Expanding the Role of Diffusion Models for Robust Classifier Training

Pin-Han Huang [1] **Shang-Tse Chen** [1 †] **Hsuan-Tien Lin** [1 †]

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 adversarially perturbs the input images during training, remains one of the most effective approaches for achieving robustness on standard benchmarks such as RobustBench (Croce et al., 2021).

Previous work has shown that AT suffers from robust overfitting (Rice et al., 2020), where robustness on test set degrades during training despite stable accuracy on clean images and decreasing training loss. 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), which leverages large amounts of high-quality synthetic data generated by diffusion models.

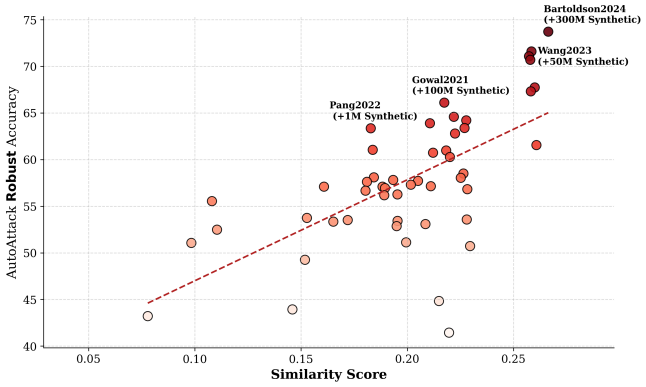


Figure 1. We plot robust accuracy and representation similarity scores (Huh et al., 2024) for CIFAR-10 ℓ_∞-robust models from RobustBench (Croce et al., 2021). Similarity scores are measured with respect to representations extracted from the diffusion model. Implementation details and discussion are in Appendix A.

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 to improve AT for robust classifier training. More broadly, most of the existing efforts to improve AT have centered on this synthetic data paradigm (Wang et al., 2023b; Ouyang et al., 2023; Bartoldson et al., 2024; Cui et al., 2024; Wu et al., 2025). However, it is known that diffusion models can produce meaningful intermediate representations (Yang & Wang, 2023; Xiang et al., 2023; Chen et al., 2025; Li et al., 2025c). Whether these representations can be additionally leveraged on top of DM-AT to improve robustness remains largely unexplored, presenting a promising opportunity beyond synthetic data generation.

In this work, we systematically investigate whether representations produced by diffusion models can enhance robust classifier training. We hypothesize that the denoising objective of diffusion models enables them to capture robust semantic features from partially corrupted images, which potentially facilitate the training of robust classifiers. Specifically, we examine whether the noisy-input intermediate activations from diffusion models, recently shown to pro-

†Co-advisors. [1]National Taiwan University. Correspondence to: Shang-Tse Chen <stchen@csie.ntu.edu.tw>, Hsuan-Tien Lin <htlin@csie.ntu.edu.tw>.

vide competitive discriminative representations (Xiang et al., 2023; Yang & Wang, 2023; Li et al., 2025c), serve as effective feature priors for improving robust classifier training.

We start our investigation with a preliminary analysis that reveals a weak correlation between robustness and the alignment with diffusion representations using such activations (Figure 1). Moreover, we observe that the extracted diffusion representations exhibit several desirable properties, encoding diverse, lower-frequency-dependent information and are less sensitive to irrelevant high-frequency noise. These characteristics suggest that diffusion representations have significant potential, in contrast to typical reconstruction-based representation learning, which is known to be more vulnerable to adversarial perturbations due to its reliance on high-frequency signals (Huang et al., 2023).

Motivated by these findings, we propose to modify 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 leverages diffusion representations as an auxiliary learning signal while enabling flexbile choices of classifier architectures for robust classification. Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet across multiple architectures and diffusion-based synthetic data settings demonstrate consistent improvements, effectively exploiting the robust semantics encoded in diffusion representations to enhance robust classifier training.

Building on recent mechanistic interpretability work (Gorton & Lewis, 2025), we further deepened our analysis by leveraging diffusion-generated synthetic data alongside diffusion representation alignment. This approach reveals that both interventions facilitate the learning of more easily disentangled representations, yet they achieve this effect through distinct underlying mechanisms.

Specifically, our analysis, guided by classification-aware dimensions (Feng et al., 2022), reveals that diffusion-generated synthetic data 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 diffusion representations and synthetic data provide complementary benefits for robust classifier training, and that combining both further enhances robustness and generalization.

