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 (Szegedy et al., 2014; Goodfellow et al., 2015), inputs perturbed by semantically imperceptible noise that can drastically alter model predictions. Among numerous proposed defenses, adversarial training (AT) (Madry et al., 2018; Zhang et al., 2019), which trains models on adversarially perturbed inputs, remains one of the most effective approaches on standard benchmarks such as RobustBench (Croce et al., 2021).

Despite its effectiveness, AT is known to suffer from robust overfitting (Rice et al., 2020), 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 (Wu et al., 2020; Chen et al., 2021; Yu et al., 2022; Wang et al., 2023a,b; Wu et al., 2024). Among them, arguably the most effective approach to date has been the diffusion model with AT (DM-AT) training recipe (Wang et al., 2023b), leveraging large amounts of high-quality synthetic data generated by diffusion models to improve robustness.

The DM-AT approach (Wang et al., 2023b), 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 (Ouyang et al., 2023; Bartoldson et al., 2024; Cui et al., 2024; Wu et al., 2025). 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 (Yang and Wang, 2023; Xiang et al., 2023; Chen et al., 2025; Li et al., 2025c). The extent

[†]Co-advisors.

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

to which these representations can further enhance robustness remains largely unexplored, suggesting a promising opportunity beyond synthetic data generation. In this work, we systematically investigate 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 improving AT.

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, in contrast to conventional reconstruction-based representation learning, which is known to be more adversarially vulnerable due to its reliance on high-frequency signals (Huang et al., 2023).

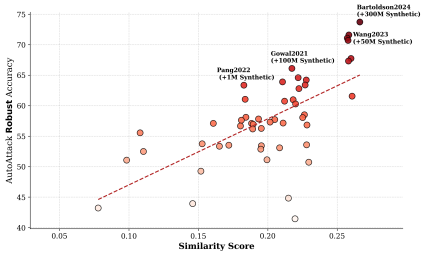


Figure 1: Robust accuracy versus similarity to diffusion representations for CIFAR-10 ℓ_∞-robust models from RobustBench. Details are in Appendix A.

Motivated by these findings, we propose to upgrade the DM-AT recipe by incorporating a simple module that aligns classifier representations with diffusion representations (Li et al., 2023b; Yu et al., 2025; Stracke et al., 2025). 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 (Gorton and Lewis, 2025), comparing the effects of synthetic data in DM-AT versus the diffusion representation alignment 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 dimension (Feng et al., 2022), 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 AT.
- 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 adversarially crafted imperceptible perturbations (Szegedy et al., 2014; Goodfellow et al., 2015). Empirical robustness is commonly assessed with AutoAttack (Croce and Hein, 2020), which is also

the main evaluation protocol for RobustBench (Croce et al., 2021). 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 (Athalye et al., 2018; Gao et al., 2022). Consequently, models competitive on RobustBench rely on adversarial training to achieve robustness. Additionally, certified defenses offer provable guarantees (Cohen et al., 2019; Carlini et al., 2023; Hu et al., 2024; Chen et al., 2024a; Lai et al., 2025), but often introduce significant inference-time cost or underperform adversarially trained models in empirical robustness, especially models that are trained with diffusion synthetic data (Wang et al., 2023b; Bartoldson et al., 2024; Wu et al., 2025). 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 (Ho et al., 2020; Dhariwal and Nichol, 2021; Song et al., 2021; Rombach et al., 2022; Karras et al., 2022; Ma et al., 2024; Yu et al., 2025; Yao et al., 2025), 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 (Xiang et al., 2023; Chen et al., 2025; Li et al., 2025c) and dense predictions (Stracke et al., 2025; Gan et al., 2025). 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 (Yu et al., 2025; Singh et al., 2026).

In this paper, we investigate whether diffusion representations encode informative and robust semantics that can benefit adversarially robust classification. Yagoda et al. (2025) 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 (Athalye et al., 2018), the robust accuracy drops substantially (Appendix D). 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 (Nie et al., 2022; Li et al., 2025b). However, such methods incur substantial inference cost, and their reliance on randomness has been shown to be vulnerable to adaptive attacks (Wang et al., 2024; Chen et al., 2024b).

