Dual Risk Minimization: Towards Next-Level Robustness in Fine-tuning Zero-Shot Models

Kaican Li[1*] **Weiyan Xie**[1*] **Yongxiang Huang**[2] **Didan Deng**[2] **Lanqing Hong**[2]
Zhenguo Li[1,2] **Ricardo Silva**[3] **Nevin L. Zhang**[1†]

[1]The Hong Kong University of Science and Technology
[2]Huawei [3]University College London

Abstract

Fine-tuning foundation models often compromises their robustness to distribution shifts. To remedy this, most robust fine-tuning methods aim to preserve the pre-trained features. However, not all pre-trained features are robust and those methods are largely indifferent to which ones to preserve. We propose dual risk minimization (DRM), which combines empirical risk minimization with worst-case risk minimization, to better preserve the core features of downstream tasks. 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 model robustness: expected performance and worst-case performance, establishing a new state of the art on various real-world benchmarks. DRM significantly improves the out-of-distribution performance of 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 robust fine-tuning. Our code is available at `https://github.com/vaynexie/DRM`.

1 Introduction

Foundation models such as CLIP (Radford et al., 2021) and ALIGN (Jia 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; Xu et al., 2024). 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 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 larger gaps between downstream in-distribution (ID) and out-of-distribution (OOD) performance (Wortsman et al., 2022).

Kumar et al. (2022) showed that fine-tuning tends to distort pre-trained 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, and then 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) investigated combining pre-trained models with their fine-tuned versions by weight averaging, which can be seen as yet another approach to recover robust features lost during fine-tuning.

[*] Equal contribution, listed in alphabetical order.
[†] Correspondence to lzhang@cse.ust.hk.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2110

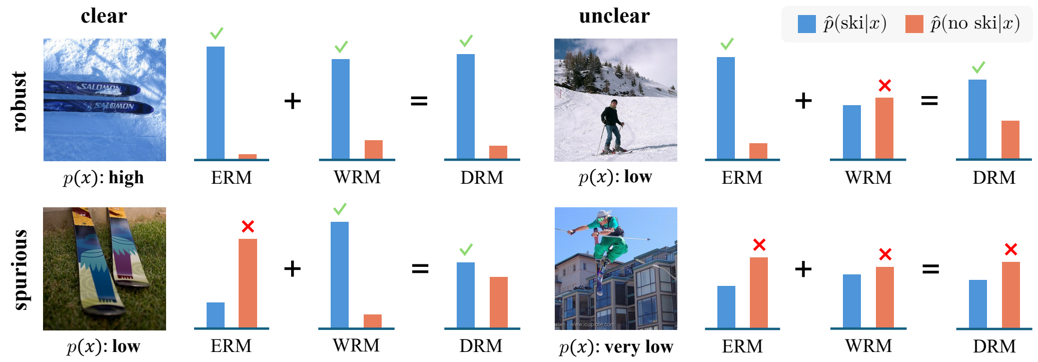


Figure 1: **Dual risk minimization (DRM) combines empirical risk minimization (ERM) and worst-case risk minimization (WRM) to complement their weaknesses.** In this simple binary classification task predicting if there are skis in a given image, (i) ERM underperforms when the core features (the appearance 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.

While the existing approaches aim to preserve pre-trained features, the fine-tuning process is still guided by empirical risk minimization (ERM; Vapnik, 1998), which favors the most predictive but not necessarily the most robust features. 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 non-core features even when the core features are clear (Geirhos et al., 2020; Shah et al., 2020). This often harms OOD performance as non-core features are generally less reliable out-of-distribution.

To better preserve the core features, we propose a new principle called *dual risk minimization* (DRM) which combines ERM with worst-case risk minimization (WRM; Wald, 1945), a common principle for domain generalization (Arjovsky et al., 2019; Sagawa et al., 2020; Cha et al., 2021; Kirichenko et al., 2023). This combination rests on our view that robustness involves *two* main aspects: the *expected* (or average) performance and the *worst-case* performance over all domains. While there is often a trade-off between these two aspects (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 applying DRM to real-world tasks is to assess of worst-case risk. To this end, we use *concept descriptions* (Pratt et al., 2023)—short texts that describe the core features of each class—obtained with GPT-4 (Achiam et al., 2023). The description for *cougar*, for instance, is "*a large, tawny cat with a muscular build and a small head.*" We feed these 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 for each training image according to the similarity scores between the image and text embeddings. The risk w.r.t. the soft labels can be seen as a proxy of the worst-case risk and is thus minimized instead. Empirically, DRM significantly outperforms the state of the art on challenging benchmarks such as ImageNet (Deng et al., 2009) and WILDS (Koh et al., 2021).

In summary, we make the following key contributions in this paper:

- We propose *dual risk minimization* (DRM), a novel approach that combines ERM and WRM to improve downstream robustness of zero-shot foundation models while addressing the intractability of WRM through innovative use of concept descriptions.
- We highlight that robustness for many real-world problems concerns both *expected* and *worst-case* performance while most previous works focus on only one. We then show that DRM offers a simple and effective way to balance these two important aspects of robustness.
- We establish a strong new state of the art on multiple real-world benchmarks, promising next-level robustness in fine-tuning zero-shot models. On CLIP ViT-L/14@336, DRM achieves a significant, over 5% relative improvement in OOD performance over the best baseline method.

2 Related work

Robust fine-tuning of pre-trained models. Prior to the work of Kumar et al. (2022); Wortsman et al. (2022); Goyal et al. (2023) which we have introduced, Li et al. (2018) proposed to restrict the L^2 distance between the parameters of pre-trained and fine-tuned models via regularization. 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 and Cohen, 1989). Apart from explicit constraints on model parameters, Ge and 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 OOD accuracy tends to improve initially but then plateaus as the fine-tuning proceeds. For more discussion on related work including concurrent ones, see Appendix B.

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 notable example is invariant risk minimization (IRM; Arjovsky et al., 2019), which aims to learn core-feature representations from multi-domain data. Such representations, under mild causal assumptions, give rise to classifiers that minimize the worst risk (Peters et al., 2016). Another key method, GroupDRO (Sagawa et al., 2020), imposes higher penalties on domains with higher empirical risks. Unlike DRM, neither IRM nor GroupDRO formulates WRM as an explicit optimization constraint for ERM. Eastwood et al. (2022) pointed out that sacrificing too much average performance for worst-case performance is not ideal for DG. Hence, they proposed to minimize the risk among the most likely domains. Lastly, our setup is partly similar to Alabdulmohsin et al. (2023) which also relies on external information.

Prompt design for zero-shot classification. To better leverage the capability of zero-shot models, various prompt designs have been proposed. Menon and Vondrick (2022); Pratt et al. (2023); Maniparambil et al. (2023) mainly explored prompts for zero-shot classification. Their prompts were generated by LLMs (Radford et al., 2019) with slightly different instructions than ours, not explicitly focusing on core features. For example, Pratt et al. (2023) used "*Describe an image from the internet of a(n) ...*", which may inadvertently introduce descriptions of 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 used LLM-generated concept descriptions to build concept bottleneck models for interpretable image classification. More recently, Mao et al. (2024) proposed to use context-aware prompts such as "*a [context] of [class name]*," while Cheng et al. (2024) used both domain-invariant and domain-specific prompts generated by LLMs. However, both methods require either image context or specific domain information to generate the prompts.

3 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* and X_n *non-core features* of (X, Y). 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 and 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 calligraphic letters such as $\mathcal{X}$ and $\mathcal{Y}$ to denote the set of possible outcomes of the random variables.

https://doi.org/10.52202/079017-2110

Ideal objective for 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 $\hat{p}_\theta(y|x)$ be a prediction model parameterized by $\theta \in \Theta$. Its risk in terms of negative log-likelihood,

$$R_d(\theta) = \mathbb{E}_{(x,y)\sim p_d}[-\log \hat{p}_\theta(y|x)], \tag{2}$$

can be seen as a measure of its performance in domain d. Let $d_{\mathrm{s}} \in \mathcal{D}$ be the training domain. For simplicity, we will omit d when it is clear from the context, e.g., $R_{d_{\mathrm{s}}}(\theta)$ will be written 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. We note that Eastwood et al. (2022) and Zhang et al. (2023b) share a similar view with us on robustness.

We formalize the above intuition as the following constrained optimization problem, namely *idealized dual risk minimization* (IDRM), which aims to minimize the empirical risk of $\hat{p}_\theta(y|x)$ while ensuring its worst-case risk is below some threshold value α:

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

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 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. Another closely related work, GroupDRO (Sagawa et al., 2020), proposes to minimize the worst training-domain risk—a more empirical flavor of WRM. Both IRM and GroupDRO rely on ideally grouped training data to capture invariance across domains. In Section 4, we will show that this is largely unnecessary for zero-shot models and provide a practical solution for IDRM with just *single-domain* data.