Our contributions are summarized as follows:

- We show that diffusion representations encode features that are partially robust and diverse, and leveraging diffusion representations as an auxiliary learning signal improves adversarial training.
- We find that the incorporation of diffusion models encourages 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. Empirical robustness is commonly assessed with AutoAttack (Croce & 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 competitve 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 & 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., 2025).

In this paper, we investigate whether diffusion representations encode informative and robust semantics that can benefit adversarially robust classification. Yagoda et al. (2025) proposes to train prediction heads on frozen unconditional diffusion representations as a lightweight robustness 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 C). Conversely, we show that using diffusion representations as an auxiliary learning signal can further strengthen adversarial training, and we analyze how this integration shapes the learned classifier 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 & 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 imbalanced datasets (Li et al., 2025a) that is free of the need to retrain prediction heads on 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 $\mathcal{S}(\mathbf{x}) = \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_p \le \varepsilon\}$ is the ℓ_p-ball of radius ε centered at $\mathbf{x}$. During training, projected gradient descent (PGD) is used to approximately solve the inner maximization by iteratively updating the adversarial example. Considering the standard ℓ_∞ adversary, the adversarial example is obtained by

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

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.

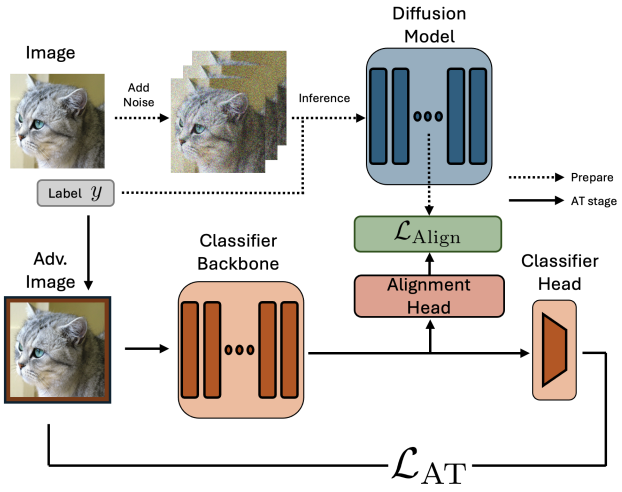
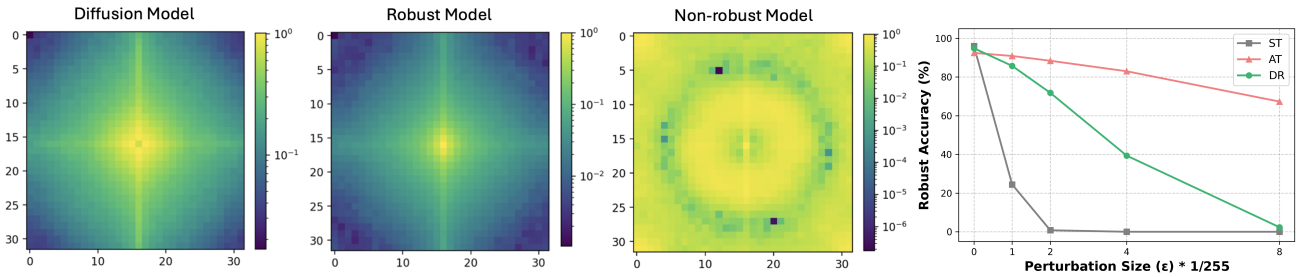


Figure 2. Overview of Diffusion Representation Alignment (DRA). We leverage an auxiliary projection head to align classifiers with the extracted diffusion representations.

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 & 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 at those timesteps (Li et al., 2025c). In this work, we follow Xiang et al. (2023); Li et al. (2025c) to extract the diffusion representations near the optimal timesteps.

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 & LeCun, 2024; Huang et al., 2023). In



(a) (b)

Figure 3. (a) The frequency saliency analysis of the linear-probed diffusion representation, adversarial trained robust model, and standard trained non-robust model. Low frequencies are being centered. (b) The CIFAR-10 robust accuracy across perturbation budgets for the linear probed diffusion representation (DR), adversarial trained robust model (AT), and standard trained non-robust model (ST).