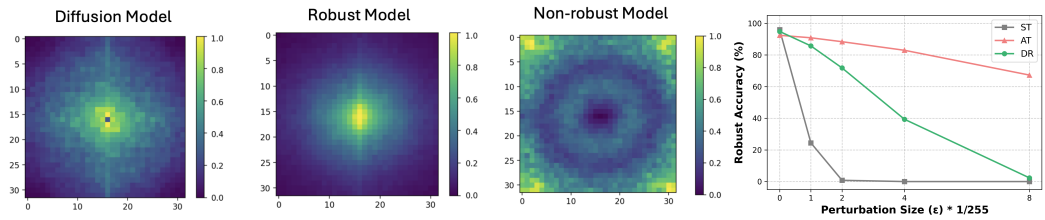
Another direction is to turn off-the-shelf diffusion models into Bayesian generative classifiers (Li et al., 2023a; Clark and Jaini, 2023; Chen et al., 2024b). At a high level, these methods add noise to an input image, then denoise it conditioned on each class, and finally select the class whose reconstruction is most similar to the original input. They exhibit desirable properties such as high error consistency with humans (Jaini et al., 2024), robustness to spurious correlations (Li et al., 2025a) that is free of the need for a balanced dataset (Kirichenko et al., 2023), and adversarial robustness (Chen et al., 2024b). These approaches can also be integrated with randomized smoothing (Cohen et al., 2019) to provide certified defenses (Chen et al., 2024a). 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 (Li et al., 2023a; Chen et al., 2024b). 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 (Madry et al., 2018) 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



(a) (b)

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

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 (Zhang et al., 2019) is often used to balance the performance on adversarial and clean images (Wang et al., 2023b; Bartoldson et al., 2024; Cui et al., 2024; Wu et al., 2025).

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 $\mathbf{x}_t$, 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 (Xiang et al., 2023), whereas for newer transformer-based diffusion models (Peebles and Xie, 2023), representations are extracted near the middle layers (Xiang et al., 2023; Chen et al., 2025). 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 within the class-relevant subspace at those timesteps (Li et al., 2025c). In this work, we extract the representations from timesteps based on linearly probed clean accuracy experiments following prior work (Xiang et al., 2023). Implementation details are provided in Appendix E.

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 (Yu et al., 2025; Balestriero and LeCun, 2024; Huang et al., 2023). 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 (Wang and Isola, 2020). Uniformity measures how evenly representations are distributed on the unit hypersphere, reflecting the information preserved in the representation space, whereas alignment measures

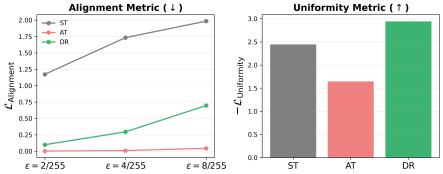


Figure 3: Representation metrics on CIFAR-10 for the standard-trained model (ST), the adversarially trained model (AT), and the diffusion representations (DR).

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.

Frequency and Robustness Behavior. Vision models are often sensitive to high-frequency input perturbations, while adversarial training typically reduces this sensitivity and places more emphasis on low-frequency components (Chan et al., 2022). 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 (Huang et al., 2023). 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 (Yagoda et al., 2025). 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 (Athalye et al., 2018), 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 (Yang and Wang, 2023; Yu et al., 2025; Stracke et al., 2025). Figure 4 illustrates the overall framework, where we leverage an auxiliary projection head to regularize the classifier representations.

