.

3.2. Feature Disentanglement with Superclass Guidance

Superclass-guided feature disentanglement. For each $(\mathbf{x}, y) \in \mathcal{D}_s$, we use a β-VAE [24] to facilitate the disentanglement of the latent feature representation $\mathbf{z}$ of $\mathbf{x}$ by maximizing

$$\mathcal{L}^{\text{Beta}}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})), \tag{1}$$

where $p_\theta(\mathbf{x}|\mathbf{z})$ is modeled by a decoder, $q_\phi(\mathbf{z}|\mathbf{x})$ approximates the posterior distribution as Gaussian $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x}))$, and the prior $p(\mathbf{z})$ is the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This objective promotes feature disentanglement by encouraging $\mathbf{z}$ to capture independent generative factors of $\mathbf{x}$.

Gradient-based visual attention have been shown to provide visual explanations by highlighting regions that the model attends to during inference [25, 26]. Meanwhile, CLIP [27] possesses strong capability in mapping semantic information from superclass text descriptions into a shared latent space with images. Therefore, to further incorporate superclass guidance from y^{super}, our insight is to leverage CLIP's attention mechanism to guide the partition $\mathbf{z} = [\mathbf{z}_{\text{rel}}; \mathbf{z}_{\text{irr}}]$, such that $\mathbf{z}_{\text{rel}}$ captures superclass-relevant information and $\mathbf{z}_{\text{irr}}$ captures superclass-irrelevant information.

Formally, for any $(\mathbf{x}, y) \in \mathcal{D}_s$ and text prompt $\mathbf{T}$, we compute a normalized gradient-based attribution map $L^{\mathbf{T}}_{\text{CLIP}}(\mathbf{x})$ that reveals the regions CLIP attends to when classifying $\mathbf{x}$ as $\mathbf{T}$ (details in Appendix A.1). Since $\mathbf{z}_{\text{rel}}$ and $\mathbf{z}_{\text{irr}}$ are intended to extract features relevant and irrelevant to the superclass, respectively, the text prompts $\mathbf{T}$ must correspond to these semantic aspects. Specifically, we use m text prompts $\mathbf{T}^1, \ldots, \mathbf{T}^m$ that are semantically aligned with the superclass label y^{super} (e.g., "`a/an [superclass]`"; see Table 31), and average their attribution maps to obtain $L^{\mathbf{T}_{\text{rel}}}_{\text{CLIP}}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} L^{\mathbf{T}^i}_{\text{CLIP}}(\mathbf{x})$, which guides the extraction of superclass-relevant features. For attention guidance of superclass-irrelevant features, given that spurious features are unknown in our setting and previous studies [28] have shown CLIP's limitations in understanding negative prompts (e.g., the meaning of the word "not"), we instead define $L^{\mathbf{T}_{\text{irr}}}_{\text{CLIP}}(\mathbf{x}) = \mathbf{J} - L^{\mathbf{T}_{\text{rel}}}_{\text{CLIP}}(\mathbf{x})$ as the attribution map, where $\mathbf{J}$ represents an all-ones matrix.

To align the CLIP attribution maps with the attribution maps derived from $\mathbf{z}_{\text{rel}}$ and $\mathbf{z}_{\text{irr}}$, we train two different classifiers ω_{rel} and ω_{irr} on $\boldsymbol{\mu}_{\text{rel}}$ and $\boldsymbol{\mu}_{\text{irr}}$ (i.e., the means of $\mathbf{z}_{\text{rel}}$ and $\mathbf{z}_{\text{irr}}$), respectively, by minimizing the cross-entropy losses $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{rel}}}(\mathbf{x}, y)$ and $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{irr}}}(\mathbf{x}, y)$. For each sample $(\mathbf{x}, y)$, we then compute gradient-based attribution maps $L_{\phi,\omega_{\text{rel}}}(\mathbf{x}, y)$ and $L_{\phi,\omega_{\text{irr}}}(\mathbf{x}, y)$ with respect to the true label y (details in Appendix A.2). Finally, superclass-guided feature disentanglement is fulfilled by minimizing the alignment regularization loss:

$$\mathcal{L}^{\text{ATT}}_{\phi,\omega_{\text{rel}},\omega_{\text{irr}}}(\mathbf{x}, y) = \left\| L^{\mathbf{T}_{\text{rel}}}_{\text{CLIP}}(\mathbf{x}) - L_{\phi,\omega_{\text{rel}}}(\mathbf{x}, y) \right\|^2_F + \left\| L^{\mathbf{T}_{\text{irr}}}_{\text{CLIP}}(\mathbf{x}) - L_{\phi,\omega_{\text{irr}}}(\mathbf{x}, y) \right\|^2_F. \tag{2}$$

The combination of $\mathcal{L}^{\text{Beta}}_{\theta,\phi}(\mathbf{x})$ and $\mathcal{L}^{\text{ATT}}_{\phi,\omega_{\text{rel}},\omega_{\text{irr}}}(\mathbf{x}, y)$ successfully achieves the feature disentanglement of $\mathbf{x}$, the separation between $\mathbf{z}_{\text{rel}}$ and $\mathbf{z}_{\text{irr}}$, as well as the gradient-based attention supervision based on superclass semantic information.

Robustness to inherent biases in the guiding model. While CLIP may exhibit inherent biases toward spurious correlations [29, 30] (see Table 30), both superclass-level guidance and feature disentanglement play crucial roles in mitigating such biases. First, since the superclass label y^{super} is shared across all samples $(\mathbf{x}, y) \in \mathcal{D}_s$, it does not provide any discriminative features that contribute to distinguishing specific task labels. This superclass-level guidance therefore avoids CLIP's severe spurious correlations that arise when conditioning on the fine-grained label $y \in \mathcal{Y}$. Second, because the β-VAE encourages independent latent factors with semantic structure [31], once a latent dimension of $\mathbf{z}$ predominantly represents a semantic component under CLIP's guidance, occasional attribution errors from CLIP are effectively overridden by the dominant semantic signal (see Figures 3, 5, and 6). Overall, this weak form of guidance from CLIP significantly reduces the impact of CLIP's own biases while granting SupER sufficient autonomy to learn robust features. Therefore, combined with the theoretically supported minimax-optimal feature-usage strategy presented in Section 3.3, as shown in Table 30, SupER's performance goes far beyond the limitations of the guiding model itself.

3.3. Principled feature-usage with Theoretical Support

Unlike existing approaches that rely on group labels to balance risks across groups [3], or on environment labels to enforce invariant feature learning [10], SupER provides a new feature-usage strategy that leverages the more intrinsic superclass label. Building on the superclass-guided feature disentanglement in Section 3.2, classifier heads ω_{rel} and ω_{irr} respectively make predictions from $\mathbf{z}_{\text{rel}}$ and $\mathbf{z}_{\text{irr}}$ by exploiting those components that are strongly correlated with the label y on $\mathcal{D}_s$. Our objective, therefore, is to determine an appropriate strategy for using different superclass-relevant

Algorithm 1 SupER Model Training

Input: $\mathcal{D}_{\mathrm{s}}$, initial model parameters $\phi, \theta, \omega_{\mathrm{rel}}, \omega_{\mathrm{irr}}$, learning rate η, epochs T, batch size B, $\lambda_1, \lambda_2, \lambda_3$
for epoch $t = 1$ **to** T **do**
  Shuffle $\mathcal{D}_{\mathrm{s}}$ into mini-batches $\{\mathcal{B}\}$ with batch size B
  for each mini-batch $\mathcal{B}$ **do**
    for each sample $(\mathbf{x}, y) \in \mathcal{B}$ **do**
      Compute $\mathcal{L}^{\text{Beta}}_{\theta,\phi}(\mathbf{x})$ according to Equation (1)
      Compute $\mathcal{L}^{\text{ATT}}_{\phi,\omega_{\mathrm{rel}},\omega_{\mathrm{irr}}}(\mathbf{x}, y)$ according to Equation (2)
      Compute cross-entropy losses $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\mathrm{rel}}}(\mathbf{x}, y)$ and $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\mathrm{irr}}}(\mathbf{x}, y)$
    end for
    Compute

$$\mathcal{L}(\mathcal{B}) = \sum_{(\mathbf{x},y)\in\mathcal{B}} \Big(\mathcal{L}^{\text{CE}}_{\phi,\omega_{\mathrm{rel}}}(\mathbf{x}, y) + \mathcal{L}^{\text{CE}}_{\phi,\omega_{\mathrm{irr}}}(\mathbf{x}, y) - \lambda_1 \mathcal{L}^{\text{Beta}}_{\theta,\phi}(\mathbf{x}) + \lambda_2 \mathcal{L}^{\text{ATT}}_{\phi,\omega_{\mathrm{rel}},\omega_{\mathrm{irr}}}(\mathbf{x}, y) + \lambda_3 \|\omega_{\mathrm{rel}}\|_2^2 \Big)$$

    Update parameters: $\phi, \theta, \omega_{\mathrm{rel}}, \omega_{\mathrm{irr}} \leftarrow \phi, \theta, \omega_{\mathrm{rel}}, \omega_{\mathrm{irr}} - \eta \nabla_{\phi,\theta,\omega_{\mathrm{rel}},\omega_{\mathrm{irr}}} \mathcal{L}(\mathcal{B})$
  end for
end for

and superclass-irrelevant features, so that the resulting predictor achieves optimal performance in the worst case in our setting.

For simplicity, we denote by $Z_1, Z_2 \in \mathbb{R}^p$ two distinct latent features of X extracted from q_ϕ, and let $Y \in \mathbb{R}$ be the label. Under Assumption 1 on the linear generative model, we analyze two specific scenarios to provide insight into how ω_{rel} and ω_{irr} should utilize different latent features for prediction.

- (1) The generating index is known to be $c = 1$. In this case, Z_1 is treated as a superclass-relevant feature that generates Y, while Z_2 is treated as a superclass-irrelevant feature. This interpretation follows from the fact that a feature unrelated to the superclass y^{super} cannot serve as a core feature.
- (2) The generating index is uncertain with $c \in \{1, 2\}$. In this case, both Z_1 and Z_2 are viewed as distinct superclass-relevant features. This reflects the general setting in Section 3.1: when both features aligned with the superclass y^{super}, it is unknown in this more general case whether their relationship to the label remains consistent across P_s and P_t.

Further, let $Z_1^s, Z_2^s \in \mathbb{R}^{n\times p}$ denote the fixed matrices obtained by stacking the n training samples of Z_1 and Z_2 row-wise, respectively. Define the empirical second moments and cross-moment under the training distribution P_s by $\hat{\Sigma}_1^s = (Z_1^s)^\top Z_1^s/n$, $\hat{\Sigma}_2^s = (Z_2^s)^\top Z_2^s/n$, and $\hat{\Sigma}_{1,2}^s = (Z_1^s)^\top Z_2^s/n = (\hat{\Sigma}_{2,1}^s)^\top$. For the test distribution P_t, let $\Sigma_1^t = \mathbb{E}_{Z_1\sim P_t}[Z_1 Z_1^\top]$, $\Sigma_2^t = \mathbb{E}_{Z_2\sim P_t}[Z_2 Z_2^\top]$, and $\Sigma_{1,2}^t = \mathbb{E}_{Z_1,Z_2\sim P_t}[Z_1 Z_2^\top] = (\Sigma_{2,1}^t)^\top$. Then, Assumptions 2- 3 formalize the notion of spurious correlation through the strong correlation between features during training (and consequently also between the features and the label), whereas this correlation vanishes under the test distribution.

Assumption 1. *There exists an index $c \in \{1, 2\}$ and $\boldsymbol{\beta}_c \in \mathbb{R}^p$ such that $Y = Z_c^\top \boldsymbol{\beta}_c + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of (Z_1, Z_2).*

Assumption 2. *Under the training distribution P_s, $\hat{\Sigma}_{1,2}^s = \rho\,(\hat{\Sigma}_1^s)^{1/2}(\hat{\Sigma}_2^s)^{1/2}$ with $|\rho| \in (0, 1)$ close to 1, while under the test distribution P_t we have $\Sigma_{1,2}^t = 0$.*

Assumption 3. *$\hat{\Sigma}_1^s, \hat{\Sigma}_2^s, \Sigma_1^t, \Sigma_2^t$ commute with each other, with positive eigenvalues $\{d_{1,i}^s\}_{i=1}^p$, $\{d_{2,i}^s\}_{i=1}^p$, $\{d_{1,i}^t\}_{i=1}^p$, and $\{d_{2,i}^t\}_{i=1}^p$ (the same i refers to the same common eigen-direction).*

Consider three prediction strategies: S_1 predicts with Z_1, S_2 with Z_2, and $S_{1,2}$ with $Z = (Z_1, Z_2)$. Then, Theorem 1 provides an intuitive characterization of the optimal feature-usage strategy. Specifically, among the features that exhibit strong correlations with Y during training, SupER should (i) discard all superclass-irrelevant features (part (1)), and (ii) exploit as diverse a set of superclass-relevant features as possible (part (2)). This selection rationale is intuitive: the superclass label

acts as a semantic prior regarding feature validity, and SupER embodies a principle of elimination under certainty and diversification under uncertainty. Consequently, SupER follows this strategy by excluding the classifier ω_{irr} during evaluation, and by adding an L_2 penalty $\|\omega_{\text{rel}}\|_2^2$ on the classifier head ω_{rel} during training, which encourages smoother and more evenly distributed weights across diverse informative superclass-relevant features [32].

To summarize, SupER effectively leverages the semantic information in the superclass label y^{super} through CLIP-based guidance, and combines superclass-guided feature disentanglement with a theoretically supported form of feature usage. As a result, under the setting of Section 3.1, SupER attains robustness to spurious correlations without relying on group signals. The detailed training algorithm is presented in Algorithm 1, and the complete pipeline is illustrated in Figure 2.

Theorem 1 (Informal result (formally in Theorems 2-3))**.** *Let $\mathcal{E}(S; c, \boldsymbol{\beta}_c)$ denote the excess risk under P_t of the ordinary least-squares predictor trained on P_s, given strategy S and generating index c. Under Assumptions 1-3:*

1. (*Discard superclass-irrelevant features despite strong correlations with Y on P_s.*) *If $c = 1$ is known, then*
$$\min_{S \in \{S_1, S_{1,2}\}} \mathcal{E}(S; 1, \boldsymbol{\beta}_1) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \frac{d_{1,i}^t}{d_{1,i}^s},$$
achieved by $S = S_1$.
2. (*Retain all superclass-relevant features with strong correlations with Y on P_s.*) *If $c \in \{1, 2\}$ is uncertain, let $\mathcal{B}_r^{(c)} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}^\top \Sigma_c^t \boldsymbol{\beta} = r\}$, then there exists a constant C such that whenever $nr > C$,*
$$\min_{S \in \{S_1, S_2, S_{1,2}\}} \max_{c \in \{1,2\}} \max_{\boldsymbol{\beta}_c \in \mathcal{B}_r^{(c)}} \mathcal{E}(S; c, \boldsymbol{\beta}_c) = \frac{\sigma^2}{n(1-\rho^2)} \sum_{i=1}^{p} \left(\frac{d_{1,i}^t}{d_{1,i}^s} + \frac{d_{2,i}^t}{d_{2,i}^s} \right)$$
achieved by $S = S_{1,2}$.

Remark 1. *Unlike prior works on spurious correlations [10, 33–35] that study how the presence of spurious and core features of X affects generalization, our analysis makes a finer partition using superclass information. The part (1) of Theorem 1 shares a similar spirit with prior findings that spurious features can harm generalization, but specifically restricts the scope to superclass-relevant features. More importantly, to our knowledge, the minimax analysis in part (2) of Theorem 1 regarding the use of different superclass-relevant features has not appeared in earlier spurious correlation theory, and this constitutes the main message conveyed by our theorem.*

4. Experiments

In this section, we empirically evaluate SupER on a diverse set of datasets characterized by distinct types of spurious features and underlying group structures. We compare the performance of SupER with different baseline models that make use of group information to different extents. Detailed experimental results are provided in Appendix D.

4.1. Datasets and Baselines

Datasets. We evaluate SupER on datasets that cover diverse spurious correlation structures caused by group imbalance, including **Waterbirds-95%** [3], **Waterbirds-100%** [36], **MetaShift** [37, 38], **Spawrious** [39], and **SpuCo Dogs** [40]. Waterbirds-95% and SpuCo Dogs exhibit a strong correlation ($\sim$95%) between background and label during training. Waterbirds-100% represents an extreme case where two groups are entirely absent during training. MetaShift consists of four subsets, each introducing different degrees of spurious correlation and testing on groups unseen in training. Spawrious contains six subsets and is used to assess performance under two correlation regimes: one-to-one, where each class correlates with a unique attribute, and many-to-many, where multiple classes correlate with multiple attributes. Detailed dataset splits are provided in Appendix D.1. We defer additional evaluation on more datasets to Appendix E due to space constraints.

Baselines and training. Given that SupER does not use any group information during training, our primary comparisons are against baseline methods that do not require group annotations. These

include **ERM**, **CVaR DRO** [41], **LfF** [5], **JTT** [6], **CnC** [7], and **GALS** [36], all of which have been widely adopted in prior work. For completeness, we additionally compare SupER with methods that explicitly leverage group labels to mitigate distribution shifts, such as **GroupDRO** [3], **UW** [3], and **DFR** [42]. We also consider multi-source environment methods such as **IRM** [10]. For SupER, we instantiate both the CLIP and β-VAE components with a ResNet-50 backbone architecture [43]. For each dataset, we train SupER following Algorithm 1. Additional training details, as well as the choices of hyperparameters and superclass labels y^{super}, are provided in Appendix D.2.

4.2. Main Results

4.2.1. Comparison of Accuracy Across Groups

We report the worst group accuracy, average accuracy, and variance of accuracy across groups for SupER and baseline methods. Tables 1-4 present selected results. Comprehensive results are available in Appendix D.3 due to space constraints.

SupER achieves strong performance on worst group accuracy. As shown in Tables 1-3, for most datasets—including Waterbirds-100%, the last five subsets of the Spawrious dataset, and the last three subsets of the MetaShift dataset—SupER's worst group accuracy exceeds that of all selected baseline methods, regardless of whether they require group information. For the remaining datasets, SupER still outperforms the majority of the baselines that do not rely on group labels.

