Expanding the Role of Diffusion Models for Robust Classifier Training

Pin-Han Huang
National Taiwan University
pinhankhuang@gmail.com

Shang-Tse Chen[†]
National Taiwan University
stchen@csie.ntu.edu.tw

Hsuan-Tien Lin[†]
National Taiwan University
htlin@csie.ntu.edu.tw

Abstract

Incorporating diffusion-generated synthetic data into adversarial training (AT) has been shown to substantially improve the training of robust image classifiers. In this work, we extend the role of diffusion models beyond merely generating synthetic data, examining whether their internal representations, which encode meaningful features of the data, can provide additional benefits for robust classifier training. Through systematic experiments, we show that diffusion models offer representations that are both diverse and partially robust, and that explicitly incorporating diffusion representations as an auxiliary learning signal during AT consistently improves robustness across settings. Furthermore, our representation analysis indicates that incorporating diffusion models into AT encourages more disentangled features, while diffusion representations and diffusion-generated synthetic data play complementary roles in shaping representations. Experiments on CIFAR-10, CIFAR-100, and ImageNet validate these findings, demonstrating the effectiveness of jointly leveraging diffusion representations and synthetic data within AT.

1 Introduction

Machine learning models are known to be vulnerable to adversarial examples [1, 2], inputs perturbed by semantically imperceptible noise that can drastically alter model predictions. Among numerous proposed defenses, adversarial training (AT) [3, 4], which trains models on adversarially perturbed inputs, remains one of the most effective approaches on standard benchmarks such as RobustBench [5].

AT is known to suffer from robust overfitting [6], where robustness on the test set degrades during training despite decreasing training loss and stable clean accuracy. Multiple methods have been proposed to understand and mitigate this issue [7, 8, 9, 10, 11, 12]. Among them, arguably the most effective approach to date has been the diffusion model with AT (DM-AT) training recipe [11], leveraging large amounts of high-quality synthetic data generated by diffusion models to improve robustness.

The DM-AT approach [11], which does not rely on additional real data to train diffusion models, mainly treats diffusion models as synthetic data generators for robust classifier training. More broadly, most of the existing efforts to improve AT have centered on this synthetic data paradigm [11, 13, 14, 15, 16]. While effective, this paradigm largely exploits diffusion models only at the level of their outputs, potentially overlooking the rich structure developed during the generative process. In particular, diffusion models are known to produce meaningful intermediate representations throughout their denoising trajectories [17, 18, 19, 20]. The extent to which these representations can further enhance robustness remains largely unexplored, suggesting a promising opportunity beyond synthetic data generation.

[†]Co-advisors.

Preprint.

This work systematically investigates the utility of diffusion representations for enhancing adversarial robustness. We hypothesize that the denoising objective encourages diffusion models to encode resilient semantic features from corrupted inputs, which can be harnessed to guide classifier training. Specifically, we examine the utility of intermediate activations computed during denoising as feature priors for enhancing adversarial robustness.

Our idea of studying the intermediate activations as feature priors is inspired by an intriguing analysis that uncovers a weak correlation between robustness and alignment with diffusion representations (Figure 1). Going further, we find that the extracted diffusion representations possess a range of highly desirable properties: they capture rich and diverse information, emphasize lower-frequency structures, and exhibit strong resilience to irrelevant high-frequency noise. These characteristics suggest that diffusion representations have significant potential, standing in sharp contrast to conventional reconstruction-based representation learning, which is known to be more adversarially vulnerable due to its reliance on high-frequency signals [21].

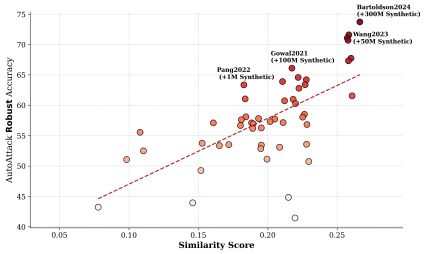


Figure 1: Robust accuracy versus similarity to diffusion representations for CIFAR-10 ℓ_∞-robust models from RobustBench. Details are in Appendix A.

Motivated by these findings, we propose to upgrade the DM-AT recipe by incorporating a simple module that aligns classifier representations with diffusion representations [22, 23, 24]. The modified recipe uses diffusion representations as an auxiliary learning signal to guide feature learning during robust classifier training, and is agnostic to the choice of classifier backbone. Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet across multiple architectures and diffusion synthetic data settings demonstrate consistent improvements in both clean and robust accuracy, showing the effectiveness of using the semantic features encoded in diffusion representations to guide feature learning during AT.

We deepen our analysis to understand the mechanism behind the empirical improvements by building on the recent framework of mechanistic interpretability [25], comparing the effects of synthetic data in DM-AT versus the representation alignments in our upgraded recipe. This approach reveals that both interventions facilitate the learning of more easily disentangled representations, yet they achieve this effect through distinct underlying mechanisms. Specifically, our analysis, guided by classification-aware dimensions [26], reveals that synthetic data in DM-AT promotes robustness and generalization by enabling the model to learn low-rank representations with strong generalization properties.

In contrast, diffusion representation alignment encourages the model to effectively leverage its representational dimensions to encode robust features, which are not necessarily low-rank. Together, these findings suggest that our proposed diffusion representation alignment complements the synthetic data used in DM-AT for robust classifier training, supporting our goal of *expanding* the role of diffusion models in robustness and generalization.

Our contributions are summarized as follows:

- We show that diffusion representations encode features that are partially robust and diverse, and leveraging diffusion representations as an auxiliary learning signal improves adversarial training.
- We find that the incorporation of diffusion models encourages models to learn representations that are easier to disentangle, with synthetic data and representation alignment playing complementary roles.
- Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet show that incorporating both diffusion representation alignment and diffusion synthetic data consistently improves robustness, offering an updated recipe to build robust classifiers.

2 Related Work

Adversarial Robustness. Adversarial robustness studies the stability of model predictions under imperceptible, adversarially crafted input perturbations [1, 2, 3]. Empirical robustness is commonly assessed with AutoAttack [27], which is also the main evaluation protocol for RobustBench [5].

RobustBench excludes defenses that rely on inference-time randomness or optimization loops, since such mechanisms are frequently broken by adaptive attacks and require more careful and costly evaluations [28, 29]. Consequently, models competitive on RobustBench rely on adversarial training to achieve robustness. Additionally, certified defenses offer provable guarantees [30, 31, 32, 33, 34], but often introduce significant inference-time cost or underperform adversarially trained models in empirical robustness, especially models that are trained with diffusion synthetic data [11, 14, 16]. In this work, we focus on improving adversarial training and adopt AutoAttack as our primary evaluation.

Diffusion Representations. Diffusion models have achieved remarkable success in image generation [35, 36, 37, 38, 39, 40, 23, 41], and studies in representation learning have shown that intermediate activations extracted from diffusion models are effective for discriminative tasks, including competitive performance compared with other self-supervised learning methods for image classification [18, 19, 20] and dense predictions [24, 42]. The latest advances in diffusion models have also leveraged this insight to accelerate the training of diffusion models by aligning the activations with large-scale self-supervised learning encoders [23, 43].

In this paper, we investigate whether diffusion representations encode informative and robust semantics that can benefit adversarially robust classification. Yagoda et al. [44] propose to train prediction heads on frozen unconditional diffusion representations as a lightweight approach, but it relies heavily on inference-time randomness and is less robust than adversarial training. Under EOT-based evaluation [28], the robust accuracy drops substantially (Appendix C). Conversely, we show that using diffusion representations as an auxiliary learning signal can further strengthen adversarial training, and analyze how this integration shapes the learned representations.

Diffusion Purification and Generative Classifiers. In addition to generating synthetic data for adversarial training, diffusion models have also been applied in adversarial purification to remove adversarial noise [45, 46]. However, such methods incur substantial inference cost, and their reliance on randomness has been shown to be vulnerable to adaptive attacks [47, 48].

Another direction is to turn off-the-shelf diffusion models into Bayesian generative classifiers [49, 50, 48]. At a high level, these methods add noise to an input image, then denoise it conditioned on each class, and finally select the class whose reconstruction is most similar to the original input. They exhibit desirable properties such as high error consistency with humans [51], robustness to imbalanced datasets [52] that is free of the need for a balanced dataset [53], and adversarial robustness [48]. These approaches can also be integrated with randomized smoothing [30] to provide certified defenses [33]. Despite these advantages, the approach incurs substantial inference overhead due to iterative denoising and class-conditional evaluation, which scales inference cost with the number of classes and limits practicality for deployment and full evaluation on datasets such as ImageNet [49, 48]. In this work, we pursue a parallel path that focuses on leveraging diffusion models to enhance robust classifier training, which is free of inference-time overhead.