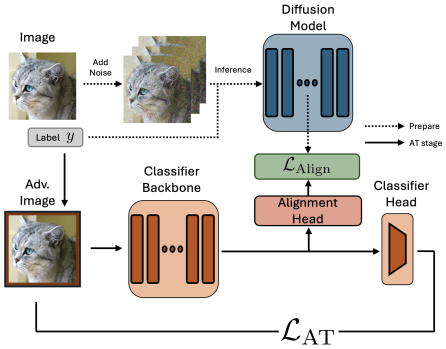


Figure 4: Diffusion Representation Alignment (DRA). We leverage an auxiliary projection head to align classifiers with the extracted diffusion representations.

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 (Section 3), 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 (Yu et al., 2025; Stracke et al., 2025). For each image, we use a fixed timestep around the optimal linearly probed timestep (Xiang et al., 2023) to extract diffusion representations. Details on extracting diffusion representations, projection head, and timestep choices are in Appendix E.

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 Setup

We mainly follow the DM-AT (Wang et al., 2023b) setup, which is the state-of-the-art adversarial training framework that is also used by Bartoldson et al. (2024); Cui et al. (2024); Wu et al. (2025). In the following, we briefly explain the setup for each dataset. Implementation details are in Appendix C.

CIFAR-10/100. For CIFAR-10/100 (Krizhevsky, 2009), 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 (Zhang et al., 2019) 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. (2023b), with DRA using the same frozen CIFAR-10 EDM diffusion model checkpoint (Karras et al., 2022) 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. (2023b) are also provided. For model choices, we conduct experiments on the WideResNets (Zagoruyko and Komodakis, 2016) WRN-28-10 and WRN-34-10, which are widely used in the adversarial training literature, as well as ViT-B/2 (Dosovitskiy et al., 2021) following Wu et al. (2025).

ImageNet. Additionally, to demonstrate the effectiveness in real-world scenarios, we conduct experiments on ImageNet (Russakovsky et al., 2015), 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 (Liu et al., 2022), initialized from the DINOv3 (Siméoni et al., 2025) pretrained checkpoints released via `timm` (Wightman, 2019). For the diffusion synthetic data, we generate 4 million 256×256 synthetic images using LightningDiT (Yao et al., 2025).

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 in Appendix N.

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 (Wu et al., 2024)	WRN-34-10	-	200	86.11	79.90	55.26	61.83
+IKL (Cui et al., 2024)	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 (Wang et al., 2023b)	WRN-28-10	1M	400	91.12	86.62	63.35	66.70
DM-AT (Wang et al., 2023b)	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 (Wang et al., 2023b)	WRN-28-10	20M	2400	92.44	88.38	67.31	70.25
+IKL (Cui et al., 2024)	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 (Wang et al., 2023b)	WRN-70-16	50M	2000	93.25	89.71	70.69	69.78
+RA (Peng et al., 2023)	RaWRN-70-16	50M	2000	93.27	89.71	71.07	68.71
DM-AT (Wu et al., 2025)	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

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 6 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 (Elhage et al., 2022). While this can be effective for compressing features that rarely co-activate, it has been suggested that superposition may be exploited by adversarial examples (Gorton and Lewis, 2025; Stevinson et al., 2025).

Moving from toy settings to real-world datasets, where the degree of superposition is difficult to quantify, Gorton and Lewis (2025) use sparse autoencoders (SAEs) (Cunningham et al., 2024) 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 (Gao et al., 2025; Gorton and Lewis, 2025).

Figure 7 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 (2025), who report that adversarial training encourages more easily disentangled 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. (2022) and examine whether the use of representation dimensions changes when diffusion synthetic data and diffusion representations are introduced. Specifically, Feng et al. (2022) propose *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 11 (Appendix H) 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.

Table 4: Classification and robust dimensions with WRN-28-10s on CIFAR-10.

