Dual Risk Minimization for Robust Fine-tuning of Zero-Shot Models

Kaican Li [*1] **Weiyan Xie** [*1] **Ricardo Silva** [2] **Nevin L. Zhang** [1]

Abstract

Fine-tuning zero-shot foundation models often compromises their robustness to downstream distribution shifts. We propose dual risk minimization (DRM) which combines empirical risk minimization with worst-case risk minimization to better preserve core features conducive to downstream robustness. In particular, we utilize core-feature descriptions generated by LLMs to induce core-based zero-shot predictions which then serve as proxies to estimate the worst-case risk. DRM balances two crucial aspects of robustness: expected and worst-case performance over all possible domains, establishing a new state of the art on various real-world benchmarks. DRM significantly improves the out-of-distribution performance of fine-tuned CLIP ViT-L/14@336 on ImageNet (75.9→77.1), WILDS-iWildCam (47.1→51.8), and WILDS-FMoW (50.7→53.1); opening up new avenues for achieving next-level robustness in fine-tuning zero-shot models.

1. Introduction

Foundation models such as CLIP (Radford et al., 2021) have revolutionized machine learning with their remarkable zero-shot and adaptive capabilities. Research has shown that such capabilities are mainly due to robust feature representations gained from large-scale training data (Fang et al., 2022; Mayilvahanan et al., 2023). The models have been proven useful in various downstream tasks (Shen et al., 2022; Zhang et al., 2022; Betker et al., 2023; Pi et al., 2024) and are the cornerstones of recent large multimodal models (Alayrac et al., 2022; Liu et al., 2023; Zhu et al., 2024).

Fine-tuning is one of the most common approaches to the downstream adaptation of foundation models (Bommasani et al., 2021; Shen et al., 2022). However, such adaptation often comes at the cost of robustness (Radford et al., 2021; Pham et al., 2023), resulting in enlarged gaps between downstream in-distribution (ID) and out-of-distribution (OOD) performance (Wortsman et al., 2022).

To address the issue, previous robust fine-tuning methods mostly aim to preserve pre-trained features during or after the course of fine-tuning (Kumar et al., 2022; Wortsman et al., 2022; Goyal et al., 2023). Nevertheless, the process is guided by the principle of empirical risk minimization (ERM; Vapnik, 1998) which favors the most predictive but not necessarily the most robust features.[1] In general, there are two kinds of robust features: *core features* which essentially define the target classes, and *non-core features* that may aid prediction when the core features are not *clear* (Gao et al., 2023). ERM models tend to exploit the generally less reliable non-core features even when the core features are clear (Geirhos et al., 2020; Shah et al., 2020).

To better preserve downstream core features, we propose dual risk minimization (DRM) which combines ERM with worst-case risk minimization (WRM; Wald, 1945), a common objective for domain generalization (Arjovsky et al., 2019; Sagawa et al., 2020; Cha et al., 2021; Kirichenko et al., 2023). More fundamentally, DRM rests on our view that robustness concerns both the *expected* (or average) performance and the *worst-case* performance over all domains. As there is often a trade-off between them (Tsipras et al., 2019; Teney et al., 2023), Figure 1 illustrates how DRM balances the trade-off to improve overall robustness.

The main challenge of DRM is to assess the worst-case risk. To this end, we use *concept descriptions* (Pratt et al., 2023)—text that describes the core features of each class—obtained with GPT-4 (Achiam et al., 2023). The description we obtained for *cougar*, for instance, is "*a large, tawny cat with a muscular build and a small head* ." We subsequently feed the descriptions to a pre-trained CLIP text encoder (Radford et al., 2021) for the text embeddings which are then used to construct soft class labels upon the similarity scores between the text embeddings and the image embedding of each training image. The risk w.r.t. the soft labels can be considered as a proxy for the worst-case risk and is optimized instead.

[*]Equal contribution, listed in alphabetical order. [1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China [2]Department of Statistical Science, University College London, United Kingdom. Correspondence to: Kaican Li <klibf@cse.ust.hk>, Weiyan Xie <wxieai@cse.ust.hk>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

[1]More discussion on related work can be found in Appendix A.

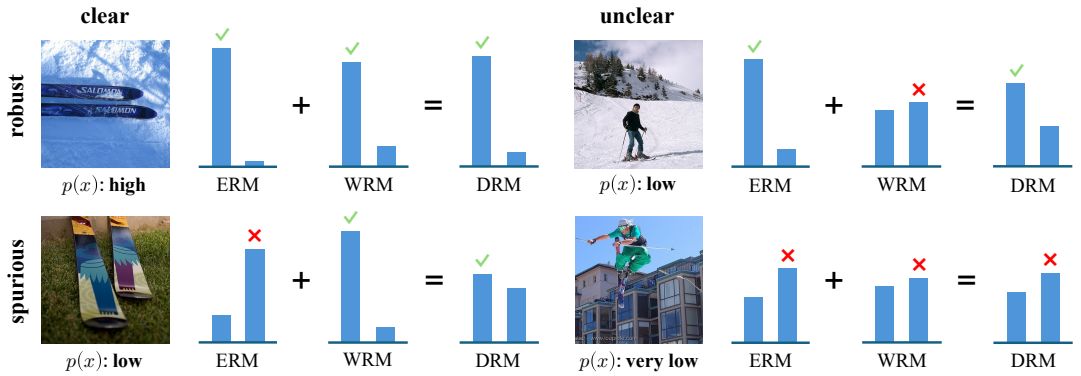


Figure 1. **Dual risk minimization (DRM) combines empirical risk minimization (ERM) and worst-case risk minimization (WRM) to complement their weaknesses.** In this simple binary classification task predicting if there are skis in a given image, either ERM or WRM is suboptimal because (i) ERM underperforms when the core features (the appearance of ski) are *clear* but the non-core features such as background/context are *spurious* (i.e. negatively correlated with ski), and (ii) WRM underperforms when the core features are *unclear* but the non-core features are *robust* (i.e. positively correlated with ski). DRM outperforms ERM and WRM under mild conditions such that the core features are not always clear and the non-core features are more often robust than not.

Empirically, DRM significantly outperforms previous baselines on challenging benchmarks such as ImageNet (Deng et al., 2009) and WILDS (Koh et al., 2021).

In summary, we make the following key contributions:

- We propose *dual risk minimization* (DRM), a novel approach that combines ERM and WRM to improve downstream robustness of zero-shot foundation models through innovative use of concept descriptions.
- We highlight that real-world robustness often concerns both *expected* and *worst-case performance* while most previous works focus on only one. We show that DRM offers a simple and effective way to balance them.
- We establish new states of the art on various real-world benchmarks. For CLIP ViT-L/14@336, DRM significantly improves the average OOD performance on those benchmarks from 57.9 to 60.7 over the best baseline.

2. Dual Risk Minimization

Data model. Let X and Y be the input and ground-truth target variables for which we adopt the following data generation model:

$$\begin{aligned} X &\leftarrow h_X(X_\mathrm{c}, X_\mathrm{n}, \varepsilon), \\ Y &\leftarrow h_Y(X_\mathrm{c}); \end{aligned} \tag{1}$$

where $(X_\mathrm{c}, X_\mathrm{n})$ are latent variables and ε is exogenous noise. We call X_c *core features* of (X, Y), and X_n *non-core features*. X_n and Y may be correlated due to hidden confounders of $(X_\mathrm{c}, X_\mathrm{n})$ and direct causal mechanisms between $(X_\mathrm{c}, X_\mathrm{n})$. Following Peters et al. (2016), we assume the causal mechanisms and the distribution of ε are invariant across domains. There are no other hidden variables or mechanisms. Similar models were widely adopted in the literature (Tenenbaum & Freeman, 1996; Mahajan et al., 2021; Mitrovic et al., 2021; Ahuja et al., 2021; Liu et al., 2021; Lv et al., 2022; Ye et al., 2022; Zhang et al., 2023a; Gao et al., 2024a) where X_c and X_n are sometimes referred to as 'content' and 'style'. We use curly letters such as $\mathcal{X}$ and $\mathcal{Y}$ to denote the space of possible values the random variables may take.

Ideal objective for maximal robustness. Let $\mathcal{D}$ be all possible domains of a task, and $\mathscr{P}$ be some natural distribution over $\mathcal{D}$. By definition, $\mathscr{P}(d) > 0$ for all $d \in \mathcal{D}$. Every domain d is associated with a data distribution $p_d(x, y, x_\mathrm{c}, x_\mathrm{n})$ consistent with (1). Let $p_\theta(y|x)$ be a prediction model parameterized by $\theta \in \Theta$. Its risk in terms of negative log-likelihood, $R_d(\theta) = \mathbb{E}_{(x,y)\sim p_d}[-\log p_\theta(y|x)]$, can be seen as a measure of its performance in domain d. Let $d_\mathrm{s} \in \mathcal{D}$ be the training domain. We will omit d when it is clear from the context, e.g., we will write $R_{d_\mathrm{s}}(\theta)$ simply as $R_\mathrm{s}(\theta)$.

For real-world applications, we argue that a *robust* model should optimize its *expected performance* over $\mathscr{P}$ while maintaining acceptable *worst-case performance* across $\mathcal{D}$. The expected performance implies how well the model would perform at the most general population level, while the worst-case performance tells us the model's performance in the worst scenario one may encounter. Similar views are shared by Eastwood et al. (2022) and Zhang et al. (2023b).