3 Preliminaries

Adversarial Training (AT). Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of image-label pairs, adversarial training [3] is formulated as

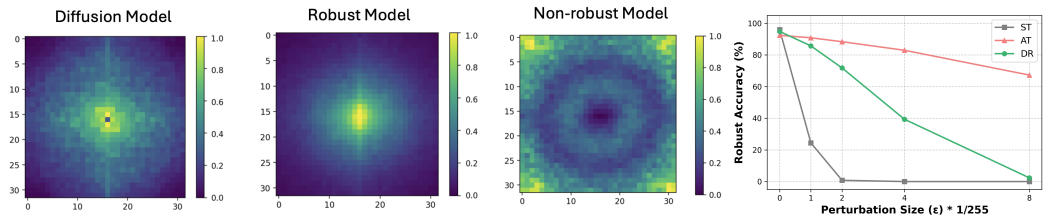
$$\min_{\theta} \sum_{i=1}^{n} \max_{\mathbf{x}_i' \in \mathcal{S}(\mathbf{x}_i)} \mathcal{L}\big(f_\theta(\mathbf{x}_i'),\, y_i\big), \tag{1}$$

where f_θ is the model parameterized by θ, $\mathcal{L}$ is the cross-entropy loss, and $\mathcal{S}(\mathbf{x}) = \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_p \leq \varepsilon\}$ is the ℓ_p-ball of radius ε centered at $\mathbf{x}$. During training, projected gradient descent (PGD) is used to approximately solve the inner maximization by iteratively updating the adversarial example. Considering the standard ℓ_∞ adversary, the adversarial example is obtained by

$$\mathbf{x}_i^{(t+1)} = \Pi_{\mathcal{S}(\mathbf{x}_i)}\Big(\mathbf{x}_i^{(t)} + \alpha \operatorname{sign}\Big(\nabla_{\mathbf{x}} \mathcal{L}\big(f_\theta(\mathbf{x}_i^{(t)}), y_i\big)\Big)\Big), \tag{2}$$

where α is the step size and $\Pi_{\mathcal{S}(\mathbf{x}_i)}(\cdot)$ denotes projection onto $\mathcal{S}(\mathbf{x}_i)$ and the valid image pixel range. The TRADES loss [4] is often used to balance the performance on adversarial and clean images [11, 14, 15, 16].

Extracting Diffusion Representations. For a diffusion model g_ϕ, it can be seen to be composed of an encoder $g_{\phi_{\text{enc}}}$ and a decoder $g_{\phi_{\text{dec}}}$. Given a denoising timestep t, the corresponding noisy image $\mathbf{x}_t$,



(a) (b)

Figure 2: (a) The frequency saliency analysis of the linearly probed diffusion representation, adversarially trained robust model, and standard non-robust model. Low frequencies are being centered. (b) The CIFAR-10 robust accuracy across perturbation budgets for the linear probed diffusion representation (DR), adversarially trained robust model (AT), and standard trained non-robust model (ST).

and optional conditions $\mathbf{c}$, we refer to the output of the encoder, $g_{\phi_{\text{enc}}}(\mathbf{x}_t, t, \mathbf{c}) = \mathbf{h}^{\text{DR}}_{\mathbf{x}_t,t,\mathbf{c}}$, as diffusion representations. In practice, for UNet-based diffusion models, representations are typically extracted from the upsampling blocks near the bottleneck layer [18], whereas for newer transformer-based diffusion models [54], representations are extracted near the middle layers [18, 19]. Additionally, the representation quality of diffusion models is often unimodal across timesteps, peaking at timesteps where the noisy image $\mathbf{x}_t$ contains a small amount of noise that removes irrelevant details for the perception task. This behavior can be explained by the high signal-to-noise ratio at those timesteps [20]. In this work, we follow Xiang et al. [18] to extract the representations from timesteps based on linearly probed clean accuracy experiments.

4 Methodology

Prior work suggests diffusion models may be less effective for representation learning and that pixel-reconstruction objectives can encourage non-informative or high-frequency features that hurt downstream adversarial training [23, 55, 21]. In this section, we empirically show that diffusion models trained via noisy-image denoising in fact learn features with desirable properties. Building on this, we leverage diffusion features as an auxiliary learning signal to improve downstream adversarial training. Additional discussion and analysis based on representation similarity are provided in Appendix A.

4.1 Observation

Representation Metrics. We posit that diffusion models encode representations that are inherently mildly robust but preserve diversity that can improve downstream adversarial training. To investigate the hypothesis, we analyze these representations using two key metrics from the representation learning literature: uniformity and alignment [56]. Uniformity measures how evenly representations are distributed on the unit hypersphere, reflecting the information preserved in the representation space, whereas alignment measures the distance between the representations of data examples with different positive views. In our setting, we form positive pairs by constructing adversarial images. As shown in Figure 3, diffusion representations are more robust than standard supervised training, while achieving richer features with a noticeably higher uniformity metric. In contrast, adversarial training increases alignment but decreases feature diversity, reflecting the difficulty of learning a robust model that preserves feature diversity. In this work, we aim to leverage diffusion representations to improve adversarial training by shifting the alignment–uniformity frontier.

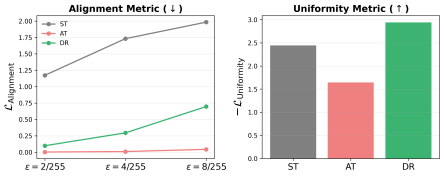


Figure 3: Representation metrics on CIFAR-10 for the standard-trained model (ST), the adversarially trained model (AT), and the diffusion representations (DR).

Frequency and Robustness Behavior. Vision models are often sensitive to high-frequency input perturbations, while adversarial training typically reduces this sensitivity and emphasizes more on low-frequency components [57]. Moreover, pixel reconstruction-based pretraining like MAE has been reported to have worse adversarially trained downstream performance than standard supervised

training, explained by a stronger reliance on mid and high-frequency features [21]. To investigate if diffusion representations exhibit similar behaviors, we conduct frequency-saliency analysis on the PGD perturbations on the CIFAR-10 linearly probed diffusion representations. Figure 2a shows that diffusion representations exhibit lower high-frequency saliency that resembles robust models. It also suggests that diffusion representations do not suffer from the same pixel-reconstruction frequency behavior reported in previous work, which is likely related to the partially noise-corrupted images seen during denoising training.

Additionally, one possibility is that frozen diffusion features are already robust enough and do not require further robust finetuning [44]. We evaluate robustness by measuring the robust accuracy of a linear probe trained on top of a frozen unconditional diffusion model. To avoid a false sense of robustness due to randomness [28], we make the noise used during diffusion feature extraction deterministic. Figure 2b shows that these representations are inherently more robust to small-budget perturbations than standard supervised finetuned models, but still require robust training for competitive robustness.

4.2 Diffusion Representation Alignment for Robust Classifier Training

To efficiently improve downstream adversarial training with diffusion representations, we propose to integrate adversarial training with representation alignment [17, 23, 24]. Figure 4 illustrates the overall framework, where we leverage an auxiliary projection head to regularize the classifier representations.

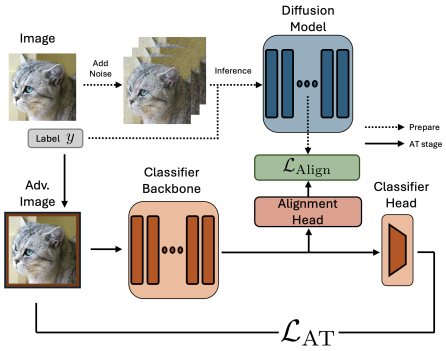


Figure 4: Diffusion Representation Alignment (DRA). We leverage an auxiliary projection head to align classifiers with the extracted diffusion representations.

Given an image-label pair $(\mathbf{x}, y)$, and the classifier $f_{\text{CLS}} = g_\theta \circ f_\theta$ consisting of an encoder f_θ and a classification head g_θ, we regularize the classifier representation $\mathbf{h}^{\text{CLS}}_{\hat{\mathbf{x}}} = f_\theta(\hat{\mathbf{x}})$ given the adversarial example $\hat{\mathbf{x}}$ computed during training. Specifically, we align classifier representations with the representations $\mathbf{h}^{\text{DR}}_{\mathbf{x}_t,t,y}$ extracted from a frozen diffusion model, using a trainable projection head g_{proj} that maps between the representation spaces. The diffusion representation alignment loss is defined as

$$\mathcal{L}_{\text{DRA}} = -\text{sim}(g_{\text{proj}}(\mathbf{h}^{\text{CLS}}_{\hat{\mathbf{x}}}), \mathbf{h}^{\text{DR}}_{\mathbf{x}_t,t,y}), \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ is the given representation similarity metric. The overall training objective becomes

$$\mathcal{L}_{\text{AT-DRA}} = \mathcal{L}_{\text{AT}} + \lambda \mathcal{L}_{\text{DRA}}, \tag{4}$$

where λ controls the regularization strength. In practice, we implement the projection head g_{proj} with an MLP module and use cosine similarity as the similarity metric, which performs well empirically and matches the prior work recommendations [23, 24]. For each image, we use a fixed timestep around the optimal linearly probed timestep [18] to extract diffusion representations. Details on extracting diffusion representations, projection head, and timestep choices are in Appendix D.

5 Experiments

In this section, we evaluate the effectiveness of Diffusion Representation Alignment (DRA) with adversarial training. The experiments across CIFAR-10, CIFAR-100, and ImageNet with different architectures and settings show consistent improvements. Lastly, we analyze the effect on classifier representations when incorporating diffusion models into adversarial training.