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 (Hu et al., 2024). Prior work has also explored using partially synthesized diffusion data as an augmentation strategy to improve robustness (Sastry et al., 2024), 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 (Wang et al., 2023b; Bartoldson et al., 2024), 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 5 shows that DRA consistently improves upon the baseline in both clean and robust accuracy. When λ becomes too large, robust-

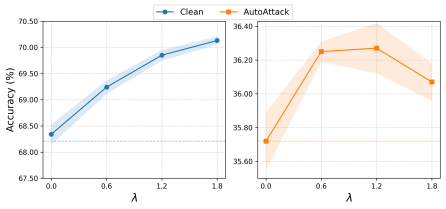


Figure 5: Clean and Robust accuracy on CIFAR-100 for DM-AT+DRA with different DRA regularization strength λ.

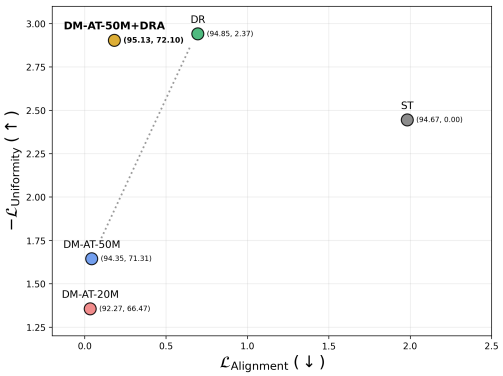


Figure 6: We plot the alignment and uniformity metrics, along with clean and robust accuracy on CIFAR-10 ($\ell_\infty = 8/255$) shown in parentheses.

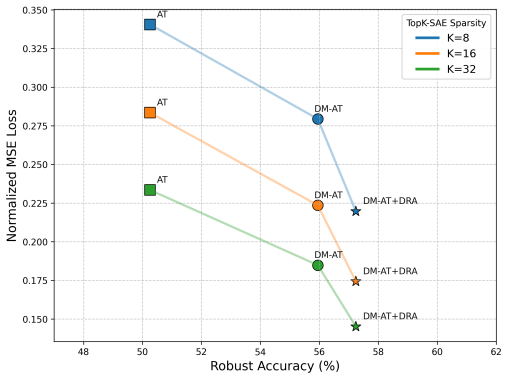


Figure 7: Normalized reconstruction losses of TopK SAEs trained on ImageNet using AT, DM-AT, and DM-AT+DRA ViT-B checkpoints.

ness improvements saturate, as the alignment term can outweigh the original adversarial training objective.

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. (2024) find 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 K).

Why Not Adversarially Finetune the Diffusion Encoder Directly? Several works in self-supervised learning that leverage diffusion representations finetune the diffusion encoder with a prediction head on top (Xiang et al., 2023; Yagoda et al., 2025). 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 depending on the available diffusion model (Xiang et al., 2023; Chen et al., 2025). In contrast, DRA better leverages the robust knowledge encoded in diffusion representations while enabling more flexible downstream classifier choices for the classification task.

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

6 Conclusions

We systematically explore whether diffusion models can improve adversarial training beyond synthetic data generation alone. We find that diffusion representations are partially robust yet diverse, and show that aligning classifiers with them as an auxiliary learning signal consistently improves AT across settings and datasets. Our analysis further shows that incorporating diffusion models encourages representations that are easier to disentangle. We hope these findings inspire further uses of diffusion models in AT beyond generating high-quality synthetic images.

Acknowledgement

We thank the anonymous reviewers and members of CLLab and AISLab for their constructive feedback. This work was supported in part by the National Science and Technology Council in Taiwan under Grants NSTC 113-2222-E-002-004-MY3, 114-2634-F-002-004, 115-2628-E-002-029, 1132634-F-002-008, 114-2221-E-002-102-MY3, NTU AI Center of Research Excellence within Taiwan Centers of Excellence, and NTU Center for Data Intelligence via NTU-114L900901. We thank to National Center for High-performance Computing (NCHC) of National Institutes of Applied Research (NIAR) in Taiwan for providing computational and storage resources. H.-T. Lin is honored to be supported by the Leap Fellowship of the Foundation for the Advancement of Outstanding Scholarship in Taiwan since 2025.

References

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.

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

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.

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

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.

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)*, 2024a.

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)*, 2024b.

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.

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.

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.

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

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

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.