We formalize the intuition as the following constrained optimization problem, namely *idealized dual risk minimization* (IDRM), which aims to minimize the empirical risk while ensuring the worst-case risk is below a threshold value α.

$$\min_{\theta\in\Theta} R_\mathrm{s}(\theta) \quad \text{subject to} \quad \max_{d\in\mathcal{D}} R_d(\theta) \leq \alpha. \tag{IDRM}$$

IDRM generalizes ERM (Vapnik, 1998) and WRM (Wald, 1945) as it reduces to ERM when α is large and to WRM when α is small. IDRM also bears some resemblance to invariant risk minimization (IRM; Arjovsky et al., 2019) which involves an implicit WRM constraint. The constraint, however, requires the classification head to be *optimal* in all training domains and thus may be too demanding in practice. The threshold α makes IDRM more flexible. Another closely related work, GroupDRO (Sagawa et al., 2020), proposes to minimize the worst risk among training domains—a more empirical flavor of WRM. Both IRM and GroupDRO rely on carefully grouped training data to capture invariance across domains. For zero-shot models, we find this is unnecessary, and provide a practical method to realize IDRM with just *single-domain* data (to be elaborated in Section 3).

Relaxation from IDRM to DRM. IDRM is equivalent to the following unconstrained problem due to strong duality.[2]

Theorem 1. *Strong duality holds between IDRM and*

$$\max_{\lambda' \geq 0} \min_{\theta \in \Theta} \left[R_{\mathrm{s}}(\theta) + \lambda' \max_{d \in \mathcal{D}} R_d(\theta) \right] - \lambda' \alpha. \tag{2}$$

The worst-case risk in (2) is closely related to the degree to which $p_\theta(y|x)$ relies on core features to make predictions because for a diverse set of domains $\mathcal{D}$, leveraging non-core features would always lead to worse performance in certain domains (Arjovsky et al., 2019; Geirhos et al., 2020).

Let $f_{\mathrm{c}}(x)$ be an oracle feature extractor that returns a faithful representation of the core features for any input x, and let $p_{\mathrm{c}}(y|x)$ be the optimal model that can be built upon $f_{\mathrm{c}}(x)$. The risk of $p_\theta(y|x)$ w.r.t. $p_{\mathrm{c}}(y|x)$ on the training domain d_{s}, i.e., $R_{\mathrm{s}}^{\mathrm{c}}(\theta) = \mathbb{E}_{x \sim p_{\mathrm{s}}} \mathbb{E}_{y \sim p_{\mathrm{c}}(y|x)}[-\log p_\theta(y|x)]$, measures the degree to which the model's prediction is based on the core features and thus can be regarded as a proxy for the worst-case risk. In summary, we relax IDRM to

$$\min_{\theta \in \Theta} R_{\mathrm{s}}(\theta) + \lambda R_{\mathrm{s}}^{\mathrm{c}}(\theta) \tag{DRM}$$

with some properly chosen $\lambda \geq 0$. Here, the risk $R_{\mathrm{s}}^{\mathrm{c}}(\theta)$ can also be seen as a regularization term for ERM. Next, we answer the key question on how to obtain a good estimate of $p_{\mathrm{c}}(y|x)$ by leveraging the zero-shot capability of foundation models such as the CLIP models (Radford et al., 2021).

3. Fine-tuning Zero-shot Models with DRM

Zero-shot models like CLIP typically consist of an image encoder f_ϕ and a text encoder g_ψ with parameters $\theta = (\phi, \psi)$. Image classification with such models is usually done by first creating a text prompt t_y for each class label $y \in \mathcal{Y}$, and then assigning a probability for each y to an image x by

$$p_\theta(y|x) = \frac{\exp(A_\theta(x, t_y)/\tau)}{\sum_{y' \in \mathcal{Y}} \exp(A_\theta(x, t_{y'})/\tau)}, \tag{3}$$

[2]The proof can be found in Appendix B.

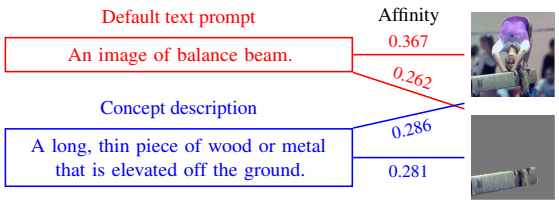


Figure 2. The affinities between the images and the default prompt are not stable w.r.t. changes in the context (non-core features). On the other hand, the changes do not significantly affect the affinities between the images and the concept description. See Appendix D for more examples and details about this empirical study.

where $A_\theta(x, t_y) = \langle f_\phi(x), g_\psi(t_y) \rangle$ and τ is the temperature. The inner product $\langle f_\phi(x), g_\psi(t_y) \rangle$ can be intuitively understood as the *affinity* between x and t_y, and we thus denote it as $A_\theta(x, t_y)$.

Estimating $p_{\mathrm{c}}(y|x)$ with concept descriptions. Recall that the oracle model $p_{\mathrm{c}}(y|x)$ is based on a faithful representation of the core features. To estimate $p_{\mathrm{c}}(y|x)$, we ask GPT-4 (Achiam et al., 2023) to describe the *core visual features* of each class, producing a set of *concept descriptions*[3], $\mathcal{T}^{\mathtt{cd}} = \{t_y^{\mathtt{cd}} \mid y \in \mathcal{Y}\}$. We then feed every training image x and every $t_y^{\mathtt{cd}}$ to a pre-trained CLIP model $\theta_0 = (\phi_0, \psi_0)$ to compute the affinities $A_\theta(x, t_y^{\mathtt{cd}})$ between them. The affinities reflect the significance of the core features of each class y in an image x, and thus can be used to construct an estimate for $p_{\mathrm{c}}(y|x)$. Figuratively, the text embedding $g_{\psi_0}(t_y^{\mathtt{cd}})$ "pulls out" the core features from the image embedding $f_{\phi_0}(x)$ via the inner product. As shown in Figure 2, the affinity $A_\theta(x, t_y)$ is indeed a good measure for the core features of the ground-truth class of an image.

Following the above analysis, a direct estimate for $p_{\mathrm{c}}(y|x)$ is (3) with $\theta = \theta_0$ and $t_y = t_y^{\mathtt{cd}}$; but there is a crucial caveat. For an image x with label y, if y' is another label whose core features are not in x, the affinity $A_\theta(x, t_y)$ should ideally be close to 0. However, this is seldom the case due to imperfections of the pre-trained model θ_0 and the concept descriptions $t_y^{\mathtt{cd}}$. These extraneous affinity values, which we call *artifact terms*, may be class-specific and lead to poor estimates of $p_{\mathrm{c}}(y|x)$. To remove the artifact, let $\xi(x, y) = \exp(A_\theta(x, t_y)/\tau)$, and $\mathcal{X}_y \subseteq \mathcal{X}$ be the training images labeled y. For every training image x and class y, we confine the comparison of the affinities within $\mathcal{X}_y$ by

$$\gamma(x, y) = \frac{\xi(x, y) - \min_{x' \in \mathcal{X}_y} \xi(x', y)}{\max_{x' \in \mathcal{X}_y} \xi(x', y) - \min_{x' \in \mathcal{X}_y} \xi(x', y)},$$

i.e., the min-max normalization of $\xi(x, y)$ over $\mathcal{X}_y$. The final estimation for $p_{\mathrm{c}}(y|x)$ we adopt is defined as

[3]See Appendix C for more details about concept descriptions.

Table 1. ID and OOD performances of DRM and the baselines methods on three datasets, using CLIP ViT-B/16, with and without WiSE-FT. Performance metrics are averaged over 5 runs with 95% confidence intervals. Best performances are highlighted in **bold**. For ImageNet, we report the average performance over its 5 OOD test sets. Results on individual test sets are provided in Appendix F.1.

	ImageNet				iWildCam				FMoW			
	w/o WiSE-FT		WiSE-FT		w/o WiSE-FT		WiSE-FT		w/o WiSE-FT		WiSE-FT	
Method	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
0-shot	68.3±0.0	58.7±0.0	-	-	8.7±0.0	11.0±0.0	-	-	20.4±0.0	18.7±0.0	-	-
LP	79.9±0.0	57.2±0.0	80.0±0.0	58.3±0.0	44.5±0.6	31.1±0.4	45.5±0.6	31.7±0.4	48.2±0.1	30.5±0.3	48.7±0.1	31.5±0.3
FT	81.4±0.1	54.8±0.1	82.5±0.1	61.3±0.1	48.1±0.5	35.0±0.5	48.1±0.5	35.0±0.5	68.5±0.1	39.2±0.7	68.5±0.1	41.5±0.5
L2-SP	81.6±0.1	57.9±0.1	82.2±0.1	58.9±0.1	48.6±0.4	35.3±0.3	48.6±0.4	35.3±0.3	68.6±0.1	39.4±0.6	68.4±0.1	40.3±0.6
LP-FT	81.8±0.1	60.5±0.1	82.1±0.1	61.8±0.1	49.7±0.5	34.7±0.4	50.2±0.5	35.7±0.4	68.4±0.2	40.4±1.0	68.5±0.2	42.4±0.7
FLYP	**82.6±0.0**	60.2±0.1	**82.9±0.0**	63.2±0.1	52.2±0.6	35.6±1.2	52.5±0.6	37.1±1.2	68.6±0.2	41.3±0.8	**68.9±0.3**	42.0±0.9
DRM	82.0±0.3	**63.2±0.2**	82.4±0.2	**64.0±0.2**	**54.1±0.5**	**40.0±0.6**	**55.3±0.4**	**41.4±0.7**	**68.7±0.3**	**45.9±1.1**	68.7±0.2	**46.1±0.8**

Table 2. DRM *vs*. FLYP on two larger CLIP ViT models.