From IDRM to DRM. IDRM can be solved as the following unconstrained optimization problem due to strong duality (proof in Appendix A).

Theorem 1. *Strong duality holds between IDRM and the following dual problem:*

$$\max_{\lambda'\ge 0}\min_{\theta\in\Theta}\left[R_{\mathrm{s}}(\theta) + \lambda' \max_{d\in\mathcal{D}} R_d(\theta)\right] - \lambda'\alpha. \tag{3}$$

Let $\lambda^\star$ be any solution of λ' to (3). By the strong duality, IDRM then reduces to $\min_{\theta\in\Theta}[R_{\mathrm{s}}(\theta) + \lambda^\star \max_{d\in\mathcal{D}} R_d(\theta)]$. The worst-case risk, i.e., $\max_{d\in\mathcal{D}} R_d(\theta)$, is still intractable in itself, but it is closely related to the degree to which the model $\hat{p}_\theta(y|x)$ relies on core features to predict y. This is 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). To minimize the worst-case risk, therefore, the model must only utilize the core features to make prediction.

Suppose there is an oracle feature extractor f_{c} that returns a faithful representation of the core features of any input x. As the core features may not always be clear, $f_{\mathrm{c}}(x)$ can be viewed as some distribution over the core features for each x. Let $p_{\mathrm{c}}(y|x)$ be the optimal model that can be built upon $f_{\mathrm{c}}(x)$. The risk of $\hat{p}_\theta(y|x)$ w.r.t. $p_{\mathrm{c}}(y|x)$ on the training domain d_{s} is given by

$$R^{\mathrm{c}}_{\mathrm{s}}(\theta) = \mathbb{E}_{x\sim p_{\mathrm{s}}}\mathbb{E}_{y\sim p_{\mathrm{c}}(y|x)}[-\log \hat{p}_\theta(y|x)]. \tag{4}$$

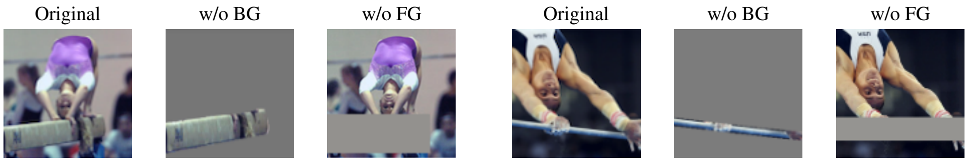
Assuming $p_{\mathrm{s}}(x)$ is fairly diverse, the risk $R^{\mathrm{c}}_{\mathrm{s}}(\theta)$ measures the degree to which the model's prediction is based on the core features and thus can be viewed as a proxy for the worst-case risk. Hence, we can replace $\max_{d\in\mathcal{D}} R_d(\theta)$ with $R^{\mathrm{c}}_{\mathrm{s}}(\theta)$ while still achieving a similar optimization effect.

In summary, we relax IDRM to the following DRM formulation:

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

with some properly chosen $\lambda \ge 0$. Now the risk $R^{\mathrm{c}}_{\mathrm{s}}(\theta)$ can also be interpreted as a regularization term for ERM to help preserve the core features. In the following section, we demonstrate how to utilize zero-shot models like CLIP models (Radford et al., 2021) to estimate $p_{\mathrm{c}}(y|x)$, and subsequently how to apply DRM to robustly fine-tune the same CLIP models.

https://doi.org/10.52202/079017-2110



	Original	w/o BG	w/o FG	Original	w/o BG	w/o FG
df	An image of balance beam.			An image of gymnastic horizontal bar.		
cd	A long, thin piece of wood or metal elevated off the ground.			Long metal or wood bar held up by upright supports.		
df	0.367	0.262 (–28.6%)	0.395 (+7.6%)	0.333	0.267 (–19.8%)	0.367 (+10.2%)
cd	0.286	0.281 (–1.7%)	0.123 (–57.0%)	0.258	0.273 (+5.8%)	0.131 (–49.2%)

Figure 2: **Concept descriptions better capture core features than default prompts.** The affinities between images and default prompts (df) are not stable w.r.t. changes in image background (BG) containing non-core features and are insensitive to changes in image foreground (FG) containing core features, as indicated by the relative changes (gray numbers in parentheses) w.r.t. the affinities of the original images. In contrast, the affinities between images and concept descriptions (cd) are stable w.r.t. to changes in BG while being highly responsive to changes in FG, making them a good detector for core features. See Appendix D.1 for more examples and a full quantitative study on this.

4 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

$$\hat{p}_\theta(y|x) = \frac{\exp(A_\theta(x, t_y)/\tau)}{\sum_{y' \in \mathcal{Y}} \exp(A_\theta(x, t_{y'})/\tau)}, \tag{5}$$

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 $A_\theta(x, t_y)$.

The classifier (5) was originally introduced by Radford et al. (2021) for zero-shot classification. We follow Goyal et al. (2023) to directly fine-tune this classifier, viewing the text embeddings $g_\psi(t_y)$ as the weights of a standard linear classification head for the image embeddings $f_\phi(x)$. Unlike standard classifiers, however, (5) additionally depends on the text prompt t_y of which the design is important.

4.1 Dual prompts for fine-tuning zero-shot models

Let $\mathcal{T} = \{t_y \,|\, y \in \mathcal{Y}\}$ be the set of text prompts used to construct the classifier $p_\theta(y|x)$ according to Eq. (5). For such zero-shot classifiers, the general DRM objective (DRM) becomes

$$\min_{\theta \in \Theta} R_{\mathrm{s}}(\theta; \mathcal{T}) + \lambda R_{\mathrm{s}}^{\mathrm{c}}(\theta; \mathcal{T}), \tag{6}$$

where the risks R_{s} and $R_{\mathrm{s}}^{\mathrm{c}}$ not only depend on the model parameters θ but also on the prompts $\mathcal{T}$. For R_{s}, usually a set of *default prompts*, $\mathcal{T}^{\mathtt{df}} = \{t_y^{\mathtt{df}} \,|\, y \in \mathcal{Y}\}$, like "*an image of [class name]*" is used (Goyal et al., 2023; Oh et al., 2023). However, such prompts may not be suitable for $R_{\mathrm{s}}^{\mathrm{c}}$ as they are not specifically designed to bind with core features. In fact, as we will show, default prompts elicit representation that is biased towards non-core features. It is also know that non-visual and spurious descriptions contribute significantly to CLIP's representation (Esfandiarpoor et al., 2024).

To generate better prompts for $R_{\mathrm{s}}^{\mathrm{c}}$, we ask GPT-4 (Achiam et al., 2023) to describe the *core visual features* of each class, producing a set of *concept descriptions*, $\mathcal{T}^{\mathtt{cd}} = \{t_y^{\mathtt{cd}} \,|\, y \in \mathcal{Y}\}$. For example, to generate a concept description for *cougar*, we prompt GPT-4 with

> "*Q: Generate a short sentence that describes the visual features of Cougar. 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].*"

and the concept description returned is

> "*a large, tawny cat with a muscular build and a small head.*" [3]

[3] More details about concept description generation can be found in Appendix C.

Figuratively, the text embedding $g_\psi(t_y^{\mathtt{cd}})$ of the concept description represents the core features of the class from the text side. We use it to “pull out” the core features from the image embedding $f_\phi(x)$ via inner-product. As illustrated in Figure 2, the affinity between an image and its concept description is indeed a much better measure of the significance of core visual features. Regarding this point, a full quantitative study can be found in Appendix D.1.

Together, the two sets of prompts naturally give rise to the following objective:

$$\min_{\theta\in\Theta} R_{\mathrm{s}}(\theta;\mathcal{T}^{\mathtt{df}}) + \lambda R_{\mathrm{s}}^{\mathrm{c}}(\theta;\mathcal{T}^{\mathtt{cd}}), \tag{7}$$

where we use default prompts $\mathcal{T}^{\mathtt{df}}$ for ERM and concept descriptions $\mathcal{T}^{\mathtt{cd}}$ for WRM. We do not use concept descriptions for ERM because it would erode the text-side core feature representation elicited by concept descriptions. This is supported by empirical evidence from our ablation study (Section 5.3) showing that (7) works best among various alternatives.

The dual prompts elicit separate predictions from the same model for the two sub-objectives of DRM. The ERM part, $R_{\mathrm{s}}(\theta;\mathcal{T}^{\mathtt{df}})$, is supervised by regular one-hot labels. The WRM part, $R_{\mathrm{s}}^{\mathrm{c}}(\theta;\mathcal{T}^{\mathtt{cd}})$, is supervised by $p_{\mathrm{c}}(y|x)$ as in (4). Next, we show how the same set of concept descriptions $\mathcal{T}^{\mathtt{cd}}$ can be used to obtain a good estimate of $p_{\mathrm{c}}(y|x)$.