SupER exhibits superior robustness to complex spurious correlations, especially in the presence of unseen groups at test time. Spawrious and Metashift provide chances to investigate model performance under various levels of spurious correlations. Results in Tables 1-3 show that, despite increasing complexity of spurious correlations, the standard deviation of worst group accuracy across the six Spawrious datasets is 2.7%, and across the four MetaShift datasets is 3.7%, significantly lower than other baseline methods. Moreover, Waterbirds-100%, Spawrious, and MetaShift all contain certain test groups that are entirely absent during training, and they induce different types of spurious correlations. SupER performs particularly well on these datasets. For instance, compared to all selected baselines, SupER's worst group accuracy exceeds the best competing baseline by 17.6% on Waterbirds-100%, 11.9% on MetaShift (d), 25.8% on Spawrious M2M-hard, and 7.4% on Spawrious O2O-hard.

SupER shows smaller generalization gaps among groups. Table 4 shows the variance of accuracy across groups for SupER and other baseline methods on selected datasets. SupER exhibits more consistent test accuracy across different groups within the same dataset. This indicates that, under

Table 1: Mean $\pm$ std of worst and average group accuracy (%) for Waterbirds datasets. As a "ceiling" reference with spurious features fully removed, we include results from ERM trained and evaluated on a bird-only region with backgrounds removed. It achieves $85.6_{\pm 0.4}$% worst group accuracy on Waterbirds-95% and $83.0_{\pm 0.5}$% on Waterbirds-100%. Several reported baselines exceed this reference on Waterbirds-95%, suggesting potential benchmark overfitting. **Bold** indicates the best across all baselines; Underlined indicates the best among methods without group information.

Method	Group Info	Train Twice	Waterbirds-95%		Waterbirds-100%	
			Worst	Avg	Worst	Avg
ERM	×	×	$64.9_{\pm 1.5}$	$90.7_{\pm 1.0}$	$46.4_{\pm 6.9}$	$74.8_{\pm 3.0}$
CVaR DRO	×	×	$73.1_{\pm 7.1}$	$90.7_{\pm 0.7}$	$58.0_{\pm 2.2}$	$79.0_{\pm 1.2}$
LfF	×	×	$79.1_{\pm 2.5}$	$\underline{91.9}_{\pm 0.7}$	$61.5_{\pm 2.8}$	$80.6_{\pm 1.2}$
GALS	×	×	$75.4_{\pm 2.2}$	$89.0_{\pm 0.5}$	$55.0_{\pm 5.5}$	$79.7_{\pm 0.4}$
JTT	×	✓	$86.4_{\pm 1.0}$	$89.5_{\pm 0.5}$	$61.3_{\pm 5.5}$	$79.7_{\pm 3.0}$
CnC	×	✓	$\underline{86.5}_{\pm 5.9}$	$91.0_{\pm 0.5}$	$62.1_{\pm 0.9}$	$81.9_{\pm 1.5}$
SupER (Ours)	×	×	$84.4_{\pm 2.3}$	$87.3_{\pm 0.6}$	$\underline{\mathbf{79.7}}_{\pm 1.7}$	$\underline{\mathbf{85.0}}_{\pm 1.4}$
UW	✓	×	$89.3_{\pm 1.5}$	$\mathbf{94.5}_{\pm 0.9}$	$56.4_{\pm 2.3}$	$78.6_{\pm 0.8}$
IRM	✓	×	$76.2_{\pm 6.3}$	$89.4_{\pm 0.9}$	$57.0_{\pm 5.4}$	$80.5_{\pm 5.0}$
GroupDRO	✓	×	$87.2_{\pm 1.3}$	$93.2_{\pm 0.4}$	$56.5_{\pm 1.4}$	$79.4_{\pm 0.3}$
DFR	✓	✓	$\mathbf{89.7}_{\pm 2.4}$	$93.6_{\pm 0.6}$	$48.2_{\pm 0.4}$	$76.4_{\pm 0.2}$

Table 2: Mean ± std of worst group accuracy (%) on Spawrious. Due to space limitations, we compare methods that use group information, as they typically outperform methods that do not use group labels. The final column reports the mean ± std of worst group accuracy across all six subsets. **Bold** indicates the best among these methods.

Method	Group Info	One–To–One			Many–To–Many			Average
		Easy	Medium	Hard	Easy	Medium	Hard	
ERM	×	$78.4_{\pm 1.8}$	$63.4_{\pm 2.3}$	$71.1_{\pm 3.7}$	$72.9_{\pm 1.3}$	$52.7_{\pm 2.9}$	$50.7_{\pm 1.0}$	$64.9_{\pm 11.3}$
SupER (Ours)	×	$82.7_{\pm 2.0}$	$\mathbf{80.3}_{\pm 4.6}$	$\mathbf{83.8}_{\pm 3.4}$	$\mathbf{87.4}_{\pm 1.3}$	$\mathbf{83.4}_{\pm 2.3}$	$\mathbf{79.9}_{\pm 4.7}$	$\mathbf{82.9}_{\pm 2.7}$
UW	✓	$\mathbf{87.4}_{\pm 1.1}$	$67.9_{\pm 2.1}$	$75.9_{\pm 2.9}$	$72.9_{\pm 1.3}$	$52.7_{\pm 2.9}$	$50.7_{\pm 1.0}$	$67.9_{\pm 14.1}$
IRM	✓	$78.4_{\pm 1.0}$	$64.5_{\pm 3.2}$	$64.9_{\pm 2.2}$	$77.9_{\pm 3.7}$	$57.1_{\pm 2.9}$	$50.7_{\pm 1.1}$	$65.6_{\pm 11.1}$
GroupDRO	✓	$86.7_{\pm 1.2}$	$67.2_{\pm 0.7}$	$76.4_{\pm 2.2}$	$74.3_{\pm 0.9}$	$55.7_{\pm 1.4}$	$49.9_{\pm 0.8}$	$68.3_{\pm 13.7}$
DFR	✓	$79.1_{\pm 5.2}$	$64.3_{\pm 1.9}$	$70.0_{\pm 1.9}$	$76.4_{\pm 1.9}$	$58.7_{\pm 2.2}$	$54.1_{\pm 2.2}$	$67.1_{\pm 9.9}$

Table 3: Mean ± std of worst group accuracy (%) for the MetaShift dataset using baselines that do not require group information. A larger value of d indicates a greater distribution shift. The final column reports the mean ± std of worst group accuracy across the four datasets. **Bold** indicates the best among these methods.

Method	Group Info	Train Twice	MetaShift Subsets				Average
			(a) $d = 0.44$	(b) $d = 0.71$	(c) $d = 1.12$	(d) $d = 1.43$	
ERM	×	×	$78.8_{\pm 1.0}$	$75.8_{\pm 0.8}$	$61.9_{\pm 5.9}$	$52.6_{\pm 2.6}$	$67.3_{\pm 12.2}$
CVaR DRO	×	×	$77.8_{\pm 2.5}$	$72.5_{\pm 2.8}$	$65.1_{\pm 0.2}$	$54.7_{\pm 3.2}$	$67.5_{\pm 10.0}$
LfF	×	×	$77.2_{\pm 1.7}$	$73.9_{\pm 0.6}$	$69.5_{\pm 1.0}$	$59.5_{\pm 3.1}$	$70.0_{\pm 7.7}$
GALS	×	×	$74.8_{\pm 3.9}$	$68.8_{\pm 2.0}$	$70.6_{\pm 2.2}$	$50.0_{\pm 0.9}$	$66.0_{\pm 11.0}$
JTT	×	✓	$76.7_{\pm 2.3}$	$73.2_{\pm 0.8}$	$67.1_{\pm 4.6}$	$53.0_{\pm 1.6}$	$67.5_{\pm 10.4}$
CnC	×	✓	$\mathbf{81.1}_{\pm 1.4}$	$71.4_{\pm 2.4}$	$65.4_{\pm 6.8}$	$49.6_{\pm 1.6}$	$66.9_{\pm 13.2}$
SupER (Ours)	×	×	$79.8_{\pm 3.6}$	$\mathbf{78.4}_{\pm 1.9}$	$\mathbf{77.6}_{\pm 2.1}$	$\mathbf{71.4}_{\pm 2.1}$	$\mathbf{76.8}_{\pm 3.7}$

the guidance of superclass information, the model consistently focuses on features with semantic meaning and becomes less influenced by spurious features.

4.2.2. Visualization Analysis of Feature Attention

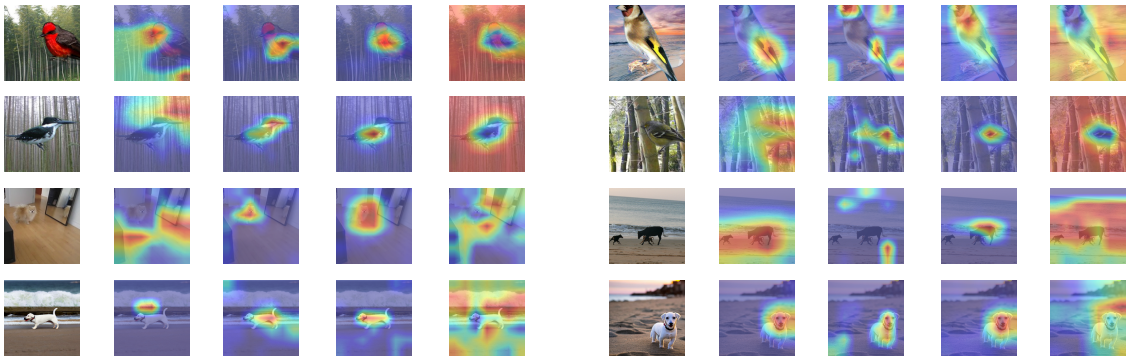
In this section, we analyze the visualized gradient-based attribution maps from different test samples across ERM, CLIP, and SupER to better understand each model's focus areas and feature disentanglement quality. More visualizations are provided in Appendix D.4.

SupER achieves effective disentanglement of superclass-relevant and irrelevant features. As shown in the left five columns of Figure 3, while ERM tends to rely on spurious features for prediction, the attribution maps derived from CLIP can be considered as suitable guidance for superclass semantic information. Furthermore, in SupER, ω_{rel} and ω_{irr} exhibit clear attention to superclass-relevant and superclass-irrelevant features respectively, which validates our approach.

SupER can adjust internal biases in CLIP and significantly outperform CLIP. While CLIP's attention in the left-hand (c) column of Figure 3 can provide general guidance for superclass information, occasional cases from the right-hand (c) column reveal that internal biases in CLIP may lead it to focus on incorrect or incomplete features of the superclass. However, as shown in the right-hand (d) column, SupER is able to reduce these biases and redirect attention toward more accurate and comprehensive superclass-relevant features. Furthermore, Table 30 shows that SupER achieves substantially higher accuracy than the CLIP teacher. This robustness and improvement stems from the superclass-level guidance, the disentanglement ability of the β-VAE, and our principled feature-usage mechanism (see the final paragraph of Section 3.2 for a detailed discussion).

4.3. Ablation Study

Text prompt. To examine the impact of superclass guidance, we vary the text prompts provided to CLIP in two aspects. First, although our main experiments are based on a single text prompt, our general framework in Section 3 allows for multiple prompts. Second, we are interested in the effect of prompt specificity, particularly in terms of superclass hierarchy. In Table 5, we report SupER's performance on the Spawrious datasets under different prompt configurations: (1) increasing the number of prompts, and (2) changing the superclass label y^{super} from "dog" to the more general



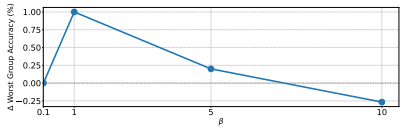
(a) Original (b) ERM (c) CLIP (d) ω_{rel} (e) ω_{irr} (a) Original (b) ERM (c) CLIP (d) ω_{rel} (e) ω_{irr}

Figure 3: Visualization of GradCAM maps across different models and datasets. Rows: (1) Waterbirds-95%, (2) Waterbirds-100%, (3) MetaShift, (4) Spawrious. Each group of five columns ((a)–(e)) shows: original image, GradCAM maps of ERM, CLIP, ω_{rel}, and ω_{irr}. In the left group, guided by superclass information from CLIP, SupER's classifier ω_{rel} successfully learns to focus on superclass-relevant features. The right group shows occasional cases where CLIP exhibits internal bias; nonetheless, ω_{rel} is able to focus on the correct and more complete superclass-relevant features through effective feature disentanglement.

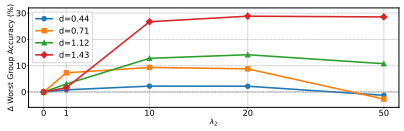
Table 4: Variance of accuracy across groups (%) . **Bold** indicates the smallest across all baselines.

Method	Waterbirds-95%	Waterbirds-100%
ERM	245.9	778.1
CVaR DRO	154.6	528.8
LfF	89.2	442.1
GALS	126.7	516.5
JTT	6.1	405.0
CnC	16.3	347.6
SupER (Ours)	**6.0**	**16.0**
UW	12.7	536.2
IRM	127.2	479.8
GroupDRO	28.0	495.1
DFR	14.2	573.0

Figure 4: Ablation of feature disentanglement (β) and superclass guidance (λ_2) strength.



(a) Effect of β on SpuCo Dogs relative to the $\beta = 0.1$ setting.



(b) Effect of λ_2 on MetaShift relative to the $\lambda_2 = 0$ setting.

"animal". Results show that using multiple prompts generally hurts performance. This may occur because attention maps from different prompts could highlight distinct non-superclass regions due to imperfect guidance, and averaging them mixes biases from each prompt. Moreover, replacing the prompt "dog" with "animal" leads to a drop in accuracy, likely due to the coarser semantic alignment between the generalized superclass and the visual features. (Additional results and more ablation experiments are provided in Appendix D.5.)

Feature disentanglement strength. We study how the strength of feature disentanglement, controlled by the β coefficient in the β-VAE objective, affects model performance. We vary β to observe its impact on SupER's worst group accuracy. Figure 4(a) shows results on the SpuCo Dogs dataset. As β increases, the worst group accuracy initially rises and then declines. This trend indicates that moderate feature disentanglement benefits semantic feature extraction and superclass-relevant feature utilization, whereas overly strong disentanglement can distort task-relevant information.

Degree of superclass guidance. We study the effect of varying the weight λ_2 of the alignment loss $\mathcal{L}^{\text{ATT}}_{\phi,\omega_{\text{rel}},\omega_{\text{irr}}}(\mathbf{x}, y)$ in Algorithm 1, which governs the strength of superclass guidance from CLIP. Results on the four MetaShift subsets in Figure 4(b) indicate that as λ_2 increases, worst group accuracy initially rises and then declines. The effectiveness of superclass guidance becomes more pronounced as distribution shift intensifies (larger d), with larger optimal values of λ_2. These results clearly reveal a trade-off between external guidance and model autonomy: excessive reliance on superclass

Table 5: Ablation results on Spawrious under different prompt configurations. All values indicate the change in worst group accuracy (%) relative to the setting $m = 1$, superclass = `dog`.

#Prompts	Superclass	O2O-Easy	O2O-Medium	O2O-Hard	M2M-Easy	M2M-Medium	M2M-Hard	Average
1	`dog`	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	`dog`	+0.7	-2.2	-2.3	+0.8	-9.6	-2.4	-2.5
5	`dog`	+0.3	-2.5	-2.9	-1.1	-9.4	+1.1	-2.4
1	`animal`	-4.0	-1.1	-5.6	-4.9	-7.5	-5.6	-4.8

guidance may prevent the model from learning discriminative features, while ignoring guidance altogether increases the risk of learning spurious correlations between background and labels.

5. Conclusion

In this work, we propose SupER, a group-label-free framework that leverages superclass-level semantics as a more intrinsic and reliable signal for mitigating spurious correlations. SupER integrates feature disentanglement with superclass guidance, together with a principled feature-usage mechanism supported by theory. Across multiple domain generalization benchmarks, SupER consistently delivers strong performance, and is particularly effective when the training and test groups are not identical or when spurious correlations are highly complex. Beyond spurious correlation mitigation, our framework also illustrates a broader paradigm in which a teacher (e.g., CLIP) provides weak and coarse supervision that enables a student (e.g., SupER) not only to reduce the teacher's biases but also significantly surpass the teacher's performance.

Acknowledgments

QL acknowledges support of NSF DMS-2523382. YS was supported in part by the National Institutes of Health (grant no. 1R01EB036530-01A1).

References

[1] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

[2] Chenruo Liu, Kenan Tang, Yao Qin, and Qi Lei. Bridging distribution shift and ai safety: Conceptual and methodological synergies. *arXiv preprint arXiv:2505.22829*, 2025.

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

[4] Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. *Advances in neural information processing systems*, 36:1390–1402, 2023.

[5] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

[6] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[7] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

[8] Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. *arXiv preprint arXiv:2404.13815*, 2024.

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

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

[11] Jingchao Ni, Wei Cheng, Zhengzhang Chen, Takayoshi Asakura, Tomoya Soma, Sho Kato, and Haifeng Chen. Superclass-conditional gaussian mixture model for learning fine-grained embeddings. In *International Conference on Learning Representations*, 2021.

[12] Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, and Judy Hoffman. Facts: First amplify correlations and then slice to discover bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4794–4804, 2023.

[13] Aahlad Puli, Lily H Zhang, Eric K Oermann, and Rajesh Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. *arXiv preprint arXiv:2107.00520*, 2021.

[14] Aahlad Puli, Nitish Joshi, Yoav Wald, He He, and Rajesh Ranganath. Nuisances via negativa: Adjusting for spurious correlations via data augmentation. *arXiv preprint arXiv:2210.01302*, 2022.

[15] Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[17] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021.

[18] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8024–8034, 2022.

[19] Wanqian Yang, Polina Kirichenko, Micah Goldblum, and Andrew G Wilson. Chroma-vae: Mitigating shortcut learning with generative classifiers. *Advances in Neural Information Processing Systems*, 35:20351–20365, 2022.

[20] Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, pages 18171–18206. PMLR, 2023.

[21] Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. *Advances in Neural Information Processing Systems*, 36:27682–27698, 2023.

[22] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16761–16772, 2022.

[23] Matteo Pagliardini, Martin Jaggi, François Fleuret, and Sai Praneeth Karimireddy. Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202.04414*, 2022.

[24] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[26] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018.

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

[28] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[30] Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *International Conference on Machine Learning*, pages 39365–39379. PMLR, 2023.

[31] Xin Wang, Hong Chen, Zihao Wu, Wenwu Zhu, et al. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[32] Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.

[33] Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pages 12857–12867. PMLR, 2021.

[34] Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, and William Yang Wang. Causal balancing for domain generalization. *arXiv preprint arXiv:2206.05263*, 2022.

