JOINT DISCRIMINATIVE-GENERATIVE MODELING VIA DUAL ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Simultaneously achieving robust classification and high-fidelity generative modeling within a single framework presents a significant challenge. Hybrid approaches, such as Joint Energy-Based Models (JEM), interpret classifiers as EBMs but are often limited by the instability and poor sample quality inherent in SGLD-based training. We address these limitations by proposing a novel training framework that integrates adversarial training (AT) principles for both discriminative robustness and stable generative learning. The proposed method introduces three key innovations: (1) the replacement of SGLD-based JEM learning with a stable, AT-based approach that optimizes the energy function by discriminating between real data and PGD-generated contrastive samples using the BCE loss; (2) synergistic adversarial training for the discriminative component that enhances classification robustness while eliminating the need for explicit gradient penalties; and (3) a two-stage training procedure to resolve the incompatibility between batch normalization and EBM training. Experiments on CIFAR-10, CIFAR-100, and ImageNet demonstrate that our method substantially improves adversarial robustness over existing hybrid models while maintaining competitive generative performance. On ImageNet, when optimized for generative modeling, our model's generative fidelity surpasses that of BigGAN and approaches diffusion models, representing the first MCMC-based EBM approach to achieve high-quality generation on complex, high-resolution datasets. Our approach addresses key stability issues that have limited JEM scaling and demonstrates that adversarial training can serve as an effective foundation for unified frameworks capable of generating and robustly classifying visual data.

1 INTRODUCTION

Deep learning models have traditionally been developed with either discriminative or generative objectives in mind, rarely excelling at both simultaneously. Discriminative models are optimized for classification or regression tasks but lack the ability to model data distributions, while generative models can synthesize new data samples but may underperform on downstream classification tasks. Recent research has explored unifying these approaches through joint discriminative-generative modeling frameworks that aim to combine the predictive power of discriminative approaches with the rich data understanding of generative models.

Among these unification efforts, Energy-Based Models (EBMs) have emerged as a promising framework due to their flexibility and theoretical connections to both paradigms. In particular, Joint Energy-Based Models (JEM) (Grathwohl et al., 2019) demonstrated that standard classifier architectures could be reinterpreted to simultaneously function as EBMs, enabling both high-accuracy classification and reasonable sample generation. However, a critical limitation of JEM and similar approaches is their reliance on Markov Chain Monte Carlo (MCMC) methods such as Stochastic Gradient Langevin Dynamics (SGLD) for training the generative component. SGLD-based EBM learning suffers from significant training instabilities, computational inefficiency, and often produces poor-quality samples (Grathwohl et al., 2019; Duvenaud et al., 2021; Du & Mordatch, 2019; Zhao et al., 2020; Gao et al., 2018; Nijkamp et al., 2019), limiting the practical adoption of these hybrid models.

We address these limitations by introducing **Dual Adversarial Training (DAT)**, a novel framework that leverages adversarial training (AT) principles for both discriminative robustness and stable generative learning within a unified JEM-based architecture. Our approach employs a dual application of adversarial training: (1) standard AT for the discriminative component to achieve robustness against adversarial perturbations, and (2) an AT-based energy function learning strategy for the generative component that replaces unstable SGLD-based JEM learning.

Our key technical contributions include:

1. **A stable AT-based alternative to SGLD-based JEM learning.** We replace the unstable SGLD-based JEM learning with an adversarial training approach that optimizes the energy function through Binary Cross-Entropy loss using PGD-generated contrastive samples. This fundamentally addresses the training instabilities that have plagued JEM, enabling reliable convergence and significantly improved sample quality.
2. **Adversarial training with synergistic effects.** We incorporate adversarial training for the discriminative component, which not only enhances classification robustness but also eliminates the need for explicit R_1 gradient penalty required by previous AT-EBMs frameworks (Yin et al., 2022), simplifying the training procedure and avoiding constraints on model expressiveness.
3. **Two-stage training for batch normalization compatibility.** We introduce a two-stage training strategy that reconciles the conflicting requirements of batch normalization for discriminative training and its incompatibility with EBM sampling, eliminating the need for alternative normalization techniques.

Experiments across datasets of increasing complexity, from CIFAR-10 to ImageNet, demonstrate that our approach scales effectively. Compared to existing hybrid models, out approach achieves substantially improved adversarial robustness while maintaining competitive generative performance. This unique combination of strong robustness and generative capabilities enables higher-quality counterfactual explanations—our model creates examples substantially more faithful to target class characteristics compared to non-robust and robustness-only methods. Furthermore, on ImageNet, when optimized for generative modeling, our model's generative fidelity surpasses that of strong, specialized models like BigGAN and approaches the quality of diffusion models, demonstrating that EBM-based approaches can compete with other generative models. These results highlight the flexibility of our framework, demonstrating its capacity to function as both a state-of-the-art hybrid model and a powerful, standalone generative model with potential applications in a broader range of image synthesis tasks (Santurkar et al., 2019).

2 Related work

Joint discriminative-generative modeling The pursuit of joint discriminative-generative modeling, or hybrid modeling, aims to combine the predictive power of discriminative approaches with the rich data understanding of generative models within a single framework. This line of research is motivated by the potential to improve classifier robustness, calibration, and out-of-distribution detection, while also enabling tasks like sample generation (e.g., for counterfactual explanation) and semi-supervised learning. A significant thrust in this area involves Energy-Based Models (EBMs). Early work by Xie et al. (2016) showed how generative ConvNets could be derived from discriminative ones, framing them as EBMs. Du & Mordatch (2019) demonstrated that implicitly generative EBMs can achieve strong performance on discriminative tasks like adversarially robust classification and out-of-distribution detection, while addressing scalable EBM training challenges. Grathwohl et al. (2019) introduced Joint Energy-Based Models (JEM), which explicitly reinterpret standard classifiers as EBMs over the joint distribution of data and labels $p(x, y)$, allowing simultaneous classification and generation. Yang et al. (2023) incorporated sharpness-aware minimization (SAM) to smooth energy landscapes and removed data augmentation from the EBM loss term to improve both classification accuracy and generation quality of JEM. Guo et al. (2023) proposed EGC, which employs Fisher divergence within a diffusion framework to learn an unconditional score function $\nabla \log p(x)$ and a conditional classifier $p(y|x)$ for unified classification and generation, thereby circumventing the computational challenges of traditional energy-based model training.

Alternative architectural approaches have also been explored for joint modeling. Rather than energy-based formulations, joint diffusion models (Deja et al., 2023) attach classifiers directly to diffusion

model UNet encoders for joint end-to-end training. Another distinct approach is "introspective learning," where a single model functions as both a generator and a discriminator through an iterative self-evaluation process, developed across works by Lazarow et al. (2017), Jin et al. (2017), and Lee et al. (2018). Flow-based models have also been explored for hybrid tasks; for instance, Residual Flows (Chen et al., 2019) utilized invertible ResNet and showed competitive performance in joint generative and discriminative settings, offering an alternative to EBMs by allowing exact likelihood computation. These diverse approaches underscore the continued effort to create models that synergistically leverage both discriminative and generative learning.

Joint Energy-Based Models (JEM) A significant step towards unifying discriminative and generative modeling within a single framework was presented by Grathwohl et al. (2019) with their Joint Energy-Based Model (JEM). Their key insight was to reinterpret the logits produced by a standard discriminative classifier, typically used to model $p(y|x)$, as defining an energy function for the joint distribution $p(x, y)$. Specifically, they defined the energy $E_\theta(x, y)$ as the negative of the logit corresponding to class y, $E_\theta(x, y) = -f_\theta(x)[y]$. This formulation allows for the recovery of the standard conditional distribution $p(y|x)$ via softmax normalization over y, while also yielding an unnormalized probability density $p(x)$ by marginalizing out y, effectively using the negative LogSumExp of the logits as the energy function for $p(x)$. They proposed a hybrid training objective that combines the standard cross-entropy loss for $p(y|x)$ with an EBM-based objective for $p(x)$ optimized using Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011). Grathwohl et al. (2019) demonstrated that this joint training approach allows JEM to achieve strong performance on both classification and generative tasks, while simultaneously improving classifier calibration, out-of-distribution detection capabilities, and robustness against adversarial examples compared to standard discriminative training.

Our work builds upon these foundations by incorporating adversarial training principles into the joint modeling framework. We draw particularly from recent advances in adversarial training for EBMs (Yin et al., 2022) and methods for achieving robustness on both in-distribution and out-of-distribution data (Augustin et al., 2020). The connection between robust classifiers and generative capabilities has also been explored, with Santurkar et al. (2019) demonstrating that robust classifiers can perform various image synthesis tasks through gradient-based optimization. For a comprehensive discussion of these approaches, see Appendix A.1.

3 METHOD

3.1 JOINT ENERGY-BASED MODEL

Our approach builds upon the Joint Energy-Based Model (JEM) framework introduced by Grathwohl et al. (2019), which reinterprets the outputs of a standard discriminative classifier as an energy-based model (EBM) over the joint distribution of data x and labels y. Given a classifier network that produces logits $f_\theta(x) \in \mathbb{R}^K$ for K classes, JEM defines the joint energy function as:

$$E_\theta(x, y) = -f_\theta(x)[y] \tag{1}$$

where $f_\theta(x)[y]$ is the logit corresponding to class y. This energy function can be normalized to obtain a joint probability density:

$$p_\theta(x, y) = \frac{\exp(-E_\theta(x, y))}{Z(\theta)} = \frac{\exp(f_\theta(x)[y])}{Z(\theta)} \tag{2}$$

where $Z(\theta)$ is an intractable global normalizing constant. By marginalizing out the label y, a marginal density over the input data x can be obtained:

$$p_\theta(x) = \sum_y p_\theta(x, y) = \frac{\sum_y \exp(f_\theta(x)[y])}{Z(\theta)} \tag{3}$$

Thus, a valid energy function for $p_\theta(x)$ is given by:

$$E_\theta(x) = -\log \sum_y \exp(f_\theta(x)[y]) \tag{4}$$

This energy is related to the marginal density by $p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z(\theta)}$.

A JEM is trained by maximizing the joint log-likelihood $\log p_\theta(x, y)$ over labeled training datapoints (x, y) drawn from an empirical joint distribution $p_{\text{data}}(x, y)$. The joint log-likelihood is typically factorized as $\log p_\theta(y|x) + \log p_\theta(x)$. The conditional term $\log p_\theta(y|x)$ can be maximized by minimizing the standard cross-entropy classification loss. The marginal term $\log p_\theta(x)$ is optimized using the EBM gradient:

$$\nabla_\theta \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log p_\theta(x)] = \mathbb{E}_{x\sim p_{\text{data}}(x)}[-\nabla_\theta E_\theta(x)] - \mathbb{E}_{x\sim p_\theta(x)}[-\nabla_\theta E_\theta(x)] \tag{5}$$

where $p_{\text{data}}(x)$ is the empirical marginal distribution obtained by marginalizing y from $p_{\text{data}}(x, y)$. This gradient decreases the energy of real data samples while increasing the energy of model-generated samples. At equilibrium when $p_\theta(x) = p_{\text{data}}(x)$, these terms balance and the gradient becomes zero.