5.1 Setups

We mainly follow the DM-AT [11] setup, which is the state-of-the-art adversarial training framework that is also used by Bartoldson et al. [14], Cui et al. [15], Wu et al. [16]. In the following, we briefly explain the setup for each dataset. Implementation details are in Appendix B.

CIFAR-10/100. For CIFAR-10/100 [58], we follow the DM-AT recipe and perform AT with ℓ_∞ perturbation budget $\varepsilon = 8/255$ and step size $\alpha = 2/255$, using 10 PGD steps to adversarially augment the training images. The TRADES loss [4] is used as the adversarial training objective. To demonstrate the effectiveness of the proposed method across different training budget setups, CIFAR-10 experiments are conducted with synthetic data containing 1 million, 20 million, and 50 million synthetic images released by Wang et al. [11], with DRA using the same frozen CIFAR-10 EDM diffusion model checkpoint [39] that was used to generate the synthetic images. Additionally, experiments with CIFAR-100 using 1 million synthetic images and the EDM model released by Wang et al. [10] are also provided. For model choices, we conduct experiments on the WideResNets [59] WRN-28-10 and WRN-34-10, which are widely used in the adversarial training literature, as well as ViT-B/2 [60] following Wu et al. [16].

ImageNet. Additionally, to demonstrate the effectiveness in real-world scenarios, we conduct experiments on ImageNet [61], with models initialized from strong pretrained checkpoints. We set the perturbation budget to $\varepsilon = 4/255$ and evaluate with an input resolution of 224×224 following the RobustBench standard. For model selection, we train and evaluate ViT-B/16 and ConvNeXt-B [62], initialized from the DINOv3 [63] pretrained checkpoints released via `timm` [64]. For the diffusion synthetic data, we generate 4 million 256×256 synthetic images using LightningDiT [41].

5.2 Diffusion Representations Improve AT

Table 1 summarizes the results on CIFAR-10, CIFAR-100, and ImageNet when combining DRA with the state-of-the-art DM-AT training recipe. The results indicate that diffusion representations provide an effective feature prior for robust learning, leading to consistent gains in both clean accuracy and adversarial robustness.

Table 1: Clean and Robust Accuracy incorporating DRA on CIFAR-10, CIFAR-100, and ImageNet.

Dataset	Model	Method	Clean Acc.	AutoAttack Acc.
CIFAR-10 ($\ell_\infty = 8/255$)	WRN-28-10	DM-AT	92.44	67.31
		DM-AT + DRA	**93.17**	**67.86**
	ViT-B/2	DM-AT	94.35	71.31
		DM-AT + DRA	**95.13**	**72.03**
CIFAR-100 ($\ell_\infty = 8/255$)	WRN-28-10	DM-AT	68.34	35.72
		DM-AT + DRA	**69.85**	**36.27**
	ViT-B/2	DM-AT	68.53	36.52
		DM-AT + DRA	**69.95**	**37.43**
ImageNet ($\ell_\infty = 4/255$)	ConvNeXt-B	DM-AT	74.40	54.34
		DM-AT + DRA	**76.30**	**56.42**
	ViT-B/16	DM-AT	74.67	54.88
		DM-AT + DRA	**76.86**	**55.53**

We further evaluate CIFAR-10 across varying amounts of synthetic data (Table 2), from none up to 50 million images. While scaling synthetic data remains important, incorporating diffusion representations as an auxiliary learning signal more effectively leverages the robust knowledge encoded in diffusion models. Moreover, compared to methods targeting the robust-overfitting regime (e.g., IKL, ADR), which are reported to trade clean accuracy for robustness when training with diffusion synthetic data, DRA improves DM-AT robustness without sacrificing clean accuracy.

Finally, we scale up to ImageNet (Table 3) to verify that DRA remains effective when robust finetuning is applied to strong self-supervised foundation models on a downstream dataset, with or without diffusion synthetic data. Additional experiments with models trained from scratch are provided in Appendix M.

5.3 Diffusion Representation Alignment Improves Representation Quality

To understand if DRA actually improves representation quality, we plot the uniformity and alignment metrics on CIFAR-10 trained checkpoints, along with the corresponding clean and robust accuracy. Figure 7 presents the results, showing that DRA effectively leverages the diverse features encoded in diffusion representations, contributing to the improved clean and robust accuracy.

5.4 Diffusion Models Encourage Learning of More Easily Disentangled Representations

Recent mechanistic interpretability work hypothesizes that models may rely on feature superposition to encode more features than the available representation dimensions, which involves representing features as non-orthogonal directions in the activation space [66]. While this can be effective for compressing features that rarely co-activate, it has been suggested that superposition may be exploited by adversarial examples [25, 67].

Table 2: Comparison with state-of-the-art AT methods across different settings on CIFAR-10.

Method	Architecture	Syn.	Epoch	Clean Acc.	AutoAttack Acc.		
					$\ell_\infty = 2/255$	$\ell_\infty = 8/255$	$\ell_2 = 0.5$
AT	WRN-34-10	-	200	84.62	78.58	55.27	62.28
+ADR [12]	WRN-34-10	-	200	86.11	79.90	55.26	61.83
+IKL [15]	WRN-34-10	-	200	85.31	79.91	57.15	61.86
+DRA (Ours)	WRN-34-10	-	200	$\mathbf{89.18}_{\pm.11}$	$\mathbf{83.55}_{\pm.10}$	$\mathbf{57.71}_{\pm.20}$	$\mathbf{66.10}_{\pm.44}$
DM-AT [11]	WRN-28-10	1M	400	91.12	86.62	63.35	66.70
DM-AT [11]	WRN-28-10	1M	800	91.43	87.17	63.96	66.72
+DRA (Ours)	WRN-28-10	1M	400	$\mathbf{92.32}_{\pm.12}$	$\mathbf{87.90}_{\pm.09}$	$\mathbf{64.09}_{\pm.04}$	$\mathbf{69.19}_{\pm.05}$
DM-AT [11]	WRN-28-10	20M	2400	92.44	88.38	67.31	70.25
+IKL [15]	WRN-28-10	20M	2400	92.16	88.32	67.75	69.61
+DRA (Ours)	WRN-28-10	20M	2400	$\mathbf{93.17}_{\pm.04}$	$\mathbf{89.42}_{\pm.06}$	$\mathbf{67.86}_{\pm.03}$	$\mathbf{71.77}_{\pm.25}$
DM-AT	ViT-B/2	20M	500	92.21	88.21	66.41	70.08
+DRA (Ours)	ViT-B/2	20M	500	$\mathbf{93.26}_{\pm.09}$	$\mathbf{89.54}_{\pm.14}$	$\mathbf{67.68}_{\pm.06}$	$\mathbf{72.52}_{\pm.29}$
DM-AT [11]	WRN-70-16	50M	2000	93.25	89.71	70.69	69.78
+RA [65]	RaWRN-70-16	50M	2000	93.27	89.71	71.07	68.71
DM-AT [16]	ViT-B/2	50M	2000	94.35	91.10	71.31	72.00
+DRA (Ours)	ViT-B/2	50M	2000	$\mathbf{95.13}_{\pm.06}$	$\mathbf{91.74}_{\pm.02}$	$\mathbf{72.03}_{\pm.18}$	$\mathbf{73.68}_{\pm.14}$

Table 3: Clean and Robust Accuracy incorporating DRA on ImageNet.

Method	Model	Clean Acc.	AutoAttack Acc.		
			$\ell_\infty = 2/255$	$\ell_\infty = 4/255$	$\ell_2 = 2.0$
AT ($\ell_\infty = 4/255$)	ConvNext-B	$71.91_{\pm.48}$	$61.31_{\pm.54}$	$49.40_{\pm.44}$	$40.01_{\pm.81}$
	+DRA	$\mathbf{74.53}_{\pm.10}$	$\mathbf{64.17}_{\pm.07}$	$\mathbf{51.77}_{\pm.12}$	$\mathbf{44.65}_{\pm.76}$
	Δ	+2.62	+2.86	+2.37	+4.64
	ViT-B/16	$73.75_{\pm.23}$	$63.22_{\pm.09}$	$51.62_{\pm.02}$	$53.26_{\pm.23}$
	+DRA	$\mathbf{76.23}_{\pm.04}$	$\mathbf{65.64}_{\pm.01}$	$\mathbf{52.95}_{\pm.03}$	$\mathbf{55.41}_{\pm.05}$
	Δ	+2.48	+2.42	+1.33	+2.15
DM-AT ($\ell_\infty = 4/255$)	ConvNext-B	$74.40_{\pm.09}$	$65.26_{\pm.26}$	$54.34_{\pm.09}$	$43.54_{\pm.18}$
	+DRA	$\mathbf{76.30}_{\pm.18}$	$\mathbf{67.48}_{\pm.16}$	$\mathbf{56.42}_{\pm.23}$	$\mathbf{47.64}_{\pm.23}$
	Δ	+1.90	+2.22	+2.08	+4.10
	ViT-B/16	$74.67_{\pm.05}$	$65.65_{\pm.05}$	$54.88_{\pm.16}$	$50.75_{\pm1.03}$
	+DRA	$\mathbf{76.86}_{\pm.06}$	$\mathbf{67.44}_{\pm.12}$	$\mathbf{55.53}_{\pm.22}$	$\mathbf{55.81}_{\pm.92}$
	Δ	+2.19	+1.79	+0.65	+5.06

Moving from toy settings to real world datasets, where the degree of superposition is difficult to quantify, Gorton and Lewis [25] use sparse autoencoders (SAEs) [68] as a proxy for understanding the effect of robust training on classifier representations. SAEs learn a set of wide but sparse and interpretable features that aim to reconstruct model activations, with lower reconstruction loss reflecting that the representation is easier to disentangle into the set of sparse features. To investigate the effect of incorporating diffusion models into adversarial training, we train TopK-SAEs on classifier representations with different sparsity levels, $K \in \{8, 16, 32\}$, using model checkpoints trained on ImageNet, and compare the normalized SAE reconstruction loss [69, 25].