[35] Yipei Wang and Xiaoqian Wang. On the effect of key factors in spurious correlation: A theoretical perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 3745–3753. PMLR, 2024.

[36] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. On guiding visual attention with language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18092–18102, 2022.

[37] Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*, 2022.

[38] Hoang Phan, Andrew Gordon Wilson, and Qi Lei. Controllable prompt tuning for balancing group distributional robustness. *arXiv preprint arXiv:2403.02695*, 2024.

[39] Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.

[40] Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating more challenging spurious correlations: A benchmark & new datasets. *arXiv preprint arXiv:2306.11957*, 2023.

[41] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[42] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[43] 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*, pages 770–778, 2016.

[44] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

[45] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[46] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[47] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[48] 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 *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

[49] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. *Advances in Neural Information Processing Systems*, 34:16846–16859, 2021.

[50] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36, 2024.

[51] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. *arXiv preprint arXiv:2302.04269*, 2023.

[52] Zhiqiang Kou, Jing Wang, Yuheng Jia, Biao Liu, and Xin Geng. Instance-dependent inaccurate label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[53] Zhiqiang Kou, Jing Wang, Yuheng Jia, and Xin Geng. Inaccurate label distribution learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10237–10249, 2024.

[54] Zhiqiang Kou, Jing Wang, Yuheng Jia, and Xin Geng. Progressive label enhancement. *Pattern Recognition*, 160:111172, 2025.

[55] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji

Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[56] Jongin Lim, Youngdong Kim, Byungjai Kim, Chanho Ahn, Jinwoo Shin, Eunho Yang, and Seungju Han. Biasadv: Bias-adversarial augmentation for model debiasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3832–3841, 2023.

Appendix

Contents

A. Algorithms for Gradient-based Attribution Maps

A.1. Gradient-based Attribution Map for CLIP

Algorithm 2 Gradient-based Attribution Map for CLIP

Input: Image $\mathbf{x}$, text $\mathbf{T}$, pre-trained ResNet50-based CLIP
Output: Normalized attribution map $L^{\mathbf{T}}_{\text{CLIP}}(\mathbf{x})$
Pass $\mathbf{x}$ through CLIP's vision encoder to get the feature vector $\mathbf{z}$ and K feature maps $\mathbf{A}_k \in \mathbb{R}^{h \times w}$ for $k = 1, 2, \ldots, K$, from the last convolutional layer of ResNet50
Pass $\mathbf{T}$ through CLIP's text encoder to get text embedding $\mathbf{t}$
Compute similarity score:

$$s(\mathbf{x}, \mathbf{T}) = \frac{\mathbf{z} \cdot \mathbf{t}}{\|\mathbf{z}\|\|\mathbf{t}\|}$$

for $k = 1$ to K **do**
  for each $(i, j) \in \{1..h\} \times \{1..w\}$ **do**
    Calculate gradient $\dfrac{\partial s(\mathbf{x}, \mathbf{T})}{\partial \mathbf{A}_k^{ij}}$ for spatial location (i, j)
  end for
  Compute importance weight $\alpha_k^{\mathbf{T}}$ through global average pooling:

$$\alpha_k^{\mathbf{T}} = \frac{1}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} \frac{\partial s(\mathbf{x}, \mathbf{T})}{\partial \mathbf{A}_k^{ij}}$$

end for
Combine feature maps weighted by importance: $L^{\mathbf{T}}_{\text{CLIP}}(\mathbf{x}) = \text{ReLU}\left(\sum_{k=1}^{K} \alpha_k^{\mathbf{T}} \mathbf{A}_k\right)$
Normalize $L^{\mathbf{T}}_{\text{CLIP}}(\mathbf{x})$ to the range $[0, 1]$ using min-max normalization

A.2. Gradient-based Attribution Map for SupER's ω_{rel} and ω_{irr}

Algorithm 3 Gradient-based Attribution Map for SupER's ω_{rel} and ω_{irr}

Input: Image $\mathbf{x}$, true label y, ResNet50-based encoder ϕ, classifiers ω_{rel} (for $\mathbf{z}_{\text{rel}}$) and ω_{irr} (for $\mathbf{z}_{\text{irr}}$)
Output: Normalized attribution maps $L_{\phi,\omega_{\text{rel}}}(\mathbf{x}, y)$ and $L_{\phi,\omega_{\text{irr}}}(\mathbf{x}, y)$
Pass $\mathbf{x}$ through encoder to obtain latent feature $\mathbf{z} = [\mathbf{z}_{\text{rel}}; \mathbf{z}_{\text{irr}}]$ with mean $\boldsymbol{\mu} = [\boldsymbol{\mu}_{\text{rel}}; \boldsymbol{\mu}_{\text{irr}}]$, and K feature maps $\mathbf{A}_k \in \mathbb{R}^{h \times w}$ for $k = 1, 2, \ldots, K$, from the last convolutional layer of ResNet50
Compute logits $g_1 = \omega_{\text{rel}}(\boldsymbol{\mu}_{\text{rel}})$ and $g_2 = \omega_{\text{irr}}(\boldsymbol{\mu}_{\text{irr}})$
for $l = 1$ to 2 **do**
  Let $s_l(\mathbf{x}, y) = g_l[y]$
  for $k = 1$ to K **do**
    for each $(i, j) \in \{1..h\} \times \{1..w\}$ **do**
      Calculate gradient $\dfrac{\partial s_l(\mathbf{x}, y)}{\partial \mathbf{A}_k^{ij}}$ for spatial location (i, j)
    end for
    Compute importance weight α_k^l through global average pooling:

$$\alpha_k^l = \frac{1}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} \frac{\partial s_l(\mathbf{x}, y)}{\partial \mathbf{A}_k^{ij}}$$

  end for
  Combine feature maps weighted by importance:

$$L_{\phi,\omega_l}(\mathbf{x}, y) = \text{ReLU}\left(\sum_{k=1}^{K} \alpha_k^l \mathbf{A}_k\right)$$

  Normalize $L_{\phi,\omega_l}(\mathbf{x}, y)$ to the range $[0, 1]$ using min-max normalization
end for

B. Theoretical Results

Theorem 2 (Fixed-design setting: formal restatement of Theorem 1)**.** *Assume Assumptions 1–3 hold. For any generating index $c \in \{1,2\}$ with coefficient vector $\boldsymbol{\beta}_c \in \mathbb{R}^p$, and any strategy $S \in \{S_1, S_2, S_{1,2}\}$, let $\hat{f}_{S,c,\boldsymbol{\beta}_c}$ be the ordinary least-squares (OLS) predictor obtained by fitting on the training data using the feature(s) specified by S. Define the excess risk under the test distribution by*

$$\mathcal{E}(S; c, \boldsymbol{\beta}_c) \;:=\; \mathbb{E}_{Y^s}\mathbb{E}_{(Z_1,Z_2)\sim P_t}\Big[\,\big(\hat{f}_{S,c,\boldsymbol{\beta}_c}(Z_1,Z_2) - Z_c^\top \boldsymbol{\beta}_c\big)^2\,\Big].$$

Then:

1. (*Superclass-irrelevant feature should not be used.*) *When the generating index is known to be $c = 1$, restricting to strategies $\{S_1, S_{1,2}\}$, for all $\boldsymbol{\beta}_1 \in \mathbb{R}^p$,*

$$\min_{S\in\{S_1,S_{1,2}\}} \mathcal{E}(S; 1, \boldsymbol{\beta}_1) \;=\; \frac{\sigma^2}{n}\sum_{i=1}^{p}\frac{d_{1,i}^t}{d_{1,i}^s}, \qquad \text{with equality achieved by } S = S_1.$$

2. (*All superclass-relevant features should be used.*) *When the generating index c is uncertain and may be either 1 or 2, restricting to strategies $\{S_1, S_2, S_{1,2}\}$, fix any $r > 0$ and define the set*

$$\mathcal{B}_r^{(c)} \;:=\; \big\{\,\boldsymbol{\beta} \in \mathbb{R}^p :\; \boldsymbol{\beta}^\top \Sigma_c^t \boldsymbol{\beta} \;=\; r\,\big\}.$$

There exists a constant $C > 0$, independent of both n and r, such that whenever $nr > C$,

$$\min_{S\in\{S_1,S_2,S_{1,2}\}} \max_{c\in\{1,2\}} \max_{\boldsymbol{\beta}_c\in\mathcal{B}_r^{(c)}} \mathcal{E}(S; c, \boldsymbol{\beta}_c) \;=\; \frac{\sigma^2}{n\,(1-\rho^2)}\sum_{i=1}^{p}\left(\frac{d_{1,i}^t}{d_{1,i}^s} \;+\; \frac{d_{2,i}^t}{d_{2,i}^s}\right),$$

with equality achieved by $S = S_{1,2}$.

Proof of Part (1). By Assumption 3, there exists an orthogonal matrix $U \in \mathbb{R}^{p\times p}$ such that

$$\hat{\Sigma}_1^s = U\,\mathrm{diag}\big(d_1^s\big)\,U^\top, \quad \hat{\Sigma}_2^s = U\,\mathrm{diag}\big(d_2^s\big)\,U^\top, \quad \Sigma_1^t = U\,\mathrm{diag}\big(d_1^t\big)\,U^\top, \quad \Sigma_2^t = U\,\mathrm{diag}\big(d_2^t\big)\,U^\top,$$

for vectors $d_1^s, d_2^s, d_1^t, d_2^t \in \mathbb{R}_+^p$. We work under the generating model $Y = Z_1^\top \boldsymbol{\beta}_1 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of (Z_1, Z_2), and fixed Z_1^s, Z_2^s. The OLS estimator using Z_1 is

$$\hat{\boldsymbol{\beta}}_1 = \big((Z_1^s)^\top Z_1^s\big)^{-1}(Z_1^s)^\top Y^s = \boldsymbol{\beta}_1 + \big((Z_1^s)^\top Z_1^s\big)^{-1}(Z_1^s)^\top \boldsymbol{\varepsilon}.$$

Then,

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) = \sigma^2\big((Z_1^s)^\top Z_1^s\big)^{-1} = \frac{\sigma^2}{n}\,(\hat{\Sigma}_1^s)^{-1}.$$

Therefore the excess risk is

$$\mathcal{E}(S_1; 1, \boldsymbol{\beta}_1) = \mathbb{E}_{Y^s}\mathbb{E}_{Z_1\sim P_t}\big[(Z_1^\top(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1))^2\big] = \mathrm{tr}\big(\Sigma_1^t\,\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)\big) = \frac{\sigma^2}{n}\,\mathrm{tr}\big(\Sigma_1^t(\hat{\Sigma}_1^s)^{-1}\big).$$

Using the simultaneous diagonalization by U,

$$\mathcal{E}(S_1; 1, \boldsymbol{\beta}_1) = \frac{\sigma^2}{n}\,\mathrm{tr}\big(\mathrm{diag}(d_1^t)\,\mathrm{diag}(d_1^s)^{-1}\big) = \frac{\sigma^2}{n}\sum_{i=1}^{p}\frac{d_{1,i}^t}{d_{1,i}^s}.$$

Furthermore, let $\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1 \\ 0\end{bmatrix} \in \mathbb{R}^{2p}$ be the true parameter in the joint model and $Z^s = [\,Z_1^s\;\; Z_2^s\,]$. The OLS estimator satisfies

$$\hat{\boldsymbol{\beta}} = \big((Z^s)^\top Z^s\big)^{-1}(Z^s)^\top Y^s = \boldsymbol{\beta} + \big((Z^s)^\top Z^s\big)^{-1}(Z^s)^\top \boldsymbol{\varepsilon}.$$

Hence,

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \frac{\sigma^2}{n}\,\big(\hat{\Sigma}^s\big)^{-1}, \quad \text{where } \hat{\Sigma}^s = \frac{1}{n}(Z^s)^\top Z^s = \begin{bmatrix}\hat{\Sigma}_1^s & \hat{\Sigma}_{1,2}^s \\ \hat{\Sigma}_{2,1}^s & \hat{\Sigma}_2^s\end{bmatrix}.$$

Using the training-time correlation $\hat{\Sigma}^s_{1,2} = \rho\,(\hat{\Sigma}^s_1)^{1/2}(\hat{\Sigma}^s_2)^{1/2}$ and the shared eigen-basis $U' = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}$, we can write

$$\hat{\Sigma}^s = U' \begin{bmatrix} \operatorname{diag}\big(d^s_1\big) & \rho \operatorname{diag}\big(\sqrt{d^s_1 d^s_2}\big) \\ \rho \operatorname{diag}\big(\sqrt{d^s_1 d^s_2}\big) & \operatorname{diag}\big(d^s_2\big) \end{bmatrix} U'^{\top}, \quad \Sigma^t = U' \begin{bmatrix} \operatorname{diag}\big(d^t_1\big) & 0 \\ 0 & \operatorname{diag}\big(d^t_2\big) \end{bmatrix} U'^{\top},$$

where we used $\Sigma^t_{1,2} = 0$. The inverse of $\hat{\Sigma}^s$ can be computed as

$$(\hat{\Sigma}^s)^{-1} = \frac{1}{1-\rho^2} U' \begin{bmatrix} \operatorname{diag}\big(\frac{1}{d^s_1}\big) & -\operatorname{diag}\Big(\frac{\rho}{\sqrt{d^s_1 d^s_2}}\Big) \\ -\operatorname{diag}\Big(\frac{\rho}{\sqrt{d^s_1 d^s_2}}\Big) & \operatorname{diag}\big(\frac{1}{d^s_2}\big) \end{bmatrix} U'^{\top}.$$

Therefore,

$$\mathcal{E}(S_{1,2}; 1, \boldsymbol{\beta}_1) = \mathbb{E}_{Y^s}\, \mathbb{E}_{(Z_1,Z_2)\sim P_t}\big[(Z^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2\big] = \operatorname{tr}\big(\Sigma^t \operatorname{Cov}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big) = \frac{\sigma^2}{n}\, \operatorname{tr}\big(\Sigma^t (\hat{\Sigma}^s)^{-1}\big),$$

which yields

$$\mathcal{E}(S_{1,2}; 1, \boldsymbol{\beta}_1) = \frac{\sigma^2}{n\,(1-\rho^2)} \sum_{i=1}^{p} \left(\frac{d^t_{1,i}}{d^s_{1,i}} + \frac{d^t_{2,i}}{d^s_{2,i}} \right).$$

Subtracting the two expressions, we have $\mathcal{E}(S_{1,2}; 1, \boldsymbol{\beta}_1) - \mathcal{E}(S_1; 1, \boldsymbol{\beta}_1) > 0$, for $\rho \in [-1, 1] \setminus \{0\}$. Hence

$$\min_{S \in \{S_1, S_{1,2}\}} \mathcal{E}(S; 1, \boldsymbol{\beta}_1) = \mathcal{E}(S_1; 1, \boldsymbol{\beta}_1) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \frac{d^t_{1,i}}{d^s_{1,i}},$$

with equality attained by $S = S_1$. □

Proof of Part (2). Case $S = S_{1,2}$. By the same blockwise computation used in the proof of part (1), we have

$$\mathcal{E}(S_{1,2}; 1, \boldsymbol{\beta}_1) = \frac{\sigma^2}{n\,(1-\rho^2)} \sum_{i=1}^{p} \left(\frac{d^t_{1,i}}{d^s_{1,i}} + \frac{d^t_{2,i}}{d^s_{2,i}} \right).$$

By symmetry (interchanging indices 1 and 2 in the same calculation),

$$\mathcal{E}(S_{1,2}; 2, \boldsymbol{\beta}_2) = \frac{\sigma^2}{n\,(1-\rho^2)} \sum_{i=1}^{p} \left(\frac{d^t_{1,i}}{d^s_{1,i}} + \frac{d^t_{2,i}}{d^s_{2,i}} \right).$$

Therefore, for any $r > 0$ and any $\boldsymbol{\beta}_c \in \mathcal{B}_r^{(c)}$,

$$\max_{c \in \{1,2\}} \max_{\boldsymbol{\beta}_c \in \mathcal{B}_r^{(c)}} \mathcal{E}(S_{1,2}; c, \boldsymbol{\beta}_c) \;=\; \frac{\sigma^2}{n\,(1-\rho^2)} \sum_{i=1}^{p} \left(\frac{d^t_{1,i}}{d^s_{1,i}} + \frac{d^t_{2,i}}{d^s_{2,i}} \right).$$

Case $S = S_1$. When $c = 1$, by part (1),

$$\mathcal{E}(S_1; 1, \boldsymbol{\beta}_1) = \frac{\sigma^2}{n} \sum_{i=1}^{p} \frac{d^t_{1,i}}{d^s_{1,i}}.$$

When $c = 2$, let

$$\hat{\boldsymbol{\beta}}_1 = \big((Z^s_1)^\top Z^s_1\big)^{-1} (Z^s_1)^\top Y^s, \qquad Y^s = Z^s_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

Then

$$\mathcal{E}(S_1; 2, \boldsymbol{\beta}_2) = \mathbb{E}_{Y^s}\, \mathbb{E}_{(Z_1,Z_2)\sim P_t}[(Z_1^\top \hat{\boldsymbol{\beta}}_1 - Z_2^\top \boldsymbol{\beta}_2)^2] = r + \mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1^\top \Sigma^t_1 \hat{\boldsymbol{\beta}}_1],$$

because $\Sigma^t_{1,2} = 0$ and $\boldsymbol{\beta}_2^\top \Sigma^t_2 \boldsymbol{\beta}_2 = r$ for $\boldsymbol{\beta}_2 \in \mathcal{B}_r^{(2)}$. Decompose

$$\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1^\top \Sigma^t_1 \hat{\boldsymbol{\beta}}_1] = \big(\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1]\big)^\top \Sigma^t_1 \big(\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1]\big) \;+\; \operatorname{tr}\big(\Sigma^t_1 \operatorname{Var}(\hat{\boldsymbol{\beta}}_1)\big).$$

With zero-mean noise,

$$\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1] = \big((Z_1^s)^\top Z_1^s\big)^{-1}(Z_1^s)^\top Z_2^s\,\boldsymbol{\beta}_2 = (\hat{\Sigma}_1^s)^{-1}\hat{\Sigma}_{1,2}^s\,\boldsymbol{\beta}_2 = \rho\,(\hat{\Sigma}_1^s)^{-1/2}(\hat{\Sigma}_2^s)^{1/2}\boldsymbol{\beta}_2.$$