To approximate the expectation $\mathbb{E}_{x\sim p_\theta(x)}[\cdot]$, samples are drawn from $p_\theta(x)$ using Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011). SGLD generates samples x starting from an initial distribution $p_0(x)$ (e.g., uniform noise) and iteratively applies the update rule:

$$x_{t+1} = x_t - \frac{\alpha}{2}\nabla_x E_\theta(x_t) + \xi_t, \quad \text{where } \xi_t \sim \mathcal{N}(0, \alpha) \tag{6}$$

Here, α is the step size, and $\nabla_x E_\theta(x_t)$ is the gradient with respect to the marginal energy function.

3.2 Learning JEM with adversarial training

The JEM framework successfully integrates generative modeling into classifiers, but its reliance on SGLD and EBM gradient (Eq. 5) causes significant training instabilities (Grathwohl et al., 2019; Duvenaud et al., 2021) and results in poor sample quality. We address these limitations by replacing the SGLD-based JEM with an adversarial training (AT) approach inspired by AT-EBMs (Yin et al., 2022).

Specifically, we replace the standard EBM gradient (Eq. 5) with a stabilized formulation:

$$\begin{aligned} &\mathbb{E}_{x\sim p_{\text{data}}(x)}[-\nabla_\theta E_\theta(x)] - \mathbb{E}_{x\sim p_\theta(x)}[-\nabla_\theta E_\theta(x)] \\ \implies &\mathbb{E}_{x\sim p_{\text{data}}(x)}[-\alpha(x)\nabla_\theta E_\theta(x)] - \mathbb{E}_{x\sim p_\theta(x)}[-\beta(x)\nabla_\theta E_\theta(x)] \end{aligned} \tag{7}$$

where $\alpha(x) = 1 - \sigma(-E_\theta(x))$ and $\beta(x) = \sigma(-E_\theta(x))$ are data-dependent scaling factors, and σ denotes the logistic sigmoid function. This formulation preserves the structural form of Eq. 5 while introducing adaptive scaling factors that modulate gradient contributions according to the model's current energy assignments. According to Yin et al. (2022), these scaling factors stabilize training by providing automatic gradient regularization: as $-E_\theta(x)$ increases for p_{data} samples, the corresponding scaling factor $\alpha(x) = 1 - \sigma(-E_\theta(x))$ approaches zero, thereby attenuating the gradient contribution from such samples and preventing numerical overflow; conversely, when $-E_\theta(x)$ becomes very negative for contrastive samples, $\beta(x) = \sigma(-E_\theta(x))$ approaches zero, preventing numerical underflow. In contrast, the standard EBM gradient (Eq. 5) is unconstrained and permits $-E_\theta(x)$ to achieve arbitrarily large or small magnitudes, resulting in numerical instability during optimization. This gradient formulation stabilizes training at the cost of limiting the EBM to modeling the support of p_{data} rather than learning the full density.

In addition to the above gradient reformualtion, the sampling required to estimate $\mathbb{E}_{x\sim p_\theta(x)}[\cdot]$ is performed using the PGD attack (Madry et al., 2017) instead of SGLD. Specifically, the contrastive samples x from the model distribution are generated by initializing from an auxiliary out-of-distribution dataset p_{ood} (e.g., the 80 million tiny images dataset for CIFAR-10 training) and performing T iterations of gradient ascent on the negative energy function $-E_\theta(x)$:

$$x_{t+1} = x_t + \eta \frac{\nabla_x(-E_\theta(x_t))}{||\nabla_x(-E_\theta(x_t))||_2}, \quad t = 0, 1, \ldots, T-1 \tag{8}$$

where $E_\theta(x)$ is the marginal energy function defined in Eq. 4, η is the step size, and T is the total number of PGD steps. Using the update direction suggested by Eq. 7 is equivalent to minimizing the Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{\text{BCE}}(\theta) = -\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(\sigma(-E_\theta(x)))] - \mathbb{E}_{x\sim p_\theta(x)}[\log(1 - \sigma(-E_\theta(x)))] \tag{9}$$

Minimizing this $\mathcal{L}_{\text{BCE}}$ implicitly trains the energy function $E_\theta(x)$ to assign low energy to data samples from $p_{\text{data}}(x)$ and high energy to the contrastive samples computed using the PGD attack.

We find this AT-based approach effectively addresses JEM's training stability issues and produces high quality samples.

3.3 CLASSIFIER ROBUSTNESS AND IMPLICIT REGULARIZATION

Classifier robustness. While our AT-based approach improves the generative capabilities of JEM, the original JEM's discriminative component still exhibits weak adversarial robustness compared to dedicated adversarially trained classifiers. To address this limitation, we complement our generative improvements by incorporating adversarial training for the discriminative term $p_\theta(y|x)$.

For each input sample x with label y, we find an adversarial example x_{adv} within an ϵ-ball $B(x, \epsilon)$ around x that maximizes the classification loss:

$$x_{adv} = \underset{x' \in B(x,\epsilon)}{\arg\max}\, \mathcal{L}_{\text{CE}}(\theta; x', y) \tag{10}$$

where $\mathcal{L}_{\text{CE}}(\theta; x', y)$ is the standard cross-entropy loss and $B(x, \epsilon)$ is an L_p-norm ball. Similar to our generative component, we approximate this optimization using the PGD attack (Madry et al., 2017), generating adversarial examples through iterative gradient steps within the constraint set. The classification term is then defined as:

$$\mathcal{L}_{\text{AT-CE}}(\theta) = \mathbb{E}_{(x,y)\sim p_{\text{data}}(x,y)} \left[-\log p_\theta(y|x_{adv}) \right] \tag{11}$$

Implicit regularization. Incorporating AT for the classifier not only ensures robust accuracy but also yields a synergistic benefit for the generative component. Specifically, we find empirically that it eliminates the need for the R_1 gradient penalty (Mescheder et al., 2018), an explicit regularization required by the original AT-EBMs framework (Yin et al., 2022) that can constrain model expressiveness. Our analysis reveals that AT provides implicit regularization that encompasses R_1-style penalties through its first-order penalty on gradient norms (see Appendix A.2). While these two approaches operate at different scales and through different objectives, our experiments suggest that the local smoothness induced by AT provides sufficient implicit regularization for stable EBM training, reducing the need for explicit gradient penalties.

3.4 DUAL AT FOR JOINT MODELING

Our complete model integrates adversarial training principles for both the generative and discriminative components, resulting in the combined objective:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{AT-CE}}(\theta) + \mathcal{L}_{\text{BCE}}(\theta) \tag{12}$$

where $\mathcal{L}_{\text{AT-CE}}(\theta)$ is the robust classification loss from Eq. 11, and $\mathcal{L}_{\text{BCE}}(\theta)$ is the AT-based generative loss from Eq. 9. This DAT approach simultaneously enhances the model's discriminative robustness and generative capabilities, addressing the key limitations of the original JEM framework; full algorithmic details are provided in Appendix A.3.

Our approach shares conceptual similarities with RATIO (Augustin et al., 2020), which also combines adversarially robust classification with adversarial perturbations applied to out-of-distribution data:

$$\mathcal{L}_{\text{RATIO}}(\theta) = \mathcal{L}_{\text{AT-CE}}(\theta) + \lambda \mathbb{E}_{x\sim p_{\text{ood}}(x)} \left[\max_{x' \in B(x,\epsilon_o)} \mathcal{L}_{\text{CE}}(\theta; x', \mathbf{1}/K) \right] \tag{13}$$

where $\mathbf{1}$ is the vector of all ones and K is the number of classes. Despite this structural similarity, the approaches differ fundamentally in their objectives. RATIO's secondary term attacks OOD samples to maximize classifier confidence, then penalizes this confidence via cross-entropy against a uniform distribution, explicitly targeting robust OOD detection. In contrast, our $\mathcal{L}_{\text{BCE}}(\theta)$ leverages AT-based energy function learning (Yin et al., 2022), using PGD to generate contrastive samples from OOD data and employing BCE loss to shape the energy landscape. While RATIO focuses primarily on reducing confidence in OOD regions, our approach prioritizes learning a stable and effective energy function that enables high-quality generative modeling alongside robust classification.

3.5 TWO-STAGE TRAINING

A fundamental challenge in training joint models is the use of batch normalization (BN) (Ioffe & Szegedy, 2015). While BN is highly beneficial for stabilizing and speeding up standard deep network training, it is often found to interfere with the learning dynamics of EBMs and their sampling

procedures (Grathwohl et al., 2019; Yin et al., 2022; Zhao et al., 2020; 2016). Consistent with these findings, we observe that enabling BN during joint training destabilizes the optimization of the generative modeling term $\mathcal{L}_{\text{BCE}}$, leading to oscillating losses and failure to converge. This instability arises from BN's reliance on batch-dependent normalization statistics: as negative samples (e.g., from Eq. 6) change throughout training, the BN statistics S_t continuously drift, effectively transforming the energy function from a fixed parameterization $E_\theta(x)$ into a time-varying one $E_{\theta,S_t}(x)$. In our experiments, we find that disabling BN removes this dependence and restores stable optimization of $\mathcal{L}_{\text{BCE}}$, confirming that the failure mode arises from computing BN statistics on the evolving negative samples.

However, simply disabling BN from the start would negatively impact the initial training of the robust classifier backbone. To reconcile these conflicting requirements, we implement a two-stage training strategy:

- **Stage 1 discriminative training.** In this initial stage, we train the network with BN enabled, optimizing only the robust classification objective $\mathcal{L}_{\text{AT-CE}}$ (Eq. 11). This stage is equivalent to standard adversarial training and leverages BN to achieve faster convergence and strong robust classification performance. Notably, this stage can be skipped by initializing from pretrained off-the-shelf robust classifiers, making our approach immediately applicable to existing robust models.
- **Stage 2 joint training.** After robust discriminative training, we disable BN throughout the network by setting the BN modules in eval mode. We then continue training by optimizing the complete objective function $\mathcal{L}(\theta) = \mathcal{L}_{\text{AT-CE}}(\theta) + \mathcal{L}_{\text{BCE}}(\theta)$ (Eq. 12).

While alternative approaches such as spectral normalization and virtual batch normalization have been considered for stabilizing EBM training (Zhao et al., 2020; Miyato et al., 2018; Zhao et al., 2016), our experiments demonstrate that this two-stage approach effectively addresses the BN incompatibility without requiring such alternatives. Disabling BN in Stage 2 enables stable generative loss convergence and dramatically improves generative performance, with minimal impact on the robust accuracy established in Stage 1 (see Appendix A.12 for training curves).

3.6 Data augmentation

Strong data augmentations are necessary for achieving robust classification (Rebuffi et al., 2021; Gowal et al., 2020) but can distort the data distribution in ways detrimental to generative modeling. We therefore follow Yang et al. (2023) and apply separate augmentation strategies to the discriminative and generative components. While Yang et al. (2023) concludes that augmentations like random cropping with padding should be excluded from generative training to avoid artifacts like black borders, we find this is not a limitation in our framework. Notably, even with random cropping and padding applied, our generated samples do not inherit these artifacts, allowing us to improve robustness without degrading sample quality (see Appendix A.4.2). We therefore apply strong augmentations to the discriminative term $\mathcal{L}_{\text{AT-CE}}$ and mild augmentations to the generative term $\mathcal{L}_{\text{BCE}}$.

4 Experiments

4.1 Training setup

Datasets and architectures. We evaluate our approach on CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and ImageNet (Deng et al., 2009). For CIFAR-10/100 experiments, we use WRN-34-10 (Zagoruyko & Komodakis, 2016) following the official RATIO implementation. For ImageNet experiments, we employ ResNet-50 (He et al., 2016) and WRN-50-4 (Zagoruyko & Komodakis, 2016).