		ImageNet		iWildCam		FMoW	
	Method	ID	OOD	ID	OOD	ID	OOD
L/14	FLYP	84.6±0.3	73.4±0.1	56.0±1.1	41.9±0.7	71.2±0.5	48.2±0.5
	+WiSE-FT	85.1±0.2	75.1±0.1	57.2±0.7	42.1±0.5	**72.0**±0.4	49.1±0.6
	DRM	85.0±0.2	75.5±0.2	**61.8**±0.5	49.2±0.4	70.9±0.8	**51.3**±0.7
	+WiSE-FT	**86.2**±0.1	**76.2**±0.2	61.6±0.3	**49.8**±0.4	71.4±0.5	**51.3**±0.7
L/14@336	FLYP	85.4±0.2	75.0±0.3	58.7±0.6	45.4±1.0	72.5±0.3	50.5±0.5
	+WiSE-FT	86.1±0.2	75.9±0.2	60.5±0.5	47.1±1.2	72.6±0.3	50.7±0.6
	DRM	85.9±0.1	76.0±0.2	**62.8**±0.6	51.4±0.5	**73.8**±0.5	52.5±0.9
	+WiSE-FT	**87.4**±0.0	**77.1**± 0.2	62.5±0.4	**51.8**±0.5	**73.8**±0.3	**53.1**±0.6

$$\tilde{p}_{\mathrm{c}}(y|x) = \begin{cases} \gamma(x,y), & y = y_x; \\ [1-\gamma(x,y_x)] \cdot \frac{\gamma(x,y)}{\sum_{y' \neq y_x} \gamma(x,y')}, & y \neq y_x; \end{cases} \quad (4)$$

where y_x is the ground-truth label of x. The overall impact of artifact terms is now much reduced since $\tilde{p}_{\mathrm{c}}(y|x)$ is independent of variations in the artifact terms.

Practical improvements to DRM. Let $\mathcal{T}$ be the set of text prompts used to construct the classifier $p_\theta(y|x)$ defined by (3). For such classifiers, the original DRM objective can be re-expressed as $\min_{\theta\in\Theta} R_{\mathrm{s}}(\theta;\mathcal{T}) + \lambda R_{\mathrm{s}}^{\mathrm{c}}(\theta;\mathcal{T})$ where R_{s} and $R_{\mathrm{s}}^{\mathrm{c}}$ not only depend on θ but also on $\mathcal{T}$. Typically, a set of default prompts $\mathcal{T}^{\mathtt{df}}$ like "*an image of [class name]*" is used. With the estimation of $p_{\mathrm{c}}(y|x)$ also incorporated, this gives rise to the following vanilla DRM for fine-tuning zero-shot models:

$$\min_{\theta\in\Theta} R_{\mathrm{s}}(\theta;\mathcal{T}^{\mathtt{df}}) + \lambda \tilde{R}_{\mathrm{s}}^{\mathrm{c}}(\theta;\mathcal{T}^{\mathtt{df}}), \quad (5)$$

where $\tilde{R}_{\mathrm{s}}^{\mathrm{c}}$ stands for $R_{\mathrm{s}}^{\mathrm{c}}$ with $p_{\mathrm{c}}(y|x)$ replaced by its estimation $\tilde{p}_{\mathrm{c}}(y|x)$ derived from the concept descriptions $\mathcal{T}^{\mathtt{cd}}$.

We can further improve vanilla DRM (5) by using $\mathcal{T}^{\mathtt{cd}}$ instead of $\mathcal{T}^{\mathtt{df}}$ to construct the classifier $p_\theta(y|x)$ for minimizing $\tilde{R}_{\mathrm{s}}^{\mathrm{c}}$. This leads to the final DRM objective,

$$\min_{\theta\in\Theta} R_{\mathrm{s}}(\theta;\mathcal{T}^{\mathtt{df}}) + \lambda \tilde{R}_{\mathrm{s}}^{\mathrm{c}}(\theta;\mathcal{T}^{\mathtt{cd}}), \quad (6)$$

which is used to fine-tune CLIP models in our main experiments. The adjustment proves to be beneficial (in ablation study) as it ensures both the target $\tilde{p}_{\mathrm{c}}(y|x)$ and the classifier $p_\theta(y|x)$ are based on the same text prompts $\mathcal{T}^{\mathtt{cd}}$. It makes $\tilde{R}_{\mathrm{s}}^{\mathrm{c}}$ more effective in limiting the divergence of θ from θ_0 which helps better preserve pre-trained features compared to learning with a classifier based on different prompts.

Inference. The fine-tuning process (6) involves two classifiers: the ERM classifier $p_\theta^{\mathtt{df}}(y|x)$ induced by $\mathcal{T}^{\mathtt{df}}$, and the WRM classifier $p_\theta^{\mathtt{cd}}(y|x)$ induced by $\mathcal{T}^{\mathtt{cd}}$. While either alone can be used for inference, we find that their mixture,

$$p_\theta^{\mathtt{dual}}(y|x) = \beta \cdot p_\theta^{\mathtt{df}}(y|x) + (1-\beta) \cdot p_\theta^{\mathtt{cd}}(y|x) \quad (7)$$

where $\beta \in (0,1)$, usually performs the best. This is perhaps unsurprising as (7) essentially combines ERM with WRM as depicted in Figure 1. By default, we set $\beta = 1/(1+\lambda)$ so to be as consistent with (6) as possible.

4. Experiments

We evaluate DRM on real-world DG benchmarks against existing robust fine-tuning methods such as LP-FT (Kumar et al., 2022) and FLYP (Goyal et al., 2023). Additionally, we examine combining WiSE-FT (Wortsman et al., 2022), a method that averages parameters between pre-trained and fine-tuned models, with DRM and all baseline methods. Training-domain validation is adopted for all methods. More implementation details are in Appendix E.

Table 1 shows that on CLIP ViT-B/16, DRM consistently outperforms the baselines in OOD performance across all datasets, with and without WiSE-FT. Further results in Table 2 on two larger models confirm that DRM consistently

beats FLYP. The previous best OOD scores of CLIP ViT-L/14@336 for iWildCam and FMoW were 47.1 and 50.7 by FLYP+WiSE-FT. DRM improves them to 51.8 and 53.1, marking relative improvements of 10.0% and 4.7%.

We also compare DRM with some more recently proposed robust fine-tuning methods, and the results are reported in Appendix F.3. It is shown in Appendix F.6 that DRM is also effective in fine-tuning ImageNet pre-trained CNNs.

Finally, we have conducted a thorough ablation study to assess how various aspects of DRM including both training and inference affect the model's final performance. We have also examined the impact of the hyperparameter λ of DRM. These results can be found in Appendix F.4 and F.5.

Acknowledgement

Research on this paper was supported in part by Hong Kong Research Grants Council under grant 16204920. Kaican Li and Weiyan Xie were supported in part by the Huawei PhD Fellowship Scheme. We are grateful for the generous GPU support from the HKUST SuperPOD system during its free pilot use period. We thank Yongxiang Huang, Didan Deng, Lanqing Hong and Zhenguo Li for valuable discussions.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ahuja, K., Wang, J., Dhurandhar, A., Shanmugam, K., and Varshney, K. R. Empirical or invariant risk minimization? a sample complexity perspective. In *ICLR*, 2021.

Alabdulmohsin, I., Chiou, N., D'Amour, A., Gretton, A., Koyejo, S., Kusner, M. J., Pfohl, S. R., Salaudeen, O., Schrouff, J., and Tsai, K. Adapting to latent subgroup shifts via concepts and proxies. In *AISTATS*, pp. 9637–9661. PMLR, 2023.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pp. 23716–23736, 2022.

Andreassen, A., Bahri, Y., Neyshabur, B., and Roelofs, R. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv:1907.02893*, 2019.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *NeurIPS*, 32, 2019.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*, volume 28. Princeton university press, 2009.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8, 2023.

Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. In *NeurIPS*, 2011.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. In *NeurIPS*, volume 34, pp. 22405–22418, 2021.

Cha, J., Lee, K., Park, S., and Chun, S. Domain generalization by mutual-information regularization with pre-trained models. In *ECCV*, pp. 440–457. Springer, 2022.

Cheng, D., Xu, Z., Jiang, X., Wang, N., Li, D., and Gao, X. Disentangled prompt representation for domain generalization. In *CVPR*, pp. 23595–23604, 2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Eastwood, C., Robey, A., Singh, S., Von Kügelgen, J., Hassani, H., Pappas, G. J., and Schölkopf, B. Probable domain generalization via quantile risk minimization. In *NeurIPS*, volume 35, pp. 17340–17358, 2022.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., and Schmidt, L. Data determines distributional robustness in contrastive language image pre-training (clip). In *ICML*, pp. 6216–6234. PMLR, 2022.