4.2 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 core features. Since we have demonstrated that concept descriptions bind well with core features on pre-trained CLIP models, a direct estimate for $p_{\mathrm{c}}(y|x)$ can be obtained via (5) with $t_y \leftarrow t_y^{\mathtt{cd}}$ and $\theta \leftarrow \theta_0$ where $\theta_0 = (\phi_0, \psi_0)$ denote the pre-trained CLIP parameters.

However, there is a crucial caveat. To illustrate, consider an image x of class y. For another class y' whose core features are *not* present in x, the affinity $A_{\theta_0}(x, t_{y'}^{\mathtt{cd}})$ should ideally be very small. In practice, however, we find this is seldom the case. These extraneous affinity values, which we call *artifact terms*, often vary among classes and lead to poor estimates of $p_{\mathrm{c}}(y|x)$ with high entropy.

To mitigate the impact of artifact terms, we perform a simple min-max normalization on $\xi(x,y) = \exp(A_{\theta_0}(x, t_y^{\mathtt{cd}})/\tau)$ w.r.t. all training images $\mathcal{X}_y \subseteq \mathcal{X}$ labeled the same class y, as follows:

$$\gamma(x,y) = \frac{\xi(x,y) - \min_{x'\in\mathcal{X}_y}\xi(x',y)}{\max_{x'\in\mathcal{X}_y}\xi(x',y) - \min_{x'\in\mathcal{X}_y}\xi(x',y)}. \tag{8}$$

This effectively adjusts the affinity range of each class, reducing the difference in the artifact terms of different classes. Based on the normalization, the final estimation we propose for $p_{\mathrm{c}}(y|x)$ is

$$\tilde{p}_{\mathrm{c}}(y|x) = \begin{cases} \gamma(x,y), & y = y_x; \\ [1-\gamma(x,y_x)]\cdot\frac{\gamma(x,y)}{\sum_{y'\neq y_x}\gamma(x,y')}, & y \neq y_x; \end{cases} \tag{9}$$

where y_x is the ground-truth label of x. This ensures that for every class y, there exists at least one $x \in \mathcal{X}_y$ for which $\tilde{p}_{\mathrm{c}}(y|x) = 1$, promoting balanced learning. When $\tilde{p}_{\mathrm{c}}(y|x) < 1$ for $y = y_x$, the remaining probability is distributed to other classes $y \neq y_x$ according to the relative scale of the respective affinities.

We pre-compute the estimate $\tilde{p}_{\mathrm{c}}(y|x)$ with the pre-trained CLIP model θ_0 before fine-tuning, which is now the learning target of $R_{\mathrm{s}}^{\mathrm{c}}(\theta;\mathcal{T}^{\mathtt{cd}})$. Since the computation requires the class labels, $\tilde{p}_{\mathrm{c}}(y|x)$ can only be used for training (not for inference). Intuitively, learning from $\tilde{p}_{\mathrm{c}}(y|x)$ can be seen as a form of self-distillation targeted at core features. In comparison, standard self-distillation methods (Furlanello et al., 2018; Zhang et al., 2019; Ji et al., 2021; Zhang et al., 2021) are largely indifferent to what specific information should be distilled.

4.3 Inference

The fine-tuning objective (7) involves two classifiers: the ERM classifier $\hat{p}_\theta^{\mathtt{df}}(y|x)$ induced by $\mathcal{T}^{\mathtt{df}}$, and the WRM classifier $\hat{p}_\theta^{\mathtt{cd}}(y|x)$ induced by $\mathcal{T}^{\mathtt{cd}}$. While either alone can be used for inference, we find that their mixture,

$$\hat{p}_\theta^{\mathtt{dual}}(y|x) = \beta\cdot\hat{p}_\theta^{\mathtt{df}}(y|x) + (1-\beta)\cdot\hat{p}_\theta^{\mathtt{cd}}(y|x), \tag{10}$$

where $\beta \in (0,1)$, performs the best. This is expected as (10) essentially combines ERM with WRM as depicted in Figure 1. By default, we set $\beta = 1/(1+\lambda)$ so to be as consistent with (7) as possible.

https://doi.org/10.52202/079017-2110

5 Experiments

In this section, we evaluate DRM on multiple real-world benchmarks and conduct ablation studies to assess the impacts of various design choices. We conduct our experiments on three varying sizes of pre-trained CLIP models: ViT-B/16, ViT-L/14 and ViT-L/14@336 (Radford et al., 2021). Finally, we analyze the reliability of LLM-generated concept descriptions and the impact of λ on performance.

5.1 Setup

Datasets. IMAGENET (Deng et al., 2009) comprises over a million natural images across 1,000 classes. We use the training set for fine-tuning and the validation set for assessing ID accuracy. For OOD evaluation, we consider ImageNet variants: IMAGENET-V2 (Recht et al., 2019), IMAGENET-R (Hendrycks et al., 2021a), IMAGENET-SKETCH (Wang et al., 2019), IMAGENET-A (Hendrycks et al., 2021b), and OBJECTNET (Barbu et al., 2019). We report accuracy for both ID/OOD performance.

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 for land use prediction. The dataset is split into training, validation, and testing domains based on the year of collection. There is also a notable shift between different continents. We report the ID testing accuracy and the worst-region OOD testing accuracy.

DOLLAR STREET-DA and GEOYFCC-DA (Prabhu et al., 2022) are datasets for testing model generalization from images in specific countries to new ones. For Dollar Street-DA, training images are from North America and Europe, with testing images from other continents. GeoYFCC-DA has a similar setup. Model effectiveness is measured by accuracy in seen and unseen countries.

Baseline methods. The key baseline we compare our DRM method with is **FLYP** (Goyal et al., 2023). Following the FLYP paper, we include several baselines that do not utilize the text encoder. These methods are **LP** (linear probing), **FT** (fine-tuning), **L2-SP** (Li et al., 2018), and **LP-FT** (Kumar et al., 2022). In addition, we incorporate some more recent fine-tuning methods for zero-shot vision models. We also consider combining the weight-space averaging method, **WiSE-FT** (Wortsman et al., 2022), with DRM and the baselines. For more introduction to these methods, see Appendix B.

Implementation details. We update both the image encoder and text encoder during fine-tuning, following FLYP (Goyal et al., 2023). Furthermore, FLYP uses the CLIP contrastive loss (Radford et al., 2021) instead of the standard cross-entropy loss. We adopt this approach for the ERM part of DRM to facilitate comparison. In short, when the hyperparameter $\lambda = 0$ in the DRM objective (7), the WRM loss term vanishes and DRM reduces to exactly FLYP.

We choose all hyperparameters of DRM and baseline methods based on the performance on the ID validation set, i.e., training-domain validation (Gulrajani and Lopez-Paz, 2021). The hyperparameter λ of DRM is picked from $\{1, 2, 3, 4, 5\}$. More implementation details are presented in Appendix F.

5.2 Main results

We report the main results on IMAGENET, IWILDCAM, and FMOW in Table 1 and 2. The results on DOLLAR STREET-DA and GEOYFCC-DA can be found in Appendix E.2. We also compare DRM with some concurrent methods in Appendix E.3. All performance statistics, except some reported by previous papers (which we simply reuse), are averaged over 5 runs with different random seeds. The 95% confidence intervals over the 5 runs are reported.

Table 1 shows the results on CLIP ViT-B/16, the smallest of the three CLIP models. DRM achieves consistently better OOD performance than the baselines across all datasets, with and without WiSE-FT. Without WiSE-FT, DRM attains 5.0%, 12.4%, and 11.1% relative improvements over the best baseline method, FLYP, on the three benchmarks respectively. With WiSE-FT, the improvements remain significant at 1.9%, 11.6%, and 9.8% respectively. In terms of ID performance, DRM is roughly on par with FLYP, with a notable advantage on IWILDCAM.

https://doi.org/10.52202/079017-2110

Table 1: ID and OOD performances of DRM and baselines methods on CLIP ViT-B/16, with and without WiSE-FT. 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 E.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: ID and OOD performances of DRM and FLYP on two larger CLIP models.

		IMAGENET		IWILDCAM		FMOW	
Pre-trained model	Method	ID	OOD	ID	OOD	ID	OOD
CLIP ViT-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
	FLYP+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
	DRM+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
CLIP ViT-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
	FLYP+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
	DRM+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

Table 2 compares the performance of DRM and FLYP on two larger CLIP models. DRM again consistently outperforms FLYP in all cases. The previous state-of-the-art OOD performance for IWILDCAM and FMOW are 47.1 and 50.6 respectively, both achieved by FLYP+WiSE-FT with CLIP ViT-L/14@336. DRM improves those scores by 10.0% and 5.0% to 51.8 and 53.1 respectively.