Hence

$$\big(\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1]\big)^\top \Sigma_1^t \big(\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1]\big) = \rho^2\,\boldsymbol{\beta}_2^\top(\hat{\Sigma}_2^s)^{1/2}(\hat{\Sigma}_1^s)^{-1/2}\Sigma_1^t(\hat{\Sigma}_1^s)^{-1/2}(\hat{\Sigma}_2^s)^{1/2}\boldsymbol{\beta}_2.$$

Move to the common eigenbasis U and write $\boldsymbol{\beta}_2 = U\gamma$. Using $\hat{\Sigma}_\ell^s = U\,\mathrm{diag}(d_\ell^s)U^\top$ and $\Sigma_\ell^t = U\,\mathrm{diag}(d_\ell^t)U^\top$, we get

$$\big(\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1]\big)^\top \Sigma_1^t \big(\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1]\big) = \rho^2 \sum_{i=1}^p \gamma_i^2\,\frac{d_{2,i}^s\,d_{1,i}^t}{d_{1,i}^s}, \qquad \text{subject to} \quad \sum_{i=1}^p \gamma_i^2\,d_{2,i}^t = r.$$

Maximizing under this weighted ℓ_2-constraint concentrates mass on

$$i^\star \in \arg\max_i\ \frac{d_{2,i}^s\,d_{1,i}^t}{d_{1,i}^s\,d_{2,i}^t},$$

hence

$$\max_{\boldsymbol{\beta}_2 \in \mathcal{B}_r^{(2)}} \big(\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1]\big)^\top \Sigma_1^t \big(\mathbb{E}_{Y^s}[\hat{\boldsymbol{\beta}}_1]\big) = \rho^2\,r\ \max_i \frac{d_{2,i}^s\,d_{1,i}^t}{d_{1,i}^s\,d_{2,i}^t}.$$

Furthermore, since $\mathrm{Var}(\hat{\boldsymbol{\beta}}_1) = \sigma^2\big((Z_1^s)^\top Z_1^s\big)^{-1} = \sigma^2(\hat{\Sigma}_1^s)^{-1}/n$, we have

$$\mathrm{tr}\big(\Sigma_1^t\ \mathrm{Var}(\hat{\boldsymbol{\beta}}_1)\big) = \frac{\sigma^2}{n}\,\mathrm{tr}\big(\Sigma_1^t(\hat{\Sigma}_1^s)^{-1}\big) = \frac{\sigma^2}{n}\sum_{i=1}^p \frac{d_{1,i}^t}{d_{1,i}^s}\ =\ \mathcal{E}(S_1;1,\boldsymbol{\beta}_1).$$

Therefore,

$$\max_{\boldsymbol{\beta}_2 \in \mathcal{B}_r^{(2)}}\ \mathcal{E}(S_1;2,\boldsymbol{\beta}_2) = r\left(1+\rho^2\ \max_i \frac{d_{2,i}^s\,d_{1,i}^t}{d_{1,i}^s\,d_{2,i}^t}\right) + \frac{\sigma^2}{n}\sum_{i=1}^p \frac{d_{1,i}^t}{d_{1,i}^s}\ \geq\ \mathcal{E}(S_1;1,\boldsymbol{\beta}_1).$$

Hence

$$\max_{c\in\{1,2\}}\ \max_{\boldsymbol{\beta}_c \in \mathcal{B}_r^{(c)}}\ \mathcal{E}(S_1;c,\boldsymbol{\beta}_c) = r\left(1+\rho^2\ \max_i \frac{d_{2,i}^s\,d_{1,i}^t}{d_{1,i}^s\,d_{2,i}^t}\right) + \frac{\sigma^2}{n}\sum_{i=1}^p \frac{d_{1,i}^t}{d_{1,i}^s}.$$

Case $S = S_2$. By symmetry with the case $S = S_1$ after interchanging indices $1 \leftrightarrow 2$, we obtain

$$\max_{c\in\{1,2\}}\ \max_{\boldsymbol{\beta}_c \in \mathcal{B}_r^{(c)}}\ \mathcal{E}(S_2;c,\boldsymbol{\beta}_c) = r\left(1+\rho^2\ \max_i \frac{d_{1,i}^s\,d_{2,i}^t}{d_{2,i}^s\,d_{1,i}^t}\right) + \frac{\sigma^2}{n}\sum_{i=1}^p \frac{d_{2,i}^t}{d_{2,i}^s}.$$

Let

$$C\ =\ \frac{\sigma^2}{(1-\rho^2)}\cdot\frac{1}{C'}\sum_{i=1}^p\left(\frac{d_{1,i}^t}{d_{1,i}^s}+\frac{d_{2,i}^t}{d_{2,i}^s}\right),$$

where

$$C'\ =\ 1+\rho^2\ \min\left\{\max_i \frac{d_{1,i}^s\,d_{2,i}^t}{d_{2,i}^s\,d_{1,i}^t},\ \max_i \frac{d_{2,i}^s\,d_{1,i}^t}{d_{1,i}^s\,d_{2,i}^t}\right\}.$$

Then whenever $n > C/r$, we have

$$\mathcal{E}(S_{1,2};1,\boldsymbol{\beta}_1),\ \ \mathcal{E}(S_{1,2};2,\boldsymbol{\beta}_2)\ <\ \min_{S\in\{S_1,S_2\}}\ \max_{c\in\{1,2\}}\ \max_{\boldsymbol{\beta}_c\in\mathcal{B}_r^{(c)}}\ \mathcal{E}(S;c,\boldsymbol{\beta}_c),$$

and hence the claim holds. □

In addition, we provide further theoretical analysis under the random-design setting.

Assumption 4. *Under the training distribution P_s, $Z = (Z_1, Z_2)$ follows a Gaussian distribution*

$$Z \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_1^s & \Sigma_{1,2}^s \\ \Sigma_{2,1}^s & \Sigma_2^s \end{bmatrix}\right),$$

where $\Sigma_1^s = \mathbb{E}_{Z_1 \sim P_s}[Z_1 Z_1^\top]$, $\Sigma_2^s = \mathbb{E}_{Z_2 \sim P_s}[Z_2 Z_2^\top]$, and $\Sigma_{1,2}^s = \mathbb{E}_{(Z_1,Z_2)\sim P_s}[Z_1 Z_2^\top] = (\Sigma_{2,1}^s)^\top$.

Theorem 3 (Random-design setting)**.** *Assume Assumptions 1 and 4 hold. In addition, suppose the population counterparts of Assumptions 2 and 3 hold; namely, $\Sigma_{1,2}^s = \rho\,(\Sigma_1^s)^{1/2}(\Sigma_2^s)^{1/2}$, and $\Sigma_1^s, \Sigma_2^s, \Sigma_1^t, \Sigma_2^t$ commute with positive eigenvalues $\{d_{1,i}^s\}_{i=1}^p$, $\{d_{2,i}^s\}_{i=1}^p$, $\{d_{1,i}^t\}_{i=1}^p$, and $\{d_{2,i}^t\}_{i=1}^p$. For any generating index $c \in \{1,2\}$ with coefficient vector $\boldsymbol{\beta}_c \in \mathbb{R}^p$, and any strategy $S \in \{S_1, S_2, S_{1,2}\}$, let $\hat{f}_{S,c,\boldsymbol{\beta}_c}$ be the OLS predictor trained on n i.i.d. samples drawn from P_s using the feature(s) specified by S. Define the random-design excess risk*

$$\mathcal{E}^{\mathrm{rd}}(S; c, \boldsymbol{\beta}_c) \;:=\; \mathbb{E}_{(Z_1,Z_2)\sim P_t}\Big[\,\big(\hat{f}_{S,c,\boldsymbol{\beta}_c}(Z_1, Z_2) - Z_c^\top \boldsymbol{\beta}_c\big)^2\,\Big],$$

Then:

1. (*Superclass-irrelevant feature should not be used.*) *When the generating index is known to be $c = 1$, restricting to strategies $\{S_1, S_{1,2}\}$, for all $\boldsymbol{\beta}_1 \in \mathbb{R}^p$, if $n \gg p$ with p sufficiently large, we have, with high probability,*

$$\min_{S\in\{S_1,S_{1,2}\}} \mathcal{E}^{\mathrm{rd}}(S; 1, \boldsymbol{\beta}_1) \quad \textit{is attained by} \quad S = S_1.$$

2. (*All superclass-relevant features should be used.*) *When the generating index c is uncertain and may be either 1 or 2, restricting to strategies $\{S_1, S_2, S_{1,2}\}$, fix any $r > 0$ and define*

$$\mathcal{B}_r^{(c)} \;:=\; \big\{\, \boldsymbol{\beta} \in \mathbb{R}^p :\; \boldsymbol{\beta}^\top \Sigma_c^t \boldsymbol{\beta} \;=\; r \,\big\}.$$

There exists a constant $C > 0$ independent of both n and r, such that whenever $nr > C$, we have with high probability,

$$\min_{S\in\{S_1,S_2,S_{1,2}\}} \max_{c\in\{1,2\}} \max_{\boldsymbol{\beta}_c \in \mathcal{B}_r^{(c)}} \mathcal{E}^{\mathrm{rd}}(S; c, \boldsymbol{\beta}_c) \quad \textit{is attained by} \quad S = S_{1,2}.$$

Proof of Part (1). Fix any $\boldsymbol{\beta}_1 \in \mathbb{R}^p$. By Lemma 1, we have a high-probability upper bound on $\mathcal{E}^{\mathrm{rd}}(S_1; 1, \boldsymbol{\beta}_1)$. By Lemma 2, we have a high-probability lower bound on $\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1)$.

When $n \gg p$, both bounds satisfy that $\epsilon(\delta_2)$ can be made arbitrarily small, and when p is sufficiently large, the logarithmic terms (multiplied by c_1) are $o(1)$ relative to the spectral sums. In this regime, the leading term of the lower bound on $\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1)$ dominates the leading term of the upper bound on $\mathcal{E}^{\mathrm{rd}}(S_1; 1, \boldsymbol{\beta}_1)$, therefore

$$\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1) \;\geq\; \mathcal{E}^{\mathrm{rd}}(S_1; 1, \boldsymbol{\beta}_1) \quad \text{with high probability.}$$

Therefore, the minimum over $S \in \{S_1, S_{1,2}\}$ is attained by $S = S_1$ with high probability. □

Proof of Part (2). First, by Corollary 1 and by symmetry between the cases $c = 1$ and $c = 2$, with probability at least $(1-\delta_1)(1-\delta_2)$, the strategy $S_{1,2}$ satisfies the uniform bound (independent of c and $\boldsymbol{\beta}_c$)

$$\max_{c\in\{1,2\}} \max_{\boldsymbol{\beta}_c \in \mathcal{B}_r^{(c)}} \mathcal{E}^{\mathrm{rd}}(S_{1,2}; c, \boldsymbol{\beta}_c) \;\leq\; \frac{\sigma^2}{n\,(1-\rho^2)\,(1-\varepsilon(\delta_2))} \left(\sum_{i=1}^{p} \left(\frac{d_{1,i}^t}{d_{1,i}^s} + \frac{d_{2,i}^t}{d_{2,i}^s} \right) + \frac{1}{2} c_1 g_\rho \log \frac{1}{\delta_1} \right).$$

Next, consider $S = S_1$ under $c = 2$. For any $\boldsymbol{\beta}_2 \in \mathcal{B}_r^{(2)}$,

$$\mathcal{E}^{\mathrm{rd}}(S_1; 2, \boldsymbol{\beta}_2) = \mathbb{E}_{(Z_1,Z_2)\sim P_t}\big[(Z_1^\top \hat{\boldsymbol{\beta}}_1 - Z_2^\top \boldsymbol{\beta}_2)^2\big] = r + \hat{\boldsymbol{\beta}}_1^\top \Sigma_1^t \hat{\boldsymbol{\beta}}_1 \;\geq\; r,$$

since $\Sigma_{1,2}^t = 0$ and $\Sigma_1^t \succeq 0$. Hence

$$\max_{c\in\{1,2\}} \max_{\boldsymbol{\beta}_c \in \mathcal{B}_r^{(c)}} \mathcal{E}^{\mathrm{rd}}(S_1; c, \boldsymbol{\beta}_c) \;\geq\; r.$$

By symmetry, the same lower bound holds for $S = S_2$. Choose

$$C \;=\; \frac{\sigma^2\left(\sum_{i=1}^{p}(d_{1,i}^t/d_{1,i}^s + d_{2,i}^t/d_{2,i}^s) + 1/2\cdot c_1 g_\rho \log(1/\delta_1)\right)}{(1-\rho^2)\,(1-\varepsilon(\delta_2))}.$$

Then whenever $n > C/r$, with high probability (at least $(1-\delta_1)(1-\delta_2)$),

$$\max_{c\in\{1,2\}} \max_{\boldsymbol{\beta}_c\in\mathcal{B}_r^{(c)}} \mathcal{E}^{\mathrm{rd}}(S_{1,2};c,\boldsymbol{\beta}_c) \;\le\; r \;\le\; \min\left\{ \max_{c\in\{1,2\}} \max_{\boldsymbol{\beta}_c\in\mathcal{B}_r^{(c)}} \mathcal{E}^{\mathrm{rd}}(S_1;c,\boldsymbol{\beta}_c),\; \max_{c\in\{1,2\}} \max_{\boldsymbol{\beta}_c\in\mathcal{B}_r^{(c)}} \mathcal{E}^{\mathrm{rd}}(S_2;c,\boldsymbol{\beta}_c)\right\}.$$

Therefore,

$$\min_{S\in\{S_1,S_2,S_{1,2}\}} \max_{c\in\{1,2\}} \max_{\boldsymbol{\beta}_c\in\mathcal{B}_r^{(c)}} \mathcal{E}^{\mathrm{rd}}(S;c,\boldsymbol{\beta}_c) \quad \text{is attained by} \quad S = S_{1,2}.$$

□

Lemma 1. *Under the conditions of Theorem 3 with generating index $c = 1$, there exist universal constants $c_1, c_2 > 0$ such that for all $0 < \delta_1, \delta_2 < 1$ chosen suitably small, with probability at least $(1-\delta_1)(1-\delta_2)$ over the draw of the n training samples,*

$$\mathcal{E}^{\mathrm{rd}}(S_1;1,\boldsymbol{\beta}_1) \;\le\; \frac{\sigma^2}{n\,(1-\epsilon(\delta_2))}\left(\sum_{i=1}^{p}\frac{d_{1,i}^t}{d_{1,i}^s} \;+\; c_1\, g\left(\{d_{1,i}^s\}_{i=1}^p, \{d_{1,i}^t\}_{i=1}^p\right)\log\frac{1}{\delta_1}\right),$$

where

$$\epsilon(\delta_2) \;=\; c_2\max\left\{\sqrt{\frac{p+\log\left(\frac{2}{\delta_2}\right)}{n}}\;,\; \frac{p+\log\left(\frac{2}{\delta_2}\right)}{n}\right\}, \qquad g\left(\{d_{1,i}^s\}_{i=1}^p, \{d_{1,i}^t\}_{i=1}^p\right) \;=\; \max_i \frac{d_{1,i}^t}{d_{1,i}^s}.$$

Proof. First, there exists an orthogonal matrix $U \in \mathbb{R}^{p\times p}$ such that

$$\Sigma_1^s = U\,\mathrm{diag}\big(d_1^s\big)\,U^\top, \quad \Sigma_2^s = U\,\mathrm{diag}\big(d_2^s\big)\,U^\top, \quad \Sigma_1^t = U\,\mathrm{diag}\big(d_1^t\big)\,U^\top, \quad \Sigma_2^t = U\,\mathrm{diag}\big(d_2^t\big)\,U^\top,$$

for vectors $d_1^s, d_2^s, d_1^t, d_2^t \in \mathbb{R}_+^p$. For the strategy S_1 (using Z_1 only), the proof proceeds in three steps.

Step 1. We show that there exists a function $B(Z_1^s, \delta_1)$ depending on the training data Z_1^s and δ_1 such that

$$\mathbb{P}\Big(\mathcal{E}^{\mathrm{rd}}(S_1;1,\boldsymbol{\beta}_1) \;\le\; B(Z_1^s,\delta_1)\;\Big|\; Z_1^s\Big) \;\ge\; 1-\delta_1.$$

Step 2. We show the existence of $\epsilon(\delta_2)$ such that

$$\mathbb{P}\Big((1-\epsilon(\delta_2))\,\Sigma_1^s \;\preceq\; \hat{\Sigma}_1^s \;\preceq\; (1+\epsilon(\delta_2))\,\Sigma_1^s\Big) \;\ge\; 1-\delta_2.$$

Step 3. Let

$$E \;=\; \big\{\, Z_1^s : (1-\epsilon(\delta_2))\,\Sigma_1^s \preceq \hat{\Sigma}_1^s \preceq (1+\epsilon(\delta_2))\,\Sigma_1^s\big\}.$$

On the event E, we choose a constant $B(\delta_1,\delta_2)$, independent of Z_1^s, such that $B(Z_1^s,\delta_1) \le B(\delta_1,\delta_2)$. Hence,

$$\begin{aligned}
\mathbb{P}\Big(\mathcal{E}^{\mathrm{rd}}(S_1;1,\boldsymbol{\beta}_1) \le B(\delta_1,\delta_2)\Big) &= \mathbb{E}_{Z_1^s}\Big[\mathbb{P}\big(\mathcal{E}^{\mathrm{rd}}(S_1;1,\boldsymbol{\beta}_1) \le B(\delta_1,\delta_2)\;\big|\; Z_1^s\big)\Big] \\
&\ge \mathbb{E}_{Z_1^s}\Big[\mathbf{1}_E(Z_1^s)\,\mathbb{P}\big(\mathcal{E}^{\mathrm{rd}}(S_1;1,\boldsymbol{\beta}_1) \le B(Z_1^s,\delta_1)\;\big|\; Z_1^s\big)\Big] \\
&\ge (1-\delta_1)\,\mathbb{E}_{Z_1^s}[\mathbf{1}_E(Z_1^s)] = (1-\delta_1)\,\mathbb{P}(Z_1^s\in E) \;\ge\; (1-\delta_1)(1-\delta_2).
\end{aligned}$$

Details for Step 1. Condition on $Z_1^s \in \mathbb{R}^{n\times p}$. The OLS error satisfies

$$\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 = \big((Z_1^s)^\top Z_1^s\big)^{-1}(Z_1^s)^\top\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon}\sim\mathcal{N}(0,\sigma^2 I_n),$$

so $\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \sim \mathcal{N}\big(0, \sigma^2(\hat{\Sigma}_1^s)^{-1}/n\big)$, with $\hat{\Sigma}_1^s = (Z_1^s)^\top Z_1^s/n$. Thus

$$\mathcal{E}^{\mathrm{rd}}(S_1; 1, \boldsymbol{\beta}_1) = \mathbb{E}_{Z_1 \sim P_t}\big[(Z_1^\top(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1))^2\big] = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)^\top \Sigma_1^t (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \overset{d}{=} \frac{\sigma^2}{n}\, v^\top A v,$$

where $v \sim \mathcal{N}(0, I_p)$ and $A := (\hat{\Sigma}_1^s)^{-1/2}\, \Sigma_1^t\, (\hat{\Sigma}_1^s)^{-1/2}$. By the Hanson–Wright inequality [44], there exist absolute constants $c > 0$ such that for any $t > 0$,

$$\mathbb{P}\big(v^\top A v - \mathrm{tr}(A) \geq t \,\big|\, Z_1^s\big) \leq \exp\!\left(-c\, \min\!\left\{ \frac{t^2}{\|A\|_F^2},\, \frac{t}{\|A\|_2} \right\}\right).$$

Taking $c_1 = 1/c$ and $t = c_1 \|A\|_2 \log(1/\delta_1)$, when δ_1 is small such that the second regime applies, we have with probability at least $1 - \delta_1$,

$$\mathcal{E}^{\mathrm{rd}}(S_1; 1, \boldsymbol{\beta}_1) \;\leq\; \frac{\sigma^2}{n}\Big(\mathrm{tr}(A) + c_1 \|A\|_2 \,\log\frac{1}{\delta_1}\Big) =: B(Z_1^s, \delta_1).$$

Details for Step 2. By concentration inequalities for covariance matrices [45], there exists a universal constant $c_2 > 0$ such that, for any $\delta_2 \in (0,1)$, with probability at least $1 - \delta_2$,

$$(1 - \epsilon(\delta_2))\, \Sigma_1^s \;\preceq\; \hat{\Sigma}_1^s \;\preceq\; (1 + \epsilon(\delta_2))\, \Sigma_1^s,$$

where

$$\epsilon(\delta_2) = c_2 \max\!\left\{ \sqrt{\frac{p + \log\big(\frac{2}{\delta_2}\big)}{n}}\,,\; \frac{p + \log\big(\frac{2}{\delta_2}\big)}{n} \right\},$$

and we assume n is large enough that $\epsilon(\delta_2) < 1$.