Gao, H., Li, K., Xie, W., Zhi, L., Huang, Y., Wang, L., Cao, C. C., and Zhang, N. L. Consistency regularization for domain generalization with logit attribution matching. In *UAI*, 2024a.

Gao, I., Sagawa, S., Koh, P. W., Hashimoto, T., and Liang, P. Out-of-domain robustness via targeted augmentations. In *ICML*, 2023.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024b.

Ge, W. and Yu, Y. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *CVPR*, pp. 1086–1095, 2017.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

Goyal, S., Kumar, A., Garg, S., Kolter, Z., and Raghunathan, A. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, pp. 19338–19347, 2023.

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., and Feris, R. Spottune: transfer learning through adaptive fine-tuning. In *CVPR*, pp. 4805–4814, 2019.

Han, J., Lin, Z., Sun, Z., Gao, Y., Yan, K., Ding, S., Gao, Y., and Xia, G.-S. Anchor-based robust finetuning of vision-language models. In *CVPR*, pp. 26919–26928, 2024.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *CVPR*, pp. 15262–15271, 2021b.

Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Zhao, T. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 2023.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, pp. 5637–5664. PMLR, 2021.

Kumar, A., Raghunathan, A., Jones, R. M., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022.

Li, X., Grandvalet, Y., and Davoine, F. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, pp. 2825–2834. PMLR, 2018.

Liu, C., Sun, X., Wang, J., Tang, H., Li, T., Qin, T., Chen, W., and Liu, T.-Y. Learning causal semantic representation for out-of-distribution prediction. In *NeurIPS*, 2021.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, volume 36, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lv, F., Liang, J., Li, S., Zang, B., Liu, C. H., Wang, Z., and Liu, D. Causality inspired representation learning for domain generalization. In *CVPR*, pp. 8046–8056, 2022.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. In *ICML*, pp. 7313–7324. PMLR, 2021.

Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K., and O'Connor, N. E. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *ICCV*, pp. 262–271, 2023.

Mao, X., Chen, Y., Jia, X., Zhang, R., Xue, H., and Li, Z. Context-aware robust fine-tuning. *IJCV*, 132(5):1685–1700, 2024.

Mayilvahanan, P., Wiedemer, T., Rusak, E., Bethge, M., and Brendel, W. Does clip's generalization performance mainly stem from high train-test similarity? In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Menon, S. and Vondrick, C. Visual classification via description from large language models. In *ICLR*, 2022.

Mitrovic, J., McWilliams, B., Walker, J. C., Buesing, L. H., and Blundell, C. Representation learning via invariant causal mechanisms. In *ICLR*, 2021.

Moayeri, M., Singla, S., and Feizi, S. Hard imagenet: Segmentations for objects with strong spurious cues. In *NeurIPS*, volume 35, pp. 10068–10077, 2022.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *ICML*, 2013.

Nam, G., Heo, B., and Lee, J. Lipsum-ft: Robust fine-tuning of zero-shot models using random text guidance. In *ICLR*, 2024.

Neuhaus, Y., Augustin, M., Boreiko, V., and Hein, M. Spurious features everywhere-large-scale detection of harmful spurious features in imagenet. In *ICCV*, pp. 20235–20246, 2023.

Oh, C., Kim, M., Lim, H., Park, J., Jeong, E., Cheng, Z.-Q., and Song, K. Towards calibrated robust fine-tuning of vision-language models. *arXiv preprint arXiv:2311.01723*, 2023.

Peters, J., Bühlmann, P., and Meinshausen, N. 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.

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

Pi, R., Yao, L., Han, J., Liang, X., Zhang, W., and Xu, H. Ins-detclip: Aligning detection model to follow human-language instruction. In *ICLR*, 2024.

Prabhu, V., Selvaraju, R. R., Hoffman, J., and Naik, N. Can domain adaptation make object recognition work for everyone? In *CVPR*, pp. 3981–3988, 2022.

Pratt, S., Covert, I., Liu, R., and Farhadi, A. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pp. 15691–15701, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *ICML*, pp. 5389–5400. PMLR, 2019.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *ICLR*, 2020.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, volume 33, pp. 9573–9585, 2020.

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. How much can clip benefit vision-and-language tasks? In *ICLR*, 2022.

Shu, Y., Guo, X., Wu, J., Wang, X., Wang, J., and Long, M. Clipood: Generalizing clip to out-of-distributions. In *ICML*, pp. 31716–31731. PMLR, 2023.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tenenbaum, J. and Freeman, W. Separating style and content. In *NeurIPS*, volume 9, 1996.

Teney, D., Lin, Y., Oh, S. J., and Abbasnejad, E. Id and ood performance are sometimes inversely correlated on real-world datasets. In *NeurIPS*, volume 36, 2023.

Tian, J., He, Z., Dai, X., Ma, C.-Y., Liu, Y.-C., and Kira, Z. Trainable projected gradient method for robust fine-tuning. In *CVPR*, pp. 7836–7845, 2023a.

Tian, J., Liu, Y.-C., Smith, J. S., and Kira, Z. Fast trainable projection for robust fine-tuning. *NeurIPS*, 36, 2023b.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *ICLR*, 2019.

Vapnik, V. *Statistical Learning Theory*. Wiley, 1998.

Wald, A. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, pp. 265–280, 1945.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In *CVPR*, pp. 7959–7971, 2022.

Yan, A., Wang, Y., Zhong, Y., Dong, C., He, Z., Lu, Y., Wang, W. Y., Shang, J., and McAuley, J. Learning concise and descriptive attributes for visual recognition. In *ICCV*, pp. 3090–3100, 2023.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, pp. 19187–19197, 2023.

Ye, N., Li, K., Bai, H., Yu, R., Hong, L., Zhou, F., Li, Z., and Zhu, J. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *CVPR*, pp. 7947–7958, 2022.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *ICML*, pp. 3987–3995. PMLR, 2017.

Zhang, J. O., Sax, A., Zamir, A., Guibas, L., and Malik, J. Side-tuning: a baseline for network adaptation via additive side networks. In *ECCV*, pp. 698–714. Springer, 2020.

Zhang, N. L., Li, K., Gao, H., Xie, W., Lin, Z., Li, Z., Wang, L., and Huang, Y. A causal framework to unify common domain generalization approaches. *arXiv preprint arXiv:2307.06825*, 2023a.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., and Li, H. Pointclip: Point cloud understanding by clip. In *CVPR*, pp. 8552–8562, 2022.

Zhang, Y., Shen, Y., Wang, D., Gu, J., and Zhang, G. Connecting unseen domains: Cross-domain invariant learning in recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1894–1898, 2023b.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. Freelb: Enhanced adversarial training for natural language understanding. In *ICLR*, 2020.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.

A. Related Work

Robust fine-tuning of pre-trained models. Kumar et al. (2022) showed that fine-tuning tends to distort pre-trained robust features, and the distortion is exacerbated by randomly initialized heads which would significantly alter the pre-trained features to fit ID examples. The proposed remedy, LP-FT, first learns a linear probe (LP) on frozen features, followed by regular fine-tuning (FT). Goyal et al. (2023) took this idea further by reusing the pre-trained text encoder of CLIP as the classification head for fine-tuning. This method improves LP-FT and is colloquially known as "fine-tune like you pre-train" (FLYP). WiSE-FT (Wortsman et al., 2022) combines pre-trained models with their fine-tuned versions by weight averaging, yet another way to recover robust features lost during fine-tuning.

Recently, some new robust fine-tuning methods that regularize the difference in model outputs between pre-trained and fine-tuned models have been proposed. The context-aware robust fine-tuning method (CAR-FT) (Mao et al., 2024) and the method proposed in Cheng et al. (2024) seeks to reduce the distance in context distributions generated by pre-trained and fine-tuned CLIP models. These methods however require prior knowledge of image contexts, such as background and viewpoint, limiting their practical use. In contrast, without needing prior information, Lipsum-ft (Nam et al., 2024) introduces a regularization term that minimizes the L^2 distance between the inner products of training image embeddings and random text embeddings generated by the pre-trained CLIP model and those generated by the fine-tuned CLIP model.

In addition to WiSE-FT which combines the model parameters of pre-trained and fine-tuned models, several newer methods also apply regularization directly to the parameter spaces. CLIPood (Shu et al., 2023) utilizes a beta moving average for updating parameters during training. The calibrated robust fine-tuning (CaRot) method (Oh et al., 2023) focuses on regularizing singular value distributions and incorporates an exponential moving average for parameter updates. In addition, the trainable projected gradient method (TPGM) (Tian et al., 2023a) autonomously determines the most effective constraint for each layer's parameters, thereby achieving a more fine-grained regularization. The fast trainable projection (FTP) method (Tian et al., 2023b) enhances TPGM by enabling a more efficient learning of layer-specific projection constraints.

Another recent method is the anchor-based robust fine-tuning (AFT) (Han et al., 2024) which improves the robustness of fine-tuned CLIP models by incorporating richer text information with an additional image captioner. However, using the captioner on each image is computationally costly, and the fine-tuning effectiveness highly depends on the captioner used.