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 is 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 & Isola, 2020). Uniformity measures how evenly representations are distributed on the unit hypersphere, reflecting the information preserved in the representation space, whereas alignment measures the distance between the representations of data examples with different positive views. In our setting, we form positive pairs by constructing adversarial images. As shown in Figure 4, diffusion representations are more robust than standard supervised training, while achieving richer features with noticeably higher uniformity metric. In contrast, adversarial training increases alignment but decreases feature diversity, reflecting the difficulty of learning robust model with great feature quality. 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 emphasizes more 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 linear-probed diffusion representations. In Figure 3a, it shows that diffusion representations exhibit lower high-frequency saliency that resembles robust models. It also suggest that diffusion representations does not suffer 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 fine-tuning (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 3b 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 & Wang, 2023; Yu et al., 2025; Stracke et al., 2025). Figure 2 illustrates the overall framework, where we leverage an auxiliary projection head to regularize the classifier representations.

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

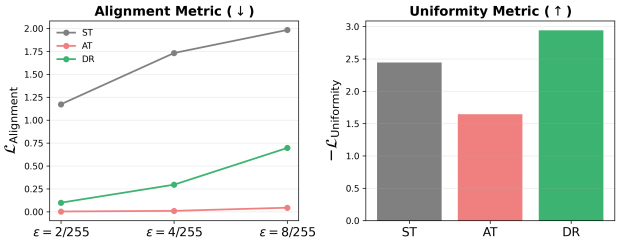


Figure 4. Alignment and uniformity metrics on CIFAR-10 for the standard-trained model (ST), the adversarially trained model (AT), and the diffusion representations (DR).

ment 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). The regularization coefficient is set to $\lambda = 1.2$ in our experiments. For each image, we use a fixed timestep around the optimal linearly probed timestep (Xiang et al., 2023) to extract diffusion representations. More details and experiments on extracting diffusion representations, projection head, and timestep choices are provided in Appendix D.

5. Experiments

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

5.1. Setups

We mainly follow the DM-AT (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. More implementation details are in Appendix B.

CIFAR-10/100. For CIFAR-10/100 (Krizhevsky, 2009), we follow the DM-AT recipe and perform AT with perturbation budget $\varepsilon = 8/255$ and step size $\alpha = 2/255$, using 10 PGD steps to adversarially augment the training images. TRADES loss (Zhang et al., 2019) is used as the adversarial training objective. To demonstrate the proposed method effectiveness 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 Diffusion Representation (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 EDM model released by Wang et al. (2023a) is also provided. For model choices, we conduct experiments on the WideResNets (Zagoruyko & 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).

Table 1. Clean and Robust Accuracy incorporating Diffusion Representation Alignment 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	**93.14**	**67.83**
	ViT-B/2	DM-AT	94.35	71.31
		DM-AT + DRA	**95.22**	**71.77**
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	74.49	54.44
		DM-AT + DRA	**76.03**	**56.07**
	ViT-B/16	DM-AT	74.62	54.64
		DM-AT + DRA	**76.87**	**55.16**

ImageNet. Additionally, we conduct experiments on ImageNet (Russakovsky et al., 2015), with models initialized from strong pretrained checkpoints to demonstrate the effectiveness in real-world scenarios. 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 recipe, which employs diffusion synthetic data. The results indicate that diffusion representations provide an effective feature prior for robust learning, leading to consis-

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

Method	Architecture	Synthetic	Batch	Epoch	Clean	AA
AT	WRN-34-10	-	128	200	84.33	55.25
AT+ADR (Wu et al., 2024)	WRN-34-10	-	128	200	86.11	55.26
AT+IKL (Cui et al., 2024)	WRN-34-10	-	128	200	84.80	57.09
AT+DRA (Ours)	WRN-34-10	-	128	200	**88.54**	**57.29**
DM-AT (Wang et al., 2023b)	WRN-28-10	1M	512	400	91.12	63.35
DM-AT (Wang et al., 2023b)	WRN-28-10	1M	1024	800	91.43	63.96
DM-AT+DRA (Ours)	WRN-28-10	1M	512	400	**92.36**	**64.12**
DM-AT (Wang et al., 2023b)	WRN-28-10	20M	2048	2400	92.44	67.31
DM-AT+IKL (Cui et al., 2024)	WRN-28-10	20M	2048	2400	92.16	67.75
DM-AT+DRA (Ours)	WRN-28-10	20M	2048	2400	**93.14**	**67.83**
DM-AT	ViT-B/2	20M	1024	500	92.27	66.47
DM-AT+DRA (Ours)	ViT-B/2	20M	1024	500	**93.36**	**67.74**
DM-AT (Wang et al., 2023b)	WRN-70-16	50M	1024	2000	93.25	70.69
DM-AT+RA (Peng et al., 2023)	RaWRN-70-16	50M	1024	2000	93.27	71.07
DM-AT (Wu et al., 2025)	ViT-B/2	50M	1024	2000	94.35	71.31
DM-AT+DRA (Ours)	ViT-B/2	50M	1024	2000	**95.22**	**71.77**

tent gains in both clean accuracy and adversarial robustness.