Two-stage training. Since Stage 1 training is equivalent to standard adversarial training, we use pretrained models when available: we use a pretrained CIFAR-10 model from RATIO (Augustin et al., 2020) and pretrained ImageNet models from Salman et al. (2020), while training our own CIFAR-100 model following Augustin et al. (2020). For Stage 2 training, we initialize from the Stage 1 model and continue joint training by setting the BN modules to eval mode (which disables BN while preserving the BN statistics computed during Stage 1). Complete training hyperparameters can be found in Appendix A.4.1.

Data augmentation. We employ separate data augmentation strategies for Stage 2 training: strong augmentations for $\mathcal{L}_{\text{AT-CE}}$ and basic transformations for $\mathcal{L}_{\text{BCE}}$ to preserve the data distribution. Detailed specifications can be found in Appendix A.4.2.

Out-of-distribution data. Same as RATIO, we use the 80 million tiny images (Torralba et al., 2008) as the OOD dataset (p_{ood}) for CIFAR-10/100 experiments. For ImageNet, as there are no established OOD datasets, we follow OpenImage-O (Wang et al., 2022) and construct an OOD dataset from Open Images training set (Krasin et al., 2016). We randomly sample 350K images, restricting our selection to those whose labels do not overlap with any ImageNet classes, yielding 300K samples for training and 50K for FID evaluation.

4.2 Evaluation metrics

We measure both classification and generative modeling performance. For classification, we report clean accuracy and robust accuracy against L_2 attacks ($\epsilon = 0.5$ for CIFAR-10/100 and $\epsilon = 3.0$ for ImageNet) computed using AutoAttack (Croce & Hein, 2020). For generative modeling, we evaluate sample diversity and visual fidelity using Fréchet Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016). We focus on conditional generation; details of the generation setup are provided in Appendix A.6.

To measure the quality of counterfactuals, we generate sets of counterfactual examples by applying targeted attacks to training samples across a range of perturbation limits. For each target class, we compute the class-wise FID score between the set of counterfactuals targeted at that class and the set of training samples from the same class. Note that counterfactuals are generated by applying PGD attacks to in-distribution training samples, whereas generative modeling samples are created by applying PGD attacks to OOD inputs.

4.3 Results

4.3.1 Classification and generative modeling

Comparison with hybrid models. Most existing hybrid models are not explicitly optimized for adversarial robustness, achieving significantly lower robust accuracy than standard AT—for example, JEM (40.5%) and SADA-JEM (31.93%) versus standard AT (75.73%) on CIFAR-10 (Table 1), and EGC (13.56%) versus standard AT (34.44%/40.28%) on ImageNet (Table 2). Our DAT approach addresses this limitation by incorporating adversarial training directly into the joint objective, achieving robust accuracy comparable to standard AT. RATIO is the only existing hybrid that explicitly targets robustness, but it does not optimize for generation quality. Our approach achieves comparable robustness to RATIO while substantially improving generative performance. These results demonstrate that our approach bridges the gap between robustness-focused and generation-focused hybrid models, achieving strong performance in both objectives simultaneously. This synergy between robustness and generative capabilities yields practical benefits, as shown in Section 4.3.2, where our model produces substantially higher-quality counterfactuals through PGD-based generation.

Comparison with generative models. On ImageNet, our DAT approach demonstrates competitive performance relative to dedicated generative models. Our best generative configuration (WRN-50-4 with $T = 65$) achieves an FID of 5.39, outperforming BigGAN-deep (6.95) and approaching recent diffusion models such as ADM-G (4.59) and LDM-G (3.60), while requiring significantly less sampling steps. To our knowledge, this represents the first MCMC-based EBM approach to achieve such high-quality generation on complex, high-resolution datasets like ImageNet. Additionally, our model attains relatively strong IS performance, likely due to the PGD-based sampling explicitly optimizing for classifier confidence—a property that aligns with the Inception network-based evaluation metric. We show ImageNet generated samples in Figure 8.

Trading off generative and discriminative performance. Our experiments reveal that the number of PGD training steps T in Eq. 8 serves as a mechanism for controlling the balance between discriminative and generative objectives. On CIFAR-10, increasing T from 40 to 50 improves FID from 9.07 to 7.57 at the cost of standard and robust accuracy. A similar trend is observed on CIFAR-100 and ImageNet, where increasing T consistently improves generation quality while reducing classification performance.

Effect of model capacity. Our experiments on ImageNet demonstrate the benefits of increased model capacity. Scaling from ResNet-50 (26M parameters) to WRN-50-4 (223M parameters) yields consistent improvements across both discriminative and generative metrics.

Qualitative results. Figures 5, 6, and 7 show generated samples produced by our approach, RATIO, and standard AT. We observe that our method produces visually superior samples with fewer artifacts compared to RATIO and standard AT.

Table 1: Classification and generative modeling results on CIFAR-10 and CIFAR-100.

Method	Acc% ↑	Robust Acc% ↑	IS ↑	FID ↓
CIFAR-10 hybrid models				
Residual Flow (Chen et al., 2019)	70.3	–	3.6	46.4
Glow (Kingma & Dhariwal, 2018)	67.6	–	3.92	48.9
IGEBM (Du & Mordatch, 2019)	49.1	–	8.3	37.9
JEM (Grathwohl et al., 2019)	92.9	40.5	8.76	38.4
SADA-JEM (Yang et al., 2023)	95.5	31.93	8.77	9.41
EGC (Guo et al., 2023)	95.9	–	9.43	3.30
Joint-Diffusion (Deja et al., 2023)	96.4	–	–	6.4
RATIO (Augustin et al., 2020)	92.23	76.25	9.61	21.96
Standard AT (Augustin et al., 2020)	92.43	75.73	9.58	28.41
DAT ($T = 40$)	91.86	75.66	9.96	9.07
DAT ($T = 50$)	90.72	74.65	9.86	7.57
CIFAR-10 conditional generative models				
SNGAN (Miyato et al., 2018)	–	–	8.59	25.5
BigGAN (Brock et al., 2018)	–	–	9.22	14.73
StyleGAN2 (Karras et al., 2020b)	–	–	9.53	6.96
StyleGAN2 ADA (Karras et al., 2020a)	–	–	10.24	3.49
EDM (Karras et al., 2022)	–	–	–	1.79
CIFAR-100 hybrid models				
Joint-Diffusion (Deja et al., 2023)	77.6	–	–	16.8
SADA-JEM (Yang et al., 2023)	75.0	–	11.63	14.4
EGC (Guo et al., 2023)	77.9	–	11.50	4.88
RATIO (Augustin et al., 2020)	71.58	47.74	9.28	24.17
Standard AT (Augustin et al., 2020)	72.16	47.78	9.54	23.59
DAT ($T = 45$)	65.55	45.97	10.83	10.70
DAT ($T = 50$)	62.57	45.97	11.51	9.96

Table 2: Classification and generative modeling results on ImageNet (Standard AT and DAT use 224×224 generation; all other methods use 256×256 generation).

Method	Acc% ↑	Robust Acc% ↑	FID ↓	IS ↑	Params	Steps
Hybrid models						
Standard AT (Salman et al., 2020)	62.83	34.44	15.97	274.9	26M (ResNet-50)	13
DAT ($T = 15$)	57.88	34.84	6.64	339.6	26M (ResNet-50)	13
DAT ($T = 30$)	53.46	32.42	5.88	344.0	26M (ResNet-50)	16
Standard AT (Salman et al., 2020)	69.67	40.28	37.25	228.6	223M (WRN-50-4)	11
DAT ($T = 30$)	62.09	39.68	6.35	347.6	223M (WRN-50-4)	17
DAT ($T = 65$)	54.77	34.14	5.39	341.9	223M (WRN-50-4)	20
EGC (Guo et al., 2023)	78.90	13.56	6.05	231.3	543M	1000
Conditional generative models						
BigGAN-deep (Brock et al., 2018)	–	–	6.95	203.6	340M	1
ADM-G (Dhariwal & Nichol, 2021)	–	–	5.44	–	608M	25
ADM-G (Dhariwal & Nichol, 2021)	–	–	4.59	186.7	608M	250
LDM-G (Rombach et al., 2022)	–	–	3.60	247.7	400M	250
VAR (Tian et al., 2024)	–	–	1.73	350.2	2.0B	10

4.3.2 COUNTERFACTUAL GENERATION, OOD DETECTION, AND CALIBRATION

Counterfactual generation. Figure 1 compares counterfactual quality across different models while accounting for classifier confidence. Our approach consistently generates counterfactuals with lower FIDs than baseline methods when achieving similar target class confidence. For instance, when the RATIO baseline reaches approximately 0.89 confidence in the target class (at $\epsilon = 8$), its corresponding FID is 43.18. Our DAT model achieves a similar confidence level at $\epsilon = 4$ with a significantly better FID of 25.53. This demonstrates that, for a comparable level of certainty that the counterfactual represents the target class, our generated samples are substantially more faithful to the true visual characteristics of that class, indicating more plausible counterfactuals. Therefore, our model's improved generative capability directly translates to higher-quality counterfactual explanations, enhancing model explainability. We provide visualizations of counterfactuals in Appendix A.9.

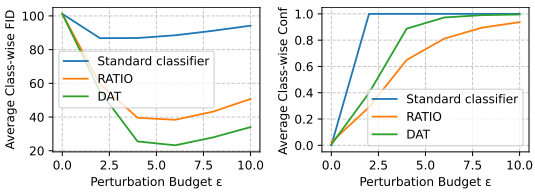


Figure 1: Counterfactual FIDs and classifier confidences under different perturbations.

OOD detection. Our approach generally underperforms RATIO on OOD detection. Ablation studies show this gap persists even when using identical aggressive augmentation for both the generative and discriminative components, indicating it stems from fundamental objective differences rather than the use of milder augmentation for the generative term: RATIO explicitly optimizes for OOD detection while our generative loss prioritizes learning accurate energy functions for generation. The complete details can be found in Appendix A.7

Calibration. Our model's calibration performance is dataset-dependent, with detailed results provided in Appendix A.8. While the model is well-calibrated on CIFAR-10, outperforming the standard AT and RATIO baselines, it exhibits higher overconfidence on CIFAR-100 and ImageNet. The results suggest that prioritizing generative quality may come at the cost of calibration.

4.3.3 ADDITIONAL ANALYSES

In Appendix A.5 we conduct additional analyses including component ablation of DAT, OOD dataset effects, and loss weighting mechanisms. Component ablation reveals that both the generative loss and decoupled augmentation contribute to the improved generative quality compared to a standard AT baseline (Appendix A.5.1). OOD dataset analysis demonstrates notable data efficiency of our approach, suggesting strong performance is achievable even with limited auxiliary OOD data (Appendix A.5.2). Loss weighting analysis confirms that re-weighting the two loss terms provides an alternative mechanism for controlling the generative-discriminative trade-off beyond varying PGD steps (Appendix A.5.3).

5 CONCLUSION

We addressed the challenge of developing models that excel simultaneously at robust classification and high-fidelity generative modeling. While JEM offers a promising foundation, it suffers from training instability and poor generation quality. Our DAT approach integrates adversarial training principles for both components: replacing unstable SGLD-based JEM learning with an AT-based approach for the generative component, while incorporating standard AT for classification robustness. Experiments across multiple datasets demonstrate that our approach achieves substantially better adversarial robustness over existing hybrid models while maintaining competitive generative capabilities; when optimized for generative performance on ImageNet, it achieves generation quality comparable to dedicated generative models such as GANs and diffusion models. Future work could advance this approach in several directions: improving training efficiency with persistent markov chains, scaling the framework to higher-capacity models to test performance limits; improving secondary tasks like out-of-distribution detection by developing hybrid objectives that combine our generative loss with RATIO's; and applying the model to broader image synthesis tasks such as those demonstrated by Santurkar et al. (2019).