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.

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.

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

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

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.

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.

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.

Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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.

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.

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.

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

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

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.

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

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

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.

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

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

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)*, 2022.

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

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.

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

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.

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)*, 2023a.

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

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)*, 2023b.

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)*, 2025b.

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)*, 2025c.

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.

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.

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.

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.

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

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

Sheng Yun 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.

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.

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

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.

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.

Chandramouli S. Sastry, Sri Harsha Dumpala, and Sageev Oore. DiffAug: A diffuse-and-denoise augmentation for training robust classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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. *arXiv preprint arXiv:2508.10104*, 2025.

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.

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.

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

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.

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.

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.

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

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)*, 2023a.

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)*, 2023b.

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

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

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.

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.

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.

Mika Yagoda, Shady Abu-Hussein, and Raja Giryes. Diffusion models are robust pretrainers. *IEEE Signal Processing Letters*, 2025.

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

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.

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.

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.

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

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.

Appendix

A Representation Similarity Analysis **16**

B Motivational Frequency Analysis Implementation and Additional Results **16**

C Experiment Setup Details **17**

D Additional Discussion for Related Work **17**

E Representation Alignment Implementation Details **17**

F Computational Cost of DRA **18**

G Additional Analysis on SAE Features **18**

H Classification Dimension Visualization **19**

I SSL Features as Surrogate Representations in DRA **19**

J Applicability of DRA with Foundation Diffusion Models **20**

K Noisy-Image Discriminative Pretrained Representations **20**

L Ablation Study on the Sensitivity of Timesteps **20**

M Additional Sweep of Regularization Strength **21**

N Additional ImageNet Experiments **21**

O Limitations **21**

P Broader Impacts **21**

Q Reproducibility **22**

R Licenses for Existing Assets and LLM Usage Declaration **22**

A Representation Similarity Analysis

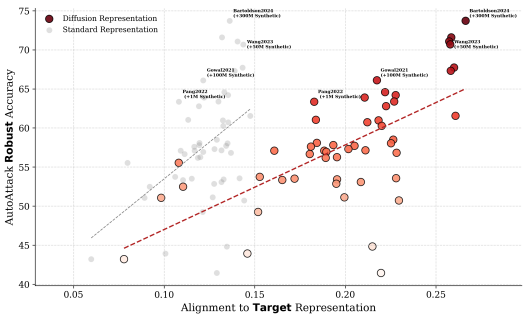


Figure 8: Similarity scores are measured with respect to representations extracted from the diffusion model (red) and from the RobustBench standard-trained model (gray).

As a motivational experiment, following the work of Huh et al. (2024) on analyzing representation similarities between models trained with different learning objectives, we extract the representations from the CIFAR-10 EDM diffusion model (Karras et al., 2022) used in DM-AT (Wang et al., 2023b), and measure the representation similarity score CKNNA (Huh et al., 2024) with the CIFAR-10 ℓ_∞-robust models from RobustBench (Croce et al., 2021). 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 8). The observed correlation motivates us to investigate whether explicitly leveraging diffusion representations as an auxiliary learning signal is beneficial for robust classifier training.

B Motivational Frequency Analysis Implementation and Additional Results

Implementation. We perform the frequency analysis on ℓ_∞ PGD-20 input perturbations ($\epsilon = 8/255$ on CIFAR-10 and $4/255$ on ImageNet). Following prior frequency-saliency analyses (Huang et al., 2023; Chan et al., 2022), we compute the magnitude of the 2-D discrete Fourier transform of each perturbation, average it over images and channels, shift the low frequencies to the center, and normalize each spectrum to $[0, 1]$ so that models can be compared.

In addition to Figure 2a, Figure 9 reports the same analysis on ImageNet for a linear probe on frozen LightningDiT representations, a standard-trained non-robust ViT-B/16, and an adversarially trained robust ViT-B/16. The standard-trained model exhibits less high-frequency reliance than its CIFAR-10 counterpart, but it still shows the strongest mid-frequency reliance of the three. Note that the diffusion model in this setting is a latent diffusion model, so the perturbation is also affected by the VA-VAE encoder, which may contribute to the frequency reliance in this setting.