Moreover, the experiments on ImageNet also show the effectiveness of leveraging diffusion representations when strong self-supervised pre-trained vision foundation models are available to be robust fintetuned to the downstream dataset.

Lastly, we evaluate CIFAR-10 with varying amounts of synthetic data (Table 2), ranging from no synthetic images to 50 million. While scaling adversarial training with diffusion synthetic data remains important for improving robustness, we show that incorporating diffusion representations as an auxiliary learning signal can more effectively leverage the robust knowledge encoded in diffusion models.

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 5 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 Model Encourages to Learn Representations that are Easier to Disentangle

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 & Lewis, 2025; Stevinson et al., 2025).

Moving from toy settings to real world datasets, where the degree of superposition is difficult to quantify, Gorton & Lewis (2025) uses Sparse AutoEncoders (SAEs) (Huben et al., 2024; Gao et al., 2025) 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 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 & Lewis, 2025).

Figure 6 presents the results. We find that incorporating diffusion synthetic data and diffusion representation alignment improves robustness while also reducing the normalized TopK-SAE reconstruction loss, suggesting that the learned representations become easier to decompose into sparse features. This complements the observations of Gorton & Lewis (2025), which report that adversarial training encourages more disentanglable representations, with larger perturbation budgets further amplifying this effect. Our results show that incorporating diffusion models into adversarial training can improve robustness and also encourages to learn representations that are easier to disentangle.

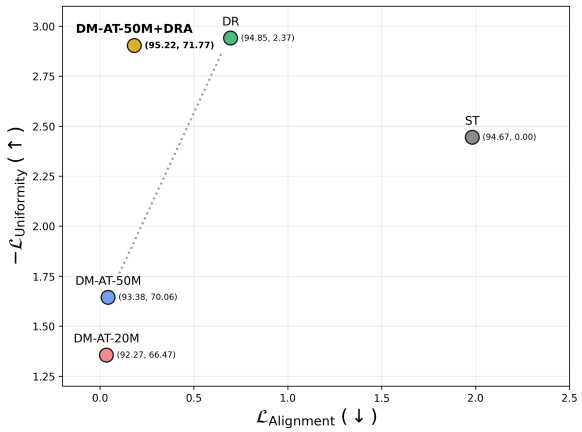


Figure 5. We plot the alignment and uniformity metrics, along with clean and robust accuracy on CIFAR-10 ($\ell_\infty = 8/255$) shown in parentheses.

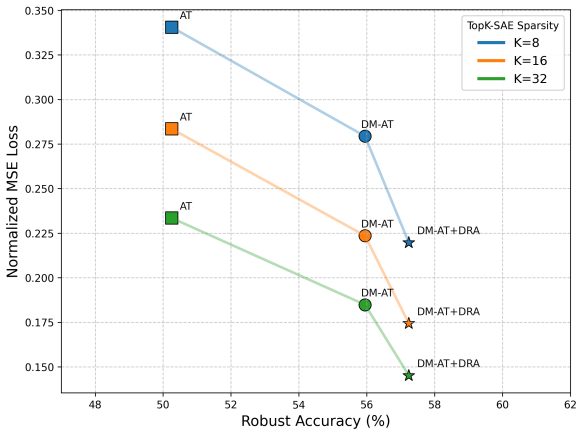


Figure 6. We train Top-K SAEs on ImageNet with sparsity $K \in \{8, 16, 32\}$, using ViT-B checkpoints trained with AT, DM-AT, and DM-AT+DRA. Normalized SAE reconstruction losses are reported for comparison (Gao et al., 2025).

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-based representations are introduced. Specifically, Feng et al. (2022) proposed *classification dimension* as a measure of the intrinsic dimensionality of the feature space, approximated by the minimum number of principal components needed to preserve high classification accuracy.