REPRODUCIBILITY STATEMENT

We have made substantial efforts to ensure reproducibility of our results. Complete training hyperparameters for both stages across all datasets are provided in Tables 3 to 5 and Appendix A.4.1. The two-stage training procedure is detailed in Section 3.5 with the complete algorithm provided in Appendix A.3. Data augmentation strategies for each component are specified in Appendix A.4.2. All evaluation metrics, generation parameters, and experimental settings are documented in Section 4.3 and Appendix A.6. Ablation study configurations are detailed in Appendix A.5. We use standard publicly available datasets (CIFAR-10/100, ImageNet) and follow established evaluation protocols to facilitate comparison and reproduction. Source code and model checkpoints for reproducing the results are provided as supplementary material.

REFERENCES

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pp. 228–245. Springer, 2020.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 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*, pp. 2206–2216. PMLR, 2020.

Kamil Deja, Tomasz Trzciński, and Jakub M Tomczak. Learning data representations with joint diffusion models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 543–559. Springer, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

David Duvenaud, Jacob Kelly, Kevin Swersky, Milad Hashemi, Mohammad Norouzi, and Will Grathwohl. No mcmc for me: Amortized samplers for fast and stable training of energy-based models. In *International Conference on Learning Representations (ICLR)*, 2021.

Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. Learning generative convnets via multi-grid modeling and sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9155–9164, 2018.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Qiushan Guo, Chuofan Ma, Yi Jiang, Zehuan Yuan, Yizhou Yu, and Ping Luo. Egc: Image generation and classification via a diffusion energy-based model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22952–22962, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Long Jin, Justin Lazarow, and Zhuowen Tu. Introspective classification with convolutional nets. *Advances in Neural Information Processing Systems*, 30, 2017.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020b.

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*, volume 35, pp. 26565–26577, 2022.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, and Gal Chechik. Openimages: A public dataset for large-scale multi-label and multi-class image classification. 2016. Dataset available from `https://storage.googleapis.com/openimages/web/index.html`.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Justin Lazarow, Long Jin, and Zhuowen Tu. Introspective neural networks for generative modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2774–2783, 2017.

Kwonjoon Lee, Weijian Xu, Fan Fan, and Zhuowen Tu. Wasserstein introspective neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3702–3711, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pp. 3481–3490. PMLR, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 32, 2019.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

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*, pp. 10684–10695, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems*, 32, 2019.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International conference on machine learning*, pp. 2635–2644. PMLR, 2016.

Xiulong Yang, Qing Su, and Shihao Ji. Towards bridging the performance gaps of joint energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15732–15741, 2023.

Xuwang Yin, Shiying Li, and Gustavo K Rohde. Learning energy-based models with adversarial training. In *European Conference on Computer Vision*, pp. 209–226. Springer, 2022.

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

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Yang Zhao, Jianwen Xie, and Ping Li. Learning energy-based generative models via coarse-to-fine expanding and sampling. In *International Conference on Learning Representations*, 2020.

A SUPPLEMENTARY MATERIAL

A.1 EXTENDED DISCUSSION ON RELATED WORK

Learning EBMs with adversarial training Yin et al. (2022) explored an alternative approach to learning EBMs by leveraging the mechanism of Adversarial Training (AT). They established a connection between the objective of binary AT (discriminating real data from adversarially perturbed out-of-distribution data) and the SGLD-based maximum likelihood training commonly used for EBMs. Specifically, they showed that the binary classifier learned via AT implicitly defines an energy function that models the support of the data distribution, assigning low energy to in-distribution regions and high energy to out-of-distribution (OOD) regions. The PGD attack used in AT to generate adversarial samples from OOD data was interpreted as a non-convergent sampler that produces contrastive data, analogous to MCMC sampling in EBM training. Although the resulting energy function can only capture the support rather than recover the exact density, their model achieves competitive image generation performance compared to explicit EBMs. Notably, this AT-based EBM learning approach is more stable than traditional MCMC-based EBM training and demonstrated strong performance in worst-case out-of-distribution detection, similar to methods like RATIO (Augustin et al., 2020).

In- and out-distribution adversarial robustness Addressing the multifaceted challenge of creating models that are simultaneously accurate, robust, and reliable on out-of-distribution (OOD) data, Augustin et al. (2020) proposed RATIO (Robustness via Adversarial Training on In- and Out-distribution). Their approach combines standard adversarial training (AT) on the in-distribution data, aimed at improving robustness against adversarial examples, with a form of AT on OOD data, which enforces low and uniform confidence predictions within a neighborhood around OOD samples. The combined objective trains the model to maintain correct, robust classifications for in-distribution data while actively discouraging high-confidence predictions for OOD inputs, even under adversarial manipulation. Augustin et al. (2020) demonstrated that RATIO achieves state-of-the-art L_2 robustness on datasets like CIFAR-10, often with less degradation in clean accuracy compared to standard AT alone. Furthermore, they showed that RATIO yields reliable OOD detection performance, particularly in worst-case scenarios where OOD samples are adversarially perturbed to maximize confidence. Their work also highlighted that the L_2 robustness fostered by RATIO enables the generation of meaningful visual counterfactual explanations directly in pixel space, where optimizing confidence towards a target class results in the emergence of corresponding class-specific visual features.

Robust classifiers for image synthesis and manipulation Santurkar et al. (2019) demonstrated that adversarially robust classifiers can serve as powerful primitives for diverse image synthesis tasks. The core insight of their work is that the process of adversarial training—which optimizes the worst-case loss over an ℓ_2 perturbation set rather than expected loss—compels a model to learn more perceptually aligned and human-interpretable feature representations by preventing reliance on imperceptible artifacts. Based on this insight, Santurkar et al. (2019) showed that simple gradient ascent on class scores from such robust classifiers enables a unified framework for image generation, inpainting, image-to-image translation, super-resolution, and interactive manipulation—tasks typically requiring specialized GAN architectures or complex generative models.

A.2 Intuitive connection between R_1 regularization and adversarial training

We provide mathematical intuition for why adversarial training can serve as implicit regularization in place of explicit R_1 gradient penalties. For classification tasks, consider a vector-valued function $f : \mathbb{R}^d \to \mathbb{R}^K$ producing logits for K classes.

R_1 regularization directly penalizes large gradients of the true class logit:

$$\mathcal{L}_{R_1} = \mathbb{E}_{(x,y)\sim p_{\text{data}}} \left[\|\nabla_x f_y(x)\|_2^2 \right] \tag{14}$$

In practice, adversarial training uses cross-entropy loss to enforce consistent predictions:

$$\mathcal{L}_{\text{AT}} = \mathbb{E}_{(x,y)\sim p_{\text{data}}} \left[\max_{\|\delta\|_2 \leq \epsilon} L(f(x+\delta), y) \right] \tag{15}$$

To understand how adversarial training provides implicit gradient regularization, we analyze the first-order behavior of this cross-entropy adversarial objective.

First-order expansion of the adversarial objective. Let $z = f(x) \in \mathbb{R}^K$, $p = \text{softmax}(z)$, and $J_f(x) \in \mathbb{R}^{K\times d}$ be the input-Jacobian of the logits. For cross-entropy loss $L(z, y) = -\log p_y$, a first-order Taylor expansion gives:

$$L(f(x+\delta), y) = L(f(x), y) + \nabla_x L(f(x), y)^T \delta + O(\|\delta\|_2^2) \tag{16}$$

$$\max_{\|\delta\|_2 \leq \epsilon} L(f(x+\delta), y) \approx L(f(x), y) + \epsilon \|\nabla_x L(f(x), y)\|_2 \tag{17}$$

This approximation is valid for sufficiently small ϵ relative to the local curvature of L, such that higher-order terms remain negligible over the L_2 constraint ball $\{\delta : \|\delta\|_2 \leq \epsilon\}$.

This shows that adversarial training implicitly adds a penalty term proportional to $\|\nabla_x L\|_2$ (the first power of the gradient norm). To understand what this gradient represents, we apply the chain rule:

$$\nabla_x L = J_f(x)^T \nabla_z L = J_f(x)^T (p - e_y) \tag{18}$$

where e_y is the one-hot label vector. Let $g_k(x) := \nabla_x f_k(x) \in \mathbb{R}^d$ denote the per-class input gradients (the rows of J_f). Then we can write:

$$\nabla_x L = \sum_{k=1}^{K} (p_k - \delta_{ky}) g_k = -(1 - p_y) g_y + \sum_{k\neq y} p_k g_k \tag{19}$$

This decomposition reveals that the cross-entropy gradient is a weighted combination of per-class gradients, where the true class gradient g_y appears with negative weight $(1 - p_y)$ and competitor gradients g_k appear with positive weights p_k.

Expansion into R_1-style components. To understand how this relates to standard R_1 regularization, we can expand the squared gradient norm by substituting the final expression for $\nabla_x L$. Although the actual adversarial penalty is proportional to $\|\nabla_x L\|_2$, examining $\|\nabla_x L\|_2^2$ provides useful analytical insight:

$$\|\nabla_x L\|_2^2 = \left\| -(1 - p_y) g_y + \sum_{k\neq y} p_k g_k \right\|_2^2 \tag{20}$$

Expanding this expression:

$$\begin{aligned} \|\nabla_x L\|_2^2 = \; & \underbrace{(1 - p_y)^2 \|g_y\|_2^2}_{\text{down-weighted true-class } R_1} + \underbrace{\sum_{k\neq y} p_k^2 \|g_k\|_2^2}_{\text{competitor } R_1 \text{ terms}} \\ & \underbrace{-2(1 - p_y) \sum_{k\neq y} p_k \langle g_y, g_k \rangle}_{\text{true vs competitor alignment}} + \underbrace{2 \sum_{\substack{i<j \\ i,j\neq y}} p_i p_j \langle g_i, g_j \rangle}_{\text{competitor-competitor alignment}} \end{aligned} \tag{21}$$

This expansion decomposes the adversarial penalty into interpretable components:

1. **True-class R_1 regularization**: $(1 - p_y)^2 \|g_y\|_2^2$, which is the standard R_1 penalty on the true class, down-weighted by confidence
2. **Competitor R_1 terms**: $\sum_{k \neq y} p_k^2 \|g_k\|_2^2$, providing R_1-style regularization on competing classes weighted by their predicted probabilities
3. **Gradient alignment terms**: Cross-class inner products $\langle g_i, g_j \rangle$ that discourage competitor–competitor alignment (favoring orthogonality), while encouraging the true-class gradient to align with the competitor average

The above decomposition shows that $\mathbb{E}_{(x,y)}[\|\nabla_x L\|_2^2]$ includes the core R_1 regularization term $(1 - p_y)^2 \|g_y\|_2^2$ alongside additional smoothness constraints, suggesting that adversarial training may provide comparable regularization to explicit R_1 penalties in contexts where gradient control is important. However, there are important distinctions: adversarial training down-weights high-confidence points through the $(1 - p_y)^2$ factor, has competitor R_1 terms, and introduces cross terms that encourage specific gradient alignment patterns absent in standard R_1. Additionally, the analysis relies on the first-order approximation being valid, which may not hold for the ϵ values commonly used in practice, and the actual adversarial penalty is proportional to $\|\nabla_x L\|_2$ rather than its square. Despite these caveats, this mathematical framework provides insights for why adversarial training's comprehensive implicit regularization may offer sufficient smoothness constraints for stable energy-based model training, without requiring explicit gradient penalties.