Details for Step 3. On E we have $(\hat{\Sigma}_1^s)^{-1} \preceq \frac{1}{1-\epsilon(\delta_2)} (\Sigma_1^s)^{-1}$. Hence,

$$\mathrm{tr}(A) = \mathrm{tr}\big((\hat{\Sigma}_1^s)^{-1}\Sigma_1^t\big) \leq \frac{1}{1 - \epsilon(\delta_2)}\; \mathrm{tr}\big((\Sigma_1^s)^{-1}\Sigma_1^t\big) = \frac{1}{1 - \epsilon(\delta_2)} \sum_{i=1}^p \frac{d_{1,i}^t}{d_{1,i}^s},$$

Moreover,

$$\|A\|_2 = \big\|(\Sigma_1^t)^{1/2}(\hat{\Sigma}_1^s)^{-1}(\Sigma_1^t)^{1/2}\big\|_2 \leq \frac{1}{1 - \epsilon(\delta_2)}\, \big\|(\Sigma_1^t)^{1/2}(\Sigma_1^s)^{-1}(\Sigma_1^t)^{1/2}\big\|_2 = \frac{1}{1 - \epsilon(\delta_2)}\; \max_i \frac{d_{1,i}^t}{d_{1,i}^s}.$$

Consequently, on E,

$$B(Z_1^s, \delta_1) \leq \frac{\sigma^2}{n\,(1 - \epsilon(\delta_2))} \left(\sum_{i=1}^p \frac{d_{1,i}^t}{d_{1,i}^s} \;+\; c_1 \max_i \frac{d_{1,i}^t}{d_{1,i}^s}\, \log\frac{1}{\delta_1} \right) := B(\delta_1, \delta_2).$$

Combining the three steps yields the stated high-probability bound with $g\big(\{d_{1,i}^s\}_{i=1}^p, \{d_{1,i}^t\}_{i=1}^p\big) = \max_i d_{1,i}^t / d_{1,i}^s$. □

Lemma 2. *Under the conditions of Theorem 3 with generating index $c = 1$, there exist universal constants $c_1, c_2 > 0$ such that for all $0 < \delta_1, \delta_2 < 1$ chosen suitably small, with probability at least $(1 - \delta_1)(1 - \delta_2)$ over the draw of the n training samples,*

$$\begin{aligned}\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1) \;\geq\; \frac{\sigma^2}{n\,(1-\rho^2)} \Bigg(& \frac{1}{1 + \epsilon(\delta_2)} \sum_{i=1}^p \left(\frac{d_{1,i}^t}{d_{1,i}^s} + \frac{d_{2,i}^t}{d_{2,i}^s} \right) \\ & - \frac{c_1}{2(1 - \epsilon(\delta_2))}\, g_\rho\big(\{d_{1,i}^s\}_{i=1}^p, \{d_{2,i}^s\}_{i=1}^p, \{d_{1,i}^t\}_{i=1}^p, \{d_{2,i}^t\}_{i=1}^p\big) \log\frac{1}{\delta_1} \Bigg).\end{aligned}$$

where

$$\epsilon(\delta_2) \;=\; c_2 \max\!\left\{ \sqrt{\frac{2p + \log\big(\frac{2}{\delta_2}\big)}{n}}\,,\; \frac{2p + \log\big(\frac{2}{\delta_2}\big)}{n} \right\},$$

and

$$g_\rho\big(\{d_{1,i}^s\}_{i=1}^p, \{d_{2,i}^s\}_{i=1}^p, \{d_{1,i}^t\}_{i=1}^p, \{d_{2,i}^t\}_{i=1}^p\big) \;=\; \max_i \left(\frac{d_{1,i}^t}{d_{1,i}^s} + \frac{d_{2,i}^t}{d_{2,i}^s} + \sqrt{\left(\frac{d_{1,i}^t}{d_{1,i}^s} - \frac{d_{2,i}^t}{d_{2,i}^s} \right)^2 + 4\rho^2\, \frac{d_{1,i}^t}{d_{1,i}^s}\, \frac{d_{2,i}^t}{d_{2,i}^s}} \right).$$

Proof. First, there exists an orthogonal matrix $U \in \mathbb{R}^{p\times p}$ such that

$$\Sigma_1^s = U\,\mathrm{diag}\big(d_1^s\big)\,U^\top, \quad \Sigma_2^s = U\,\mathrm{diag}\big(d_2^s\big)\,U^\top, \quad \Sigma_1^t = U\,\mathrm{diag}\big(d_1^t\big)\,U^\top, \quad \Sigma_2^t = U\,\mathrm{diag}\big(d_2^t\big)\,U^\top,$$

for vectors $d_1^s, d_2^s, d_1^t, d_2^t \in \mathbb{R}_+^p$. For the strategy $S_{1,2}$ (using $Z = (Z_1, Z_2)$), the proof proceeds in three steps.

Step 1. We show there exists a function $C(Z^s, \delta_1)$ depending on the full training data $Z^s = [\,Z_1^s\ Z_2^s\,] \in \mathbb{R}^{n\times 2p}$ and δ_1 such that

$$\mathbb{P}\Big(\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1)\ \geq\ C(Z^s, \delta_1)\ \Big|\ Z^s\Big)\ \geq\ 1-\delta_1.$$

Step 2. We show the existence of $\epsilon(\delta_2)$ such that

$$\mathbb{P}\Big((1-\epsilon(\delta_2))\,\Sigma^s\ \preceq\ \hat{\Sigma}^s\ \preceq\ (1+\epsilon(\delta_2))\,\Sigma^s\Big)\ \geq\ 1-\delta_2,$$

where $\Sigma^s = \mathbb{E}_{(Z_1,Z_2)\sim P_s}[ZZ^\top]$ and $\hat{\Sigma}^s = (Z^s)^\top Z^s/n$ are the population and empirical covariances of $Z = (Z_1, Z_2)$.

Step 3. Let

$$E\ =\ \big\{\, Z^s :\ (1-\epsilon(\delta_2))\,\Sigma^s \preceq \hat{\Sigma}^s \preceq (1+\epsilon(\delta_2))\,\Sigma^s\big\}.$$

On the event E, we choose a constant $C(\delta_1, \delta_2)$, independent of Z^s, such that $C(Z^s, \delta_1)\ \geq\ C(\delta_1, \delta_2)$. Hence

$$\begin{aligned}
\mathbb{P}\Big(\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1) \geq C(\delta_1,\delta_2)\Big) &= \mathbb{E}_{Z^s}\Big[\mathbb{P}\big(\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1) \geq C(\delta_1,\delta_2)\ \big|\ Z^s\big)\Big] \\
&\geq \mathbb{E}_{Z^s}\Big[\mathbf{1}_E(Z^s)\,\mathbb{P}\big(\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1) \geq C(Z^s,\delta_1)\ \big|\ Z^s\big)\Big] \\
&\geq (1-\delta_1)\,\mathbb{E}_{Z^s}\big[\mathbf{1}_E(Z^s)\big] = (1-\delta_1)\,\mathbb{P}(Z^s \in E) \\
&\geq (1-\delta_1)(1-\delta_2).
\end{aligned}$$

Details for Step 1. Let $\hat{\boldsymbol{\beta}}$ be the OLS estimator in the joint model trained on Z^s. The standard OLS formula yields

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \big((Z^s)^\top Z^s\big)^{-1}(Z^s)^\top \boldsymbol{\varepsilon}, \qquad \boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1 \\ 0\end{bmatrix},$$

hence $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \mathcal{N}\big(0, \sigma^2(\hat{\Sigma}^s)^{-1}/n\big)$. So

$$\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1) = \mathbb{E}_{(Z_1,Z_2)\sim P_t}\big[(Z^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2\big] = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma^t (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\ \overset{d}{=}\ \frac{\sigma^2}{n}\, v^\top A v,$$

where $v \sim \mathcal{N}(0, I_{2p})$ and $A := (\hat{\Sigma}^s)^{-1/2}\,\Sigma^t\,(\hat{\Sigma}^s)^{-1/2}$. By the Hanson–Wright inequality [44], there exists an absolute constant $c > 0$ such that, for any $t > 0$,

$$\mathbb{P}\big(v^\top A v - \mathrm{tr}(A) \leq -t\ \big|\ Z^s\big) \leq \exp\left(-c\,\min\left\{\frac{t^2}{\|A\|_F^2},\ \frac{t}{\|A\|_2}\right\}\right).$$

Similar to the proof in Lemma 1, when δ_1 is small, there exists $c_1 > 0$ such that with probability at least $1-\delta_1$,

$$\mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1)\ \geq\ \frac{\sigma^2}{n}\Big(\mathrm{tr}(A) - c_1\|A\|_2\,\log\frac{1}{\delta_1}\Big) =: C(Z^s, \delta_1).$$

Details for Step 2. By concentration inequalities for covariance matrices [45], there exists a universal constant $c_2 > 0$ such that, for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$,

$$(1-\epsilon(\delta_2))\,\Sigma^s\ \preceq\ \hat{\Sigma}^s\ \preceq\ (1+\epsilon(\delta_2))\,\Sigma^s,$$

where

$$\epsilon(\delta_2) = c_2\,\max\left\{\sqrt{\frac{2p + \log\big(\frac{2}{\delta_2}\big)}{n}}\ ,\ \frac{2p + \log\big(\frac{2}{\delta_2}\big)}{n}\right\},$$

and we assume n is large enough that $\epsilon(\delta_2) < 1$.

Details for Step 3. On E, $(\hat{\Sigma}^s)^{-1} \succeq \frac{1}{1+\epsilon(\delta_2)}(\Sigma^s)^{-1}$ and $(\hat{\Sigma}^s)^{-1} \preceq \frac{1}{1-\epsilon(\delta_2)}(\Sigma^s)^{-1}$. Thus, by Loewner monotonicity of trace and spectral norm,

$$\mathrm{tr}(A) = \mathrm{tr}\big((\hat{\Sigma}^s)^{-1}\Sigma^t\big) \;\geq\; \frac{1}{1+\epsilon(\delta_2)}\;\mathrm{tr}\big((\Sigma^s)^{-1}\Sigma^t\big), \qquad \|A\|_2 \;\leq\; \frac{1}{1-\epsilon(\delta_2)}\,\big\|(\Sigma^t)^{1/2}(\Sigma^s)^{-1}(\Sigma^t)^{1/2}\big\|_2.$$

We further simultaneously diagonalize with $U' = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}$ and write

$$(\Sigma^s)^{-1} = \frac{1}{1-\rho^2}\,U'\begin{bmatrix} \mathrm{diag}\big(\frac{1}{d_1^s}\big) & -\mathrm{diag}\big(\frac{\rho}{\sqrt{d_1^s d_2^s}}\big) \\ -\mathrm{diag}\big(\frac{\rho}{\sqrt{d_1^s d_2^s}}\big) & \mathrm{diag}\big(\frac{1}{d_2^s}\big) \end{bmatrix} U'^{\top}, \qquad \Sigma^t = U'\begin{bmatrix} \mathrm{diag}(d_1^t) & 0 \\ 0 & \mathrm{diag}(d_2^t) \end{bmatrix} U'^{\top}.$$

Therefore,

$$\mathrm{tr}\big((\Sigma^s)^{-1}\Sigma^t\big) = \frac{1}{1-\rho^2}\sum_{i=1}^{p}\left(\frac{d_{1,i}^t}{d_{1,i}^s} + \frac{d_{2,i}^t}{d_{2,i}^s}\right),$$

and

$$(\Sigma^t)^{1/2}(\Sigma^s)^{-1}(\Sigma^t)^{1/2} = \frac{1}{1-\rho^2}\,U'\begin{bmatrix} \mathrm{diag}\big(\frac{d_1^t}{d_1^s}\big) & -\rho\,\mathrm{diag}\Big(\frac{\sqrt{d_1^t d_2^t}}{\sqrt{d_1^s d_2^s}}\Big) \\ -\rho\,\mathrm{diag}\Big(\frac{\sqrt{d_1^t d_2^t}}{\sqrt{d_1^s d_2^s}}\Big) & \mathrm{diag}\big(\frac{d_2^t}{d_2^s}\big) \end{bmatrix} U'^{\top} := C.$$

By a suitable simultaneous permutation of rows and columns, C is similar to $\bigoplus_{i=1}^{p} C_i$ with

$$C_i = \begin{bmatrix} \frac{d_{1,i}^t}{d_{1,i}^s} & -\rho\sqrt{\frac{d_{1,i}^t d_{2,i}^t}{d_{1,i}^s d_{2,i}^s}} \\ -\rho\sqrt{\frac{d_{1,i}^t d_{2,i}^t}{d_{1,i}^s d_{2,i}^s}} & \frac{d_{2,i}^t}{d_{2,i}^s} \end{bmatrix}.$$

Hence

$$\big\|(\Sigma^t)^{1/2}(\Sigma^s)^{-1}(\Sigma^t)^{1/2}\big\|_2 = \frac{1}{1-\rho^2}\,\max_i \lambda_{\max}(C_i) \;\leq\; \frac{1}{2(1-\rho^2)}\,g_\rho\big(\{d_{1,i}^s\}_{i=1}^p, \{d_{2,i}^s\}_{i=1}^p, \{d_{1,i}^t\}_{i=1}^p, \{d_{2,i}^t\}_{i=1}^p\big),$$

where $g_\rho(\cdot)$ is as in the lemma statement. Combining the bounds on $\mathrm{tr}(A)$ and $\|A\|_2$ on E yields

$$\begin{aligned} C(Z^s, \delta_1) \;\geq\; & \frac{\sigma^2}{n\,(1-\rho^2)}\left(\frac{1}{1+\epsilon(\delta_2)}\sum_{i=1}^{p}\left(\frac{d_{1,i}^t}{d_{1,i}^s} + \frac{d_{2,i}^t}{d_{2,i}^s}\right)\right. \\ & \left. -\;\frac{c_1}{2(1-\epsilon(\delta_2))}\,g_\rho\big(\{d_{1,i}^s\}_{i=1}^p, \{d_{2,i}^s\}_{i=1}^p, \{d_{1,i}^t\}_{i=1}^p, \{d_{2,i}^t\}_{i=1}^p\big)\log\frac{1}{\delta_1}\right) := C(\delta_1, \delta_2). \end{aligned}$$

which completes the proof. □

Corollary 1. *Under the conditions of Theorem 3 with generating index $c = 1$, there exist universal constants $c_1, c_2 > 0$ such that for all $0 < \delta_1, \delta_2 < 1$ chosen suitably small, with probability at least $(1-\delta_1)(1-\delta_2)$ over the draw of the n training samples,*

$$\begin{aligned} \mathcal{E}^{\mathrm{rd}}(S_{1,2}; 1, \boldsymbol{\beta}_1) \;\leq\; & \frac{\sigma^2}{n\,(1-\rho^2)\,(1-\varepsilon(\delta_2))}\left(\sum_{i=1}^{p}\left(\frac{d_{1,i}^t}{d_{1,i}^s} + \frac{d_{2,i}^t}{d_{2,i}^s}\right)\right. \\ & \left. +\;\frac{1}{2}c_1 g_\rho\big(\{d_{1,i}^s\}_{i=1}^p, \{d_{2,i}^s\}_{i=1}^p, \{d_{1,i}^t\}_{i=1}^p, \{d_{2,i}^t\}_{i=1}^p\big)\,\log\frac{1}{\delta_1}\right), \end{aligned}$$

where

$$\varepsilon(\delta_2) \;=\; c_2\max\left\{\sqrt{\frac{2p + \log\big(\frac{2}{\delta_2}\big)}{n}}\;,\;\frac{2p + \log\big(\frac{2}{\delta_2}\big)}{n}\right\},$$

and $g_\rho(\cdot)$ is as in Lemma 2.