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

Figure 7 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 3 presents the results for CIFAR-10 adversarially trained models with the number of principal components required to recover the clean accuracy performance, and the robust-aware dimension that achieves the highest robust accuracy. The results indicate that diffusion representation alignment encourages the model to more effectively leverage representation dimensions to encode robust features, whereas diffusion synthetic data leads to lower-rank representations.

While it is intuitive that representation alignment promotes diverse and robust feature encoding, the mechanism behind the lower-rank effect of diffusion 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. As described in Section 4.2, we set the alignment regularization strength to $\lambda = 1.2$ in all experiments. We select this value based on a sweep using WRN-28-10 on CIFAR-100, and then fix the same coefficient for all remaining settings. Figure 8 shows that DRA consistently improves upon the baseline in both clean and robust accuracy. When λ becomes too large, robustness improvements saturate, as the alignment term can outweigh the original adversarial training objective.

Does noisy input training alone learn good enough features? Given that representations extracted from diffusion

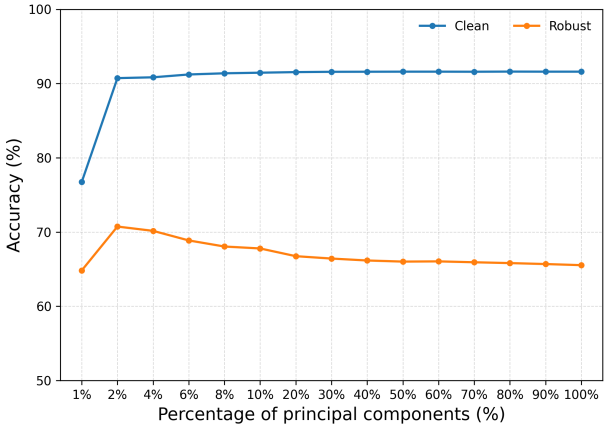


Figure 7. Clean and robust accuracy evaluated with a CIFAR-10 robust WRN-28-10 model when projecting features from natural or adversarial images onto the top-K principal components before the classification head.

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 2. Results show that replacing the diffusion target in a WRN-28-10 trained with DM-AT+DRA (1M synthetic) reduces robust accuracy from 64.12% to 62.62%, which is inferior to the vanilla DM-AT baseline. This indicates that noisy-input training alone are 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) found that diffusion-like noisy discriminative pre-training increases shape bias but hurts OOD accuracy. As a complementary finding, we analyze the frequency behavior of leveraging diffusion representations or the noisy-input pretrained discriminative features, and found that the latter have a undesirable effect of increasing the sensitivity to mid-high frequency components (Appendix E).

Why not adversarially finetune the diffusion encoder directly? Several works in self-supervised learning that leverage diffusion representations finetune the diffusion encoder end-to-end with a prediction head on top (Xiang et al., 2023; Li et al., 2025c; Yagoda et al., 2025). We evaluated 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 diffusion encoder at inference time can only provide unconditional representations, which can result in slight decrease in representation quality (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 that are better suited to the classification task.

Table 3. We evaluate *Classification Dimension*, with CLS-95 and CLS-99 refering to the number of components required to recover 95% and 99% of the original classification performance, and Robust-aware dimension, the number of principal components where robust accuracy peaks, across different training methods on CIFAR-10 with WRN-28-10.

Method	CLS-95 Dim	CLS-99 Dim	Robust Dim
AT	9	14	9
AT+DRA	15	42	22
DM-AT	10	11	11
DM-AT+DRA	12	15	23

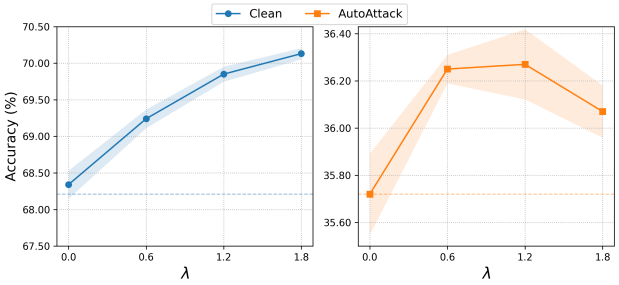


Figure 8. Clean and Robust accuracy on CIFAR-100 for DM-AT+DRA with different DRA regularization strength λ.

6. Conclusions

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

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

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

Bartoldson, B. R., Diffenderfer, J., Parasyris, K., and Kailkhura, B. Adversarial robustness limits via scaling-law and human-alignment studies. In *International Conference on Machine Learning (ICML)*, 2024.