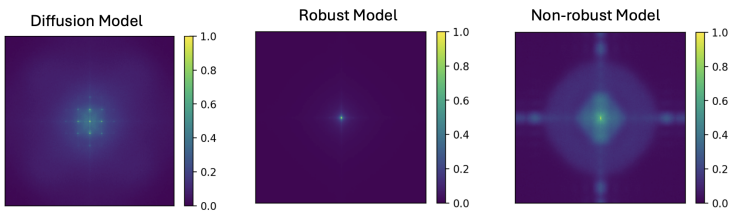


Figure 9: The frequency saliency analysis of the linearly probed diffusion representation, adversarially trained robust model, and standard non-robust model. Low frequencies are centered.

C Experiment Setup Details

Training Setup. For CIFAR-10/100 WideResNet training, we follow the exact same setup as DM-AT (Wang et al., 2023b): PGD-10 TRADES (Zhang et al., 2019) 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/2, we follow Wu et al. (2025) and use the Lion optimizer (Chen et al., 2023) 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 evaluate them with the released checkpoints. For our experiments, statistics over three runs are reported.

For the SAE experiments, we follow prior works (Gorton and Lewis, 2025; Gao et al., 2025) 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 (Kingma and Ba, 2015) for a total of 50000 steps.

Computational Resources. The experiments are mostly conducted on machines with 4 GPUs (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.

D Additional Discussion for Related Work

Frozen Diffusion Features for Robustness. Yagoda et al. (2025) 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 (Athalye et al., 2018). For example, the configuration of Attention Head with hyperparameters b=8 and 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 (Stracke et al., 2025) leverages self-distillation across timesteps to produce clean, timestep-independent features for dense prediction tasks; DreamTeacher (Li et al., 2023b) and RepFusion (Yang and Wang, 2023) distill diffusion priors into discriminative visual models for downstream discriminative tasks; REPA (Yu et al., 2025) 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.

E Representation Alignment Implementation Details

Diffusion Representation Extraction. For CIFAR-10/100, we use the same EDM diffusion model checkpoint as Wang et al. (2023b) 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. (2025). 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 (Fang et al., 2025; Gan et al., 2025), which reduces their effectiveness as a learning signal. To mitigate this, we apply the method introduced in Gan et al. (2025), 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.

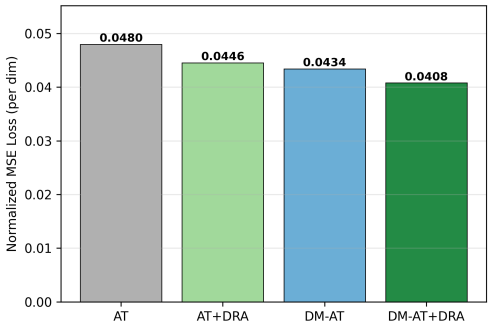
Lastly, for CIFAR-10 models trained with more than one million synthetic images, we extract an additional representation per image by sampling an extra timestep using the same sampling function the EDM model originally used during training, and take the average of the cosine similarity loss. This results in a slight improvement, likely by better leveraging the representations across timesteps (Stracke et al., 2025; Li et al., 2025c).

Alignment Module. We follow Yu et al. (2025) 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 (Yu et al., 2025; Stracke et al., 2025).

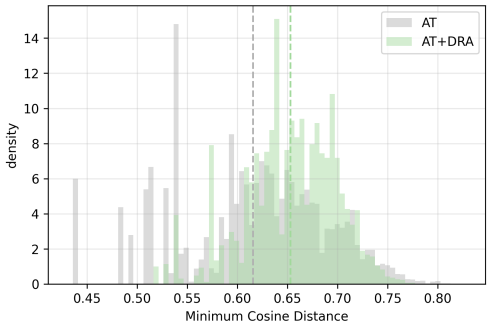
F Computational Cost of DRA

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.