Prior to the work which we have introduced above, Li et al. (2018) proposed to use the L^2 norm of the difference between the parameters of pre-trained and fine-tuned models as a regularization penalty to help preserve pre-trained features. Some other work explored updating only a small number of (pre-trained/add-on) parameters (Guo et al., 2019; Zhang et al., 2020; Gao et al., 2024b). Similar ideas (Kirkpatrick et al., 2017; Zenke et al., 2017) were also discussed in continual learning to mitigate catastrophic forgetting (McCloskey & Cohen, 1989). Without explicit constraints on model parameters, Ge & Yu (2017) turned to the source of robust features and proposed to incorporate a subset of pre-trained data for fine-tuning, while Cha et al. (2022) aimed to enhance the mutual information between pre-trained and fine-tuned features. Jiang et al. (2019); Zhu et al. (2020) added smoothness constraints on model predictions for adversarial examples (Szegedy et al., 2013) to help retain robust features. Andreassen et al. (2021) showed that the OOD accuracy tends to improve initially but then plateaus as the fine-tuning proceeds.

Worst-case risk minimization. The study of worst-case risk minimization (WRM) dates back to the work of Wald (1945), which has gradually evolved into what we know as robust optimization today (Ben-Tal et al., 2009). More recently, WRM has been considered (by many) a basic principle for domain generalization (DG; Blanchard et al., 2011; Muandet et al., 2013). A seminal work in this direction is invariant risk minimization (IRM; Arjovsky et al., 2019) which aims to learn core-feature representations from multi-domain training data. Such representations, under certain causal invariant assumptions, give rise to classifiers that minimize the worst risk (Peters et al., 2016). Another key paper introduces GroupDRO (Sagawa et al., 2020) which applies higher penalties to domains with higher empirical risks. Neither IRM nor GroupDRO formulates the worst-case risk as an explicit optimization constraint for ERM. Eastwood et al. (2022) pointed out that average and worst-case performance are both important. Their proposed objective, namely probable DG, aims to minimize the risk among the most likely domains, with little restriction on the worst-case risk. Our setup also has some similarities with the work of Alabdulmohsin et al. (2023) on latent subgroup shift, which also relies on external sources of information but under stronger causal invariance assumptions.

Prompt design for zero-shot models. To better leverage the capability of zero-shot models, various methods have been proposed for designing better prompts. Menon & Vondrick (2022); Pratt et al. (2023); Maniparambil et al. (2023) mainly

explored prompt designs for zero-shot classification. Their prompts are generated by large language models (LLMs; Radford et al., 2019) with slightly different instructions than ours, not explicitly focusing on the core features. For example, Pratt et al. (2023) used "*Describe an image from the internet of a(n) ...*", which may include some descriptions of the non-core features in the resulting prompts. The prompts considered by Yang et al. (2023); Yan et al. (2023) are closer to ours in this respect, where they aimed to use the LLM-generated concept descriptions to build concept bottleneck models for interpretable image classification.

B. Proofs

Lemma 1. *Let p and q be two probability distributions over $\mathcal{X} \times \mathcal{Y}$. The cross-entropy between $p(y|x)$ and $q(y|x)$ over $p(x)$, i.e., $H_p(q) = \mathbb{E}_{(x,y)\sim p}[-\log q(y|x)]$, is convex with respect to q.*

Proof. It suffices to show that for any pair of (q_1, q_2) and $\alpha \in [0, 1]$ we have $H_p(\alpha q_1 + (1-\alpha)q_2) \leq \alpha H_p(q_1) + (1-\alpha)H_p(q_2)$.

$$\begin{aligned} H_p(\alpha q_1 + (1-\alpha)q_2) &= \mathbb{E}_{(x,y)\sim p}[-\log(\alpha q_1(y|x) + (1-\alpha)q_2(y|x))] \\ &\leq \mathbb{E}_{(x,y)\sim p}[-\alpha \log q_1(y|x) - (1-\alpha)\log q_2(y|x)] \\ &= \alpha H_p(q_1) + (1-\alpha)H_p(q_2). \end{aligned} \tag{8}$$

□

Theorem 1. *Strong duality holds between IDRM and*

$$\max_{\lambda' \geq 0} \min_{\theta \in \Theta} \left[R_{\mathrm{s}}(\theta) + \lambda' \max_{d \in \mathcal{D}} R_d(\theta) \right] - \lambda' \alpha. \tag{2}$$

Proof. Recall that IDRM aims to solve for

$$\min_{\theta \in \Theta} R_{\mathrm{s}}(\theta) \quad \text{subject to} \quad \max_{d \in \mathcal{D}} R_d(\theta) \leq \alpha. \tag{IDRM}$$

Here, $R_d(\theta) = H_{p_d}(p_\theta(y|x)) = \mathbb{E}_{(x,y)\sim p_d}[-\log p_\theta(y|x)]$ is the cross-entropy between $p_d(y|x)$ and $p_\theta(y|x)$ over $p_d(x)$. It follows from Lemma 1 that $R_d(\theta)$ is convex w.r.t. $p_\theta(y|x)$ for all $d \in \mathcal{D}$.

Since the point-wise maximum of multiple convex functions is also convex, $\max_{d\in\mathcal{D}} R_d(\theta)$ is convex and therefore IDRM is a convex optimization problem w.r.t. $p_\theta(y|x)$. By Slater's condition, strong duality holds between IDRM and the Lagrangian dual of IDRM:

$$\max_{\lambda' \geq 0} \min_{\theta \in \Theta} R_{\mathrm{s}}(\theta) + \lambda' \left[\max_{d \in \mathcal{D}} R_d(\theta) - \alpha \right], \tag{9}$$

with feasible α that is guaranteed to exist. □

C. Details on Concept Description Creation

We utilized GPT-4 (Achiam et al., 2023), a leading-edge language model developed by OpenAI, to create concept descriptions. We supplied GPT-4 with the class name and tailored prompts designed to guide the model to focus on the essential visual features of the class while omitting unrelated contextual information.

Specifically, for the iWildCam dataset, our prompt to GPT-4 was:

> Q: Generate a short sentence that describes the visual features of a given animal. Do not include its function, its surroundings, or the environment it usually inhabits. The sentence should be concise. For example, [goldfish: a long, golden body with back fins].

After that, we asked it to generate a concept description for each wildlife class. For example:

> Q: Now the given animal is *Cougar*.
>
> A: [*Cougar*: *a large, tawny cat with a muscular build and a small head.*]

Another example is the FMoW dataset. We found it is challenging to encapsulate the core visual features of land used for specific purposes in just one short sentence. Consequently, we employed GPT-4 to generate six sentences for each type of land use. When we implemented DRM on this dataset, we slightly modified the image classifier to use the average of the six text embeddings for each class as the classification head.

For this dataset, our initial prompt to GPT-4 was:

> Q: The Functional Map of the World (fMoW) dataset is designed to identify the functions of buildings and land use by examining satellite images over time.
>
> Your task is to generate six sentences, each describing the visual features likely to be seen in these satellite images, given the functional purpose of buildings and land use. The features you describe should be visible from an aerial perspective and consistent across various countries. Each sentence should clearly and comprehensively describe a visual feature, focusing solely on its appearance without including details of the surrounding area or the purpose of the features, such as "Helipad for urgent aerial access".

After that, we asked it to generate concept description for each land use. For example:

> Q: Now the given land use is *Road Bridge*.
>
> A: [*Road Bridge*:
>
> *Visibly characterized by a straight or slightly curved linear structure that spans across natural or man-made obstacles.*
>
> *Constructed from materials that appear robust in satellite images, often showing as grey or dark tones against the landscape.*
>
> *The bridge deck is typically marked with parallel lines, indicating the lanes for vehicular traffic visible from above.*
>
> *Elevation above the surrounding area is noticeable, with shadows often highlighting the height and structure of the bridge.*
>
> *Support features like piers or columns are regularly spaced, visible as distinct vertical elements that support the span.*
>
> *End points of the bridge integrate with road networks, appearing as seamless transitions from elevated to ground-level roads.*]

The full list of concept descriptions generated by GPT-4 for each class of different datasets considered in this work will be publicly available online later.

D. Empirical Study on Concept Descriptions

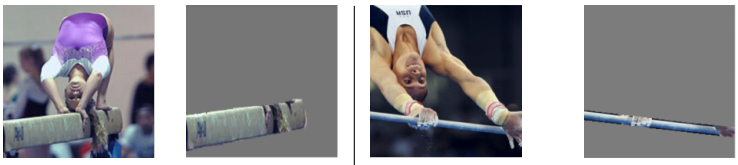
Intuitively, $g_{\psi_0}(t_y^{\texttt{cd}})$ is a representation of the core features of the class y from the text side. We use it to "pull out" the core features from the image embedding. To verify this intuition, we conducted the following empirical study.

To be more specific, we conducted experiments to check how robust the affinity $\langle f_{\phi_0}(x), g_{\psi_0}(t_{y_x}^{\texttt{cd}})\rangle$ is to changes in the context information (non-core features) while keeping the core visual features of y_x unchanged (y_x is the ground-truth label of x). It also tells us how well $g_{\psi_0}(t_y^{\texttt{cd}})$ can represent the core features of the class y_x from the text side. The default prompt "*an image of [class name].*" was used as the baseline to compare with the concept descriptions.