C. Additional Related Work

Spurious correlations under domain generalization. A broad literature studies how to mitigate models' reliance on spurious features for improved domain generalization, including invariant learning [10, 46, 47], distributionally robust optimization [3, 48], causal relationship studies [15, 49], fine-tuning methods [42, 50], contrastive learning [7], and the use of vision-language models [12, 36, 51]. All aforementioned methods except [36] become ineffective when spurious correlations cannot be identified based on training groups or domains, or no auxiliary information on groups, domains, or spurious features is available. While [36] leverages CLIP to inform visual attention, their approach does not perform feature disentanglement, nor does it incorporate a principled feature-usage strategy. As a result, it becomes less effective in handling more complex spurious correlation structures and in correcting biases inherited from CLIP guidance. In contrast, as shown in the main text, our model can overcome all these limitations.

Label distribution learning. Label distribution learning treats supervision as a distribution over labels rather than a one-hot target, which can model supervision uncertainty and inter-label structure and has been shown to improve representation robustness [52–54]. Although studied in problem settings different from spurious correlation mitigation, these works share a related high-level motivation with SupER in leveraging structured, distribution-aware supervision signals. In our case, the CLIP attribution maps provide a weak, spatially distributed form of guidance induced by superclass semantics.

D. Additional Experimental Details

D.1. Dataset Statistics

Waterbirds-95% statistics: Label set $\mathcal{Y} = \{\text{waterbird}, \text{landbird}\}$. Attribute set $\mathcal{Z} = \{\text{water}, \text{land}\}$.

Table 6: Dataset statistics for Waterbirds-95%.

Split	(waterbird, water)	(waterbird, land)	(landbird, water)	(landbird, land)
Train	1,057	56	184	3,498
Validation	133	133	466	467
Test	642	642	2,255	2,255

Waterbirds-100% statistics: Label set $\mathcal{Y} = \{\text{waterbird}, \text{landbird}\}$. Attribute set $\mathcal{Z} = \{\text{water}, \text{land}\}$.

Table 7: Dataset statistics for Waterbirds-100%.

Split	(waterbird, water)	(waterbird, land)	(landbird, water)	(landbird, land)
Train	1,101	0	0	3,694
Validation	133	133	466	467
Test	642	642	2,255	2,255

SpuCo Dogs statistics: Label set $\mathcal{Y} = \{\text{small dog}, \text{big dog}\}$. Attribute set $\mathcal{Z} = \{\text{indoor}, \text{outdoor}\}$.

Table 8: Dataset statistics for SpuCo Dogs.

Split	(big dog, indoor)	(big dog, outdoor)	(small dog, indoor)	(small dog, outdoor)
Train	500	10,000	10,000	500
Validation	25	500	500	25
Test	500	500	500	500

MetaShift statistics: Label set $\mathcal{Y} = \{\text{cat}, \text{dog}\}$. Attribute set

$$\mathcal{Z} = \{\text{sofa}, \text{bed}, \text{shelf}, \text{cabinet}, \text{bag}, \text{box}, \text{bench}, \text{bike}, \text{boat}, \text{surfboard}\}.$$

We consider four subsets in [37], each differing only in the two attributes paired with dog in the training data. According to the distances to (dog, shelf) reported in [37], these subsets are:

$$\begin{array}{lll} \text{(a)} & \{\text{cabinet, bed}\} & d = 0.44, \\ \text{(b)} & \{\text{bag, box}\} & d = 0.71, \\ \text{(c)} & \{\text{bench, bike}\} & d = 1.12, \\ \text{(d)} & \{\text{boat, surfboard}\} & d = 1.43. \end{array}$$

Larger d indicates a more challenging spurious correlation. We partition a portion of the test set into a validation set following a $15 : 85$ ratio, as in [38].

Table 9: Data statistics for MetaShift subset (a): cabinet & bed ($d = 0.44$).

Split	(cat, sofa)	(cat, bed)	(dog, cabinet)	(dog, bed)	(cat, shelf)	(dog, shelf)
Train	231	380	314	244	0	0
Validation	0	0	0	0	34	47
Test	0	0	0	0	201	259

Table 10: Data statistics for MetaShift subset (b): bag & box ($d = 0.71$).

Split	(cat, sofa)	(cat, bed)	(dog, bag)	(dog, box)	(cat, shelf)	(dog, shelf)
Train	231	380	202	193	0	0
Validation	0	0	0	0	34	47
Test	0	0	0	0	201	259

Table 11: Data statistics for MetaShift subset (c): bench & bike ($d = 1.12$).

Split	(cat, sofa)	(cat, bed)	(dog, bench)	(dog, bike)	(cat, shelf)	(dog, shelf)
Train	231	380	145	367	0	0
Validation	0	0	0	0	34	47
Test	0	0	0	0	201	259

Table 12: Data statistics for MetaShift subset (d): boat & surfboard ($d = 1.43$).

Split	(cat, sofa)	(cat, bed)	(dog, boat)	(dog, surfboard)	(cat, shelf)	(dog, shelf)
Train	231	380	459	318	0	0
Validation	0	0	0	0	34	47
Test	0	0	0	0	201	259

Spawrious statistics: Label set $\mathcal{Y} = \{\text{Bulldog}, \text{Dachshund}, \text{Corgi}, \text{Labrador}\}$. Attribute set

$$\mathcal{Z} = \{\text{Beach}, \text{Desert}, \text{Dirt}, \text{Jungle}, \text{Mountain}, \text{Snow}\}.$$

The Spawrious dataset includes two modes of spurious correlation: (1) One-to-one (O2O): each class is associated with exactly one attribute during training. At test time, the model encounters novel label–attribute combinations. (2) Many-to-many (M2M): a subset of classes is correlated with a subset of attributes during training, and this correlation is permuted in the test environment.

Each mode is divided into three subsets labeled as "easy," "medium," and "hard" following the original paper's naming convention, resulting in six subsets in total. For each subset, the original Spawrious dataset provides two training domains and one test domain. To align with the setup of other datasets, we merge the two training domains into a single training set, and for each group in the test domain, we split 10% of the test samples into a validation set.

Table 13: Data statistics for Spawrious subset: O2O–Easy

	Train I		Train II		Test
Bulldog	3,072 Desert	96 Beach	2,756 Desert	412 Beach	3,168 Dirt
Dachshund	3,072 Jungle	96 Beach	2,756 Jungle	412 Beach	3,168 Snow
Corgi	3,072 Snow	96 Beach	2,756 Snow	412 Beach	3,168 Jungle
Labrador	3,072 Dirt	96 Beach	2,756 Dirt	412 Beach	3,168 Desert

Table 14: Data statistics for Spawrious subset: O2O–Medium

	Train I		Train II		Test
Bulldog	3,072 Mountain	96 Desert	2,756 Mountain	412 Desert	3,168 Jungle
Dachshund	3,072 Beach	96 Desert	2,756 Beach	412 Desert	3,168 Dirt
Corgi	3,072 Jungle	96 Desert	2,756 Jungle	412 Desert	3,168 Snow
Labrador	3,072 Dirt	96 Desert	2,756 Dirt	412 Desert	3,168 Beach

Table 15: Data statistics for Spawrious subset: O2O–Hard

	Train I		Train II		Test
Bulldog	3,072 Jungle	96 Beach	2,756 Jungle	412 Beach	3,168 Mountain
Dachshund	3,072 Mountain	96 Beach	2,756 Mountain	412 Beach	3,168 Snow
Corgi	3,072 Desert	96 Beach	2,756 Desert	412 Beach	3,168 Jungle
Labrador	3,072 Snow	96 Beach	2,756 Snow	412 Beach	3,168 Desert

Table 16: Data statistics for Spawrious subset: M2M-Easy

	Train I	Train II	Test	
Bulldog	3,168 Desert	3,168 Mountain	3,168 Dirt	3,168 Jungle
Dachshund	3,168 Mountain	3,168 Desert	3,168 Dirt	3,168 Jungle
Corgi	3,168 Jungle	3,168 Dirt	3,168 Desert	3,168 Mountain
Labrador	3,168 Dirt	3,168 Jungle	3,168 Desert	3,168 Mountain

Table 17: Data statistics for Spawrious subset: M2M-Medium

	Train I	Train II	Test	
Bulldog	3,168 Beach	3,168 Snow	3,168 Desert	3,168 Mountain
Dachshund	3,168 Snow	3,168 Beach	3,168 Desert	3,168 Mountain
Corgi	3,168 Desert	3,168 Mountain	3,168 Beach	3,168 Snow
Labrador	3,168 Mountain	3,168 Desert	3,168 Beach	3,168 Snow

Table 18: Data statistics for Spawrious subset: M2M-Hard

	Train I	Train II	Test	
Bulldog	3,168 Dirt	3,168 Jungle	3,168 Snow	3,168 Beach
Dachshund	3,168 Jungle	3,168 Dirt	3,168 Snow	3,168 Beach
Corgi	3,168 Beach	3,168 Snow	3,168 Dirt	3,168 Jungle
Labrador	3,168 Snow	3,168 Beach	3,168 Dirt	3,168 Jungle

D.2. Hyperparameter Selection

SupER. Our SupER model employs a β-VAE encoder built upon the ResNet50 backbone architecture. For consistency, the CLIP model also uses ResNet50. We perform a grid search to assess the performance of SupER under different hyperparameter configurations and select the optimal values

for each dataset as summarized in Table 19. Specifically, the hyperparameters are as follows: β denotes the weighting factor of β-VAE; λ_1 is the weight for the loss $\mathcal{L}^{\text{Beta}}_{\theta,\phi}(\mathbf{x})$; λ_2 is the weight for the loss $\mathcal{L}^{\text{ATT}}_{\phi,\omega_{\text{rel}},\omega_{\text{irr}}}(\mathbf{x}, y)$; λ_3 controls the L_2 regularization term $||\omega_{\text{rel}}||_2^2$, where n_1 denotes the number of parameters in ω_{rel}; η is the learning rate; B is the batch size; T is the number of epochs; γ denotes the weight decay coefficient used in the Adam optimizer; and d specifies the dimensionality of features $\mathbf{z}_{\text{rel}}$ and $\mathbf{z}_{\text{irr}}$. Early stopping is adopted when applicable, and training is terminated once the worst-group accuracy on the validation set reaches its maximum. Note that our method does not require any group information; this criterion is used only for fair comparison with previous work. For the number of superclass-specific text prompts m, unless stated otherwise, we set $m = 1$. The text prompts used for each dataset are detailed in Table 20.

Table 19: SupER hyperparameter settings across datasets.

Dataset	β	λ_1	λ_2	λ_3	η	B	T	γ	d
Waterbirds-95%	1	1	40	$1000/n_1$	10^{-5}	32	50	10^{-4}	256
Waterbirds-100%	1	1	40	$1000/n_1$	10^{-5}	32	50	10^{-4}	256
SpuCo Dogs	1	1	40	$100/n_1$	10^{-6}	32	30	10^{-2}	256
MetaShift (a)	5	1	1	$100/n_1$	10^{-5}	32	100	10^{-2}	256
MetaShift (b)	5	1	1	$100/n_1$	10^{-5}	32	100	10^{-2}	256
MetaShift (c)	5	1	20	$100/n_1$	10^{-5}	32	100	10^{-2}	256
MetaShift (d)	10	1	20	$100/n_1$	10^{-5}	32	100	10^{-2}	256
Spawrious O2O–Easy	10	1	10	$100/n_1$	10^{-6}	32	30	10^{-4}	256
Spawrious O2O–Medium	1	1	80	$100/n_1$	10^{-6}	32	30	10^{-4}	256
Spawrious O2O–Hard	1	1	80	$100/n_1$	10^{-6}	32	30	10^{-4}	256
Spawrious M2M–Easy	10	1	50	$100/n_1$	10^{-6}	32	30	10^{-4}	256
Spawrious M2M–Medium	1	1	50	$100/n_1$	10^{-6}	32	30	10^{-4}	256
Spawrious M2M–Hard	1	1	50	$100/n_1$	10^{-6}	32	30	10^{-4}	256

Table 20: Superclass text prompts for each dataset

Dataset	Prompt
Waterbirds-95%	`a bird`
Waterbirds-100%	`a bird`
SpuCo Dogs	`a dog`
MetaShift	`a cat or a dog`
Spawrious	`a dog`

Baselines. For baseline methods considered in our experiments, we similarly employ ResNet50 backbone architectures and determine their optimal hyperparameters via grid search. We specifically evaluate learning rates $\eta \in \{10^{-6}, 10^{-5}, 10^{-4}\}$ and weight decay $\gamma \in \{10^{-4}, 10^{-2}\}$, with the batch size and number of training epochs for each dataset as specified in Table 19. Note that for all the above configurations, as well as additional model-specific hyperparameters, we directly use the values provided or recommended in the original papers whenever available.

D.3. Full Worst Group Accuracy, Average Accuracy, and Group Accuracy Variance for All Datasets

Worst group and average accuracy. Tables 21, 22, 23, 24, 25, and 26 summarize the worst group accuracy and average accuracy for all datasets and selected baseline methods. **Bold** indicates the best across all selected baselines; Underlined indicates the best among methods without group information; "–" indicates omitted result due to consistently subpar or unstable performance, even after comprehensive hyperparameter tuning using the original codebase.

Table 21: Worst and average group accuracy (%) for Waterbirds-95% and Waterbirds-100%.

Method	Group Info	Train Twice	Waterbirds-95%		Waterbirds-100%	
			Worst	Avg	Worst	Avg
ERM	×	×	$64.9_{\pm 1.5}$	$90.7_{\pm 1.0}$	$46.4_{\pm 6.9}$	$74.8_{\pm 3.0}$
CVaR DRO	×	×	$73.1_{\pm 7.1}$	$90.7_{\pm 0.7}$	$58.0_{\pm 2.2}$	$79.0_{\pm 1.2}$
LfF	×	×	$79.1_{\pm 2.5}$	$\underline{91.9}_{\pm 0.7}$	$61.5_{\pm 2.8}$	$80.6_{\pm 1.2}$
GALS	×	×	$75.4_{\pm 2.2}$	$89.0_{\pm 0.5}$	$55.0_{\pm 5.5}$	$79.7_{\pm 0.4}$
JTT	×	✓	$86.4_{\pm 1.0}$	$89.5_{\pm 0.5}$	$61.3_{\pm 5.5}$	$79.7_{\pm 3.0}$
CnC	×	✓	$\underline{86.5}_{\pm 5.9}$	$91.0_{\pm 0.5}$	$62.1_{\pm 0.9}$	$81.9_{\pm 1.5}$
SupER (Ours)	×	×	$84.4_{\pm 2.3}$	$87.3_{\pm 0.6}$	$\underline{\mathbf{79.7}}_{\pm 1.7}$	$\underline{\mathbf{85.0}}_{\pm 1.4}$
UW	✓	×	$89.3_{\pm 1.5}$	$\mathbf{94.5}_{\pm 0.9}$	$56.4_{\pm 2.3}$	$78.6_{\pm 0.8}$
IRM	✓	×	$76.2_{\pm 6.3}$	$89.4_{\pm 0.9}$	$57.0_{\pm 5.4}$	$80.5_{\pm 5.0}$
GroupDRO	✓	×	$87.2_{\pm 1.3}$	$93.2_{\pm 0.4}$	$56.5_{\pm 1.4}$	$79.4_{\pm 0.3}$
DFR	✓	✓	$\mathbf{89.7}_{\pm 2.4}$	$93.6_{\pm 0.6}$	$48.2_{\pm 0.4}$	$76.4_{\pm 0.2}$

Table 22: Worst group accuracy (%) for the six Spawrious subsets.

Method	Group Info?	Train Twice?	One–To–One			Many–To–Many			Average
			Easy	Medium	Hard	Easy	Medium	Hard	
ERM	×	×	$78.4_{\pm 1.8}$	$63.4_{\pm 2.3}$	$71.1_{\pm 3.7}$	$72.9_{\pm 1.3}$	$52.7_{\pm 2.9}$	$50.7_{\pm 1.0}$	$64.9_{\pm 11.3}$
CVaR DRO	×	×	$81.7_{\pm 0.5}$	$66.4_{\pm 1.4}$	$61.2_{\pm 1.6}$	$69.7_{\pm 0.8}$	$50.3_{\pm 3.9}$	$45.9_{\pm 0.2}$	$62.5_{\pm 13.1}$
LfF	×	×	$74.6_{\pm 7.7}$	–	$62.9_{\pm 3.6}$	$72.7_{\pm 3.5}$	$50.0_{\pm 4.0}$	$48.6_{\pm 3.7}$	–
GALS	×	×	$89.1_{\pm 1.9}$	$60.0_{\pm 5.4}$	$81.0_{\pm 3.0}$	$74.0_{\pm 4.8}$	$44.9_{\pm 0.3}$	$46.9_{\pm 2.4}$	$66.0_{\pm 18.3}$
JTT	×	✓	$80.9_{\pm 2.1}$	–	$59.7_{\pm 4.9}$	$71.2_{\pm 2.0}$	$49.7_{\pm 3.5}$	$45.2_{\pm 1.8}$	–
CnC	×	✓	$\underline{\mathbf{90.0}}_{\pm 1.4}$	$73.5_{\pm 4.6}$	$81.3_{\pm 3.1}$	$82.8_{\pm 2.1}$	$62.5_{\pm 5.2}$	$78.7_{\pm 4.9}$	$78.1_{\pm 9.4}$
SupER (Ours)	×	×	$82.7_{\pm 2.0}$	$\underline{\mathbf{80.3}}_{\pm 4.6}$	$\underline{\mathbf{83.8}}_{\pm 3.4}$	$\underline{\mathbf{87.4}}_{\pm 1.3}$	$\underline{\mathbf{83.4}}_{\pm 2.3}$	$\underline{\mathbf{79.9}}_{\pm 4.7}$	$\underline{\mathbf{82.9}}_{\pm 2.7}$
UW	✓	×	$87.4_{\pm 1.1}$	$67.9_{\pm 2.1}$	$75.9_{\pm 2.9}$	$72.9_{\pm 1.3}$	$52.7_{\pm 2.9}$	$50.7_{\pm 1.0}$	$67.9_{\pm 14.1}$
IRM	✓	×	$78.4_{\pm 1.0}$	$64.5_{\pm 3.2}$	$64.9_{\pm 2.2}$	$77.9_{\pm 3.7}$	$57.1_{\pm 2.9}$	$50.7_{\pm 1.1}$	$65.6_{\pm 11.1}$
GroupDRO	✓	×	$86.7_{\pm 1.2}$	$67.2_{\pm 0.7}$	$76.4_{\pm 2.2}$	$74.3_{\pm 0.9}$	$55.7_{\pm 1.4}$	$49.9_{\pm 0.8}$	$68.3_{\pm 13.7}$
DFR	✓	✓	$79.1_{\pm 5.2}$	$64.3_{\pm 1.9}$	$70.0_{\pm 1.9}$	$76.4_{\pm 1.9}$	$58.7_{\pm 2.2}$	$54.1_{\pm 2.2}$	$67.1_{\pm 9.9}$

Table 23: Average accuracy (%) for the six Spawrious subsets.