A.3 DAT TRAINING ALGORITHM

The complete training procedure for our combined objective (Eq. 12) is detailed in Algorithm 1. We note that to train the generative component $\mathcal{L}_{\text{BCE}}$, we sample from $p_\theta(x)$ to estimate $\mathbb{E}_{x\sim p_\theta(x)}[-\nabla_\theta E_\theta(x)]$ in Eq. 5. In the context of JEM, there are broadly two strategies for drawing samples from $p_\theta(x)$ (Grathwohl et al., 2019):

1. Direct sampling from the marginal distribution using gradient-based MCMC (e.g., SGLD or PGD) on the marginal energy $E_\theta(x) = -\log \sum_y \exp(f_\theta(x)[y])$, as implied by Eq. 8.
2. Ancestral sampling, which first draws a label $y \sim p_{\text{data}}(y)$, then samples $x \sim p_\theta(x|y)$ by running gradient-based MCMC on the joint energy $E_\theta(x, y) = -f_\theta(x)[y]$.

Although both approaches yield unbiased estimates, we find ancestral sampling to be practically superior for training stability, possibly because it leverages the classifier's existing strong class representations to provide better mode coverage and mixing properties, while direct sampling from the marginal distribution often diverges. We also find ancestral sampling (conditional generation) yields substantially better FID than directly sampling from marginal distribution (see Table 11).

Consequently, our implementation adopts ancestral sampling when generating contrastive samples (Algorithm 1). Specifically, we first sample a label $y' \sim p_{\text{data}}(y)$, then generate a contrastive sample x_T by performing T iterations of PGD on the negative joint energy function $-E_\theta(x, y')$, starting from an initial sample $x_0 \sim p_{\text{ood}}$. This class-conditional contrastive sample x_T is then used in the $\mathcal{L}_{\text{BCE}}$ objective (Eq. 9), whose gradient (Eq. 7) provide an approximation to Eq. 5.

Algorithm 1 DAT training: Given network logits f_θ, in-distribution dataset p_{data}, auxiliary out-of-distribution dataset p_{ood}, classification AT bound ϵ, PGD iterations T, PGD step size η

while not converged **do**
  Sample $(x, y) \sim p_{\text{data}}(x, y)$, apply aggressive augmentation to x
  Sample $\hat{x} \sim p_{\text{data}}(x)$, $x_0 \sim p_{\text{ood}}(x)$, apply mild augmentation to $\hat{x}$ and x_0
  Solve $x_{adv} = \arg\max_{x' \in B(x,\epsilon)} \mathcal{L}_{\text{CE}}(\theta; x', y)$ via PGD attack
  $\mathcal{L}_{\text{AT-CE}}(\theta) = -\log p_\theta(y|x_{adv})$ ▷ Robust classification loss
  Initialize $x_t \leftarrow x_0$ for $t = 0$, sample $y' \sim p_{\text{data}}(y)$
  for $t \in \{1, \ldots, T\}$ **do** ▷ Generate contrastive sample for EBM
    $g = \nabla_x(-E_\theta(x_{t-1}, y'))$ ▷ Gradient of negative energy
    $x_t \leftarrow x_{t-1} + \eta \cdot g/||g||_2$ ▷ Normalized gradient ascent step
  end for
  $\mathcal{L}_{\text{BCE}}(\theta) = -\log(\sigma(-E_\theta(\hat{x}))) - \log(1 - \sigma(-E_\theta(x_T)))$ ▷ Generative modeling loss
  $\mathcal{L}(\theta) = \mathcal{L}_{\text{AT-CE}}(\theta) + \mathcal{L}_{\text{BCE}}(\theta)$
  Compute parameter gradients $\nabla_\theta \mathcal{L}(\theta)$ and update θ
end while

A.4 MODEL TRAINING

A.4.1 TRAINING SETUP

We implement the two-stage training approach as described in Section 3.5. Table 3 summarizes the key hyperparameters used for both stages across different datasets.

For Stage 1, we utilize a pretrained CIFAR-10 model from RATIO (Augustin et al., 2020) and pretrained ImageNet models from Salman et al. (2020), while training our own CIFAR-100 model following the RATIO methodology with the hyperparameters specified in Table 3. We select the EMA model with the best robust test accuracy as the final Stage 1 model.

For Stage 2, we initialize from the Stage 1 model and continue training with batch normalization disabled by setting all BN modules to evaluation mode. During this stage, we optimize the complete objective function $\mathcal{L}(\theta) = \mathcal{L}_{\text{AT-CE}}(\theta) + \mathcal{L}_{\text{BCE}}(\theta)$ using fixed learning rates as specified in Table 3. The discriminative component $\mathcal{L}_{\text{AT-CE}}(\theta)$ continues to use the same adversarial settings as Stage 1 (see Table 4), while the generative component $\mathcal{L}_{\text{BCE}}(\theta)$ employs the parameters detailed in Table 5.

We select the Stage 2 checkpoint with the best FID score for the final evaluation reported in Section 4.3.

Table 3: Training hyperparameters for both stages.

	CIFAR-10/100	ImageNet
Architecture	WideResNet-34-10	ResNet-50/WideResNet-50-4
Optimizer	SGD with Nesterov (momentum=0.9)	SGD with Nesterov (momentum=0.9)
Weight decay	5×10^{-4}	1×10^{-4} (Stage 1), 5×10^{-4} (Stage 2)
Batch size	128	512
EMA (Gowal et al., 2020)	Yes	Yes
Learning rate (Stage 1)	0.1 (cosine schedule, 300 epochs)	0.1 (step decay at epochs 30, 60, 90)
Learning rate (Stage 2)	0.001 (CIFAR-10), 0.01 (CIFAR-100)	0.001
Batch normalization (Stage 1)	Enabled (training mode)	Enabled (training mode)
Batch normalization (Stage 2)	Disabled (eval mode)	Disabled (eval mode)

Table 4: Adversarial training parameters for $\mathcal{L}_{\text{AT-CE}}$ (identical across Stage 1 and Stage 2).

	CIFAR-10/100	ImageNet
PGD steps	10	2
PGD step size	0.1	2.0
L_2 perturbation bound	0.5	3.0

Table 5: Adversarial training parameters for $\mathcal{L}_{\text{BCE}}$ (Stage 2 only).

	CIFAR-10/100	ImageNet
Max PGD steps (T)	45	15/30/70
PGD step size	0.1	2.0
L_2 perturbation bound	None (unconstrained)	None (unconstrained)
OOD data source	80M Tiny Images (Torralba et al., 2008)	Open Images (Krasin et al., 2016)

A.4.2 DATA AUGMENTATION DETAILS

As described in Section 3.6, we implement separate data augmentation pipelines for the discriminative and generative components of our objective function. Table 6 summarizes these dataset-specific augmentation strategies. Note that augmentation strategies for $\mathcal{L}_{\text{AT-CE}}$ are identical to those used by RATIO (Augustin et al., 2020) for CIFAR-10/CIFAR-100, and identical to Salman et al. (2020) for ImageNet. The effects of CIFAR-10 augmentations are illustrated in Figure 2.

Figure 3 illustrates CIFAR-10 training curves from Stage 2 joint training with various augmentation strategies applied to $\mathcal{L}_{\text{BCE}}$ (while consistently using AutoAugment with Cutout for $\mathcal{L}_{\text{AT-CE}}$). Interestingly, the choice of augmentation for the generative component influences discriminative performance as well, as evidenced by the decline in robust test accuracy when using no augmentation. The best FID performance is achieved by no augmentation and random cropping with padding, which minimally distort the underlying data distribution p_{data}. Overall we find random cropping with padding provides the optimal balance between discriminative and generative performances.

Table 6: Data augmentation strategies for discriminative and generative components.

Dataset	Component	Augmentation Strategy
CIFAR-10/100	$\mathcal{L}_{\text{AT-CE}}$	AutoAugment + Cutout + RandomHorizontalFlip()
	$\mathcal{L}_{\text{BCE}}$	RandomCrop(32, padding=4) + RandomHorizontalFlip()
ImageNet	$\mathcal{L}_{\text{AT-CE}}$	RandomResizedCrop(224) + RandomHorizontalFlip()
	$\mathcal{L}_{\text{BCE}}$	Resize(224) + CenterCrop(224) + RandomHorizontalFlip()

Figure 2: Samples produced by different augmentations on CIFAR-10.

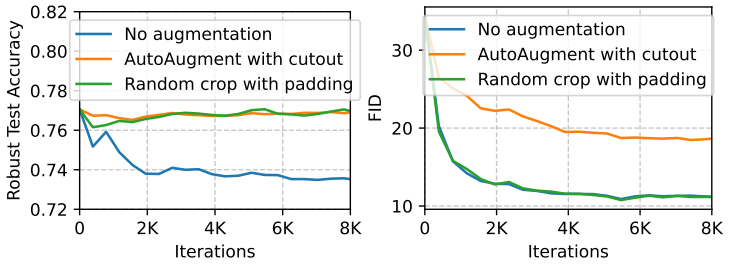


Figure 3: Training curves under different data augmentations during stage 2 joint training.

A.5 Additional analyses

A.5.1 Individual contributions of generative loss and decoupled augmentation

To analyze the individual contribution of our primary contributions, we conduct an ablation study with the following variants on CIFAR-10:

- Standard AT: A baseline adversarially trained model without a generative component.
- DAT with uniform augmentation: Our DAT approach that applies the same aggressive augmentation for both the discriminative and generative objectives.
- DAT with decoupled augmentation: Our DAT approach that applies aggressive augmentation to the discriminative term and mild augmentation for the generative term.

The results in Table 7 demonstrate the impacts of the AT-based generative loss and decoupled augmentation. Introducing the generative loss component to the standard AT baseline significantly reduces the FID from 33.04 to 15.35, while robust accuracy remains comparable. The subsequent application of a decoupled augmentation strategy yields a further reduction in FID to 9.07.

Since both our approach and RATIO extend a standard AT baseline with an objective function that leverages out-of-distribution (OOD) data, it is instructive to compare their relative efficacy in enhancing generative fidelity. The RATIO objective, which is formulated for robust OOD detection, reduces the FID from 33.04 to 21.96. In contrast, our generative objective provides a much larger improvement, lowering the FID to 15.35. This comparison confirms that for the goal of enhancing sample quality, a dedicated generative loss is more effective than an auxiliary loss designed for OOD detection.

Table 7: Effect of generative loss and augmentation on CIFAR-10.

Method	Acc% ↑	Robust Acc% ↑	FID ↓
Standard AT	92.34	75.73	33.04
DAT (uniform aug)	92.68	75.93	15.35
DAT (decoupled aug)	91.86	75.66	9.07
RATIO (Augustin et al., 2020)	92.23	76.25	21.96

A.5.2 Effect of OOD dataset size

The out-of-distribution (OOD) dataset is a critical component of our training framework, as it provides the initialization samples for computing the negative samples in the generative loss term. The influence of this dataset can be understood through the EBM learning mechanism: a more diverse OOD dataset provides better coverage of the input space, allowing the PGD attack (acting as an MCMC sampler) to discover a broader range of spurious modes in the current energy landscape. These discovered modes are then eliminated as the main objective function is optimized. Given this crucial role, the diversity and scale of the OOD dataset are expected to influence model performance.