Although DRM incurs more computational costs compared to FLYP due to an additional pass through the text encoder, the training and inference costs for DRM only increase by about 20% from FLYP (see Appendix F.3). This cost is insignificant relative to the performance gain.

Overall, our experiments suggest that DRM is highly effective in the robust fine-tuning of zero-shot models, outperforming previous methods by a large margin while maintaining scalability. As a bonus, we show in Appendix E.6 that DRM is also effective in fine-tuning ImageNet pre-trained CNNs.

5.3 Ablation study

We conduct our ablation study with CLIP ViT-L/14 on IWILDCAM. The main results are reported in Table 3 and discussed below. For additional results and details, see Appendix E.4 and E.5.

Impact of dual risks. The DRM objective (7) consists of an ERM term, $R_{\text{s}}(\theta; \mathcal{T}^{\text{df}})$, and a WRM term, $R_{\text{s}}^{\text{c}}(\theta; \mathcal{T}^{\text{cd}})$. From Table 3, we can see that, in terms of OOD performance, models fine-tuned with only the ERM term (Rows 4 & 5) significantly underperform models fine-tuned with both terms (Rows 1-3). Conversely, models fine-tuned with only the WRM term (Rows 6 & 7) have much worse ID performance. Although the OOD performance is improved, there is still a large gap from the DRM model (Row 1). Note that these hold regardless of the type of text prompts used by ERM/WRM.

Impact of dual prompts for fine-tuning. With both ERM and WRM in effect, we further investigate the impact of their prompts during fine-tuning. DRM uses two sets of prompts: default prompts $\mathcal{T}^{\text{df}}$ for ERM, and concept descriptions $\mathcal{T}^{\text{cd}}$ for WRM. Rows 8 & 9 of Table 3 show the performance of models fine-tuned using the same set of prompts for ERM and WRM. In either case, the model underperforms DRM (Row 1), validating the use of tailored prompts for specific learning targets.

https://doi.org/10.52202/079017-2110

Table 3: Results of ablation studies on DRM with CLIP ViT-L/14 performance and iWildCam. We use "`df`" and "`cd`" to denote the type of text prompts used to produce model predictions. "`dual`" refers to the mixture model (10) for inference. "–" means the corresponding loss term is not in use.

	ERM	WRM	Affinity	Inference	Performance	
Row	$R_{\mathrm{s}}(\theta;\mathcal{T})$	$R^{\mathrm{c}}_{\mathrm{s}}(\theta;\mathcal{T})$	norm.	w/ model	ID	OOD
1	`df`	`cd`	✓	`dual`	**61.8**	**49.2**
2				`df`	60.4	45.1
3				`cd`	54.8	47.2
4	`df`	–	–	`df`	56.0	41.9
5	`cd`	–	–	`cd`	56.9	43.4
6	–	`df`	✓	`df`	52.4	45.3
7	–	`cd`	✓	`cd`	51.7	46.3
8	`df`	`df`	✓	`df`	54.4	45.1
9	`cd`	`cd`	✓	`cd`	54.0	46.0
10	`df`	`cd`	✗	`dual`	32.1	24.2

Intuitively, $\mathcal{T}^{\mathtt{cd}}$ is more aligned with WRM than $\mathcal{T}^{\mathtt{df}}$ as the learning targets (9) for WRM are based on the predictions of the pre-trained model prompted by $\mathcal{T}^{\mathtt{cd}}$. So, reducing the WRM term $R^{\mathrm{c}}_{\mathrm{s}}(\theta;\mathcal{T}^{\mathtt{cd}})$ in some sense limits the divergence of the predictions between the pre-trained models and the fine-tuned models, hence helping better preserve pre-trained (core) features. This explains why using $\mathcal{T}^{\mathtt{cd}}$ for both ERM and WRM would lead to poorer outcomes (Row 1 *vs*. 9). ERM targets, typically one-hot class labels, do not capture subtle core visual differences. As a result, ERM would weaken the bond between $\mathcal{T}^{\mathtt{cd}}$ and core visual features as perceived by the model during fine-tuning, and therefore reduce the power of WRM in preserving those features.

Interestingly, comparing Row 1 & 8, we find that $\mathcal{T}^{\mathtt{cd}}$ not only improves OOD performance but also ID performance. We hypothesize that this is because the core features are better preserved with $\mathcal{T}^{\mathtt{cd}}$ and thus the fine-tuned model relies on a more diverse set of features to make predictions, reducing overfitting and improving ID generalization as well.

Impact of dual prompts for inference. After DRM fine-tuning, we obtain a new CLIP model with updated parameters θ. Since the model is fine-tuned with dual prompts, it is natural to use the same dual prompts for inference, as in (10). Indeed, we find that $\hat{p}^{\mathtt{dual}}_{\theta}$ (Row 1) outperforms $\hat{p}^{\mathtt{df}}_{\theta}$ (Row 2) and $\hat{p}^{\mathtt{cd}}_{\theta}$ (Row 3) in Table 3. In particular, while $\hat{p}^{\mathtt{df}}_{\theta}$ is better than $\hat{p}^{\mathtt{cd}}_{\theta}$ in-distribution and the latter is better out-of-distribution, they underperform $\hat{p}^{\mathtt{dual}}_{\theta}$ on both ID and OOD fronts. This is similar to the phenomenon we have observed in the fine-tuning scenario. Dual prompts generally reduce overfitting and the effect carries over from fine-tuning to inference.

Impact of affinity normalization. In Section 4.2, we pointed out a problematic issue regarding the direct estimate for $p_{\mathrm{c}}(y|x)$ obtained via (5) with $t_y \leftarrow t^{\mathtt{cd}}_y$ and $\theta \leftarrow \theta_0$. To address the issue, we proposed another estimate $\tilde{p}_{\mathrm{c}}(y|x)$ in (9) based on normalized affinities. Comparing the two approaches, Row 10 of Table 3 shows that the direct estimate leads to severe degradation in both ID and OOD performance, demonstrating the importance of affinity normalization.

5.4 Reliability of LLM-generated concept descriptions

Consistency across repeated generations. The concept descriptions used in our experiments are generated by GPT-4. Since the generation process is stochastic, it might impact the performance of DRM. To evaluate this impact, we repeatedly ask GPT-4 to generate a concept description for each iWildCam class for three times and then find the standard deviation of the resulting image-text affinities. The average standard deviation over 20,000 randomly sampled images of iWildCam is 0.0061, which is very small compared to the mean affinity, 0.2659. This suggests that the affinities are robust to the randomness in the generation process, which is therefore unlikely to have any noticeable

Table 4: Performance of fine-tuned CLIP ViT-L/14 on iWildCam with concept descriptions generated by different LLMs of various sizes.

Method	LLM (#params)	ID	OOD
FLYP	N/A	52.2	35.6
DRM	GPT-3.5 (20B?)	53.4	38.7
DRM	GPT-4 (>1T?)	54.1	40.0
DRM	Llama-3 (8B)	53.8	39.2
DRM	Llama-3 (70B)	54.0	39.9
DRM	Llama-3 (405B)	53.9	40.5

Table 5: Performance of DRM under different λ on iWildCam and ImageNet with CLIP ViT-L/14 and CLIP ViT-B/16 respectively.

	iWildCam		ImageNet	
λ	ID	OOD	ID	OOD
0.0	56.0	41.9	**82.6**	60.2
1.0	59.1	47.3	81.5	62.5
2.0	60.0	48.1	81.8	63.1
3.0	**61.8**	**49.2**	82.0	63.2
4.0	60.9	48.6	81.9	**63.4**
5.0	60.1	48.5	81.7	63.3

impact on the performance of DRM. Examples of the generated descriptions in Appendix D.2 show that the same core visual features are described quite consistently across generations.

Generation across different LLMs. We also experiment with different LLMs of various sizes (from 8B to over 400B parameters) to generate concept descriptions. The results are reported in Table 4. While larger and more advanced models lead to better performance, DRM is not sensitive to the specific choice of LLM for generating the concept descriptions. Even with a relatively small LLM, Llama-3 (8B), DRM still maintains a significant edge over FLYP on iWildCam. Examples of the concept descriptions generated by the LLMs are provided in Appendix D.3.

5.5 Study on the effect of λ in DRM

As in (7), DRM involves a hyperparameter λ that balances the weight of the empirical risk and the worst-case risk. When $\lambda = 0$, only the empirical risk is involved during fine-tuning, resulting in an ERM model. As λ increases, the influence of the empirical risk reduces, and the resulting model becomes closer to a WRM model. In practice, we choose the value of $\lambda > 0$ based on ID validation, which is often positively correlated with OOD performance (Taori et al., 2020; Miller et al., 2021). In Table 5, we show the ID and OOD performance of CLIP ViT-L/14 fine-tuned on iWildCam and CLIP ViT-B/16 fine-tuned on ImageNet under different choices of λ.