Method	Group Info?	Train Twice?	One–To–One			Many–To–Many			Average
			Easy	Medium	Hard	Easy	Medium	Hard	
ERM	×	×	$85.5_{\pm 2.6}$	$76.7_{\pm 1.3}$	$82.0_{\pm 1.0}$	$89.5_{\pm 0.6}$	$74.5_{\pm 1.4}$	$70.7_{\pm 2.1}$	$79.8_{\pm 7.1}$
CVaR DRO	×	×	$89.4_{\pm 0.1}$	$86.0_{\pm 3.7}$	$80.7_{\pm 0.6}$	$88.5_{\pm 0.6}$	$74.0_{\pm 1.0}$	$67.7_{\pm 0.6}$	$81.0_{\pm 8.7}$
LfF	×	×	$84.1_{\pm 1.5}$	–	$76.9_{\pm 1.2}$	$89.6_{\pm 0.8}$	$73.8_{\pm 2.6}$	$69.1_{\pm 1.1}$	–
GALS	×	×	$93.5_{\pm 0.9}$	$86.6_{\pm 0.9}$	$90.0_{\pm 0.4}$	$87.8_{\pm 0.2}$	$74.0_{\pm 0.3}$	$69.8_{\pm 1.9}$	$83.6_{\pm 9.5}$
JTT	×	✓	$86.1_{\pm 1.3}$	–	$77.5_{\pm 1.7}$	$89.2_{\pm 0.5}$	$72.8_{\pm 0.9}$	$66.6_{\pm 0.8}$	–
CnC	×	✓	$\underline{\mathbf{94.4}}_{\pm 1.1}$	$87.8_{\pm 2.5}$	$89.6_{\pm 0.9}$	$92.6_{\pm 1.0}$	$80.8_{\pm 4.0}$	$88.8_{\pm 1.2}$	$89.0_{\pm 4.7}$
SupER (Ours)	×	×	$90.9_{\pm 0.5}$	$\underline{\mathbf{90.1}}_{\pm 3.2}$	$\underline{\mathbf{90.5}}_{\pm 2.0}$	$\underline{\mathbf{94.9}}_{\pm 0.9}$	$\underline{\mathbf{91.6}}_{\pm 1.8}$	$\underline{\mathbf{91.4}}_{\pm 1.5}$	$\underline{\mathbf{91.6}}_{\pm 1.7}$
UW	✓	×	$93.5_{\pm 0.3}$	$82.6_{\pm 0.7}$	$86.5_{\pm 0.6}$	$89.5_{\pm 0.6}$	$74.5_{\pm 1.4}$	$70.7_{\pm 2.1}$	$82.9_{\pm 8.8}$
IRM	✓	×	$87.3_{\pm 0.3}$	$76.9_{\pm 0.4}$	$82.7_{\pm 0.4}$	$90.9_{\pm 0.9}$	$76.7_{\pm 2.6}$	$71.2_{\pm 1.0}$	$80.9_{\pm 7.4}$
GroupDRO	✓	×	$92.7_{\pm 0.3}$	$89.5_{\pm 0.3}$	$86.5_{\pm 1.5}$	$89.4_{\pm 0.6}$	$77.3_{\pm 0.5}$	$68.4_{\pm 1.7}$	$84.0_{\pm 9.3}$
DFR	✓	✓	$87.5_{\pm 3.3}$	$80.9_{\pm 1.1}$	$79.4_{\pm 1.3}$	$89.4_{\pm 0.4}$	$75.1_{\pm 0.1}$	$72.4_{\pm 1.9}$	$80.8_{\pm 6.7}$

Table 24: Worst group accuracy (%) for the four MetaShift subsets.

Method	Group Info?	Train Twice?	MetaShift Subsets				Average
			(a) $d = 0.44$	(b) $d = 0.71$	(c) $d = 1.12$	(d) $d = 1.43$	
ERM	×	×	$78.8_{\pm 1.0}$	$75.8_{\pm 0.8}$	$61.9_{\pm 5.9}$	$52.6_{\pm 2.6}$	$67.3_{\pm 12.2}$
CVaR DRO	×	×	$77.8_{\pm 2.5}$	$72.5_{\pm 2.8}$	$65.1_{\pm 0.2}$	$54.7_{\pm 3.2}$	$67.5_{\pm 10.0}$
LfF	×	×	$77.2_{\pm 1.7}$	$73.9_{\pm 0.6}$	$69.5_{\pm 1.0}$	$59.5_{\pm 3.1}$	$70.0_{\pm 7.7}$
GALS	×	×	$74.8_{\pm 3.9}$	$68.8_{\pm 2.0}$	$70.6_{\pm 2.2}$	$50.0_{\pm 0.9}$	$66.0_{\pm 11.0}$
JTT	×	✓	$76.7_{\pm 2.3}$	$73.2_{\pm 0.8}$	$67.1_{\pm 4.6}$	$53.0_{\pm 1.6}$	$67.5_{\pm 10.4}$
CnC	×	✓	$\underline{\mathbf{81.1}}_{\pm 1.4}$	$71.4_{\pm 2.4}$	$65.4_{\pm 6.8}$	$49.6_{\pm 1.6}$	$66.9_{\pm 13.2}$
SupER (Ours)	×	×	$79.8_{\pm 3.6}$	$\underline{\mathbf{78.4}}_{\pm 1.9}$	$\underline{\mathbf{77.6}}_{\pm 2.1}$	$\underline{\mathbf{71.4}}_{\pm 2.1}$	$\underline{\mathbf{76.8}}_{\pm 3.7}$

Table 25: Average accuracy (%) for the four MetaShift subsets.

Method	Group Info?	Train Twice?	MetaShift Subsets				Average
			(a) $d = 0.44$	(b) $d = 0.71$	(c) $d = 1.12$	(d) $d = 1.43$	
ERM	×	×	$80.5_{\pm 0.8}$	$78.0_{\pm 0.2}$	$73.6_{\pm 0.5}$	$69.2_{\pm 1.3}$	$75.3_{\pm 5.0}$
CVaR DRO	×	×	$80.7_{\pm 1.2}$	$78.2_{\pm 0.3}$	$74.6_{\pm 0.5}$	$69.8_{\pm 2.1}$	$75.8_{\pm 4.7}$
LfF	×	×	$79.2_{\pm 0.9}$	$77.2_{\pm 1.4}$	$74.8_{\pm 1.1}$	$69.1_{\pm 0.7}$	$75.1_{\pm 4.4}$
GALS	×	×	$80.5_{\pm 1.8}$	$77.4_{\pm 1.2}$	$78.3_{\pm 0.7}$	$69.1_{\pm 1.3}$	$76.3_{\pm 5.0}$
JTT	×	✓	$80.8_{\pm 1.3}$	$76.4_{\pm 1.0}$	$73.2_{\pm 0.5}$	$69.3_{\pm 0.6}$	$74.9_{\pm 4.9}$
CnC	×	✓	$\underline{\mathbf{82.1}}_{\pm 1.4}$	$77.0_{\pm 2.2}$	$74.4_{\pm 1.6}$	$66.7_{\pm 1.4}$	$75.1_{\pm 6.4}$
SupER (Ours)	×	×	$81.7_{\pm 1.9}$	$\underline{\mathbf{80.5}}_{\pm 1.4}$	$\underline{\mathbf{79.2}}_{\pm 1.9}$	$\underline{\mathbf{76.6}}_{\pm 1.4}$	$\underline{\mathbf{79.5}}_{\pm 2.2}$

Table 26: Worst and average group accuracy (%) for Spuco Dogs.

Method	Group Info?	Train Twice?	Spuco Dogs	
			Worst	Avg
ERM	×	×	$54.5_{\pm 1.3}$	$77.4_{\pm 1.6}$
CVaR DRO	×	×	$56.3_{\pm 3.1}$	$78.5_{\pm 2.2}$
LfF	×	×	$52.6_{\pm 2.5}$	$77.1_{\pm 1.9}$
GALS	×	×	–	–
JTT	×	✓	$50.4_{\pm 0.2}$	$77.9_{\pm 0.1}$
CnC	×	✓	$65.6_{\pm 0.7}$	$\underline{82.0}_{\pm 0.5}$
SupER (Ours)	×	×	$\underline{69.7}_{\pm 4.4}$	$76.0_{\pm 2.3}$
UW	✓	×	$\mathbf{84.7}_{\pm 2.0}$	$87.4_{\pm 0.5}$
IRM	✓	×	$50.0_{\pm 5.5}$	$75.2_{\pm 5.7}$
GroupDRO	✓	×	$83.8_{\pm 0.4}$	$\mathbf{87.6}_{\pm 0.5}$
DFR	✓	✓	$71.3_{\pm 4.4}$	$83.3_{\pm 2.8}$

Variance of accuracy across groups. Tables 27, 28, and 29 summarize the variance of accuracy across groups for all datasets and selected baseline methods. **Bold** indicates the smallest across all selected baselines; Underlined indicates the smallest among methods without group information. ; "–" indicates omitted result due to consistently subpar or unstable performance, even after comprehensive hyperparameter tuning using the original codebase.

Table 27: Variance of accuracy across groups (%) for Waterbirds-95%, Waterbirds-100%, and SpuCo Dogs.

Method	Waterbirds-95%	Waterbirds-100%	SpuCo Dogs
ERM	245.9	778.1	603.9
CVaR DRO	154.6	528.8	558.5
LfF	89.2	442.1	582.9
GALS	126.7	516.5	–
JTT	6.1	405.0	621.4
CnC	16.3	347.6	261.7
SupER (Ours)	**6.0**	**16.0**	28.4
UW	12.7	536.2	**6.5**
IRM	127.2	479.8	776.7
GroupDRO	28.0	495.1	10.0
DFR	14.2	573.0	282.9

Table 28: Variance of accuracy across groups (%) for the four MetaShift subsets.

Method	(a) $d = 0.44$	(b) $d = 0.71$	(c) $d = 1.12$	(d) $d = 1.43$
ERM	10.3	14.2	411.8	722.1
CVaR DRO	21.3	97.5	237.7	599.7
LfF	15.3	30.3	73.3	258.8
GALS	82.1	197.7	157.7	955.6
JTT	31.2	24.9	128.6	699.8
CnC	**2.3**	71.5	262.5	769.6
SupER	9.3	**13.1**	**7.7**	**49.4**

Table 29: Variance of accuracy across groups (%) for the six Spawrious subsets.

Method	O2O-Easy	O2O-Medium	O2O-Hard	M2M-Easy	M2M-Medium	M2M-Hard
ERM	50.9	109.8	101.1	92.6	246.3	373.8
CVaR DRO	63.3	261.0	241.4	109.4	323.0	469.0
LfF	88.1	–	254.2	80.9	290.2	413.8
GALS	**12.6**	490.8	61.3	161.6	563.8	584.9
JTT	31.4	–	310.2	108.3	293.2	473.2
CnC	19.2	169.2	62.9	48.1	129.1	**47.1**
SupER (Ours)	64.7	**92.1**	**41.2**	**22.8**	**33.8**	58.0
UW	29.3	109.8	99.6	92.6	246.3	373.8
IRM	76.5	107.8	183.8	70.0	291.2	383.4
GroupDRO	33.5	221.7	85.5	79.0	190.7	397.1
DFR	70.3	359.1	149.3	68.0	293.1	327.5

CLIP guidance. The goal of SupER is fundamentally different from extracting or replicating CLIP's features. Instead, CLIP only provides superclass guidance and does not contribute any information useful for distinguishing class labels, since the superclass is shared across different class labels. Moreover, in Table 30 we report the performance of directly using CLIP for prediction compared to SupER. Directly applying CLIP leads to a noticeable drop in accuracy, which suggests that CLIP itself may also rely on spurious correlations. Therefore, using CLIP as superclass guidance can both give SupER enough autonomy to learn features on its own, and avoid the spurious correlations that CLIP might exploit for fine-grained class prediction.

Table 30: Comparison of worst group accuracy (%) between CLIP and SupER on Waterbirds. CLIP (zero-shot) means directly using the pretrained CLIP model for classification. CLIP (fine-tuned) denotes standard fine-tuning of CLIP on the downstream dataset.

Method	Waterbirds-95%	Waterbirds-100%
CLIP (zero-shot)	41.6	47.9
CLIP (fine-tuned)	70.2	48.8
SupER	84.4	79.7

D.4. Visualization Results

SupER achieves effective disentanglement of superclass-relevant and irrelevant features. Figures 5 illustrates gradient-based attention visualizations from one representative samples per subset across all datasets. For each sample, we present the original image, GradCAM attribution maps from the ERM baseline, CLIP, SupER's ω_{rel} and ω_{irr}. The results show that SupER consistently succeeds in separating superclass-relevant and superclass-irrelevant features by leveraging guidance from CLIP across diverse datasets.

SupER can adjust internal biases in CLIP. Figure 6 illustrates gradient-based attention visualizations from one representative sample per subset across all datasets. Each sample includes the original image, GradCAM attribution maps from CLIP, SupER's classifiers (ω_{rel}, ω_{irr}), and an illustration of the primary issue observed in CLIP's attention (e.g., focusing on incomplete or incorrect features). The results demonstrate that SupER, by emphasizing feature disentanglement, can effectively mitigate internal biases in CLIP's attention.

D.5. Ablation Results

In this section, we examine the contributions of the core components of SupER. We focus on (i) text prompts, (ii) the strength of feature disentanglement, and (iii) the degree of superclass guidance, as these constitute the primary design elements of the method. We also study the impact of L_2 regularization and the relative strength of the two classifiers. To better isolate the effect of each factor, we keep all other hyperparameters fixed during each ablation study. This includes adopting a consistent protocol for random-seed selection across repeated trials, while in Appendix D.3, we do not enforce fixed random seeds across runs.

Text prompt. We evaluate the impact of text prompt configurations across all datasets. Tables 32, 33, 34, and 35 present the change in worst group accuracy relative to the reference setting for the Spawrious, MetaShift, Waterbirds, and SpuCo Dogs datasets, respectively. The exact text prompts used are listed in Table 31. Overall, performance tends to degrade as the number of prompts increases and as the superclass becomes more abstract.

Beyond the above prompt ablations, we further study how SupER behaves when the superclass prompt is intentionally mismatched with our definition in Section 3.1. In Table 36, on Waterbirds we intentionally replace the reference superclass prompt "bird" with "waterfowl" and "songbird" such that they only cover a strict subset of the intended superclass. We observe a consistent but relatively small drop in worst-group accuracy compared to the reference setting, likely because the CLIP attribution maps induced by "waterfowl" and "songbird" remain visually close to those induced by "bird".

We also explore whether superclasses can be extracted automatically from fine-grained class names using a large language model, which would further reduce manual prompt engineering. Specifically, we use *Qwen2.5-7B-Instruct* [55] and query it with the following prompt:

```
We have fine-grained class names: {fine-grained class label}. Give ONE
English word that is their shared superclass. Return ONLY the single word,
lowercase, no punctuation.
```

The extracted superclasses are highly reasonable across datasets (e.g., "bird" for Waterbirds, "dog" for Spawrious, and "animal" for MetaShift), suggesting that LLM-based extraction is a practical way to define superclasses and may help avoid inadvertent human specification errors.

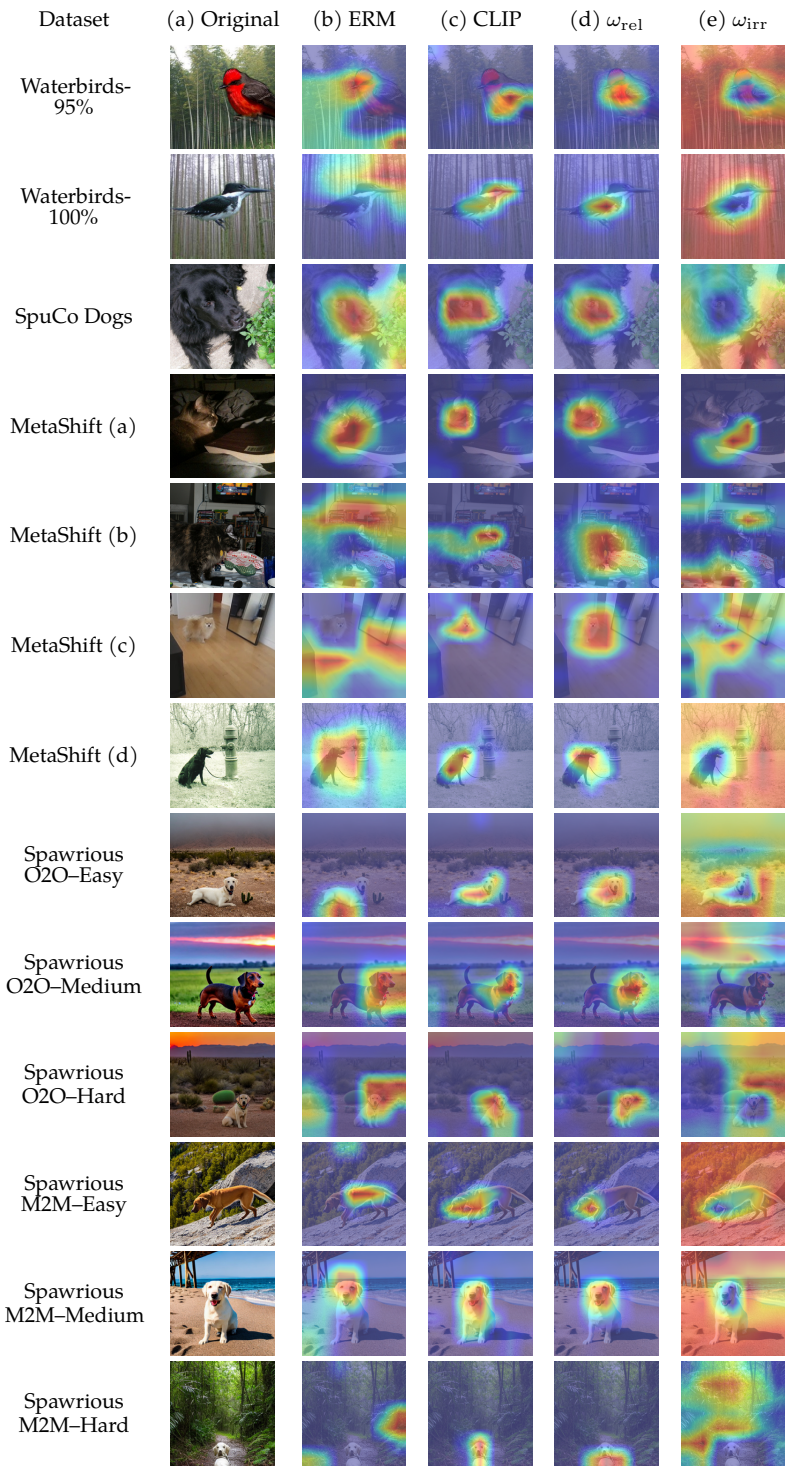


Figure 5: Visualization of GradCAM maps across all datasets to assess feature disentanglement. Each row corresponds to one representative sample per dataset subset. Columns (a)–(e) show: the original image, GradCAM maps from ERM, CLIP, SupER's classifier ω_{rel} (superclass-relevant), and ω_{irr} (superclass-irrelevant).