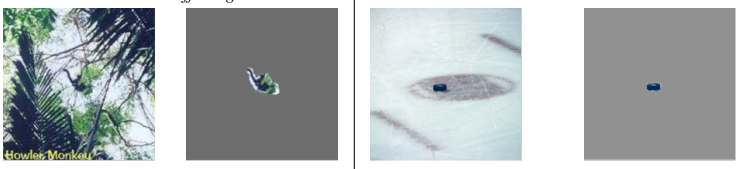
Concept descriptions yield soft labels being more robust to change of context information. Since we have no access to the training data of the zero-shot CLIP models and cannot control the image-text pairs that include the class name in its pre-training, it is possible that, compared to the affinities $\langle f_{\phi_0}(x), g_{\psi_0}(t_{y_x}^{\texttt{cd}})\rangle$ based on the concept descriptions, the affinities $\langle f_{\phi_0}(x), g_{\psi_0}(t_{y_x}^{\texttt{df}})\rangle$ associated with default text prompts could be heavily influenced by context information rather than reflecting the core visual features of class y_x as intended.

To verify this hypothesis, we conducted experiments with 15 ImageNet classes identified by Hard ImageNet (Moayeri et al., 2022) as having strong spurious cues. These classes include *Baseball Player*, which is typically depicted in a baseball field, *Volleyball*, frequently shown alongside a volleyball player, and *Balance Beam*, where images commonly feature a gymnast in action. Using the segmentations provided in Moayeri et al. (2022), we observed how the affinities from the CLIP ViT-B/16 model changed when the background of an image was removed.

Using default prompts, we observed that the average affinities $\langle f_{\phi_0}(x), g_{\psi_0}(t_{y_x}^{\texttt{df}})\rangle$ changed by **0.108** when the background pixels were substituted with mean pixel values. In contrast, the affinities from concept description prompts $\langle f_{\phi_0}(x), g_{\psi_0}(t_{y_x}^{\texttt{cd}})\rangle$ were relatively unaffected, exhibiting a small decrease of **0.036** on average. This indicates that default prompts are



df affinity:	0.367	0.262	0.333	0.267
cd affinity:	0.286	0.281	0.258	0.273
df prompt:	*An image of balance beam.*		*An image of gymnastic horizontal bar.*	
cd prompt:	*A long, thin piece of wood or metal that is elevated off the ground.*		*Long metal or wood bar held up by upright supports.*	



df affinity:	0.395	0.265	0.344	0.264
cd affinity:	0.261	0.250	0.254	0.272
df prompt:	*An image of howler monkey.*		*An image of hockey puck.*	
cd prompt:	*Four-limbed silhouette with a long tail and a large throat.*		*A small, hard, round black rubber disc.*	

Figure 3. Concept descriptions (cd) yield affinities that are more robust to the change of context information than the affinities yielded by the default text prompts (df).

(a) Volleyball court or volleyball player without the volleyball (b) Baseball field or baseball bat without players

Figure 4. Images showing typical contexts associated with specific sports without the presence of the class object itself.

more sensitive to changes in background context, whereas concept description prompts provide more stable and context-independent results. Four examples demonstrating this difference in affinity changes are presented in Figure 3.

To further understand this point, we created a variant of these 15 hard ImageNet classes where the usual context is present but the class object itself is absent. For example, this included images of volleyball courts without the *volleyball*, and images of baseball fields without the *baseball player* (examples are given in Figure 4). The development of this dataset variant was inspired by the concept of Spurious ImageNet (Neuhaus et al., 2023), and some images were from this work.

In such images, affinities that accurately reflect the presence of the core visual features should be low, given the absence of class-specific objects. Our findings indicate that while concept descriptions consistently yield low average affinities to these images, default text prompts still produce relatively high affinities. For instance, in the group of images depicted in Figure 4 (a), the default text prompt "*An image of volleyball.*" results in an average affinity of 0.292, whereas the more descriptive prompt "*A round, inflated ball.*" yields a significantly lower affinity of 0.158. Similarly, for images in Figure 4 (b), the default prompt "*An image of baseball player.*" results in an affinity of 0.262, while the concept description prompt "*An athlete in uniform playing baseball.*" yields an average affinity of 0.152. This result highlights the importance of using concept descriptions generated by LLMs to reduce the impact of contextual biases when creating the proxy model.

E. Experiment Details

E.1. Datasets

ImageNet (Deng et al., 2009) comprises a large-scale dataset with over a million images across 1000 categories. The training set is utilized for model fine-tuning and the validation set for assessing ID accuracy. For OOD evaluation, we use ImageNet variants: **ImageNet-V2** (Recht et al., 2019) (images from a later decade), **ImageNet-R** (Hendrycks et al., 2021a) and **ImageNet-Sketch** (Wang et al., 2019) (art variations), **ImageNet-A** (Hendrycks et al., 2021b) (objects in unusual contexts), and **ObjectNet** (Barbu et al., 2019) (uncommon orientations and contexts).

WILDS-iWildCam (iWildCam) (Koh et al., 2021) contains camera trap images for wildlife classification, with training images from 200 locations and OOD images from different locations. Both ID and OOD performances are measured using macro F1 scores.

WILDS-FMoW (FMoW) (Koh et al., 2021) is a dataset of satellite images from different years and continents, used for land-use prediction. The dataset is split into training, validation, and test domains based on the year of collection. There is also a notable shift in the proportion of images from different continents between different domains. Its ID performance is measured by the ID test accuracy, while OOD performance is evaluated by the worst-region accuracy on the OOD test set.

Dollar Street-DA and **GeoYFCC-DA** (Prabhu et al., 2022) contain images collected in different continents and countries. For Dollar Street-DA, training images are from North America and Europe, with test images from other continents. GeoYFCC-DA has a similar setup. Model effectiveness is measured by accuracy in seen and unseen countries.

E.2. Baseline Methods

The key baseline we compare our DRM method with is **FLYP** (Goyal et al., 2023). As explained in Section 3, both methods utilize the text encoder when performing image classification tasks. They differ only in the loss functions they use when fine-tuning the CLIP model. Conceptually, the FLYP loss is just the first term of our DRM loss. The first term of our DRM loss is originally the standard cross-entropy loss. It has been shown by Goyal et al. (2023) that replacing the standard cross-entropy loss with the CLIP contrastive loss (Radford et al., 2021) leads to superior image classification performance. We do the same for DRM in all our experiments to ease comparison.

Besides FLYP, we include several other baselines that do not utilize the text encoder. They add a linear classification head to the image encoder and fine-tune the resulting classifier. **LP** (linear probing) fine-tunes the classification only while keeping the image encoder frozen. **FT** (fine-tuning) trains both the classification head and the image encoder. **L2-SP** (Li et al., 2018) is a variant of FT that applies regularization to limit the divergence of the model under fine-tuning from the pre-trained model. **LP-FT** (Kumar et al., 2022) starts with LP and then proceeds with full fine-tuning.

We also include a weight-space averaging method, **WiSE-FT** (Wortsman et al., 2022), which interpolates parameters of a pre-trained model and that of a fine-tuned model using $\theta_{\text{wise-ft}} = \rho \cdot \theta_{\text{zs}} + (1 - \rho) \cdot \theta_{\text{ft}}$. We consider the combination of WiSE-FT with DRM and all the baselines. The hyperparameter ρ is chosen from the range 0.1 to 0.9 via ID validation.

E.3. Hyperparameter Settings

As listed in Table 3, we primarily adopted the hyperparameter settings from the code released by FLYP (Goyal et al., 2023). AdamW optimizer (Loshchilov & Hutter, 2017) is used for all datasets. In particular, for ImageNet, a batch size of 256 is used for CLIP ViT-L/14@336, and a batch size of 512 is used for smaller models.

Table 3. Hyperparameter settings.

	ImageNet	iWildCam	FMoW
Epochs	10	20	20
Learning rate	1e-5	1e-5	1e-5
Batch size	256/512	256	256
Weight decay	0.1	0.2	0.2
λ	1.0	3.0	3.0

The value of λ for DRM was picked from $\{0.5, 1, 2, 3, 4, 5\}$ based on the performance on the ID validation set. For datasets lacking a publicly available validation split, we partitioned the training dataset into a training split and a validation split in the ratio of 4:1. Following Goyal et al. (2023), we also implemented early stopping based on the ID validation performance.

In all experiments, we adopted the standard CLIP image pre-processing including resizing, center cropping, and normalization. We also adopted the standard CLIP text pre-processing including the tokenization of texts into a series of integers, each representing a unique series of characters.

E.4. Computation Resources

All experiments were conducted on a high-performance computing cluster equipped with NVIDIA DGX H800 nodes. Two H800 GPUs with 80 GB memory were utilized to fine-tune CLIP ViT-B/16 and CLIP ViT-L/14, while four H800 GPUs were employed to fine-tune CLIP ViT-L/14@336. Using the machines, fine-tuning on iWildCam requires approximately 8 GPU hours, on FMoW approximately 6 GPU hours, and on ImageNet between 24 to 30 GPU hours.

F. Additional Experiment Results

F.1. Detailed Performance on ImageNet OOD Test Sets

The average accuracy across the five ImageNet OOD test sets has been presented in Table 1. We report the detailed results for each OOD test set in Table 4. Without WiSE-FT, DRM substantially outperforms the previous best fine-tuning results by FLYP on ImageNet-R and ImageNet-A, with increases from 71.4 to 77.8 and from 48.1 to 53.3, respectively. Meanwhile, the ID performance is at a comparable level. With WiSE-FT, the improvements remain significant, rising from 76.0 to 79.5 on ImageNet-R and from 53.0 to 54.2 on ImageNet-A.