Figure 8 presents the results. We find that incorporating diffusion synthetic data and diffusion representation alignment improves robustness while also reducing the reconstruction loss, suggesting that the learned representations become easier to decompose into sparse features. This complements the observations of Gorton and Lewis [25], which report that adversarial training encourages more disentanglable representations, with larger perturbation budgets further amplifying this effect. Our results show that incorporating diffusion models into adversarial training can improve robustness and also encourage models to learn representations that are easier to disentangle.

5.5 Distinct Roles of Diffusion Models

Diffusion synthetic data augments the pool of input examples for adversarial training, while diffusion representation alignment provides an auxiliary learning signal from the mildly robust and diverse feature prior. In this section, we further investigate how incorporating these methods into robust training affects the resulting robust classifiers.

To understand how they shape classifier representations, we build on the methodology of Feng et al. [26] and examine whether the use of representation dimensions changes when diffusion synthetic

data and diffusion representations are introduced. Specifically, Feng et al. [26] proposed *classification dimension* as a measure of the intrinsic dimensionality of the feature space, approximated by the minimum number of principal components needed to preserve high classification accuracy.

Concretely, we first apply PCA to the classifier representations to obtain eigenvectors. We then modify the forward pass by projecting the representation onto the top-K eigenvectors before feeding it into the classification head, and measure the resulting accuracy. As a robust-aware variant, we additionally measure robust accuracy by computing eigenvectors from clean representations, and projecting adversarial representations onto the subspace.

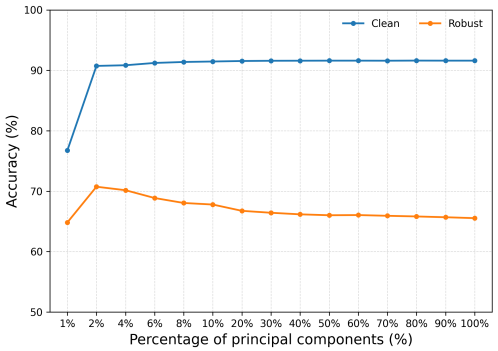


Figure 5: Clean and robust accuracy when projecting CIFAR-10 representations onto the top-K principal components.

Figure 5 shows an example on a robust model trained on CIFAR-10. As expected, classification accuracy gradually recovers to the original performance as more eigenvectors are included. For robust accuracy, however, we observe that performance first improves and then degrades as K increases, suggesting that adversarial perturbations may disproportionately exploit less important principal components. Furthermore, Table 4 presents the results for CIFAR-10 adversarially trained models with the minimum number of principal components required to recover 99% of the original clean accuracy (CLS Dim), and the robust-aware dimension that achieves the highest robust accuracy (Robust Dim). The results suggest that synthetic data concentrates predictive information in fewer principal components, while DRA increases the number needed to preserve clean accuracy.

To provide theoretical intuition for these results, we analyze a simplified linear model in which several concepts share a limited representation space (Appendix G). The analysis suggests that encoding fewer concepts or using more encoding dimensions helps limit the fraction of the classifier margin that perturbations can remove.

Table 4: Classification and robust dimensions with WRN-28-10s on CIFAR-10. CLS Dim: minimum number of components needed to recover 99% of the original clean accuracy.

Method	CLS Dim	Robust Dim
AT	14	9
AT+DRA	42	22
DM-AT	11	11
DM-AT+DRA	15	23

While it is intuitive that representation alignment may promote diverse and robust feature encoding, the reduction in the number of components needed for classification when trained with synthetic data may be related to the observation that diffusion synthetic examples are easier to classify than the original real data [32]. Prior work has also explored using partially synthesized diffusion data as an augmentation strategy to improve robustness [70], though not specifically in the context of enhancing adversarial training recipes. Although previous work has largely attributed the benefits of diffusion-based synthetic data to improved image quality [11, 14], it might not be the most essential factor behind its effectiveness. We leave this viewpoint as a potential direction for further improving adversarial training.

5.6 Ablation Studies

Regularization Strength.

In our experiments, the alignment regularization strength is set to $\lambda = 1.2$. We select this value based on an initial sweep using WRN-28-10 on CIFAR-100, and then fix the same coefficient for the remaining settings to reduce computational cost. Figure 6 shows that DRA consistently improves upon the baseline in both clean and robust accuracy. When λ becomes too large, robustness improvements saturate, as the alignment term can outweigh the original adversarial training objective.

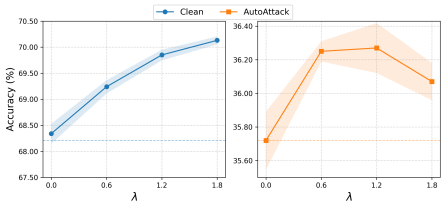


Figure 6: Clean and Robust accuracy on CIFAR-100 for DM-AT+DRA with different DRA regularization strength λ.

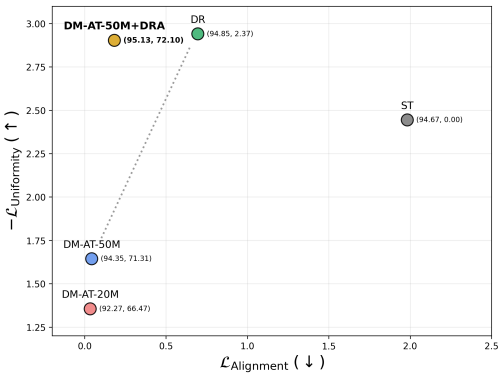


Figure 7: We plot the alignment and uniformity metrics, along with clean and robust accuracy on CIFAR-10 ($\ell_\infty = 8/255$) shown in parentheses.

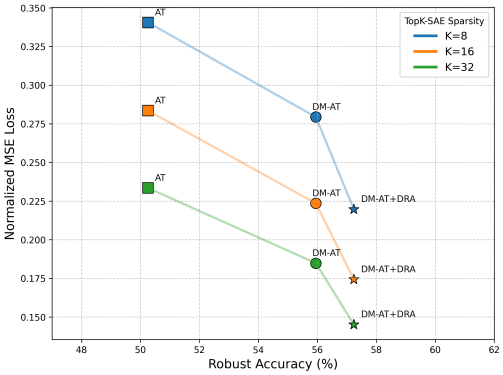


Figure 8: Normalized reconstruction losses of Top-K SAEs trained on ImageNet using AT, DM-AT, and DM-AT+DRA ViT-B checkpoints.

Does Noisy-Input Training Alone Learn Good Enough Features? Given that representations extracted from diffusion models serve as effective feature priors for robust learning, a natural question is whether standard training coupled with the same noisy-input procedure is sufficient to achieve similar benefits. We explore this hypothesis by applying noisy-input discriminative pre-training to the same UNet encoder architecture used in the EDM diffusion model, using this encoder as the alignment target in Figure 4. Results show that replacing the diffusion target in a WRN-28-10 trained with DM-AT+DRA (1M synthetic) reduces robust accuracy from 64.09% to 62.62%, which is inferior to the vanilla DM-AT baseline. This indicates that noisy-input training alone is insufficient; rather, the generative training objective of diffusion models is a critical factor in producing feature priors that benefit downstream robust training. Relatedly, Jaini et al. [51] found that diffusion-like noisy discriminative pre-training increases shape bias but hurts OOD accuracy. As a complementary finding, we analyze the frequency behavior of leveraging diffusion representations or the noisy-input pretrained discriminative features, and find that the latter have an undesirable effect of increasing the sensitivity to mid-high frequency components (Appendix J).

Why Not Adversarially Finetune the Diffusion Encoder Directly? Several works in self-supervised learning that leverage diffusion representations finetune the diffusion encoder end-to-end with a prediction head on top [18, 20, 44]. We evaluate this approach in our initial exploration under the same AT setting as in Table 2. It achieves 87.77% clean accuracy and 55.76% robust accuracy. While this improves over the WRN-34-10 AT baseline, it is still below WRN-34-10 AT+DRA and is less training-efficient (1.35× training time per epoch). Additionally, diffusion models are often trained with class conditioning for data generation; however, deploying the diffusion encoder at inference time can only provide unconditional representations, which can result in a slight decrease in representation quality [18, 19]. In contrast, DRA better leverages the robust knowledge encoded in diffusion representations while enabling more flexible downstream classifier choices that are better suited to the classification task.

Further ablations and analyses, covering DRA computational cost, alternative SSL feature alignment targets, and SAE-based representation analysis, are provided in the Appendix.