To examine the impact of OOD dataset size, we conducted an ablation study on ImageNet using our DAT ResNet-50 ($T = 15$) model with varying OOD dataset sizes: 1K, 10K, 100K, and the full 300K samples. As shown in Table 8, the FID score improves modestly from 6.96 with 1K samples to 6.64 with 300K samples. Classification accuracy remains stable across all dataset sizes, with similar robustness levels, indicating that the OOD dataset size primarily affects generation quality rather than discriminative performance.

These results demonstrate notable data efficiency, with only modest improvements when scaling from 1K to 300K OOD samples. A contributing factor to this efficiency is data augmentation: we employ RandomResizedCrop with scale=(0.08, 1.0) and aspect ratio=(0.75, 1.33), which can crop as little as 8% of the original image with varying aspect ratios, potentially amplifying the effective diversity of each sample. To investigate the contribution of augmentation, we include a baseline using 1K OOD samples without data augmentation. While augmentation provides clear benefits—improving FID from 8.00 to 6.96—even without augmentation, our approach substantially outperforms standard AT

in generation quality (FID 8.00 vs. 15.97). This indicates the data efficiency can be largely attributed to the AT-based EBM learning mechanism itself, where the PGD attack can effectively explore the energy landscape even from limited initialization points.

Table 8: Impact of OOD dataset size on ImageNet performance for DAT ResNet-50 ($T = 15$).

Method	Acc% ↑	Robust Acc% ↑	FID ↓	IS ↑
Standard AT	62.83	34.44	15.97	274.90
DAT 1K w/o aug	57.50	33.80	8.00	320.64
DAT 1K	57.56	34.22	6.94	324.23
DAT 10K	57.82	34.70	6.84	320.78
DAT 100K	58.19	34.88	6.70	322.10
DAT 300K	57.88	34.84	6.64	339.55

A.5.3 GENERATIVE-DISCRIMINATIVE TRADE-OFF VIA LOSS WEIGHTING

Our training objective, $\mathcal{L}(\theta) = \mathcal{L}_{\text{AT-CE}}(\theta) + \mathcal{L}_{\text{BCE}}(\theta)$, is a composite of a discriminative and a generative loss. This structure naturally raises the question of whether it is possible to trade off between these two capabilities by adjusting the relative weight of each component. To investigate this possibility, we perform experiments on CIFAR-10 with three distinct weighting configurations:

- Standard loss (equal weighting): $\mathcal{L}(\theta) = \mathcal{L}_{\text{AT-CE}}(\theta) + \mathcal{L}_{\text{BCE}}(\theta)$
- Emphasize generative modeling: $\mathcal{L}(\theta) = 0.6 \cdot \mathcal{L}_{\text{AT-CE}}(\theta) + 1.4 \cdot \mathcal{L}_{\text{BCE}}(\theta)$
- Emphasize classification: $\mathcal{L}(\theta) = 1.4 \cdot \mathcal{L}_{\text{AT-CE}}(\theta) + 0.6 \cdot \mathcal{L}_{\text{BCE}}(\theta)$

The results in Table 9 confirm that the balance between generative and discriminative performance can be tuned by adjusting the loss term weights. Emphasizing the generative component improves FID at the cost of slightly reduced classification performance, while emphasizing classification achieves the opposite effect. However, we note that our standard, unweighted loss corresponds to the natural factorization of the joint log-likelihood in the original JEM formulation: $\log p_\theta(x, y) = \log p_\theta(y|x) + \log p_\theta(x)$. This suggests that equal weighting is a principled default that performs well without requiring additional hyperparameter tuning.

Table 9: Trading off generative and discriminative performance by weighting loss terms.

Method	Acc% ↑	Robust Acc% ↑	FID ↓
Standard loss	91.88	75.73	9.09
Emphasize generative modeling	91.16	75.11	8.77
Emphasize classification	92.52	75.97	10.02

A.6 GENERATIVE PERFORMANCE EVALUATION

We evaluate generative performance using Fréchet Inception Distance (FID) and Inception Score (IS). FID is computed between 50K class-balanced generated samples and the full training set, while IS is computed on the same set of 50K generated samples.

Conditional generation. We generate an equal number of samples for each class. To generate samples for a given class y, we first sample an OOD data point x from the corresponding OOD data source, and then perform T steps of PGD attack according to:

$$x_{t+1} = x_t + \eta \frac{\nabla_x(-E_\theta(x_t, y))}{||\nabla_x(-E_\theta(x_t, y))||_2} \tag{22}$$

where T is the number of PGD steps and η is the corresponding step size (see Table 10).

Unconditional generation. For unconditional generation, we directly sample from the marginal distribution using PGD according to Eq. 8:

$$x_{t+1} = x_t + \eta \frac{\nabla_x(-E_\theta(x_t))}{||\nabla_x(-E_\theta(x_t))||_2}$$

The FID results for both conditional and unconditional generation across all datasets are presented in Table 11. We find conditional generation consistently outperforms unconditional generation across all the datasets.

Table 10: Sample generation parameters for FID and IS evaluation. The number of PGD steps for each model and dataset combination is determined through grid search.

Model	Dataset	PGD steps (T)	Step size	OOD data source
DAT	CIFAR-10 ($T = 40$)	33	0.2	80M Tiny Images
	CIFAR-10 ($T = 50$)	35	0.2	80M Tiny Images
	CIFAR-100 ($T = 45$)	32	0.2	80M Tiny Images
	CIFAR-100 ($T = 50$)	33	0.2	80M Tiny Images
	ImageNet (ResNet-50, $T = 15$)	13	8.0	Open Images
	ImageNet (ResNet-50, $T = 30$)	16	8.0	Open Images
	ImageNet (WRN-50-4, $T = 30$)	17	8.0	Open Images
	ImageNet (WRN-50-4, $T = 65$)	20	8.0	Open Images
RATIO	CIFAR-10	31	0.2	80M Tiny Images
	CIFAR-100	14	0.2	80M Tiny Images
Standard AT	CIFAR-10	22	0.2	80M Tiny Images
	CIFAR-100	15	0.2	80M Tiny Images
	ImageNet (ResNet-50)	13	8.0	Open Images
	ImageNet (WRN-50-4)	11	8.0	Open Images

Table 11: FIDs of conditional and unconditional generation of our approach.

	CIFAR-10	CIFAR-100	ImageNet
Conditional generation	9.07	10.70	6.64
Unconditional generation	20.57	13.56	18.67

A.7 OUT-OF-DISTRIBUTION DETECTION

We evaluate both standard out-of-distribution (OOD) detection performance and worst-case OOD detection under adversarial perturbations. For standard OOD detection, we measure the AUROC scores between in-distribution test samples and unmodified OOD samples. For worst-case detection, we evaluate against adversarially perturbed OOD samples specifically optimized to maximize the OOD detection function output. Results are computed using all the in-distribution test samples and 1024 out-distribution samples. For generating adversarial OOD samples, we use L_2-based perturbation limit of 1.0 for CIFAR-10/100 and 3.0 for ImageNet.

Energy-based detection. We use an energy-based function $s_\theta(x) = -E_\theta(x)$, which is proportional to $\log p_\theta(x)$ up to an additive constant. To find adversarial OOD inputs for this function, we employ a PGD attack to maximize the negative energy:

$$x_{adv} = \underset{x' \in B(x,\epsilon_o)}{\arg\max} \; -E_\theta(x') \tag{23}$$

where x is a clean OOD input and $B(x, \epsilon_o)$ represents an L_2-ball of radius ϵ_o centered at x.

Maximum confidence detection. We employ a maximum confidence function $s_\theta(x) = \max_y p_\theta(y|x)$ that uses the confidence in the most likely class (also used by RATIO (Augustin et al., 2020)). For this detection function, following RATIO (Augustin et al., 2020), we compute adversarial OOD inputs by maximizing the cross-entropy loss against a uniform distribution:

$$x_{adv} = \underset{x' \in B(x,\epsilon_o)}{\arg\max} \; \mathcal{L}_{\text{CE}}(\theta; x', \mathbf{1}/K) \tag{24}$$

where $\mathbf{1}/K$ represents a uniform distribution over all K classes. Maximizing this loss encourages the model to produce a non-uniform (confident) prediction, thereby maximizing the detection function.

Table 12 presents a comparison of the above two OOD detection functions. The results reveal complementary strengths: the energy-based function ($-E_\theta(x)$) achieves near-perfect AUROC scores on uniform noise detection, while the maximum confidence function ($\max_y p_\theta(y|x)$) demonstrates superior performance on natural image OOD datasets. Based on these findings, we adopt the maximum confidence score for subsequent comparisons with other methods.

Table 13 presents comparative results across different baselines. Notably, our DAT model achieves OOD detection performance comparable to standard AT on natural image datasets (CIFAR-100, SVHN), despite incorporating an additional OOD dataset during training. This observation suggests that our generative training component primarily enhances generation quality rather than improving OOD detection capabilities beyond those provided by standard adversarial training.

Compared to RATIO, our model exhibits lower OOD detection performance across most datasets. To investigate whether this gap stems from our use of milder augmentations for the generative component, we trained an ablation model that applies RATIO's aggressive augmentation strategy to both loss terms. The results show that this variant performs similarly to our standard DAT model and still underperforms RATIO. This finding indicates that the performance gap is not primarily caused by the augmentation strategy but rather by the fundamental differences in the training objectives: RATIO's loss is explicitly optimizes for OOD detection performance, while our generative loss prioritizes learning an accurate energy function for generation.

A potential avenue for addressing this limitation involves developing a hybrid objective that combines our generative loss with RATIO's explicit OOD detection term. This approach is theoretically motivated by the complementary nature of these objectives: our generative component learns to model the energy landscape of in-distribution data while naturally assigning low probability to out-of-distribution regions, which aligns conceptually with RATIO's strategy of enforcing low confidence predictions in neighborhoods around OOD samples. Such a hybrid formulation could potentially preserve the generative modeling benefits of our approach while recovering the superior OOD detection performance of RATIO. Future work could explore this direction by investigating appropriate weighting strategies between the generative and OOD detection terms to achieve optimal performance across both objectives.

Table 12: Comparison of OOD detection functions on CIFAR-10.

	CIFAR-100		SVHN		Uniform noise	
Method	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
DAT ($\max_y p_\theta(y\|x)$)	0.8709	0.6480	0.9609	0.8334	0.8922	0.8257
DAT ($-E_\theta(x)$)	0.8484	0.6647	0.8011	0.6046	0.9995	0.9983

Table 13: OOD detection performance (AUROC) with CIFAR-10 as ID dataset (JEM results are from Augustin et al. (2020)). All methods use the maximum confidence detection function $s_\theta(x) = \max_y p_\theta(y|x)$.

	CIFAR-100		SVHN		Uniform noise	
Method	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
JEM	0.8760	0.1920	0.8930	0.0730	0.1180	0.0250
Standard AT	0.8759	0.6364	0.9625	0.8306	0.8501	0.7902
DAT ($T = 40$, uniform aug)	0.8751	0.6261	0.9642	0.8303	0.9546	0.9254
DAT ($T = 40$)	0.8709	0.6480	0.9609	0.8334	0.8922	0.8257
RATIO	0.9157	0.7516	0.9843	0.9130	0.9999	0.9999

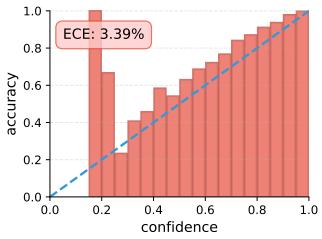
Table 14: OOD detection performance (AUROC) with CIFAR-100 as ID dataset.