Compared to FLYP ($\lambda = 0$), the results indicate that DRM maintains high-level OOD performance across λ between 1 and 5, suggesting that DRM is fairly insensitive to the choice of λ. Furthermore, it is interesting to note that ID and OOD performance improve together on iWildCam as λ increases to 3.0. Our further analysis reveals that this occurs because DRM assists in reducing overfitting—there is a decrease in training accuracy (from 88.29 to 87.41) and an increase in validation accuracy (from 81.64 to 82.43) as λ increases (from 1 to 3)—thereby enhancing both ID and OOD performance. While DRM does not always improve ID performance (e.g., on ImageNet), the result suggests that it still achieves a good trade-off between ID and OOD performance (-0.6 ID performance for +3.2 OOD performance), which in turn could lead to a better Pareto front.

6 Conclusion

In conclusion, this paper introduces dual risk minimization (DRM), a novel method that enhances the robustness of models fine-tuned from zero-shot foundation models against distribution shifts. DRM combines ERM with WRM, focusing on the preservation of core features which essentially define the target classes. To guide the fine-tuning process, DRM utilizes concept descriptions generated by LLMs like GPT-4. By balancing expected and worst-case performance, DRM overcomes the traditional limitations of ERM and achieves significant OOD performance improvements on multiple real-world benchmarks, establishing a new state of the art. Potential future directions include a deeper theoretical investigation into DRM as a general principle, using DRM to improve the general robustness of zero-shot models across a broad range of tasks, and better understanding the role of concept descriptions in vision-language modeling.

Acknowledgements

This research 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. Ricardo Silva acknowledges support of the UKRI AI programme, and the Engineering and Physical Sciences Research Council, for CHAI - EPSRC AI Hub for Causality in Healthcare AI with Real Data (grant number EP/Y028856/1).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or invariant risk minimization? a sample complexity perspective. In *ICLR*, 2021.

Ibrahim Alabdulmohsin, Nicole Chiou, Alexander D'Amour, Arthur Gretton, Sanmi Koyejo, Matt J Kusner, Stephen R Pfohl, Olawale Salaudeen, Jessica Schrouff, and Katherine Tsai. Adapting to latent subgroup shifts via concepts and proxies. In *AISTATS*, pages 9637–9661. PMLR, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pages 23716–23736, 2022.

Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

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

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *NeurIPS*, 32, 2019.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

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

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *NeurIPS*, volume 34, pages 22405–22418, 2021.

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *ECCV*, pages 440–457. Springer, 2022.

De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *CVPR*, pages 23595–23604, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

https://doi.org/10.52202/079017-2110

Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. In *NeurIPS*, volume 35, pages 17340–17358, 2022.

Reza Esfandiarpoor, Cristina Menghini, and Stephen H Bach. If clip could talk: Understanding vision-language model representations through their preferred concept descriptions. *arXiv preprint arXiv:2403.16442*, 2024.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *ICML*, pages 6216–6234. PMLR, 2022.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pages 1607–1616. PMLR, 2018.

Han Gao, Kaican Li, Weiyan Xie, Lin Zhi, Yongxiang Huang, Luning Wang, Caleb Chen Cao, and Nevin L. Zhang. Consistency regularization for domain generalization with logit attribution matching. In *UAI*, 2024a.

Irena Gao, Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Out-of-domain robustness via targeted augmentations. In *ICML*, 2023.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024b.

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

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, pages 19338–19347, 2023.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *CVPR*, pages 4805–4814, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021b.

Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, and Il-Chul Moon. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *CVPR*, pages 10664–10673, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 2023.

https://doi.org/10.52202/079017-2110

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, pages 5637–5664. PMLR, 2021.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, pages 2825–2834. PMLR, 2018.

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. In *NeurIPS*, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, volume 36, 2023.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *CVPR*, pages 8046–8056, 2022.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, pages 7313–7324. PMLR, 2021.

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *ICCV*, pages 262–271, 2023.

Xiaofeng Mao, Yufeng Chen, Xiaojun Jia, Rong Zhang, Hui Xue, and Zhao Li. Context-aware robust fine-tuning. *IJCV*, 132(5):1685–1700, 2024.

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

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

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pages 7721–7735. PMLR, 2021.

Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *ICLR*, 2021.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. *NeurIPS*, 35:10068–10077, 2022.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

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

Changdae Oh, Mijoo Kim, Hyesu Lim, Junhyeok Park, Euiseog Jeong, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. *arXiv preprint arXiv:2311.01723*, 2023.

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.

https://doi.org/10.52202/079017-2110

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

Renjie Pi, Lewei Yao, Jianhua Han, Xiaodan Liang, Wei Zhang, and Hang Xu. Ins-detclip: Aligning detection model to follow human-language instruction. In *ICLR*, 2024.

Viraj Prabhu, Ramprasaath R Selvaraju, Judy Hoffman, and Nikhil Naik. Can domain adaptation make object recognition work for everyone? In *CVPR*, pages 3981–3988, 2022.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pages 15691–15701, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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 *ICML*, pages 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, volume 33, pages 9573–9585, 2020.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *ICLR*, 2022.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *ICML*, pages 31716–31731. PMLR, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 33:18583–18599, 2020.

Joshua Tenenbaum and William Freeman. Separating style and content. In *NeurIPS*, volume 9, 1996.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. In *NeurIPS*, volume 36, 2023.

Junjiao Tian, Zecheng He, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, and Zsolt Kira. Trainable projected gradient method for robust fine-tuning. In *CVPR*, pages 7836–7845, 2023a.

Junjiao Tian, Yen-Cheng Liu, James S Smith, and Zsolt Kira. Fast trainable projection for robust fine-tuning. *NeurIPS*, 36, 2023b.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR*, 2019.

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

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

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, pages 7959–7971, 2022.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, 2024.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *ICCV*, pages 3090–3100, 2023.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, pages 19187–19197, 2023.

Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *CVPR*, pages 7947–7958, 2022.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, pages 3987–3995. PMLR, 2017.

Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *ECCV*, pages 698–714. Springer, 2020.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, pages 3713–3722, 2019.

Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *TPAMI*, 44(8):4388–4403, 2021.

Nevin L. Zhang, Kaican Li, Han Gao, Weiyan Xie, Zhi Lin, Zhenguo Li, Luning Wang, and Yongxiang Huang. A causal framework to unify common domain generalization approaches. *arXiv preprint arXiv:2307.06825*, 2023a.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, pages 8552–8562, 2022.

Yang Zhang, Yue Shen, Dong Wang, Jinjie Gu, and Guannan Zhang. 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*, pages 1894–1898, 2023b.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *ICLR*, 2020.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.

https://doi.org/10.52202/079017-2110

A 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 w.r.t. q.*

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

$$\begin{aligned} H_p(\alpha q_1 + (1-\alpha)q_2) &= \mathbb{E}_{(x,y)\sim p}[-\log(\alpha q_1(y|x) + (1-\alpha)q_2(y|x))] \\ &\leq \mathbb{E}_{(x,y)\sim p}[-\alpha \log q_1(y|x) - (1-\alpha)\log q_2(y|x)] \qquad (11) \\ &= \alpha H_p(q_1) + (1-\alpha)H_p(q_2). \end{aligned}$$

□

Theorem 1. *Strong duality holds between IDRM and the following dual problem:*

$$\max_{\lambda' \geq 0} \min_{\theta \in \Theta} \left[R_{\mathrm{s}}(\theta) + \lambda' \max_{d \in \mathcal{D}} R_d(\theta)\right] - \lambda' \alpha. \qquad (3)$$

Proof. Recall that IDRM aims to solve for

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

Here, $R_d(\theta) = H_{p_d}(\hat{p}_\theta(y|x)) = \mathbb{E}_{(x,y)\sim p_d}[-\log \hat{p}_\theta(y|x)]$ is the cross-entropy between $p_d(y|x)$ and $\hat{p}_\theta(y|x)$ over $p_d(x)$. It follows from Lemma 1 that $R_d(\theta)$ is convex w.r.t. $\hat{p}_\theta(y|x)$ 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. $\hat{p}_\theta(y|x)$. By Slater's condition, strong duality holds between IDRM and the Lagrangian dual of IDRM:

$$\max_{\lambda' \geq 0} \min_{\theta \in \Theta} R_{\mathrm{s}}(\theta) + \lambda' \left[\max_{d \in \mathcal{D}} R_d(\theta) - \alpha\right], \qquad (12)$$

for any $\alpha > \min_{\theta \in \Theta} \max_{d \in \mathcal{D}} R_d(\theta)$, i.e., when a strictly feasible solution to (IDRM) exists. □

B Introduction to baseline methods

The key baseline we compare our DRM method with is "fine-tune like you pre-train" (FLYP) (Goyal et al., 2023). While traditional fine-tuning of CLIP models adds a randomly initialized classification head on top of the image encoder, Goyal et al. (2023) demonstrated that it is more effective to simply reuse the text encoder. We therefore follow FLYP to update both the image encoder and text encoder in fine-tuning. They now differ only in the loss functions used in fine-tuning. FLYP uses the CLIP contrastive loss (Radford et al., 2021) for the ERM and showed that this is better than using the standard cross-entropy loss. We adopt this approach for the ERM component (the first term) of DRM to facilitate comparison. Thus, the only difference between DRM and FLYP is the additional regularization term in the DRM loss function.