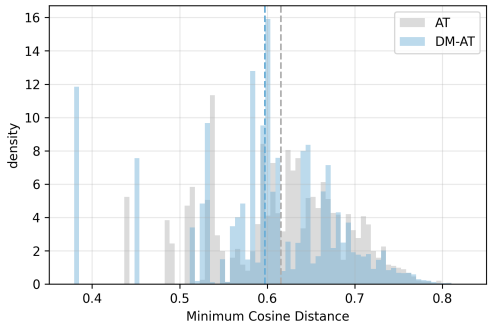
G Additional Analysis on SAE Features



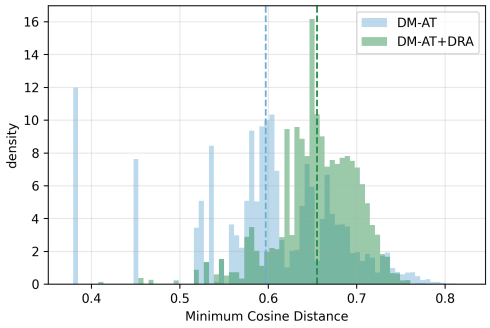
(a) Normalized TopK-SAE reconstruction loss.



(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.

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.

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.

H Classification Dimension Visualization

Figure 11 illustrates the classification dimension analysis of Section 5.5 on a CIFAR-10 adversarially trained model. Clean accuracy gradually recovers as more principal components are included, whereas robust accuracy first improves and then degrades as K increases.

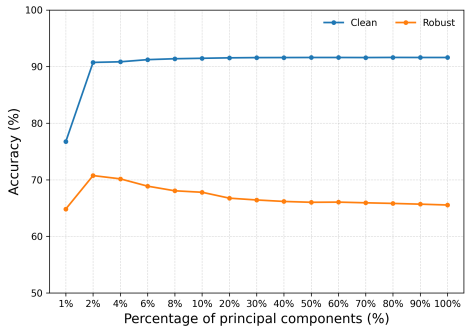


Figure 11: Clean and robust accuracy evaluated with a CIFAR-10 robust WRN-28-10 model when projecting features from natural or adversarial images onto the top-K principal components before the classification head.

I 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 (Siméoni et al., 2025), MAE (He et al., 2022), and CLIP (Radford et al., 2021) 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.

J 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) (Rombach et al., 2022) as the alignment target, and evaluate it using WRN-28-10 under the DM-AT setup with one million synthetic images.

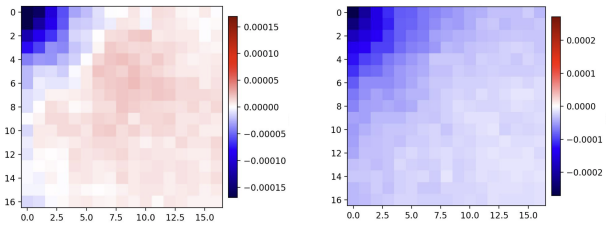
Table 5: Comparison of diffusion alignment targets on CIFAR-10 using WRN-28-10 and one million synthetic images. Accuracy is reported in percent. SD-1.5 is used without target-dataset finetuning.

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 a diffusion model trained on the downstream dataset could provide better performance.

K 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 12 shows that DRA with diffusion representations reduces the overall gradient magnitude, most strongly at low frequencies where sensitivity is concentrated, while alignment to noisy-image discriminative representations increases sensitivity to mid and high-frequency features. Our results complement the findings of Jaini et al. (2024), who find that noisy discriminative training could lead to shape bias but decreased OOD classification performance.



(a) Noisy Discriminative (b) Diffusion Representations

Figure 12: 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.

L 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 E.

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 ($\sigma_t = 0.5$)	$92.11_{\pm.13}$	$63.91_{\pm.06}$	$68.43_{\pm.12}$

M Additional Sweep of Regularization Strength

In addition to Figure 5, 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 13 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.