Table 4. Performance on ImageNet OOD variants with CLIP ViT-B/16. "OOD" stands for the average performance over the OOD datasets.

	w/o WiSE-FT							WiSE-FT						
Method	ID	Im-V2	Im-R	Im-A	Sketch	ONet	OOD	ID	Im-V2	Im-R	Im-A	Sketch	ONet	OOD
0-shot	68.3	61.9	77.7	50.0	48.3	55.4	58.7	68.3	61.9	77.7	50.0	48.3	55.4	58.7
LP	79.9	69.8	70.8	46.4	46.9	52.1	57.2	80.0	70.3	72.4	47.8	48.1	52.8	58.3
FT	81.3	71.2	66.1	37.8	46.1	53.3	54.9	82.5	72.8	74.9	48.1	51.9	59.0	61.3
L2-SP	81.7	71.8	70.0	42.5	48.5	56.2	57.8	82.2	72.9	75.1	48.6	51.4	58.9	61.4
LP-FT	81.7	72.1	73.5	47.6	50.3	58.2	60.3	82.1	72.8	75.3	50.1	51.7	59.2	61.8
FLYP	**82.6**	73.0	71.4	48.1	49.6	**58.7**	60.2	**82.9**	73.5	76.0	53.0	52.3	**60.8**	63.1
DRM	82.0	**73.4**	**77.8**	**53.3**	**52.5**	58.6	**63.2**	82.4	**73.9**	**79.5**	**54.2**	**52.8**	59.7	**64.0**

F.2. Performance on Dollar Street-DA and GeoYFCC-DA

We followed the train-test split outlined by Prabhu et al. (2022). As there was no dedicated validation set, we split 20% of the training set for validation purposes. The ID and OOD performance results are reported based on the ID performance on the validation set and the OOD performance on the test set, which consists of images from countries not included in the training and validation sets.

The results presented in Table 5 demonstrate that, compared to FLYP-trained models, DRM-trained models exhibit improved performance on images from new countries.

Table 5. ID and OOD performance on Dollar Street-DA and GeoYFCC-DA with CLIP ViT-B/16.

	Dollar Street-DA		GeoYFCC-DA	
Method	ID	OOD	ID	OOD
0-shot	64.0±0.0	53.7±0.0	56.2±0.0	52.3±0.0
FLYP	**82.4±0.3**	71.8±0.2	71.0±0.3	58.0±0.3
FLYP+WiSE-FT	**82.4±0.2**	72.7±0.2	71.2±0.3	58.7±0.2
DRM	81.4±0.2	73.9±0.3	**71.8±0.4**	62.5±0.4
DRM+WiSE-FT	82.0±0.1	**74.7±0.2**	**71.8±0.3**	**63.0±0.2**

F.3. Comparison to Some More Recent Methods

As discussed in Appendix A, there are some more recent robust fine-tuning methods. We include a comparison to some of those methods based on the results of fine-tuning CLIP ViT-B/16 on iWildCam and FMoW datasets. The results are reported in Table 6. The results clearly show that, the more recent methods still significantly lag behind DRM in term of OOD performance.

Table 6. Performance results for iWildCam and FMoW with CLIP ViT-B/16 including some more recent methods.

Method	iWildCam				FMoW			
	w/o WiSE-FT		WiSE-FT		w/o WiSE-FT		WiSE-FT	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
0-shot	8.7±0.0	11.0±0.0	-	-	20.4±0.0	18.7±0.0	-	-
LP	44.5±0.6	31.1±0.4	45.5±0.6	31.7±0.4	48.2±0.1	30.5±0.3	48.7±0.1	31.5±0.3
FT	48.1±0.5	35.0±0.5	48.1±0.5	35.0±0.5	68.5±0.1	39.2±0.7	68.5±0.1	41.5±0.5
L2-SP	48.6±0.4	35.3±0.3	48.6±0.4	35.3±0.3	68.6±0.1	39.4±0.6	68.4±0.1	40.3±0.6
LP-FT	49.7±0.5	34.7±0.4	50.2±0.5	35.7±0.4	68.4±0.2	40.4±1.0	68.5±0.2	42.4±0.7
FLYP	52.2±0.6	35.6±1.2	52.5±0.6	37.1±1.2	68.6±0.2	41.3±0.8	68.9±0.3	42.0±0.9
CLIPood	48.4±0.4	36.1±0.4	48.3±0.3	36.5±0.4	68.2±0.3	40.8±0.9	68.3±0.3	41.2±0.7
TPGM	47.5±0.3	35.9±0.4	46.8±0.3	36.2±0.3	68.4±0.3	39.6±0.8	67.8±0.2	39.9±0.7
LipSum-FT	50.7±0.8	36.6±0.7	48.4±0.5	36.9±0.6	68.4±0.3	41.3±1.0	68.1±0.3	42.0±0.5
CaRot	49.7±0.4	34.3±0.3	48.3±0.3	34.7±0.3	68.8±0.2	39.8±0.6	68.3±0.2	40.7±0.5
DRM	**54.1±0.5**	**40.0±0.6**	**55.3±0.4**	**41.4±0.7**	**68.7±0.3**	**45.9±1.1**	68.7±0.2	**46.1±0.8**

F.4. Performance of DRM under Different λ

As in (6), our DRM training involves a hyperparameter λ to balance the contribution of empirical risk and worst-case risk. When $\lambda = 0$, only empirical risk is involved in the training, resulting in an ERM-trained model. When λ is large, the effect of empirical risk is small, and the resulting model should be close to a WRM-trained model. In practical implementation, we choose the value of λ based on the performance on the ID validation set. In Table 7, we show the ID and OOD performance of CLIP ViT-L/14 fine-tuned on the iWildCam dataset with DRM objective under different choices of λ.

Table 7. ID and OOD performance for different values of λ based on CLIP ViT-L/14 model performance on iWildCam.

λ	0.0	0.1	0.5	1.0	2.0	3.0	4.0	5.0	10.0	50.0
ID	56.0	56.4	57.2	59.1	60.0	**61.8**	60.9	60.1	55.4	52.5
OOD	41.9	42.6	43.9	47.3	48.1	**49.2**	48.6	48.5	47.7	46.6

The results demonstrate that when λ is small, the model's performance is akin to that of an ERM-trained model, characterized by a relatively low OOD F1 score. As λ increases, the model progressively aligns with the characteristics of an ideal DRM

model, exhibiting enhancements in both ID and OOD performances, with the optimum performance achieved at $\lambda = 3$. However, further increases in λ, to values such as 10 or even 50, lead the model to converge towards a DRM-trained model. In this state, the model primarily relies on core visual features for making predictions. Such models can become overly restrictive for practical applications, failing to deliver robust ID and OOD performance.

F.5. Ablation Study

Setup. Given a labeled dataset $\{(x_i, y_i)\}_{i=1}^N$ sampled from the training domain d_s, the final DRM objective (6), i.e., $R_{\mathrm{s}}(\theta; \mathcal{T}^{\mathtt{df}}) + \lambda \tilde{R}_{\mathrm{s}}^{\mathrm{c}}(\theta; \mathcal{T}^{\mathtt{cd}})$, for fine-tuning zero-shot models can be expanded as

$$\frac{1}{N}\sum_{i=1}^{N}\Big[-\log p_\theta^{\mathtt{df}}(y_i|x_i) - \lambda \sum_{y'\in\mathcal{Y}} p_{\theta_0}^{\mathtt{pr}}(y'|x_i)\log p_\theta^{\mathtt{cd}}(y'|x_i)\Big], \tag{10}$$

where $p_\theta^{\mathtt{df}}(y|x)$ and $p_\theta^{\mathtt{cd}}(y|x)$ are the classifiers (3) induced by the default prompts $\mathcal{T}^{\mathtt{df}} = \{t_y^{\mathtt{df}} \mid y \in \mathcal{Y}\}$ and the concept descriptions $\mathcal{T}^{\mathtt{cd}} = \{t_y^{\mathtt{cd}} \mid y \in \mathcal{Y}\}$, respectively; and $p_{\theta_0}^{\mathtt{pr}}(y|x) = \tilde{p}_{\mathrm{c}}(y|x)$ is defined by (4) to estimate $p_{\mathrm{c}}(y|x)$. Here, we use $p_{\theta_0}^{\mathtt{pr}}(y|x)$ (where `pr` stands for 'proxy') instead of $\tilde{p}_{\mathrm{c}}(y|x)$ to ease the discussion of possible variations of DRM.

Consider the following generalized form of (10) with three varying options, `t1` and `t2` indicating the classifier types defined with different sets of text prompts, and `type` indicating the type of model used as the proxy for $p_{\mathrm{c}}(y|x)$:

$$\frac{1}{N}\sum_{i=1}^{N}\Big[-\log p_\theta^{\mathtt{t1}}(y_i|x_i) - \lambda \sum_{y'\in\mathcal{Y}} p_{\theta_0}^{\mathtt{type}}(y'|x_i)\log p_\theta^{\mathtt{t2}}(y'|x_i)\Big], \tag{11}$$

As stated in (10), our final DRM training objective (6) uses $\mathtt{t1} = \mathtt{df}$ in the ERM term, with $\mathtt{t2} = \mathtt{cd}$ and $\mathtt{type} = \mathtt{pr}$ in the regularization term. We denote this as our standard setting, (S) in short. We conduct the following ablation study with the pre-trained CLIP ViT-L/14 and fine-tune the model on the iWildCam dataset, with results presented in Table 8.