Before the introduction of FLYP, conventional fine-tuning of CLIP models involved adding a linear classification head to the image encoder. The linear probing method (LP) only fine-tunes this new classification head, keeping the image encoder fixed, whereas the full fine-tuning method (FT) trains both the head and the encoder. The LP-FT approach (Kumar et al., 2022) begins with LP and then transitions to full fine-tuning.

Besides, L2-SP (Li et al., 2018) and WiSE-FT (Wortsman et al., 2022) are two established fine-tuning variants that restrict the divergence from the pre-trained model. L2-SP specifically integrates an L^2 regularization term into the loss function to constrain the parameter shifts of the fine-tuned model relative to the pre-trained model. WiSE-FT (Wortsman et al., 2022) interpolates parameters of a pre-trained zero-shot model θ_{zs} and that of a fine-tuned model θ_{ft} using $\theta_{\text{wise-ft}} = \rho \cdot \theta_{\text{zs}} + (1-\rho) \cdot \theta_{\text{ft}}$.

Alongside L2-SP and WiSE-FT, several works concurrent with our own have also introduced robust fine-tuning methods via reducing the difference between pre-trained and fine-tuned models. For example, CAR-FT (Mao et al., 2024) and the method proposed by Cheng et al. (2024) seek to minimize the distance between the context distributions generated by pre-trained and fine-tuned CLIP

https://doi.org/10.52202/079017-2110

models. However, a significant limitation of them is that they require prior knowledge of image contexts, such as background and viewpoint, which restricts their practical applicability.

Alternatively, Lipsum-ft (Nam et al., 2024) implements a regularization strategy without requiring prior context information, focusing on minimizing the L^2 distance between the affinities of images and random texts from both pre-trained and fine-tuned models. CLIPood (Shu et al., 2023) utilizes a beta moving average for updating parameters during training, and CaRot (Oh et al., 2023) focuses on regularizing singular value distributions while incorporating an exponential moving average for parameter updates. TPGM (Tian et al., 2023a) and FTP (Tian et al., 2023b) take a more fine-grained approach by autonomously determining the most effective regularization for each layer's parameters and enabling more efficient learning of layer-specific projection regularization, respectively.

While these methods aim to prevent the fine-tuned model from deviating too far from the pre-trained one, they do not specify which pre-trained features should be preserved. In contrast, our proposed DRM focuses on preserving the dataset-related core visual features, making it a more targeted approach compared to the aforementioned methods.

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 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.

By utilizing GPT-4, we were able to automatically produce precise concept descriptions of various classes. The use of LLMs to generate class description from the class name is not new, which have been explored in Menon and Vondrick (2022); Pratt et al. (2023); Maniparambil et al. (2023) and Yang et al. (2023); Yan et al. (2023). Among them, the first three works mainly explored the use of class descriptions for zero-shot classification, while this work focuses on fine-tuning. The class descriptions generated in these studies often include contextual information due to prompts like "*Describe an image from the internet of a(n) ...*", which may not be suitable for our application of estimating worst-case risk. The descriptions in Yang et al. (2023); Yan et al. (2023) are more aligned with our approach, while they focus on using LLM-generated concept descriptions to develop concept bottleneck models for interpretable image classification.

D Empirical study on concept descriptions

D.1 Qualitative and quantitative study on the reliability of concept descriptions

Complementing Figure 2, Figure 3 provides additional examples indicating that with the pre-trained CLIP models, concept descriptions have the ability to extract core features. To further support this claim, we present a full quantitative study. The study is conducted with Hard ImageNet (Moayeri et al., 2022) and consists of two parts. First, we remove image background (BG), and observing how the image-text affinities change for default prompts (df) and concept descriptions (cd) respectively. In the second part, we do the same but with foreground (FG) removed.

Table 6: Quantitative study on the reliability of concept descriptions verse default prompts. The affinities to concept descriptions are sensitive to changes in the core features of image foregrounds (FG) and remain relatively stable against changes in the non-core features of backgrounds (BG).

	FG & BG	w/o BG	w/o FG
`df`	0.3473	0.2393 (-31.1%)	0.3407 (-1.9%)
`cd`	0.2660	0.2387 (-10.3%)	0.1180 (-55.6%)

Table 6 shows the average affinities over all 19,097 images across all 15 classes of Hard ImageNet. The percentages in the table indicate the relative changes w.r.t. the affinities of the original images (FG & BG). The result shows that the affinities of concept descriptions are much more invariant to changes in non-core features than default prompts (-10.3% vs. -31.1%). Moreover, the affinities of concept descriptions are quite responsive (-55.6%) to changes in core features. In contrast, the affinities of default prompts barely change (-1.9%) in response to the absence of core features. These results indicate that the affinities associated with concept descriptions are reliable indicators of core visual features in the images. In contrast, the affinities from default prompts are more sensitive to the changes in non-core visual features.

https://doi.org/10.52202/079017-2110

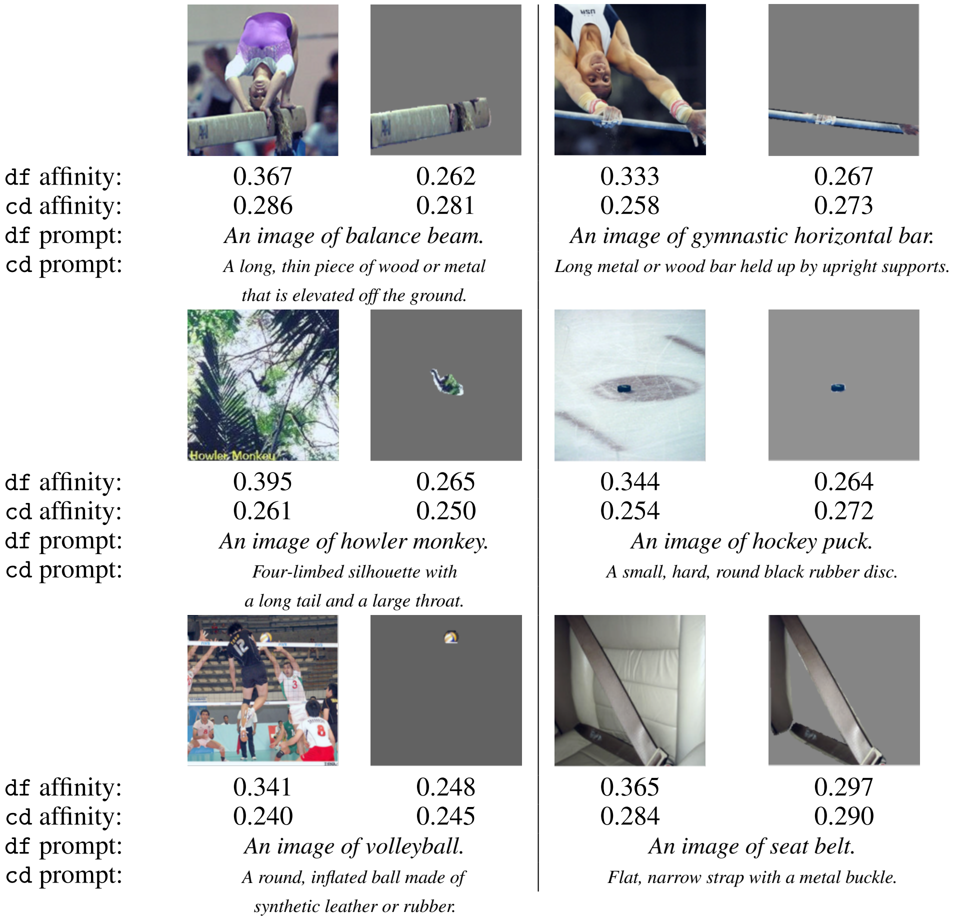

df affinity:
cd affinity:
df prompt:
cd prompt:
0.367
0.262
0.286
0.281
An image of balance beam.
A long, thin piece of wood or metal that is elevated off the ground.
0.333
0.267
0.258
0.273
An image of gymnastic horizontal bar.
Long metal or wood bar held up by upright supports.
Howler Monkey
0.395
0.265
0.261
0.250
An image of howler monkey.
Four-limbed silhouette with a long tail and a large throat.
0.344
0.264
0.254
0.272
An image of hockey puck.
A small, hard, round black rubber disc.
0.341
0.248
0.240
0.245
An image of volleyball.
A round, inflated ball made of synthetic leather or rubber.
0.365
0.297
0.284
0.290
An image of seat belt.
Flat, narrow strap with a metal buckle.