Carlini, N., Tramer, F., Dvijotham, K. D., Rice, L., Sun, M., and Kolter, J. Z. (certified!!) adversarial robustness for free! In *International Conference on Learning Representations (ICLR)*, 2023.

Chan, A., Ong, Y. S., and Tan, C. 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.

Chen, H., Dong, Y., Shao, S., Hao, Z., Yang, X., Su, H., and Zhu, J. Diffusion models are certifiably robust classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Chen, H., Dong, Y., Wang, Z., Yang, X., Duan, C., Su, H., and Zhu, J. Robust classification via a single diffusion model. In *International Conference on Machine Learning (ICML)*, 2024b.

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations (ICLR)*, 2021.

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., Lu, Y., and Le, Q. V. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chen, X., Liu, Z., Xie, S., and He, K. Deconstructing denoising diffusion models for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

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

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. In *Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Cui, J., Tian, Z., Zhong, Z., Qi, X., Yu, B., and Zhang, H. Decoupled kullback-leibler divergence loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Fang, G., Li, K., Ma, X., and Wang, X. Tinyfusion: Diffusion transformers learned shallow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Feng, R., Zheng, K., Huang, Y., Zhao, D., Jordan, M., and Zha, Z.-J. Rank diminishing in deep neural networks. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Gan, C., Tu, Y., Chen, X., Chen, T., Li, Y., Harandi, M., and Lin, W. Unleashing diffusion transformers for visual correspondence by modulating massive activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025.

Gao, Y., Shumailov, I., Fawaz, K., and Papernot, N. On the limitations of stochastic pre-processing defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hu, K., Leino, K., Wang, Z., and Fredrikson, M. A recipe for improved certifiable robustness. In *International Conference on Learning Representations (ICLR)*, 2024.

Huang, Q., Dong, X., Chen, D., Chen, Y., Yuan, L., Hua, G., Zhang, W., and Yu, N. 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.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024.

Jaini, P., Clark, K., and Geirhos, R. Intriguing properties of generative classifiers. In *International Conference on Learning Representations (ICLR)*, 2024.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Lai, B.-H., Huang, P.-H., Kung, B.-H., and Chen, S.-T. Enhancing certified robustness via block reflector orthogonal layers and logit annealing loss. In *International Conference on Machine Learning (ICML)*, 2025. Spotlight.

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Li, A. C., Kumar, A., and Pathak, D. Generative classifiers avoid shortcut solutions. In *International Conference on Learning Representations (ICLR)*, 2025a.

Li, D., Ling, H., Kar, A., Acuna, D., Kim, S. W., Kreis, K., Torralba, A., and Fidler, S. Dreamteacher: Pretraining image backbones with deep generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023b.

Li, X., Sun, W., Chen, H., Li, Q., He, Y., Shi, J., and Hu, X. ADBM: Adversarial diffusion bridge model for reliable adversarial purification. In *International Conference on Learning Representations (ICLR)*, 2025b.

Li, X., Zhang, Z., Li, X., Chen, S., Zhu, Z., Wang, P., and Qu, Q. Understanding representation dynamics of diffusion models via low-dimensional modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., and Xie, S. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, 2024.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

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

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

Peng, S., Xu, W., Cornelius, C., Hull, M., Li, K., Duggal, R., Phute, M., Martin, J., and Chau, D. H. Robust principles:

Architectural design principles for adversarially robust cnns. In *British Machine Vision Conference (BMVC)*, 2023.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Sastry, C. S., Dumpala, S. H., and Oore, S. Diffaug: A diffuse-and-denoise augmentation for training robust classifiers. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., and Xie, S. What matters for representation alignment: Global information or spatial structure?, 2025. URL https://arxiv.org/abs/2512.10794.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Stevinson, E., Prieto, L., Barsbey, M., and Birdal, T. Adversarial attacks leverage interference between features in superposition. In *Mechanistic Interpretability Workshop at NeurIPS*, 2025.