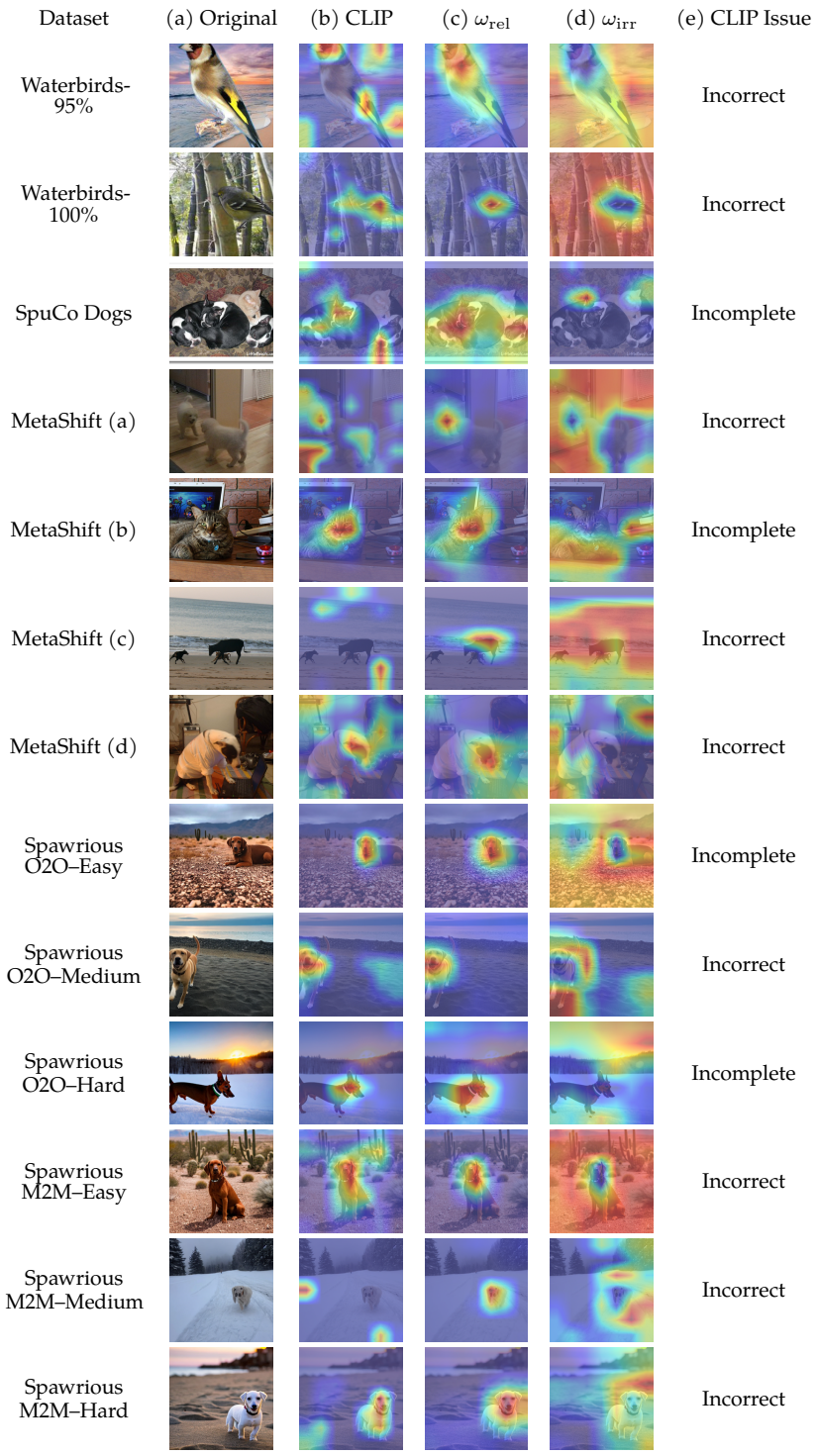


Figure 6: Visualization of GradCAM maps highlighting CLIP's internal bias and SupER's correction. Each row presents one representative sample per dataset subset. Columns (a)-(e) show: original image, GradCAM maps from CLIP, SupER's classifier ω_{rel} (superclass-relevant), ω_{irr} (superclass-irrelevant) and an illustration of the primary CLIP bias.

Table 31: Prompt variants used for different values of m. Each prompt includes the superclass placeholder, formatted as `a/an [superclass]`.

#Prompts (m)	Prompt Variant
1	`a/an [superclass]`
2	`a/an [superclass]` `a photo of a/an [superclass]`

5	`a/an [superclass]`
	`a photo of a/an [superclass]`
	`a picture of a/an [superclass]`
	`an image of a/an [superclass]`
	`a/an [superclass] photograph`

Table 32: Ablation results on Spawrious under different prompt configurations. All values indicate the change in worst group accuracy (%) relative to the setting $m = 1$, superclass = `dog`.

#Prompts	Superclass	O2O-Easy	O2O-Medium	O2O-Hard	M2M-Easy	M2M-Medium	M2M-Hard
1	`dog`	0.0	0.0	0.0	0.0	0.0	0.0
2	`dog`	+0.7	-2.2	-2.3	+0.8	-9.6	-2.4
5	`dog`	+0.3	-2.5	-2.9	-1.1	-9.4	+1.1
1	`animal`	-4.0	-1.1	-5.6	-4.9	-7.5	-5.6

Table 33: Ablation results on MetaShift under different prompt configurations. All values indicate the change in worst group accuracy (%) relative to the setting $m = 1$, superclass = `dog or cat`.

#Prompts	Superclass	(a) $d = 0.44$	(b) $d = 0.71$	(c) $d = 1.12$	(d) $d = 1.43$
1	`dog or cat`	0.0	0.0	0.0	0.0
2	`dog or cat`	-1.5	+0.4	-0.9	-2.4
5	`dog or cat`	-0.8	-0.8	-2.5	-0.1
1	`animal`	-1.1	-0.3	-8.3	-6.3

Table 34: Ablation results on Waterbirds-95% and Waterbirds-100% under different prompt configurations. All values indicate the change in worst group accuracy (%) relative to the setting $m = 1$, superclass = `bird`.

#Prompts	Superclass	Waterbirds-95%	Waterbirds-100%
1	`bird`	0.0	0.0
2	`bird`	-2.6	+3.2
5	`bird`	-1.1	-2.2
1	`animal`	-29.2	-44.8

Table 35: Ablation results on SpuCo Dogs under different prompt configurations. All values indicate the change in worst group accuracy (%) relative to the setting $m = 1$, superclass = `dog`.

#Prompts	Superclass	SpuCo Dogs
1	`dog`	0.0
2	`dog`	-0.5
5	`dog`	+0.9
1	`animal`	-14.1

Table 36: Superclass mismatch results on Waterbirds-95% and Waterbirds-100%. All values indicate the change in worst group accuracy (%) relative to the reference setting superclass = `bird`.

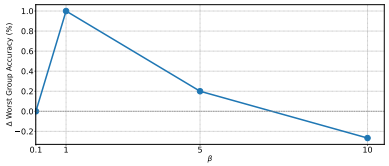
Superclass	Waterbirds-95%	Waterbirds-100%
`bird`	0.0	0.0
`waterfowl`	-0.6	-0.2
`songbird`	-2.8	-1.2

Feature disentanglement strength. We evaluate the effect of varying the feature disentanglement coefficient β across all datasets. Figure 7 shows the worst group accuracy as β changes on selected datasets. Overall, both insufficient feature disentanglement (i.e., low β) and excessive disentanglement (i.e., overly large β) can lead to degraded model performance.

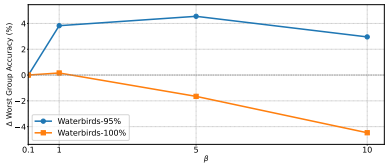
Degree of superclass guidance. We evaluate the effect of varying the superclass guidance weight λ_2 across all datasets. Figure 8 reports the worst group accuracy under different values of λ_2 on selected datasets. Overall, both insufficient guidance (i.e., low λ_2) and overly strong guidance (i.e., excessively large λ_2) can lead to degraded model performance.

L_2 regularization. We ablate the L_2 term $\|\omega_{\text{rel}}\|_2^2$ in Algorithm 1. Table 37 reports the change in worst group accuracy when removing L_2 regularization (i.e., setting $\lambda_3 = 0$) on selected datasets. Overall, removing L_2 degrades performance, which indicates that encouraging the use of all superclass-relevant features improves domain generalization.

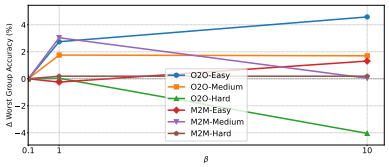
Relative strength of the two classifiers. In Algorithm 1, the losses $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{rel}}}(\mathbf{x}, y)$ and $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{irr}}}(\mathbf{x}, y)$ are, by default, weighted equally. Here we fix the coefficient of $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{rel}}}(\mathbf{x}, y)$ to 1 and vary the weight on $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{irr}}}(\mathbf{x}, y)$ to assess its effect on performance. Results in Table 38 show that either ignoring or overemphasizing ω_{irr} degrades performance, and a balanced strength between ω_{rel} and ω_{irr} better support guidance for both superclass-relevant and irrelevant features.



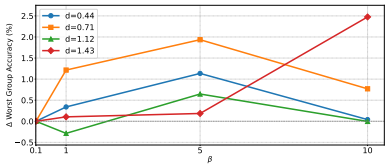
(a) Effect of β on SpuCo Dogs relative to the $\beta = 0.1$ setting.



(b) Effect of β on Waterbirds relative to the $\beta = 0.1$ setting.

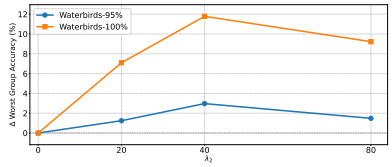


(c) Effect of β on Spawrious relative to the $\beta = 0.1$ setting.

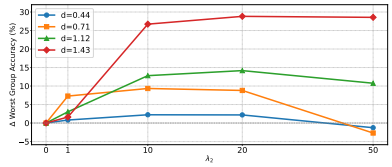


(d) Effect of β on MetaShift relative to the $\beta = 0.1$ setting.

Figure 7: Ablation of feature disentanglement strength β across all datasets.



(a) Effect of λ_2 on Waterbirds relative to the $\lambda_2 = 0$ setting.



(b) Effect of λ_2 on MetaShift relative to the $\lambda_2 = 0$ setting.

Figure 8: Ablation of the degree of superclass guidance λ_2 on Waterbirds and MetaShift.

Table 37: Ablation results when removing L_2 regularization. All values indicate the drop in worst group accuracy (%) relative to the full model. For MetaShift, we report the mean over the four subsets (a)–(d).

	Waterbirds-95%	Waterbirds-100%	MetaShift	SpuCo Dogs
$\lambda_3 = 0$	-1.2	-0.1	-1.1	-3.5

Table 38: Results on Waterbirds-95%, Waterbirds-100%, MetaShift (mean of 4 subsets), and SpuCo Dogs under different cross-entropy loss weights on ω_{irr}. The cross-entropy loss weight on ω_{rel} is fixed at 1. All values indicate the change in worst group accuracy (%) relative to the setting with weight $= 1$.

Weight	Waterbirds-95%	Waterbirds-100%	MetaShift	SpuCo Dogs
0	-3.5	-4.7	+1.1	-7.3
1	0.0	0.0	0.0	0.0
10	-1.4	-1.1	-1.3	-5.2

D.6. Compute Resources

We used a single NVIDIA A100-SXM4 GPU (40 GB VRAM), an Intel Xeon CPU @ 2.20 GHz with 12 cores, and 83 GB of system RAM. Table 39 shows the average time per epoch (in seconds) for each dataset. For epoch counts and specific hyperparameters, see Appendix D.2.

Table 39: Average time per epoch (s) for each dataset

Dataset	Time per epoch (s)
Waterbirds-95% & 100%	41
SpuCo Dogs	216
MetaShift	12
Spawrious	195

E. SupER under Internal Spurious Correlation

In Section 4, we have already demonstrated that SupER achieves significant generalization improvements under various types and degrees of spurious correlations, in particular when new groups appear at test time and when spurious features in the training data are perfectly correlated with the labels. In this section, we further consider a special case where prior knowledge indicates that spurious correlations arise entirely within the superclass. We examine this scenario because the superclass guidance from CLIP is now less dominant compared to the contributions of the β-VAE and L_2 regularization. Importantly, this does not present a contradiction: SupER is designed to address the more general setting in Section 3.1, where both the source of spurious features and the group distribution of training and test data are unknown, and Theorems 1-3 already establishes SupER as the optimal choice under this setting. On the other hand, when prior knowledge suggests that spurious correlations arise entirely within the superclass, we show in this section that SupER still exhibits relatively effective performance and can be further enhanced by integrating it with existing approaches that do not require group annotations.

Datasets. We evaluate SupER on BFFHQ [17], CelebA [9], and Color MNIST [7, 10]. Specifically, BFFHQ contains spurious correlations between age (a superclass-relevant feature) and gender labels, with the minority group ratio being only 0.5%. CelebA is a large-scale dataset exhibiting spurious correlations between hair color (a superclass-relevant feature) and gender. The data statistics of BFFHQ and CelebA are shown in Tables 40 and 41, respectively. In addition, Color MNIST introduces a spurious correlation between color (a superclass-relevant feature) and the label y. In this setting, the target label is $y \in \mathcal{Y} = \{(0,1),(2,3),(4,5),(6,7),(8,9)\}$, the spurious attribute z takes one of five

colors, and the spurious correlation ratio is 99.5% during training. For evaluation, Color MNIST adopt a regime where colors are assigned uniformly at random to each sample. Note that we use human as the superclass for both BFFHQ and CelebA, and digit as the superclass for Color MNIST.

Table 40: Dataset statistics for BFFHQ.

Split	(young, female)	(young, male)	(old, female)	(old, male)
Train	9,552	48	48	9,552
Test	250	250	250	250

Table 41: Dataset statistics for CelebA.

Split	(not blond, female)	(not blond, male)	(blond, female)	(blond, male)
Train	71,629	66,874	22,880	1,387
Test	9,767	7,535	2,480	180

Results. First, we present the performance of SupER on BFFHQ. Following [17], we report the accuracy on the bias-conflicting groups, i.e., the accuracy of two minority groups (young, male) and (old, female). As shown in Table 42, in the dataset with a strong spurious correlation (99.5%) and highly complex spurious feature (age), SupER still achieves competitive performance compared to other baselines that do not require group labels.

Table 42: Bias-conflicting group accuracy (%) on BFFHQ for SupER and baselines that do not require group labels. For baselines, we use reported results from prior work [17, 56] whenever they are stronger than our own implementations.

Method	Bias-conflicting
ERM	$57.0_{\pm 0.9}$
LfF	$62.2_{\pm 1.0}$
JTT	$62.2_{\pm 1.3}$
CnC	$63.1_{\pm 1.0}$
SupER (Ours)	$62.8_{\pm 0.9}$

Second, we show that SupER can be further improved by easily combining it with other approaches that do not require group annotations. Specifically, we combine SupER with JTT by upweighting the loss $\mathcal{L}^{\text{CE}}_{\phi,\omega_{\text{rel}}}(\mathbf{x}, y)$ for samples identified in the first step of the original JTT procedure [6], where a standard ERM model is first trained to identify potential samples with spurious correlations based on misclassification. As shown in Table 43, we evaluate both JTT and our combined SupER+JTT method on the CelebA the Color MNIST datasets. Results show that our combined method achieves higher worst group accuracy compared to JTT alone. This suggests that the identification of spurious samples by JTT complements SupER's feature disentanglement and its emphasis on leveraging all relevant superclass features for prediction. We leave further investigations on different integrations as an important future direction.

Table 43: Comparison of JTT and SupER+JTT on CelebA and Color MNIST. SupER+JTT achieves improved worst group accuracy (%) across both datasets. When applicable, shared hyperparameters are set to the same values across both methods.

Method	CelebA	Color MNIST
JTT	$80.7_{\pm 1.2}$	$83.3_{\pm 2.7}$
SupER+JTT	$83.8_{\pm 2.1}$	$84.4_{\pm 2.0}$

F. Licenses for External Assets

We use the following publicly available datasets and pretrained models in our work:

- **Pretrained models:**
  - CLIP, MIT, available at `https://github.com/openai/CLIP`.
  - ResNet50, BSD-3-Clause, available at `https://pytorch.org/vision/stable/models/generated/torchvision.models.resnet50.html`.
- **Datasets:**
  - **Waterbirds-95%** and **Waterbirds-100%**, MIT, available at `https://github.com/kohpangwei/group_DRO` and `https://github.com/spetryk/GALS`.
  - **SpuCo Dogs**, MIT, available at `https://github.com/BigML-CS-UCLA/SpuCo`.
  - **MetaShift**, MIT, available at `https://github.com/Weixin-Liang/MetaShift`.
  - **Spawrious**, CC BY 4.0, available at `https://github.com/aengusl/spawrious`.
  - **BFFHQ** (derived from FFHQ), CC BY-NC-SA 4.0 (inherited from FFHQ), FFHQ available at `https://github.com/NVlabs/ffhq-dataset`.
  - **CelebA**, Non-commercial, available at `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`.
  - **Color MNIST** (synthetic variant of MNIST), CC BY-SA 3.0 (inherited from MNIST), MNIST available via `https://keras.io/api/datasets/mnist/`.