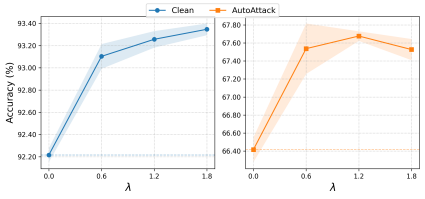


Figure 13: Sweep of the regularization strength λ on CIFAR-10 with 20M synthetic images using ViT-B/2.

N 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. 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**

O Limitations

The contribution of our work focuses on a new perspective on diffusion models through representation usage in adversarial training, viewing them as a source of diverse yet partially robust feature priors in addition to synthetic data generators. Hence, the broader empirical questions of which properties of a teacher representation make it suitable as a regularization target in adversarial training, and also what other methods could more effectively extract robust knowledge out of diffusion models to build robust discriminative classifiers, are promising directions that require future investigation. Additionally, our experiments focus on the standard ℓ_∞ adversarial robustness for image classification on CIFAR-10/100 and ImageNet; extending DRA to other modalities, tasks, or threat models for robust visual recognition is a natural direction for future work.

P Broader Impacts

Adversarial robustness contributes to safer deployment of machine learning in security-sensitive applications such as content moderation, biometric verification, and medical imaging, and our work makes diffusion-based adversarial training recipes more effective by reusing diffusion features as an

auxiliary learning signal, without additional data or substantial extra compute. On the negative side, adversarial training is computationally expensive and contributes to the energy footprint of model development; robustness measured on standard benchmarks (e.g., AutoAttack on ℓ_∞ perturbations) does not guarantee robustness under threat models that are not included in the test scenarios and may give a false sense of security.

Q Reproducibility

To ensure the reproducibility of our experiments, we have provided detailed implementation of the proposed method. Additionally, the code implementation is included in the supplementary material, enabling readers to replicate the experiments.

R Licenses for Existing Assets and LLM Usage Declaration

Table 8 lists the existing assets used in this work, including datasets, pretrained models, and code packages, together with their licenses. We use all assets in accordance with their respective licenses for non-commercial academic research.

LLM assistance. LLM tools were used for assisting code implementations and manuscript editing.

Table 8: Existing assets used in this work and their licenses.

Asset	Type	Reference	License
CIFAR-10 / CIFAR-100	Dataset	Krizhevsky (2009)	Custom (research use)
ImageNet (ILSVRC-2012)	Dataset	Russakovsky et al. (2015)	Custom (research use)
DM-AT	Dataset / Code	Wang et al. (2023b)	MIT
EDM diffusion model	Model / Code	Karras et al. (2022)	CC BY-NC-SA 4.0
LightningDiT	Model / Code	Yao et al. (2025)	MIT
DINOv3 pretrained checkpoints	Model	Siméoni et al. (2025)	DINOv3 License
`timm` library	Model / Code	Wightman (2019)	Apache 2.0
RobustBench	Model / Code	Croce et al. (2021)	MIT
AutoAttack	Code	Croce and Hein (2020)	MIT

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work in the Appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: The paper does not contain theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Implementation details are provided in the appendix, covering computational resources, experimental settings, hyperparameters. The code implementation is included in the supplementary material.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The code implementation is provided in the supplementary material, with sufficient instructions to faithfully reproduce the main experimental results.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: The training and test details, including hyperparameters and settings, are provided with sufficient instructions in the appendix. We also provide the code implementation in the supplementary material.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The mean and standard deviation over three runs are reported for the main experiments in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources needed to reproduce the experiments in the appendix.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed in the Broader Impacts section of the appendix.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets, pretrained models, and code packages used in this work are publicly available existing assets. We cite the original papers and use all assets in accordance with their respective licenses and terms of use. A complete list of assets, with their sources and licenses, is provided in Appendix R.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLM tools were used for assisting code implementations and manuscript editing.

    Guidelines:

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