	CIFAR-10		SVHN		Uniform noise	
Method	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
Standard AT	0.7430	0.4093	0.8700	0.4863	0.7858	0.5048
RATIO	0.7320	0.3795	0.8439	0.4356	0.7769	0.5881
DAT ($T = 45$)	0.7027	0.5145	0.8271	0.5823	0.4024	0.2283

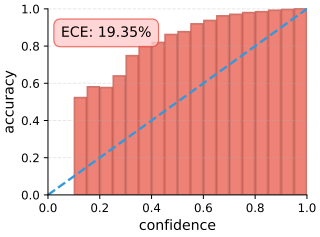
Table 15: OOD detection performance (AUROC) with ImageNet as ID dataset.

	CIFAR-10		SVHN		Uniform noise	
Method	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
Standard AT (ResNet-50)	0.7235	0.5304	0.9239	0.8089	0.8678	0.8377
DAT (ResNet-50, $T = 15$)	0.6599	0.4870	0.8813	0.7754	0.6899	0.6268

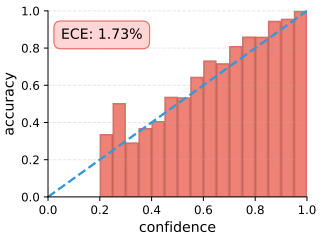
A.8 CALIBRATION



(a) Standard AT

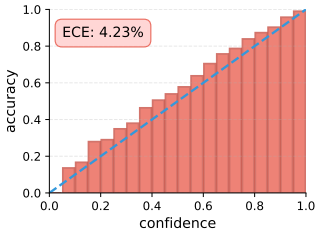


(b) RATIO

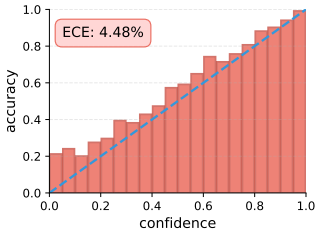


(c) DAT

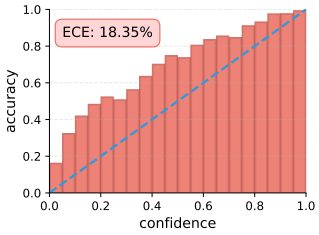
Figure 4: Calibration diagrams on CIFAR-10 (without temperature scaling).



(a) Standard AT

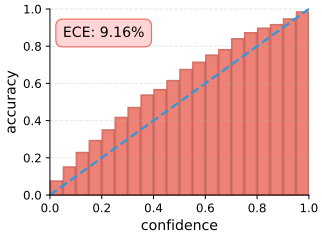


(b) RATIO

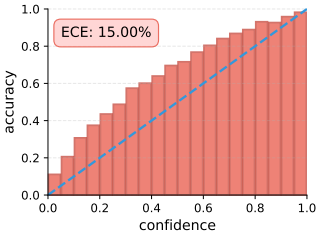


(c) DAT

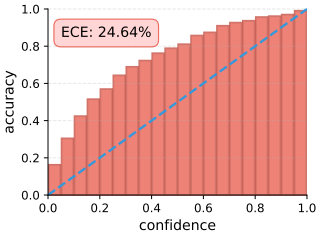
Figure 5: Calibration diagrams on CIFAR-100 (without temperature scaling).



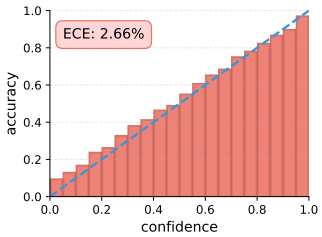
(a) Standard AT (ResNet-50)



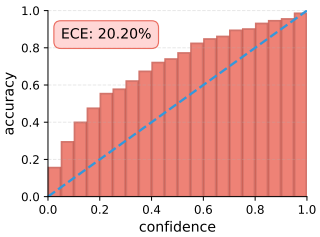
(b) DAT (ResNet-50, $K = 15$)



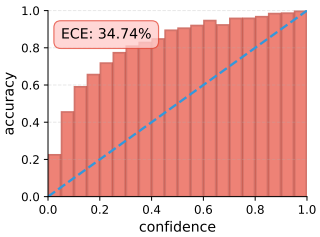
(c) DAT (ResNet-50, $K = 30$)



(d) Standard AT (WRN-50-4)



(e) DAT (WRN-50-4, $K = 30$)



(f) DAT (WRN-50-4, $K = 65$)

Figure 6: Calibration diagrams on ImageNet (without temperature scaling).

A.9 COUNTERFACTUAL GENERATION

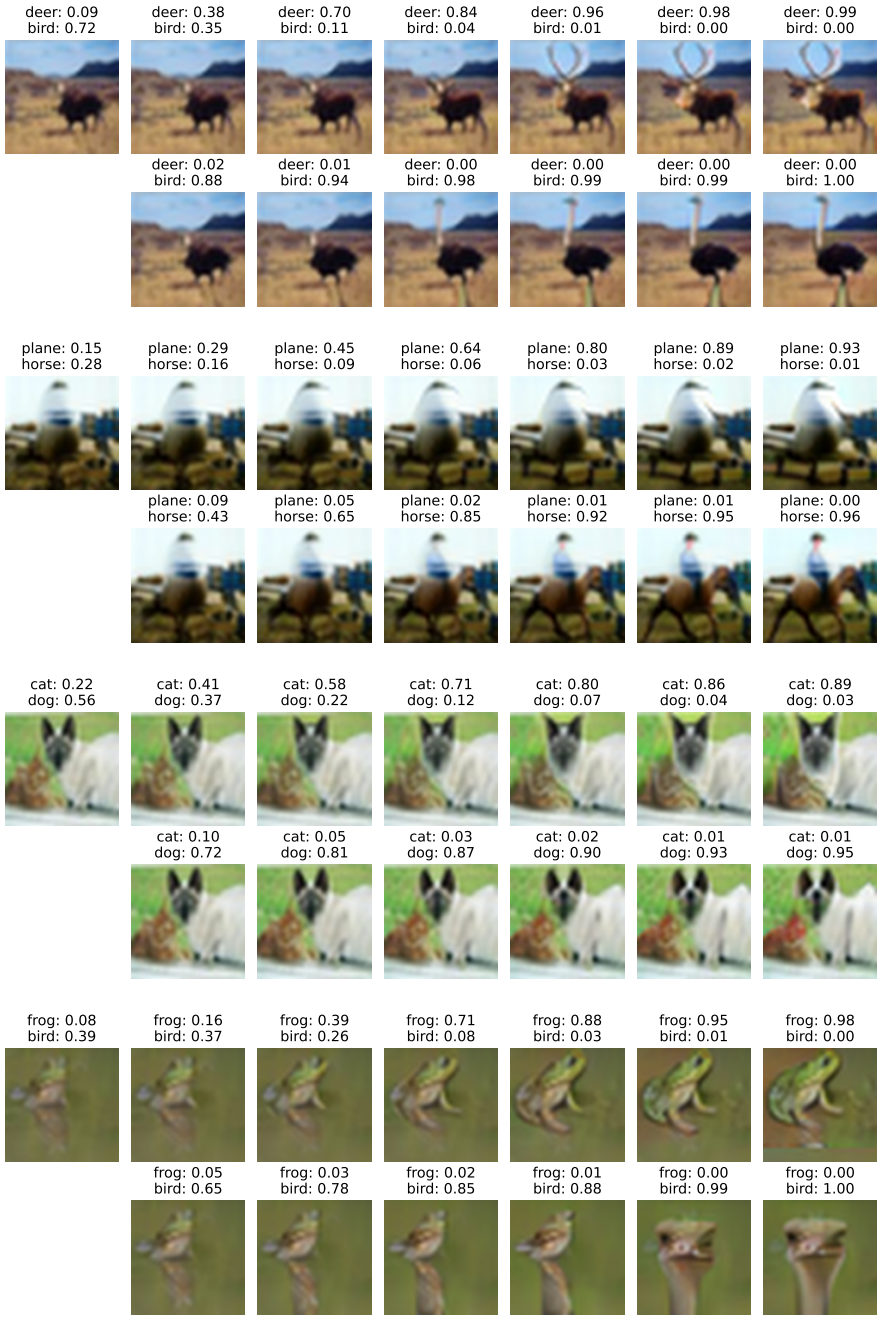


Figure 7: CIFAR-10 counterfactual examples with perturbation limits of $0.5, 1.0, 1.5, 2.0, 2.5, 3.0$. These figures display counterfactuals and corresponding classifier confidences for both the correct class (top row) and a target wrong class (bottom row). As the perturbation budget increases from left to right, the generated counterfactuals progressively resemble samples from the target class distribution while the target class confidence correspondingly increases, demonstrating that our model effectively captures the distributions of different classes and can generate meaningful class-to-class transformations.

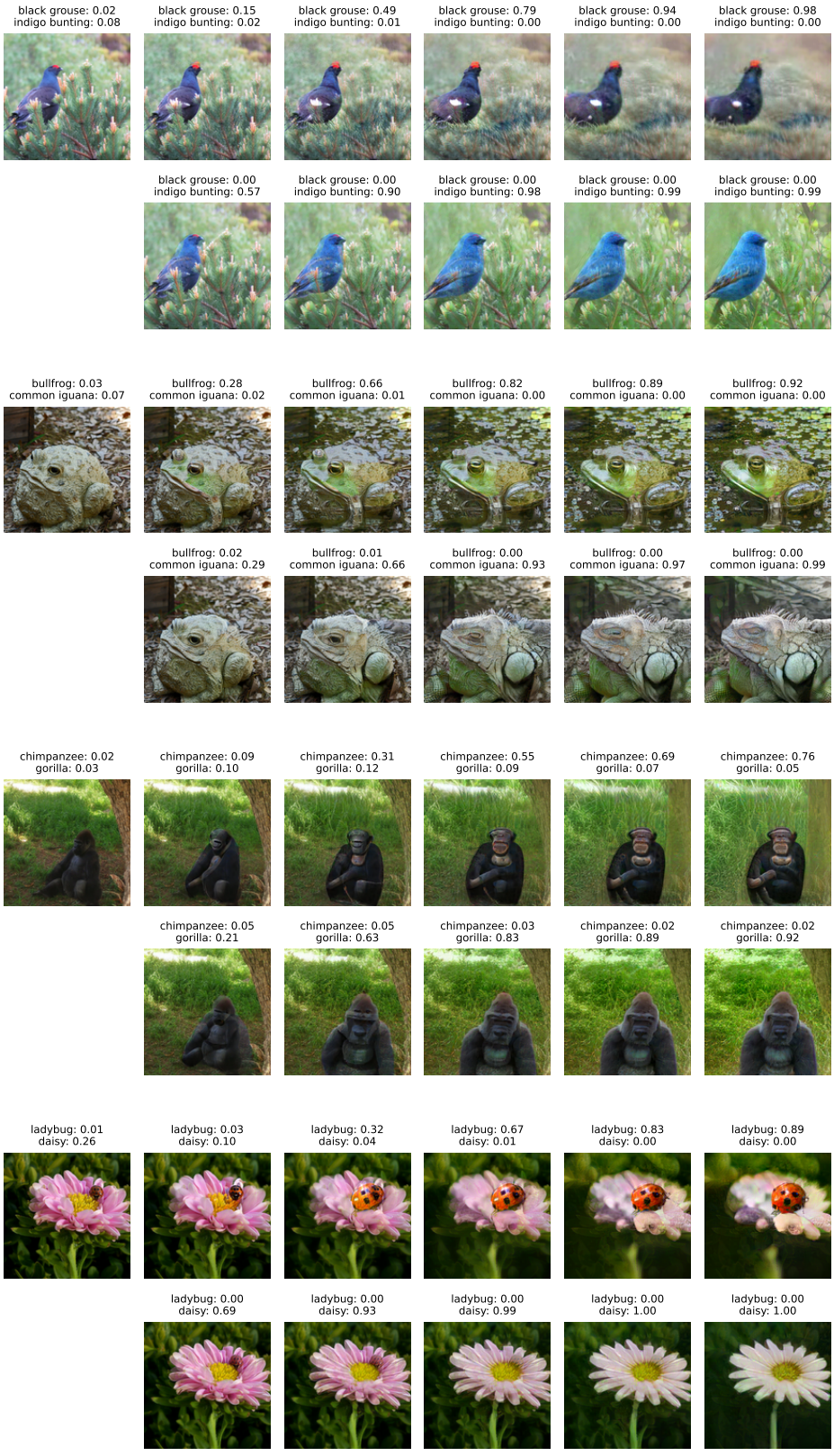
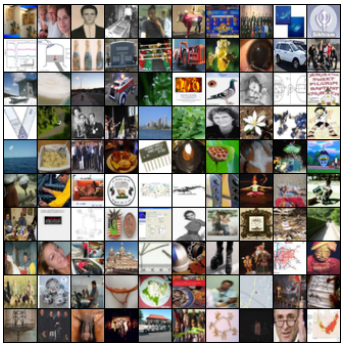
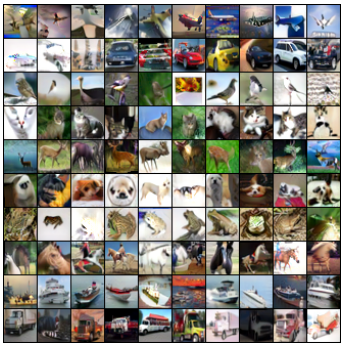