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

D.2 Examples of repeated generations of concept descriptions

To evaluate the stochasticity of LLM in generating concept descriptions, we repeatedly ask GPT-4 to generate concept descriptions for each iWildCam class. Below are some examples:

White-lipped Peccary:

- **Output 1:** *Compact, dark grey body with distinctive white markings around the mouth.*
- **Output 2:** *Compact, dark gray body with distinctive white markings around the lips.*
- **Output 3:** *A stocky body with coarse, dark hair and distinct white markings around the mouth.*

Waterbuck:

- **Output 1:** *Stocky body with long fur, a white ring on the rump, and shaggy brown coat.*
- **Output 2:** *Thick, shaggy brown coat with a white ring around the rump and long horns.*
- **Output 3:** *A robust antelope with a shaggy brown coat and a white ring around the rump.*

These examples show that while GPT-4's outputs for the same class vary slightly, they consistently highlight key visual features. The variations are primarily in the language details, such as the order of features or the terminology used. For example, for the waterbuck, the consistent visual features across all three outputs are the "stocky (thick, robust) body", "white ring on (around) the rump", and "shaggy brown coat".

D.3 Examples to generate concept descriptions with different LLMs

The examples and experiment results reported in our paper are based on concept descriptions generated by GPT-4. We now consider different LLMs to generate concept descriptions for the classes in the iWildCam dataset. We present some examples and our findings below.

Examples:

White-lipped Peccary:

- **GPT-4:** *Compact, dark grey body with distinctive white markings around the mouth.*
- **GPT-3.5-Turbo:** *Grayish body with a white stripe across the lips.*
- **Llama-3-8b:** *Reddish-brown fur with a distinctive white stripe.*
- **Llama-3-70b:** *Dark brown pig-like animal with a distinctive white stripe on the lips.*
- **Llama-3-405b:** *Stout, brown body with a distinctive white stripe on the lips and sharp tusks.*

Waterbuck:

- **GPT-4:** *Stocky body with long fur, a white ring on the rump, and shaggy brown coat.*
- **GPT-3.5-Turbo:** *Antelope species with a shaggy coat and a white ring on its rump.*
- **Llama-3-8b:** *Reddish-brown coat with a distinctive horn shape and a long, slender body.*
- **Llama-3-70b:** *Large, brown antelope with a distinctive white ring on its rump and a long, shaggy coat.*
- **Llama-3-405b:** *Large, brown body with a distinctive pair of curved horns and a white ring on the rump.*

Zebra:

- **GPT-4:** *Distinctive black and white striped coat with an upright mane and slender legs.*
- **GPT-3.5-Turbo:** *Medium-sized equid with black and white stripes.*
- **Llama-3-8b:** *Black and white striped coat with a distinctive mane and a long, slender body.*
- **Llama-3-70b:** *Black and white striped equine with a distinctive body shape and a long mane.*
- **Llama-3-405b:** *Medium-sized, black and white body with a distinctive striped pattern and a long, flowing mane.*

Findings:

1. The concept descriptions generated by different LLMs capture common key visual features for the same class, such as “white stripe” for white-lipped peccary, “white ring on the rump” for waterbuck, and “black and white stripes” for zebra.
2. Comparing outputs of GPT-3.5-Turbo and GPT-4, the outputs generated by GPT-4 cover more visual details, while those by GPT-3.5-Turbo are generally shorter.
3. As model size increases, from Llama-3-8b to Llama-3-405b, the generated outputs become more detailed. For instance, Llama-3-8b mentions “white stripe” but does not specify its location on the lips as Llama-3-70b/405b do. Additionally, Llama-3-8b sometimes makes factual errors, such as describing the incorrect color for white-lipped peccary and waterbuck.
4. GPT-3.5-Turbo sometimes deviated from the prompt instructions, generating outputs that include non-visual features, such as the "distinctive call" of the *Great Tinamou*, which are not visible.

Despite variations in the quality of concept descriptions generated by different LLMs, as discussed in Section 5.4, DRM-trained models using descriptions from any of these tested LLMs consistently show significant improvements in OOD performance compared to FLYP-trained models. This demonstrates that the effectiveness of DRM-trained models is not sensitive to the choice of the LLM.

E Additional experiment results

E.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 7. 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 7: 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**

E.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 8 demonstrate that, compared to FLYP-trained models, DRM-trained models exhibit improved performance on images from new countries.

Table 8: 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**

E.3 Comparison to some more recent methods

As discussed in Appendix B, 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 9. The results clearly show that, the more recent methods still significantly lag behind DRM in term of OOD performance.

Table 9: 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**

E.4 Full 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 (7), i.e., $R_\text{s}(\theta; \mathcal{T}^\texttt{df}) + \lambda R_\text{s}^\text{c}(\theta; \mathcal{T}^\texttt{cd})$, for fine-tuning zero-shot models can be expanded as

$$\frac{1}{N}\sum_{i=1}^{N}\Big[-\log \hat{p}_\theta^{\texttt{df}}(y_i|x_i) - \lambda \sum_{y'\in\mathcal{Y}} \hat{p}_{\theta_0}^{\texttt{pr}}(y'|x_i)\log \hat{p}_\theta^{\texttt{cd}}(y'|x_i)\Big], \tag{13}$$

where $\hat{p}_\theta^\texttt{df}(y|x)$ and $\hat{p}_\theta^\texttt{cd}(y|x)$ are the classifiers (5) induced by the default prompts $\mathcal{T}^\texttt{df} = \{t_y^\texttt{df} \,|\, y \in \mathcal{Y}\}$ and the concept descriptions $\mathcal{T}^\texttt{cd} = \{t_y^\texttt{cd} \,|\, y \in \mathcal{Y}\}$, respectively; and $\hat{p}_{\theta_0}^\texttt{pr}(y|x) = \tilde{p}_\text{c}(y|x)$ which is defined by (9) to estimate $p_\text{c}(y|x)$. Here, we use $\hat{p}_{\theta_0}^\texttt{pr}(y|x)$ (where `pr` stands for 'proxy') instead of $\tilde{p}_\text{c}(y|x)$ to ease the discussion of possible variations of DRM.

Consider the following generalized form of (13) 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 $p_\text{c}(y|x)$:

$$\frac{1}{N}\sum_{i=1}^{N}\Big[-\log \hat{p}_\theta^{\texttt{t1}}(y_i|x_i) - \lambda \sum_{y'\in\mathcal{Y}} \hat{p}_{\theta_0}^{\texttt{type}}(y'|x_i)\log \hat{p}_\theta^{\texttt{t2}}(y'|x_i)\Big], \tag{14}$$

As stated in (13), our final DRM training objective (7) uses $\texttt{t1} = \texttt{df}$ in the ERM term, with $\texttt{t2} = \texttt{cd}$ and $\texttt{type} = \texttt{pr}$ 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 10.

(a) Inference options after dual classifier training: Two classifiers are involved in our DRM training: $\hat{p}_\theta^\texttt{df}(y|x)$ and $\hat{p}_\theta^\texttt{cd}(y|x)$. As outlined in (10), we combine both classifiers for inference. An alternative is to only use one of the two classifiers for inference. We denote inference with only $\hat{p}_\theta^\texttt{df}(y|x)$ as (a1), and with only $\hat{p}_\theta^\texttt{cd}(y|x)$ as (a2). The comparison between (a1), (a2), and (S) in Table 10 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 4 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 10. The contrast between (b1) and (S) confirms our intuition: using the concept

Table 10: 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	(10)	**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	(10)	32.1	24.2
		(c2)	`df`	`cd`	`pr-df`	joint training	(10)	54.4	45.1
		(c3)	`df`	`cd`	`one-hot`	joint training	(10)	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	(10)	59.7	45.7
		(e2)	`df`	`cd`	`pr`	weight average	(10)	57.5	44.7

descriptions $\mathcal{T}^{\texttt{cd}}$ for $\hat{p}_{\theta}^{\texttt{t2}}(y|x)$, 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 $\hat{p}_{\theta}^{\texttt{t1}}(y|x)$ and $\hat{p}_{\theta}^{\texttt{t2}}(y|x)$, i.e., $\texttt{t1} = \texttt{t2} = \texttt{cd}$, slightly improves (b1). Intriguingly, (b2) is still much worse than (S) despite they both use `cd` for `t2`.