6 Conclusions

In this work, we systematically explore whether diffusion models can further improve adversarial training other than synthetic data generation. We find that diffusion models encode representations that provide partially robust but diverse features, and propose to integrate diffusion representation alignment into adversarial training. Experiments across settings and datasets show that incorporating diffusion representations effectively leverages them as an auxiliary learning signal to improve robust classifier training. Furthermore, our analysis indicates that using diffusion models improves classifier robustness and also encourages models to learn representations that are easier to disentangle. We hope our findings inspire further uses of diffusion models to improve adversarial training beyond the perspective of generating higher-quality synthetic images.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

[5] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

[6] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, 2020.

[7] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[8] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, and Tongliang Liu. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning (ICML)*, 2022.

[10] Yifei Wang, Liangchen Li, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. Balance, imbalance, and rebalance: Understanding robust overfitting from a minimax game perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[11] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning (ICML)*, 2023.

[12] Yu-Yu Wu, Hung-Jui Wang, and Shang-Tse Chen. Annealing self-distillation rectification improves adversarial training. In *International Conference on Learning Representations (ICLR)*, 2024.

[13] Yidong Ouyang, Liyan Xie, and Guang Cheng. Improving adversarial robustness through the contrastive-guided diffusion process. In *International Conference on Machine Learning (ICML)*, 2023.

[14] Brian R. Bartoldson, James Diffenderfer, Konstantinos Parasyris, and Bhavya Kailkhura. Adversarial robustness limits via scaling-law and human-alignment studies. In *International Conference on Machine Learning (ICML)*, 2024.

[15] Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled Kullback–Leibler divergence loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[16] Juntao Wu, Ziyu Song, Xiaoyu Zhang, Shujun Xie, Longxin Lin, and Ke Wang. Vision transformers beat WideResNets on small scale datasets adversarial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[17] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[18] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[19] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2025.

[20] Xiao Li, Zekai Zhang, Xiang Li, Siyi Chen, Zhihui Zhu, Peng Wang, and Qing Qu. Understanding representation dynamics of diffusion models via low-dimensional modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[21] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Yinpeng Chen, Lu Yuan, Gang Hua, Weiming Zhang, and Nenghai Yu. Improving adversarial robustness of masked autoencoders via test-time frequency-domain prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[22] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Dreamteacher: Pretraining image backbones with deep generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[23] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

[24] Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[25] Liv Gorton and Owen Lewis. Adversarial examples are not bugs, they are superposition. In *Mechanistic Interpretability Workshop at NeurIPS*, 2025.

[26] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[27] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.

[28] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

[29] Yue Gao, Ilia Shumailov, Kassem Fawaz, and Nicolas Papernot. On the limitations of stochastic pre-processing defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[30] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

[31] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! In *International Conference on Learning Representations (ICLR)*, 2023.

[32] Kai Hu, Klas Leino, Zifan Wang, and Matt Fredrikson. A recipe for improved certifiable robustness. In *International Conference on Learning Representations (ICLR)*, 2024.

[33] Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[34] Bo-Han Lai, Pin-Han Huang, Bo-Han Kung, and Shang-Tse Chen. Enhancing certified robustness via block reflector orthogonal layers and logit annealing loss. In *International Conference on Machine Learning (ICML)*, 2025.

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[36] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[39] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[40] Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, 2024.

[41] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[42] Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[43] Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *International Conference on Learning Representations (ICLR)*, 2026.

[44] Mika Yagoda, Shady Abu-Hussein, and Raja Giryes. Diffusion models are robust pretrainers. *IEEE Signal Processing Letters*, 32:4219–4223, 2025. doi: 10.1109/LSP.2025.3624151.

[45] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

[46] Xiao Li, Wenxuan Sun, Huanran Chen, Qiongxiu Li, Yingzhe He, Jie Shi, and Xiaolin Hu. ADBM: Adversarial diffusion bridge model for reliable adversarial purification. In *International Conference on Learning Representations (ICLR)*, 2025.

[47] Kaibo Wang, Xiaowen Fu, Yuxuan Han, and Yang Xiang. DiffHammer: Rethinking the robustness of diffusion-based adversarial purification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[48] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *International Conference on Machine Learning (ICML)*, 2024.

[49] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[50] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[51] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *International Conference on Learning Representations (ICLR)*, 2024.

[52] Alexander Cong Li, Ananya Kumar, and Deepak Pathak. Generative classifiers avoid shortcut solutions. In *International Conference on Learning Representations (ICLR)*, 2025.

[53] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

[54] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[55] Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *International Conference on Machine Learning (ICML)*, 2024.

[56] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

[57] Alvin Chan, Yew Soon Ong, and Clement Tan. How does frequency bias affect the robustness of neural image classifiers against common corruption and adversarial perturbations? In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[58] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[59] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[60] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[62] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[63] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL `https://arxiv.org/abs/2508.10104`.

[64] Ross Wightman. PyTorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[65] ShengYun Peng, Weilin Xu, Cory Cornelius, Matthew Hull, Kevin Li, Rahul Duggal, Mansi Phute, Jason Martin, and Duen Horng Chau. Robust principles: Architectural design principles for adversarially robust CNNs. In *British Machine Vision Conference (BMVC)*, 2023.

[66] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[67] Edward Stevinson, Lucas Prieto, Melih Barsbey, and Tolga Birdal. Adversarial attacks leverage interference between features in superposition. In *Mechanistic Interpretability Workshop at NeurIPS*, 2025.

[68] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[69] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025.

[70] Chandramouli S. Sastry, Sri Harsha Dumpala, and Sageev Oore. DiffAug: A diffuse-and-denoise augmentation for training robust classifiers. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[71] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024.

[72] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

[74] Gongfan Fang, Kunjun Li, Xinyin Ma, and Xinchao Wang. Tinyfusion: Diffusion transformers learned shallow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[75] Jonas Hübotter, Patrik Wolf, Alexander Shevchenko, Dennis Jüni, Andreas Krause, and Gil Kur. Specialization after generalization: Towards understanding test-time training in foundation models. In *International Conference on Learning Representations (ICLR)*, 2026.

[76] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[77] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Appendix

A Representation Similarity Analysis

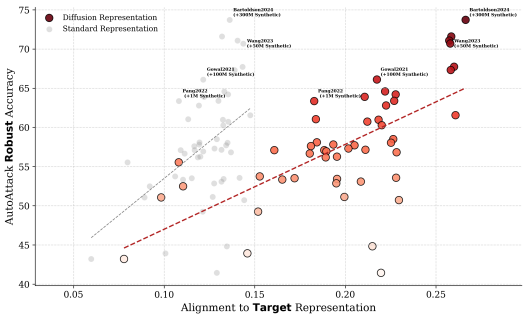


Figure 9: Similarity scores are measured with respect to representations extracted from the diffusion model (red) and from the RobustBench standard-trained model (gray).

As a motivational experiment, following the work of Huh et al. [71] on analyzing representation similarities between models trained with different learning objectives, we extract the representations from the CIFAR-10 EDM diffusion model [39] used in DM-AT [11], and measure the representation similarity score CKNNA [71] with the CIFAR-10 ℓ_∞-robust models from RobustBench [5]. Figure 1 presents the results, showing that better robust models, often trained with abundant synthetic images, already have the tendency to exhibit higher representation similarity with diffusion representations. As the tendency could also be explained by better robust models having improved natural classification performance for clean images, we also plot the representation similarity score with the standard-trained model from RobustBench (Figure 9). The observed correlation motivates us to investigate whether explicitly leveraging diffusion representations as an auxiliary learning signal is beneficial for robust classifier training.

B Experiment Setup Details

Training Setups. For CIFAR-10/100 WideResNet training, we follow the exact same setup as DM-AT [11]: PGD-10 TRADES [4] with $\beta = 5$, weight averaging with $\tau = 0.995$, SGD with momentum 0.9, weight decay 5×10^{-4}, and $lr = 0.2$ with a cosine annealing scheduler. For ViT-B, we follow Wu et al. [16] and use the Lion optimizer [72] with batch size 1024, $lr = 10^{-4}$, and weight decay 0.5. For ImageNet models, we fully finetune for 100 epochs using the Lion optimizer and the common PGD-3 TRADES with $\beta = 10$. For the baselines, we report the numbers from the original paper or evaluated with the released checkpoints. For our experiments, statistics over three runs are reported.

For the SAE experiments, we follow prior works [25, 69] using TopK SAEs trained with an expansion factor of 8, $lr = 5 \times 10^{-4}$, and $k_{aux} = 512$ with auxiliary weight 0.1, using the Adam optimizer [73] for a total of 50000 steps.

Computational Resources. The experiments are mostly conducted on machines with 4-GPU (Nvidia H100 80GB), with 12 CPU cores and 200GB of RAM per GPU. Each training run using 4 GPUs takes around 2 minutes per epoch for the CIFAR experiments, and 20 minutes per epoch for the ImageNet experiments. Due to the expensive nature of adversarial training and evaluations, an estimated total of 6000 GPU hours is needed for the experiments, including training runs, evaluations, and preliminary explorations.