Figure 8: ImageNet counterfactual examples with perturbations limits of $10., 20., 30., 40., 50.$

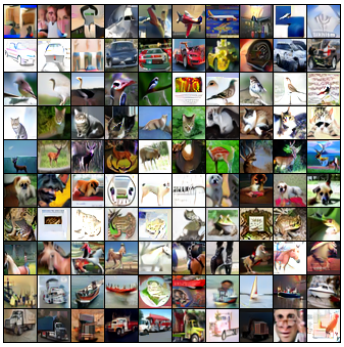
A.10 GENERATION RESULTS



(a) Seed images used for producing the generated samples.

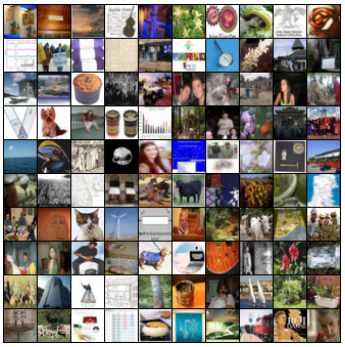


(b) Uncurated conditional samples of DAT ($T = 50$).

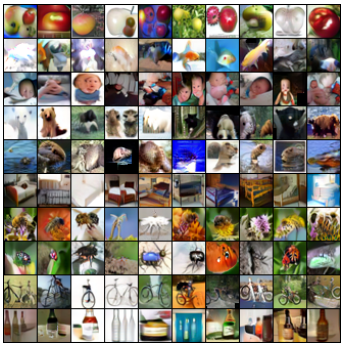


(c) Uncurated conditional samples of RATIO.

Figure 5: CIFAR-10 class-conditional generation results. Note that some samples from the RATIO baseline show potential artifacts (e.g., saturated or unnatural colors) possibly linked to the aggressive AutoAugment policy used for model training.



(a) Seed images used for producing the generated samples.

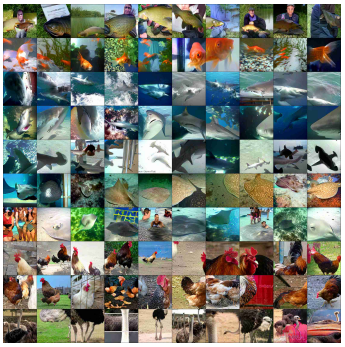


(b) Uncurated conditional samples of DAT ($T = 50$).

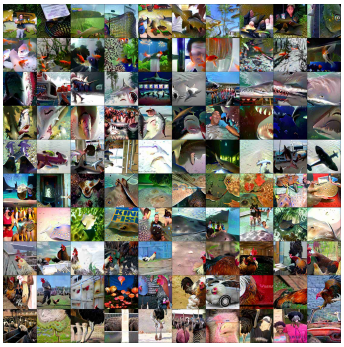
(c) Uncurated conditional samples of RATIO.

Figure 6: CIFAR-100 conditional generation results.

(a) Seed images used for producing the generated samples.



(b) Uncurated conditional samples of DAT (WRN-50-4 $T = 65$).



(c) Uncurated conditional samples of standard AT (WRN-50-4).

Figure 7: ImageNet class-conditional generation results for the first 10 classes: tench, goldfish, great white shark, tiger shark, hammerhead, electric ray, stingray, cock, hen, ostrich (images are in 224×224 resolution).

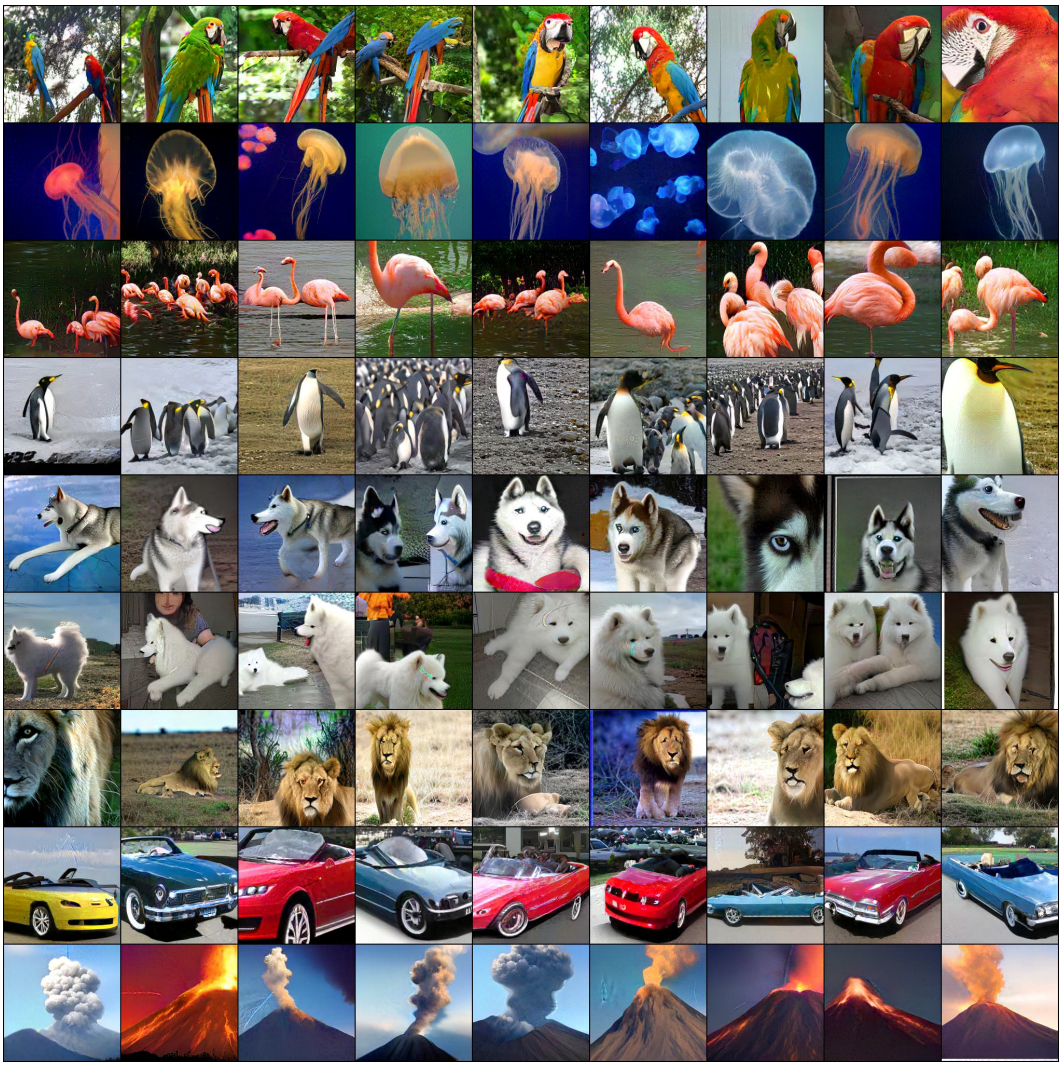


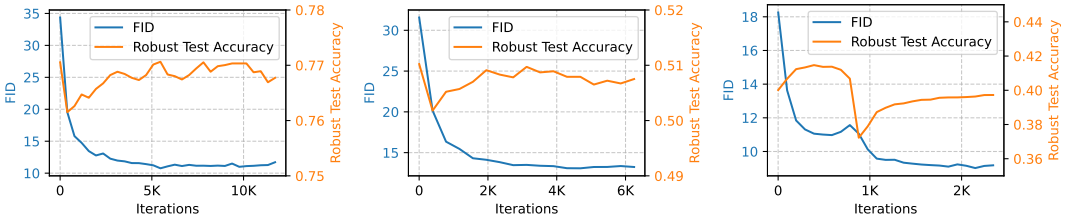
Figure 8: Selected ImageNet conditional generation results for class 88 (macaw), 107 (jellyfish), 130 (flamingo), 145 (king penguin), 248 (husky), 258 (Samoyed), 291 (lion), 511 (convertible), and 980 (volcano). Results are generated with DAT (WRN-50-4, $T = 65$) at 224×224 resolution.

A.11 VARIABILITY OF DAT PERFORMANCE ACROSS DATASETS

Table 16: Mean and standard deviation of DAT performance across datasets, computed over three independent runs with different random seeds.

Dataset	Acc% ↑	Robust Acc% ↑	IS ↑	FID ↓
CIFAR-10 ($T = 40$)	91.86 ± 0.03	75.66 ± 0.01	9.96 ± 0.02	9.07 ± 0.03
CIFAR-100 ($T = 45$)	65.55 ± 0.62	45.97 ± 0.49	10.83 ± 0.11	10.70 ± 0.22
ImageNet (ResNet-50, $T = 15$)	57.91 ± 0.09	34.87 ± 0.09	334.11 ± 8.96	6.60 ± 0.08

A.12 TRAINING CURVES



(a) CIFAR-10 ($T = 40$) (b) CIFAR-100 ($T = 45$) (c) ImageNet (ResNet-50, $T = 15$)

Figure 9: Training curves from Stage 2 joint training demonstrating substantial FID score improvements while preserving Stage 1 robust test accuracy (evaluated via PGD attacks; FID measured using 10K generated samples).