(c) Proxy model design: In our standard setting, $\texttt{type} = \texttt{pr}$. As discussed in Section 4, the proxy term $\hat{p}_{\theta_0}^{\texttt{pr}}(y|x_i)$ is based on the affinity $A_{\theta_0}(x, t_y^{\texttt{cd}}) = \langle f_{\phi_0}(x), g_{\psi_0}(t_y^{\texttt{cd}}) \rangle$ 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 4, we also mentioned the following direct estimation of the oracle model $p_{\text{c}}(y|x)$:

$$\hat{p}_{\theta_0}^{\texttt{cd}}(y|x) = \frac{\exp(A_{\theta_0}(x, t_y^{\texttt{cd}})/\tau)}{\sum_{y' \in \mathcal{Y}} \exp(A_{\theta_0}(x, t_{y'}^{\texttt{cd}})/\tau)}. \tag{15}$$

However, as discussed in Section 4, $\hat{p}_{\theta_0}^{\texttt{cd}}(y|x)$ 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 $\hat{p}_{\theta_0}^{\texttt{pr}}(y|x)$ here. As shown in Table 10, the importance of this adjustment is empirically verified by the much lower performance of (c1) compared to (S).

One can also define $\hat{p}_{\theta_0}^{\texttt{pr-df}}(y|x)$ by replacing $\mathcal{T}^{\texttt{cd}}$ with $\mathcal{T}^{\texttt{df}}$ in the formulation of $\hat{p}_{\theta_0}^{\texttt{pr}}(y|x)$. As shown by the result of (c2), this alternative still underperforms $\hat{p}_{\theta_0}^{\texttt{pr}}(y|x)$ 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 (7) 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 $\hat{p}_{\theta_{\text{ERM}}}^{\texttt{df}}(y|x)$ and a WRM model $\hat{p}_{\theta_{\text{WRM}}}^{\texttt{cd}}(y|x)$ separately. These models can be combined for inference using techniques like model ensembling or weight-space averaging. The last two rows of Table 10 show that combining (d1) and (d3) via model ensembling or weight-space averaging generally underperforms joint training (S).

E.5 Results of applying DRM with FT and LP-FT

Our experiment results reported in Section 5 are based on combining DRM with FLYP. Specifically, we follow FLYP to update both the image encoder and text encoder, and utilize the FLYP loss for the ERM component of (7). The results showed in Table 11 indicate that even without FLYP, where DRM is applied with standard full fine-tuning (Row 2), our method still outperforms FLYP (Row 1), which itself has been shown to surpass standard full fine-tuning in the ERM setting (Goyal et al., 2023). This performance advantage is demonstrated in the results of fine-tuning CLIP ViT-L/14 on iWildCam:

Table 11: Results of applying DRM under different settings.

Row	FLYP	LP-FT	DRM	ID	OOD
1	Yes	No	No	56.0	41.9
2	No	No	Yes	54.4	43.9
3	No	Yes	Yes	56.5	46.3
4	Yes	No	Yes	61.8	49.2

As the table above demonstrates, combining DRM with LP-FT (Row 3) enhances performance over just DRM with standard full fine-tuning (Row 2). Furthermore, integrating DRM with FLYP (Row 4) yields even more significant improvements. Consequently, we adopt the combination of DRM and FLYP as our default setting when fine-tuning CLIP models.

E.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 applying DRM to CNN models, we add two randomly initialized classification heads on top of the model. Analogous to the application of DRM on zero-shot models, one classification head is trained using cross-entropy loss with respect to 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 employ this approach to fine-tune an ImageNet pre-trained ResNet50 model on the iWildCam dataset, utilizing soft labels generated by the CLIP ViT-L/14 model. During inference, the outputs from the two classification heads are combined in a manner similar to that described in (10). The results are presented in Table 12. It is evident from the results that DRM significantly enhances the OOD performance of ResNet50 compared to the ERM.

Table 12: 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

F Training details

F.1 Hyperparameter settings

We primarily adopted the hyperparameter settings from the code released by FLYP (Goyal et al., 2023).

Specifically, for iWildCam, the settings were as follows: `training epochs=20`, `learning rate=1e-5`, `batch-size=256`, and `optimizer=AdamW` with `weight decay=0.2`;

For FMoW, the settings were as follows: `training epochs=20`, `learning rate=1e-5`, `batch-size=256`, and `optimizer=AdamW` with `weight decay=0.2`;

https://doi.org/10.52202/079017-2110

For ImageNet, the settings were as follows: `training epochs=10`, `learning rate=1e-5`, `batch-size=256`, and `optimizer=AdamW` with `weight decay=0.1`.

In all experiments, for the images, we applied the standard CLIP image pre-processing, which included resizing, center cropping, and normalization. For the texts, we applied the standard CLIP text pre-processing, which tokenized the texts into a series of integers, each representing a unique series of characters.

The value of λ used in our DRM training was picked from $\{1, 2, 3, 4, 5\}$ based on the performance on the ID validation set. Following Goyal et al. (2023), we also implemented early stopping based on the ID validation performance.

The hyperparameter ρ in WiSE-FT, $\theta_{\text{wise-ft}} = \rho \cdot \theta_{\text{zs}} + (1 - \rho) \cdot \theta_{\text{ft}}$, is chosen from the range 0.1 to 0.9 via ID validation.

F.2 Machines

All experiments were conducted on a high-performance computing cluster equipped with NVIDIA DGX H800 nodes. Two H800 GPUs with 80 GB VRAM were utilized for all trainings involving CLIP ViT-B/16 and CLIP ViT-L/14, while four H800 GPUs were employed for the training of CLIP ViT-L/14@336.

F.3 Analysis of computational costs

In our experiments, we implemented the ERM part of DRM with FLYP (Goyal et al., 2023). The additional computational cost of DRM, compared to FLYP, primarily arises from the preparation of soft labels for the targets of the second risk in DRM and the minimization of this second risk. In short, the training and inference cost for DRM increased by about 20% from FLYP. This additional cost is insignificant compared to the attained performance gain. In Table 13, we detail the computational costs and timing for fine-tuning CLIP ViT-L/14 models on ImageNet.

The computation time reported below is based on the setting that training batch size=256 and inference batch size=1024. There are 1,281,167 training images in ImageNet, and thus there are 5005 training batches.

Table 13: Analysis of computational costs for DRM.

Model	Generating Concept Description by LLM	Soft Label Generation (9)	Training	Inference
FLYP	N/A	N/A	In average: **58s/100 batches**, ~48 mins per training epoch	Inference on 100 batches of images takes ~**1.5s**.
DRM	We utilized the GPT-4-turbo API to generate concept descriptions for 1,000 ImageNet classes, inputting 10 classes at a time to ensure quality. The generation cost is **under 10 US Dollar** (the price of the API is 10 US Dollar/1 million prompt tokens). We are unaware of the computational cost as the model details of GPT-4 are unknown.	The primary computational cost arises from using pre-trained CLIP models to generate image and text embeddings from 1,281,167 training images and 1,000 concept descriptions. Soft labels are created using inner products between these embeddings, with some technical adjustments. The entire process takes **less than 3 minutes**.	In average: **71s/100 batches**, ~58 mins per training epoch	Inference on 100 batches of images takes ~**1.8s**.

https://doi.org/10.52202/079017-2110

G Limitations

Our research utilized 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. For instance, we found that GPT-4 generated inaccurate concept descriptions in medical imaging fields like ocular disease and breast histology.

Additionally, due to the vast number of concept descriptions 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 generating 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.

A further 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 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.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the work are discussed in Appendix G.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

https://doi.org/10.52202/079017-2110

Justification: The complete proof to proposed theorem has been provided in Appendix A.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: The specifics of our DRM training objective and its implementation are detailed in Section 4.2 and 5.1. Information regarding the generation of concept descriptions can be found in Section 4.2 and Appendix C, with sample concept descriptions included in the supplemental material. Additional training details, including the choices of hyperparameters, are provided in Appendix F.

   Guidelines:

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

https://doi.org/10.52202/079017-2110

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

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

   Answer: [Yes]

   Justification: The datasets employed in our experiments are described in Section 5.1 and are all publicly accessible. The code of our work will be made available in the given GitHub repository.

   Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: The details of data split have been provided in Section 5.1. All other training details, including the selection of hyperparameters, optimizers and others are provided in Appendix F.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: Yes. The 95% confidence intervals over five training runs are reported in Table 1 and 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-2110

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: The information is provided in Appendix F.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer:[Yes]

   Justification: We have reviewed and ensured the research conducted in this paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: We do not see particular negative societal impacts of the work performed.

https://doi.org/10.52202/079017-2110

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: We do not see any data or models in this work that have a high risk for misuse.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, have been properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

https://doi.org/10.52202/079017-2110

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

13. **New Assets**

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

    Answer: [Yes]

    Justification: The specifics of our DRM model training objective and its implementation are detailed in Section 4.2 and 5.1. A detailed documentation for running our codes will be provided in the GitHub respository.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-2110

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.