C Additional Discussion for Related Work

Frozen Diffusion Features for Robustness. Yagoda et al. [44] show that training a classification head on frozen, unconditional diffusion encoders can achieve robustness that is slightly below adversarially trained models, while being much cheaper to train. However, as their approach relies on inference-time randomness, robust evaluation should be more carefully considered [28]. For example, the configuration of Attention Head, b=8, t=50, has been reported to achieve 46.0% AutoAttack Robust Accuracy, but a simple EOT-based evaluation could reduce the reported robust accuracy to

17.3%. We also evaluate a linear probe on an EDM diffusion model on CIFAR-10 by fixing the added noise to remove inference-time randomness during adversarial evaluation (Figure 2b). The results show that diffusion representations still require robust training to provide competitive robustness.

Representation Alignment with Diffusion Models. A growing line of work inspects diffusion features for a variety of downstream tasks via feature alignment. CleanDIFT [24] leverages self-distillation across timesteps to produce clean, timestep-independent features for dense prediction tasks; DreamTeacher [22] and RepFusion [17] distill diffusion priors into discriminative visual models for downstream discriminative tasks; REPA [23] aligns diffusion intermediate features with self-supervised features to accelerate the training of diffusion transformers. These works converge on a similar recipe: a projection head with a distillation loss such as cosine similarity or L2 distance to align the feature space. In this work, we build upon the alignment mechanism, but instead study whether diffusion representations are beneficial as an auxiliary learning signal for robust classifier training in AT.

D Representation Alignment Implementation Details

Diffusion Representation Extraction. For CIFAR-10/100, we use the same EDM diffusion model checkpoint as Wang et al. [11] to generate synthetic images. We extract representations at noise scale $\sigma_t = 0.1$ from the UNet bottleneck block before the upsampling layers, and apply average pooling over spatial dimensions.

For ImageNet, we use the LightningDiT checkpoint released by Yao et al. [41]. We extract features from middle layer 14 at $t = 0.8$, where $t = 0$ corresponds to pure noise and $t = 1$ to the clean image. We also found that LightningDiT representations suffer from the large-activation issue reported in prior work [74, 42], which reduces their effectiveness as a learning signal. To mitigate this, we apply the method introduced in Gan et al. [42], which applies the zero-out-and-AdaLN-normalize method on the channels with abnormally large activations (channels 1053 and 259 in the released LightningDiT checkpoint). Finally, average pooling over spatial dimensions is applied to extract the representations.

Lastly, for CIFAR-10 models trained with more than one million synthetic images, we extract an additional representation per image by sampling an extra timestep using the same sampling function the EDM model originally used during training, and take the average of the cosine similarity loss. This results in a slight improvement, likely by better leveraging the representations across timesteps [24, 20].

Alignment Module. We follow Yu et al. [23] and use a 3-layer MLP, trained with the cosine similarity loss. In initial experiments, more sophisticated alignment heads and loss functions did not improve performance, so we retain this simple configuration as recommended in prior work [23, 24].

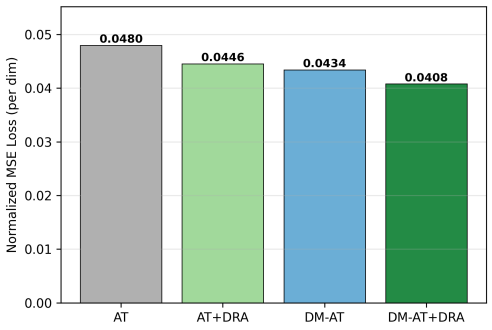
E Computational Cost of DRA

Since the computational cost is dominated by the multiple forward and backward passes required for adversarial training, extracting features from the diffusion encoder on the fly at each iteration introduces minimal training-time overhead ($\sim 1.04\times$ training time with ~ 1 GB additional VRAM per GPU on CIFAR-10/100, and ~ 3 GB additional VRAM with $\sim 1.13\times$ training time per GPU for our ImageNet experiments). In practice, one could consider further reducing the training time and memory overhead by precomputing the diffusion features during the generation of the synthetic data for DM-AT, albeit at the cost of increased disk storage.

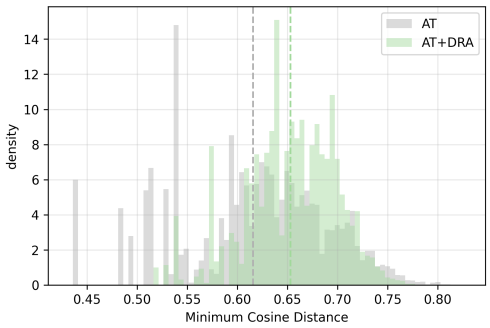
F Additional Analysis on SAE Features

Using the same CIFAR-10 checkpoints analyzed in Section 5.5, we further examine the SAE features that the classifier uses to reconstruct each input's representation.

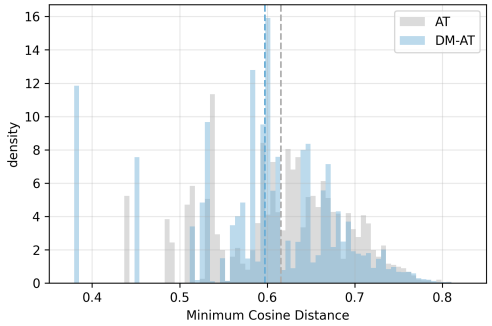
Reconstruction loss on CIFAR-10. Figure 10a presents the normalized TopK-SAE ($K = 32$) reconstruction loss on CIFAR-10 checkpoints. Similar patterns observed on ImageNet hold, where incorporating diffusion models encourages more disentangled representations.



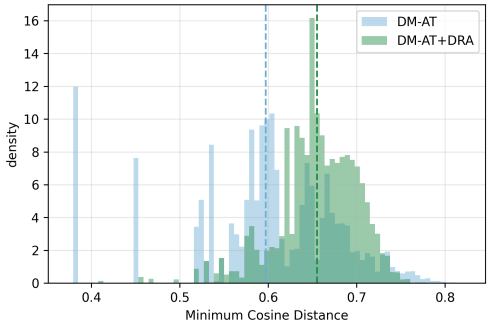
(a) Normalized TopK-SAE reconstruction loss.



(b) AT vs. AT+DRA.



(c) AT vs. DM-AT.



(d) DM-AT vs. DM-AT+DRA.

Figure 10: SAE analyses on the same CIFAR-10 checkpoints used in Section 5.5. (a) Normalized TopK-SAE reconstruction loss across AT, AT+DRA, DM-AT, and DM-AT+DRA. (b)–(d) Distributions of the minimum pairwise cosine distance among the TopK (K=32) active SAE features per input. Dashed lines mark distribution means.

Directional diversity of active features. We additionally probe the directional structure of the active features. For each input, we take the TopK (K=32) active features used to reconstruct its representation and compute the minimum pairwise cosine distance among them. A larger minimum distance indicates that the model recruits a more directionally diverse set of features to encode the representation of that input.

Figures 10b–10d report the resulting distributions. Adding DRA shifts the distribution toward larger minimum cosine distances, both on top of AT and on top of DM-AT. In contrast, switching from AT to DM-AT does not produce a comparable shift. This complements the analysis in Section 5.5, showing that using DRA encourages the per-input active SAE features to point in more distinct directions.

G Theoretical Analysis of Interference in Concept Superposition

Using diffusion synthetic data and diffusion representation alignment in AT both encourage to learn representations that are easier to be reconstructed by SAEs. However, the analysis also suggests that they may have different effects on the use of representation dimensions (Section 5.5). To provide theoretical intuition, we analyze a linear model in which n concepts are encoded in an r-dimensional representation, and ask how n and r affect the fraction of a classifier's margin that a perturbation can remove. The analysis is motivated by the observation that adversarial attacks can exploit interference between superposed features [67]. Under a uniformly random encoding subspace, Proposition G.1 shows that the mean squared relative margin loss decreases with fewer encoded concepts or a higher encoding dimension. We conclude the section by relating n and r to the analysis of Section 5.5.

Setup. Inspired by the feature-superposition framework of Elhage et al. [66], we study a linear model in which more concepts are encoded than there are independent representation dimensions. In this model, a *concept* is an assumed underlying semantic factor of the input, such as a visual attribute. For an input x, let $z(x) \in \mathbb{R}^n$ collect the coefficients of the n concepts encoded by the classifier for the task. The classifier representation is a linear encoding of these coefficients,

$$h(x) = Dz(x), \qquad D \in \mathbb{R}^{d\times n}, \qquad \operatorname{rank}(D) = r, \tag{5}$$

so column j of D is the direction assigned to concept j, d is the network width, and r is the number of dimensions that the n directions span. We call r the *encoding dimension*. When $r < n$, the n concept directions are linearly dependent, so concepts overlap in the representation.

Fix a true class and one competing class. A classifier assigns each class a logit, and the comparison between the two classes is decided by the difference between their logits. Let $\theta \in \mathbb{R}^n$ with $\|\theta\|_2 = 1$ be the concept weights of the *target logit difference* $\theta^\top z$, the logit difference that the classifier is trained to fit, with positive values favoring the true class. Let

$$S = \{j : \theta_j \neq 0\}, \qquad s = |S|, \qquad S^c = \{1, \dots, n\} \setminus S$$

denote the support of θ, its size, and its complement. Only the s concepts in S enter the target logit difference, and the concepts in S^c may matter for other class comparisons. For any vector $a \in \mathbb{R}^n$, write $a_S = (a_j)_{j\in S}$ and $a_{S^c} = (a_j)_{j\in S^c}$ for the subvectors containing entries indexed by S and S^c, respectively.