Table 8. Ablation study on DRM with CLIP ViT-L/14 (w/o WiSE-FT) on iWildCam.

General setting			Specification					Performance	
			t1	t2	type	Classifier comb.	Infer w/	ID	OOD
	Standard DRM	(S)	df	cd	pr	joint training	(7)	**61.8**	**49.2**
(a)	Infer with one classifier after dual classifier training	(a1)	df	cd	pr	joint training	df	60.4	45.1
		(a2)	df	cd	pr	joint training	cd	54.8	47.2
(b)	Vanilla DRM using one set of text prompts	(b1)	df	df	pr	joint training	df	54.4	45.1
		(b2)	cd	cd	pr	joint training	cd	54.0	46.1
(c)	Use different proxy models	(c1)	df	cd	cd	joint training	(7)	32.1	24.2
		(c2)	df	cd	pr-df	joint training	(7)	54.4	45.1
		(c3)	df	cd	one-hot	joint training	(7)	57.3	45.1
(d)	Use only one risk for training	(d1)	df	/	/	/	df	56.0	41.9
		(d2)	cd	/	/	/	cd	56.9	43.4
		(d3)	/	cd	pr	/	cd	51.7	46.3
(e)	Combine independently trained classifiers	(e1)	df	cd	pr	model ensemble	(7)	59.7	45.7
		(e2)	df	cd	pr	weight average	(7)	57.5	44.7

(a) Inference options after dual classifier training: Two classifiers are involved in our DRM training: $p_\theta^{\mathtt{df}}(y|x)$ and $p_\theta^{\mathtt{cd}}(y|x)$. As outlined in (7), we combine both classifiers for inference. An alternative is to only use one of the two classifiers for inference. We denote inference with only $p_\theta^{\mathtt{df}}(y|x)$ as (a1), and with only $p_\theta^{\mathtt{cd}}(y|x)$ as (a2). The comparison between (a1), (a2), and (S) in Table 8 shows combining both classifiers for inference enhances both ID and OOD performance compared to using either alone. This reveals that the two classifiers have a complementary effect as illustrated in Figure 1, and corroborates our view that ERM and WRM are both vital to OOD robustness.

(b) Vanilla DRM using a single set of text prompts: In our standard DRM setting (S), $\texttt{t1} = \texttt{df}$ and $\texttt{t2} = \texttt{cd}$. The vanilla DRM we discussed in Section 3 uses $\texttt{t1} = \texttt{t2} = \texttt{df}$. Alternatively, one can also consider $\texttt{t1} = \texttt{t2} = \texttt{cd}$. We experiment with these two alternative settings denoted by (b1) and (b2) in Table 8. The contrast between (b1) and (S) confirms our intuition: using the concept descriptions $\mathcal{T}^{\texttt{cd}}$ for $p_\theta^{\texttt{t2}}(y|x)$, i.e., $\texttt{t2} = \texttt{cd}$, enhances robust feature preservation and leads to better OOD performance. The other alternative (b2), which employs $\mathcal{T}^{\texttt{cd}}$ for both $p_\theta^{\texttt{t1}}(y|x)$ and $p_\theta^{\texttt{t2}}(y|x)$, i.e., $\texttt{t1} = \texttt{t2} = \texttt{cd}$, slightly improves (b1). Intriguingly, (b2) is still much worse than (S) despite they both use $\texttt{cd}$ for $\texttt{t2}$.

(c) Proxy model design: In our standard setting, $\texttt{type} = \texttt{pr}$. As discussed in Section 3, the proxy term $p_{\theta_0}^{\texttt{pr}}(y|x_i)$ is based on the affinity $A_\theta(x, t_y^{\texttt{cd}}) = \langle f_\phi(x), g_\psi(t_y^{\texttt{cd}})\rangle$ according to the pre-trained CLIP model $\theta_0 = (\phi_0, \psi_0)$ and the set of concept descriptions $\mathcal{T}^{\texttt{cd}}$. In Section 3, we also mentioned the following direct estimation of the oracle model $p_{\text{c}}(y|x)$:

$$p_{\theta_0}^{\texttt{cd}}(y|x) = \frac{\exp(A_\theta(x, t_y^{\texttt{cd}})/\tau)}{\sum_{y' \in \mathcal{Y}} \exp(A_\theta(x, t_{y'}^{\texttt{cd}})/\tau)}. \tag{12}$$

However, as discussed in Section 3, $p_{\theta_0}^{\texttt{cd}}(y|x)$ is susceptible to artifact terms. Consequently, we made a technical adjustment to mitigate the influence of these terms, resulting in the refined $\tilde{p}_{\text{c}}(y|x)$, which is denoted as $p_{\theta_0}^{\texttt{pr}}(y|x)$ here. After the adjustment, there exists at least one x in $\mathcal{X}_y$ for which $p_{\theta_0}^{\texttt{pr}}(y|x) = 1$, a condition not necessarily fulfilled by $p_{\theta_0}^{\texttt{cd}}(y|x)$. The failure to fulfill this condition may weaken the regularization effect of the second term in the DRM objective. In comparison, the first term of the DRM objective, the ERM term, always pushes $p_\theta(y|x)$ to either 1 or 0. As shown in Table 8, the importance of this adjustment is empirically verified by the much lower performance of (c1) compared to (S).

One can also define $p_{\theta_0}^{\texttt{pr-df}}(y|x)$ by replacing $\mathcal{T}^{\texttt{cd}}$ with $\mathcal{T}^{\texttt{df}}$ in the formulation of $p_{\theta_0}^{\texttt{pr}}(y|x)$. As shown by the result of (c2), this alternative still underperforms $p_{\theta_0}^{\texttt{pr}}(y|x)$ used in the standard setting. This discrepancy can be explained by the fact that the affinities between default text prompts and images are easily affected by changes in the non-core visual features instead of focusing on the core visual features, which has been discussed in Appendix D.

Another simple alternative, denoted by (c3), is to employ the ground-truth one-hot labels as the proxy. Perhaps unsurprisingly, the OOD performance of (c3) is notably inferior to (S) based on the affinities between the images and the concept descriptions.

(d) Training with either ERM or WRM: Training with only the first term in (6) results in ERM models (d1) and (d2), whereas training with only the second term leads to a WRM model (d3). Comparing them with DRM models (a1) and (a2), it is clear that models trained to minimize a single risk underperform those trained to minimize both risks, highlighting the importance of dual risk minimization.

(e) Classifier combination strategy: Our standard DRM training jointly minimizes the two risks, but one can also train an ERM model $p_{\theta_{\text{ERM}}}^{\texttt{df}}(y|x)$ and a WRM model $p_{\theta_{\text{WRM}}}^{\texttt{cd}}(y|x)$ separately. These models can be combined for inference using techniques like model ensembling or weight-space averaging. The last two rows of Table 8 show that combining (d1) and (d3) via model ensembling or weight-space averaging generally underperforms joint training (S).

F.6. Results of Applying DRM on ImageNet Pre-trained ResNet50

While this work focus on the fine-tuning of zero-shot models that are pre-trained on large-scale image-text pairs, we also explore the possibility of applying DRM on fine-tuning the ImageNet pre-trained CNN models.

When implementing DRM on the CNN models, we introduce two randomly initialized classification heads on top of the model. Similar to the application of DRM on the zero-shot model, one classification head is trained using cross-entropy loss aligned with the ground-truth labels, while the other is trained using cross-entropy loss relative to the soft labels generated by the pre-trained zero-shot model. We employed this strategy to fine-tune an ImageNet pre-trained ResNet50 model on the iWildCam dataset, using soft labels generated by the CLIP ViT-L/14. The results are presented in Table 9. It is evident from the results that DRM significantly enhances the OOD performance of ResNet50 compared to the ERM.

G. Limitations

Our research utilizes GPT-4 to generate concept descriptions for the core visual features of various classes across different domains. It is important to note that the scope of GPT-4's knowledge in certain domains might be limited, and as a result, the model may not always generate useful concept descriptions. Additionally, due to the vast number of concept descriptions

Table 9. Results of applying DRM on fine-tuning ImageNet pre-trained ResNet50 on iWildCam.

Method	ID	OOD
ERM+FT	51.6	33.7
ERM+LP-FT	50.5	36.4
DRM+LP-FT	51.0	39.1

generated, we have not been able to verify the accuracy of each generated concept description. To enhance the quality of these descriptions, potential improvements could involve engaging domain experts to review and correct errors, or design descriptions manually. Another approach could be to gather visual prototypes and use advanced multimodal LLMs such as GPT-4V (Achiam et al., 2023), LLaVA (Liu et al., 2023), or MiniGPT-4 (Zhu et al., 2024), which might yield more precise descriptions of the core visual features.

Another limitation concerns the CLIP models used in our experiments. These models may not perform optimally across all domains, particularly in less common areas, where they may lack requisite knowledge in understanding both images and text. The effectiveness of our DRM method is therefore contingent upon the breadth and depth of the pre-training data of CLIP models. Unfortunately, the specifics of the CLIP pre-training dataset have not been disclosed by OpenAI, adding an element of uncertainty to the performance of our method in niche domains.