Stracke, N., Baumann, S. A., Bauer, K., Fundel, F., and Ommer, B. Cleandift: Diffusion features without noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Wang, K., Fu, X., Han, Y., and Xiang, Y. Diffhammer: Rethinking the robustness of diffusion-based adversarial purification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Wang, Y., Li, L., Yang, J., Lin, Z., and Wang, Y. Balance, imbalance, and rebalance: Understanding robust overfitting from a minimax game perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning (ICML)*, 2023b.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Wu, J., Song, Z., Zhang, X., Xie, S., Lin, L., and Wang, K. Vision transformers beat wideresnets on small scale datasets adversarial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Wu, Y.-Y., Wang, H.-J., and Chen, S.-T. Annealing self-distillation rectification improves adversarial training. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiang, W., Yang, H., Huang, D., and Wang, Y. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yagoda, M., Abu-Hussein, S., and Giryes, R. Diffusion models are robust pretrainers. *IEEE Signal Processing Letters*, 32:4219–4223, 2025. doi: 10.1109/LSP.2025.3624151.

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

Yao, J., Yang, B., and Wang, X. 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.

Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., and Liu, T. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning (ICML)*, 2022.

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., and Xie, S. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L. E., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

A. Representation Similarity Analysis

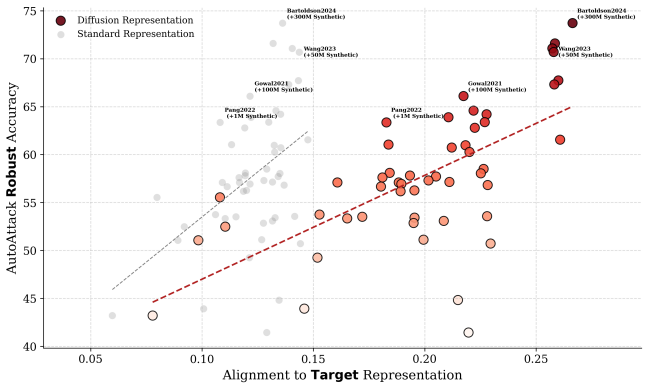


Figure 9. Similarity scores are measured with respect to representations extracted from the diffusion model (red) and from the RobustBench standard-trained model (gray).

Inspired by the work of Huh et al. (2024) analyzing representation similarities 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 9). The results indicate that the representations similarities between better robust models and diffusion representations are stronger, which further inspire us to investigate whether explicitly leveraging diffusion representations as an auxiliary learning signal is beneficial for robust classifier training.

B. Experiment Setup Details

For CIFAR-10/100 WideResNet training, we follow the exact same setup as DM-AT (Wang et al., 2023b): 10-step 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 cosine annealing scheduler. For ViT-B, 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 fine-tune for 100 epochs using the Lion optimizer using 3-step TRADES with $\beta = 10$.

C. Additional Discussion for Related Work

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 rely on inference-time randomness, robust evaluation should be more carefully considered (Athalye et al., 2018). For example, the configuration of Attention Head, b=8, t=50, have been reported to have 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 3b). The results show that diffusion representations still require robust training to provide competitive robustness.

D. Representation Alignment Implementation Details

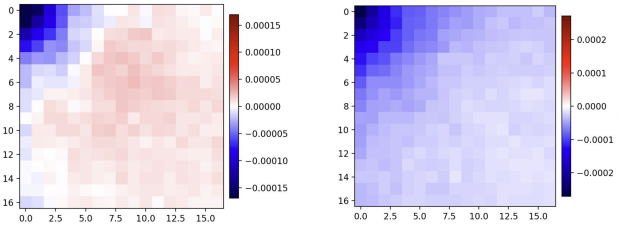
Diffusion Representation Extraction. For CIFAR-10/100, we use the same EDM diffusion model checkpoint as Wang et al. (2023b) to generate synthetic images. We extract representations at noise scale $\sigma_t = 0.1$ from the UNet bottleneck block before the upsampling layers, and apply average pooling 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 $t = 0$ orresponds to pure noise and $t = 1$ being 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 follow the same procedure to first identify channels that consistently exhibit abnormally large activation norms, channels 1053 and 259 in the released LightningDiT checkpoint, and then zero out these channels before applying AdaLN modulation.

Finally, 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. This results in a slight improvement, likely by better leveraging the representaitons 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).

E. Noisy-Image Discriminative Pretrained Representations



(a) Noisy Discriminative (b) Diffusion Representations

Figure 10. 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.

We train a classifier with the same diffusion UNet encoder on the same noisy 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 10 shows that diffusion representations reduce sensitivity to low-frequency components, while alignment to noisy-image discriminative representations increases sensitivity to mid and high-frequency features. Our results complement with the findings of Jaini et al. (2024), which found that noisy discriminative training could lead to shape-bias but decreased OOD classification performance.