Let $w \in \mathbb{R}^d$ be the difference between the two classes' learned linear classification weights. The learned logit difference, which we call the *margin*, is

$$g(z) = w^\top Dz = u^\top z, \qquad u = D^\top w \in \mathbb{R}^n, \tag{6}$$

so u_j is the classifier's effective weight on concept j, to be compared with the target weight θ_j. Ideally $u = \theta$, and the margin depends only on the concepts in S. However, u must lie in the row space of D, an r-dimensional subspace of $\mathbb{R}^n$, and when $r < n$ this subspace need not contain θ. When the row space does not contain θ, exact recovery of the target weights is impossible. The fitted weights may then acquire nonzero components on S^c, so that concepts irrelevant to the comparison move the margin. We call this effect *interference*. Under the random encoding introduced below, it occurs with probability one.

We make two simplifying assumptions. The first restricts perturbations to S^c, keeping the target logit difference $\theta^\top z$ fixed so that any margin loss is due to interference. The second assumption specifies how the encoding and the classifier are formed.

Assumption G.1 (Perturbations outside the target support)**.** *For a budget $\epsilon > 0$, the allowed perturbations $z \mapsto z + \Delta z$ satisfy $\Delta z_S = 0$ and $\|\Delta z_{S^c}\|_2 \leq \epsilon$.*

Since $g(z + \Delta z) - g(z) = u_{S^c}^\top \Delta z_{S^c}$, the Cauchy–Schwarz inequality gives the largest margin loss within this ball,

$$L(z) := g(z) - \inf_{\Delta z} g(z + \Delta z) = \epsilon \|u_{S^c}\|_2, \tag{7}$$

attained at $\Delta z_{S^c} = -\epsilon\, u_{S^c}/\|u_{S^c}\|_2$ when $u_{S^c} \neq 0$. Thus $L(z)$ measures interference through the weights on S^c. It vanishes when $u_{S^c} = 0$, as for the ideal classifier $u = \theta$, and is positive otherwise.

For a clean example with $g(z) > 0$, the *relative margin loss* is the fraction of the clean margin that the worst allowed perturbation removes:

$$R(z) := \frac{L(z)}{g(z)} = \frac{\epsilon\|u_{S^c}\|_2}{g(z)}, \qquad \inf_{\Delta z} g(z + \Delta z) = g(z)\big(1 - R(z)\big). \tag{8}$$

The ratio can exceed one, and the comparison remains in favor of the true class exactly when $R(z) < 1$.

Assumption G.2 (Random encoding and least-squares classifier)**.** *Let $1 \le s < n$ and $2 < r < n$.*

1. Isotropic concepts. $\mathbb{E}_z[zz^\top] = \sigma^2 I_n$ *for some* $\sigma > 0$.
2. Random encoding. *The row space of D, which is the set of effective weight vectors $u = D^\top w$ that the linear classifier can realize, is a uniformly random r-dimensional subspace of $\mathbb{R}^n$.*
3. Least-squares fit. *The weight vector w minimizes the population squared error $\mathbb{E}_z(\theta^\top z - w^\top Dz)^2$ over $\mathbb{R}^d$.*

Item 2 assumes an encoding with no preferred alignment to this particular class comparison, and item 3 replaces the adversarial training objective of Section 4.2 used in our experiments with a tractable least-squares fit, an approach also used by Hübotter et al. [75, Appendix C.3] to analyze superposition interference. We evaluate the clean example $z_0 = t\theta$ with $t > 0$, whose concept coefficients align with the target weights: its target logit difference is $\theta^\top z_0 = t$, and $(z_0)_{S^c} = 0$. We use this target-aligned example as an idealized probe of interference rather than as a representative draw from the data distribution. Its active concepts lie entirely in S, so any margin lost under Assumption G.1 is due to interference alone, and its alignment with θ ensures a positive fitted margin with probability one. Even on this example the clean margin falls short of the target: the proof below shows that $g(z_0) = tX$ with $X = \|\Pi\theta\|_2^2 < 1$, where Π projects onto the row space of D, which is the cost of fitting θ within an r-dimensional subspace.

Proposition G.1 (Mean squared relative margin loss)**.** *Let Assumptions G.1 and G.2 hold, with n concepts, encoding dimension r, target support size s, perturbation budget ϵ, and clean example $z_0 = t\theta$. Then $g(z_0) > 0$ with probability one over the random encoding, and*

$$\mathbb{E}_D\big[R(z_0)^2\big] = \Big(\frac{\epsilon}{t}\Big)^2 \frac{n-s}{n-1}\,\frac{n-r}{r-2}. \tag{9}$$

At fixed s and ϵ/t, the right-hand side strictly increases in n at fixed r and strictly decreases in r at fixed n. When $s \ll n$ and $r \gg 1$,

$$\mathbb{E}_D\big[R(z_0)^2\big] = \Big(\frac{\epsilon}{t}\Big)^2 \Big(\frac{n}{r} - 1\Big)\Big(1 + O\big(\tfrac{s}{n} + \tfrac{1}{r}\big)\Big). \tag{10}$$

In words, perturbing the concepts outside the target comparison removes, in root-mean-square over random encodings, a fraction of the margin that grows with the number of concepts per encoding dimension: by (10), $\sqrt{\mathbb{E}_D[R(z_0)^2]} \approx (\epsilon/t)\sqrt{n/r - 1}$. Fewer concepts and more dimensions both shrink it. Reducing n at fixed s keeps the concepts that define the target comparison and removes concepts outside it. Increasing r at fixed n represents the same concepts in more dimensions. The crowding factor $(n-r)/(r-2)$ vanishes in the full-rank limit $r = n$, where the row space of D is all of $\mathbb{R}^n$, the least-squares fit recovers $u = \theta$, and interference disappears.

Remark. The analysis considers ℓ_2 perturbations in concept space. Extending these conclusions to input-space attacks on general inputs requires further analysis and assumptions relating input-space perturbations to the concept-space budget ϵ.

Proof. Throughout the proof, $\mathbb{E}$ denotes expectation over the random encoding D. By the isotropy condition $\mathbb{E}_z[zz^\top] = \sigma^2 I_n$, the population squared error equals $\sigma^2\|\theta - D^\top w\|_2^2$. As w ranges over $\mathbb{R}^d$, $D^\top w$ ranges over the row space of D, so every minimizer yields the same effective weights $u = \Pi\theta$, where Π is the orthogonal projection onto that row space.

Let $X = \theta^\top \Pi\theta = \|\Pi\theta\|_2^2$ and $v = \Pi\theta - X\theta$. Since $\|\theta\|_2 = 1$ and $\Pi^2 = \Pi$, we have $v \perp \theta$ and $\|v\|_2^2 = X(1-X)$. Because $\theta_{S^c} = 0$, $u_{S^c} = v_{S^c}$, and the clean margin is $g(z_0) = u^\top z_0 = tX$.

We next identify the distribution of X. By rotational invariance, the distribution of $\|\Pi\theta\|_2^2$ for a uniformly random r-dimensional row space and a fixed unit vector θ equals the distribution of $\|P\theta'\|_2^2$, where P is the projection onto a fixed r-dimensional coordinate subspace and θ' is uniform on the unit sphere. Writing $\theta' = q/\|q\|_2$ with $q \sim \mathcal{N}(0, I_n)$ shows that X has the same distribution as $\sum_{j \le r} q_j^2 / \sum_{j \le n} q_j^2$, that is, $X \sim \mathrm{Beta}(r/2, (n-r)/2)$. Thus, with probability one over the random encoding D, $0 < X < 1$ and hence $g(z_0) = tX > 0$.

We now determine the direction of v given X. Let Q be any rotation of $\mathbb{R}^n$ that fixes θ. Then $Q\Pi Q^\top$ has the same distribution as Π, and replacing Π by $Q\Pi Q^\top$ leaves X unchanged while mapping v to Qv. Hence (X, v) and (X, Qv) have the same joint distribution for every such Q. Conditional on X, the distribution of v is therefore invariant under all rotations of $\theta^\perp$, and since $\|v\|_2^2 = X(1-X)$ is determined by X, the direction of v is uniform on the unit sphere of the $(n-1)$-dimensional space $\theta^\perp$. Consequently,

$$\mathbb{E}[vv^\top \mid X] = \frac{X(1-X)}{n-1}\big(I_n - \theta\theta^\top\big).$$

Summing the $n-s$ diagonal entries indexed by S^c, where $\theta_j = 0$, gives

$$\mathbb{E}\big[\|u_{S^c}\|_2^2 \mid X\big] = \frac{n-s}{n-1}\, X(1-X).$$

Combining the above,

$$\mathbb{E}\big[R(z_0)^2\big] = \frac{\epsilon^2}{t^2}\,\mathbb{E}\left[\frac{\|u_{S^c}\|_2^2}{X^2}\right] = \frac{\epsilon^2}{t^2}\,\frac{n-s}{n-1}\,\mathbb{E}\left[\frac{1-X}{X}\right] = \frac{\epsilon^2}{t^2}\,\frac{n-s}{n-1}\,\frac{n-r}{r-2},$$

where the last step uses the inverse moment $\mathbb{E}[X^{-1}] = \frac{n-2}{r-2}$ of the beta distribution, which is finite exactly when $r > 2$.

For fixed r and $s \geq 1$, $(n-s)/(n-1)$ is nondecreasing in n and $n-r$ is strictly increasing, so the right-hand side of (9) strictly increases in n. For fixed n and s, $(n-r)/(r-2)$ strictly decreases in r. Finally, $(n-s)/(n-1) = 1 + O(s/n)$ and $(n-r)/(r-2) = (n/r-1)\big(1+O(1/r)\big)$ give (10).

□

Relation to the empirical analysis. Proposition G.1 quantifies how reducing n at fixed r, or increasing r at fixed n, reduces the mean squared relative margin loss when s and ϵ/t are held fixed. In the regime $s \ll n$ and $r \gg 1$, this quantity is approximately proportional to $n/r - 1$. In Section 5.5, classification dimension serves as a qualitative proxy for the encoding dimension r. Its increase under DRA suggests that predictive information occupies more representation dimensions. In contrast, DM-AT modestly decreases this proxy, suggesting that its effect is unlikely to operate through an increase in r. Instead, its lower SAE reconstruction loss, together with the observation that diffusion synthetic images are easier to classify [32], may point to a smaller effective concept count n.

H SSL Features as Surrogate Representations in DRA

To examine whether the regularization benefit of DRA extends beyond diffusion representations, we replace the diffusion target with self-supervised features from DINOv3 [63], MAE [76], and CLIP [77] in our ImageNet setup, keeping all other components of the AT recipe fixed.

In every configuration, training was unstable and collapsed within the first few epochs, with the AT loss failing to decrease. This suggests a strong conflict between the adversarial training objective and the SSL alignment loss. We attempted to mitigate this by lowering the regularization strength from the default λ=1.2 to $\lambda \in \{0.6, 0.3, 0.1\}$, but the collapse persisted across all settings.

These results suggest that, while common SSL features can be effective as *initialization* for robust finetuning (e.g., the DINOv3-initialized ImageNet experiments described in Section 5), they do not serve as useful *regularization targets* in adversarial training. This is likely because these methods lack the denoising-based generative pretraining procedure with noisy images, and therefore do not learn representations that remain semantically stable and informative under input perturbations, which is the critical property that makes diffusion features effective regularization targets in AT. Nevertheless, DRA with diffusion representations incurs no additional pretraining cost, since diffusion models are already required for generating the synthetic data that has become critical in scaling AT. Additionally, combining DRA with strong SSL-initialized backbones also leads to further improvements.

I Applicability of DRA with Foundation Diffusion Models

To examine whether DRA can benefit from an off-the-shelf diffusion teacher, we use Stable Diffusion 1.5 (SD-1.5) [38] as the alignment targe, and evaluate it using WRN-28-10 under the DM-AT setup with one million synthetic images.

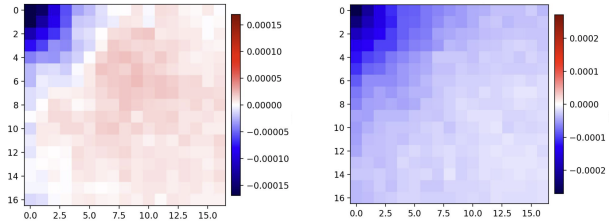
Table 5: Comparison of diffusion alignment targets on CIFAR-10 using WRN-28-10 and one million synthetic images. Accuracy is reported in percent. SD-1.5 is used without target-dataset finetuning.

Method	Clean	AA ($\ell_\infty = 8/255$)	AA ($\ell_2 = 0.5$)
DM-AT	91.12	63.35	66.70
+DRA, SD-1.5	$91.59_{\pm.04}$	$63.60_{\pm.13}$	$68.73_{\pm.07}$
+DRA, EDM	$\mathbf{92.32}_{\pm.12}$	$\mathbf{64.09}_{\pm.04}$	$\mathbf{69.19}_{\pm.05}$

Table 5 shows that SD-1.5 improves clean accuracy and robust accuracy over the baseline DM-AT. These results demonstrate that DRA can benefit from an off-the-shelf diffusion teacher without target-dataset finetuning, and using diffusion model finetuned with the downstream dataset could provide better performance.

J Noisy-Image Discriminative Pretrained Representations

We train a classifier with the same diffusion UNet encoder on the same diffusion noisy ($\sigma_t = 0.1$) inputs, using cross-entropy loss. Its accuracy is comparable to that of a linear probe trained on diffusion representations. As discussed in Section 5.6, using these noisy-image discriminative representations as the alignment target degrades robustness. To investigate the effect, we compute frequency maps on the input gradients and take the difference between DM-AT and DM-AT+DRA. Figure 11 shows that diffusion representations reduce sensitivity to low-frequency components, while alignment to noisy-image discriminative representations increases sensitivity to mid and high-frequency features. Our results complement the findings of Jaini et al. [51], which found that noisy discriminative training could lead to shape-bias but decreased OOD classification performance.



(a) Noisy Discriminative (b) Diffusion Representations

Figure 11: The input gradient frequency difference maps between DM-AT+DRA and DM-AT. (a) DRA aligned to noisy-input discriminative pretrained representations. (b) DRA aligned to the original diffusion representations. Positive values indicate increased sensitivity. Low-frequency components are in the upper-left.

K Ablation Study on the Sensitivity of Timesteps

We examine DRA's sensitivity to the diffusion noise level used for feature extraction. Using WRN-28-10 on CIFAR-10 under the DM-AT setup with one million synthetic images, we vary $\sigma_t \in \{0.05, 0.1, 0.25, 0.5\}$, including the default value $\sigma_t = 0.1$ described in Appendix D.

As shown in Table 6, all tested noise levels improve clean accuracy and robust accuracy over the DM-AT baseline. The results show that the improvement over the baseline is robust around the current default timestep.

Table 6: Sensitivity to the diffusion noise level on CIFAR-10 using WRN-28-10 and one million synthetic images. Accuracy is reported in percent.

Method	Clean	AA ($\ell_\infty = 8/255$)	AA ($\ell_2 = 0.5$)
DM-AT	91.12	63.35	66.70
+DRA ($\sigma_t = 0.05$)	$\mathbf{92.48}_{\pm.07}$	$\mathbf{64.30}_{\pm.21}$	$68.65_{\pm.05}$
+DRA ($\sigma_t = 0.1$, default)	$92.32_{\pm.12}$	$64.09_{\pm.04}$	$\mathbf{69.19}_{\pm.05}$
+DRA ($\sigma_t = 0.25$)	$92.37_{\pm.16}$	$64.07_{\pm.10}$	$68.52_{\pm.13}$
+DRA ($\sigma_t = 0.5$)	$92.11_{\pm.13}$	$63.91_{\pm.06}$	$68.43_{\pm.12}$

L Additional Sweep of Regularization Strength

In addition to Figure 6, we conduct an additional sweep of $\lambda \in \{0.0, 0.6, 1.2, 1.8\}$ for ViT-B/2 on CIFAR-10 with 20M synthetic images. Figure 12 shows that any non-zero λ improves both clean and AutoAttack robust accuracy over the DM-AT baseline, with robust accuracy peaking at λ=1.2. Due to the cost of running adversarial training, we did not sweep λ for every other architecture and synthetic-data scale; the purpose here is to show that a fixed λ=1.2 is sufficient to demonstrate the effectiveness of DRA. In practice, a small sweep of λ may further improve results in a given setting.

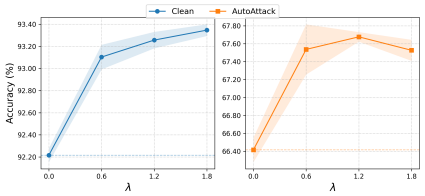


Figure 12: Sweep of the regularization strength λ on CIFAR-10 with 20M synthetic images using ViT-B/2.

M Additional ImageNet Experiments

The ImageNet experiments in Table 3 finetune from a DINOv3 checkpoint to demonstrate the effectiveness of DRA in real-world scenarios when strong pretrained features are available. To examine whether DRA also helps without such an initialization, we additionally train ViT-B/16 from scratch on ImageNet under the same AT recipe. Table 7 reports the results. DRA improves both clean and robust accuracy in both regimes. This indicates that DRA is more effective when there is no strong pretrained feature initialization, while still providing additional benefits when strong pretrained representations are available.

Table 7: ImageNet results for ViT-B/16 trained from scratch and from a DINOv3-initialized checkpoint. AutoAttack is evaluated at $\ell_\infty = 4/255$.

Model	Clean	AA
ViT-B/16 (from scratch)	70.69	48.06
+DRA	**73.77**	**50.24**
ViT-B/16 (DINOv3)	74.67	54.88
+DRA	**76.86**	**55.53**

N LLM Assistance Declaration

LLM assistance was used for manuscript editing and for mathematical checking and refinement of the theoretical analysis in this appendix.