Joint Discriminative-Generative Modeling via Dual Adversarial Training

Xuwang Yin
Independent Researcher
xuwangyin@gmail.com

Claire Zhang
MIT
clairefz@mit.edu

Julie Steele
MIT
jssteele@mit.edu

Tony T. Wang
MIT
twang6@mit.edu

Abstract

Developing models that excel simultaneously at robust classification and high-fidelity generative modeling remains a significant challenge. While hybrid approaches like Joint Energy-Based Models (JEM) offer a path by interpreting classifiers as energy-based models (EBMs), they often rely on SGLD-based training for the generative component, which suffers from instability and poor sample quality. To address this, we propose a novel training framework that integrates adversarial training principles for both discriminative robustness and stable generative learning within a unified JEM-based architecture. Our approach introduces two key innovations: (1) replacing traditional SGLD-based EBM learning with a more stable AT-based strategy that optimizes the energy function using a Binary Cross-Entropy objective discriminating real data from contrastive samples generated via PGD attacks, and (2) a two-stage training procedure with decoupled data augmentation strategies for the discriminative and generative components. Extensive experiments across CIFAR10, CIFAR100, and RestrictedImageNet datasets demonstrate that our method consistently maintains competitive robust accuracy while substantially improving generative quality compared to existing hybrid models. In addition, our model's improved generative capabilities directly transfer to producing higher quality counterfactual examples, which contributes to better model explainability. Our work presents a promising direction for building robust, stable, and high-performing joint discriminative and generative models.

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) [1] 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

Preprint. Under review.

on Markov Chain Monte Carlo (MCMC) methods such as Stochastic Gradient Langevin Dynamics (SGLD) for training the generative component. SGLD specifically suffers from significant training instabilities, computational inefficiency, and often produces poor-quality samples [1, 2, 3, 4, 5, 6], limiting the practical adoption of these hybrid models.

We address these limitations by introducing 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 [7] that replaces unstable SGLD sampling with more efficient and stable Projected Gradient Descent (PGD) attacks.

Our key technical contributions include:

1. A stable AT-based alternative to traditional SGLD-based JEM learning that optimizes the energy function through minimizing Binary Cross-Entropy using PGD-generated contrastive samples, significantly improving training stability and sample quality.
2. A decoupled data augmentation strategy that applies different transformations to samples used for discriminative and generative components, addressing the inherent conflict between augmentations that benefit robust classification and those appropriate for generative modeling.
3. A two-stage training procedure that effectively addresses the incompatibility between batch normalization and sampling-based EBM learning, enabling stable optimization across both tasks.

Extensive experiments across datasets of increasing complexity (CIFAR10, CIFAR100, and RestrictedImageNet) demonstrate that our approach scales effectively while maintaining competitive adversarial robustness and substantially improving generative performance compared to existing hybrid models. Furthermore, our model's improved generative capabilities directly translate to producing higher quality counterfactual explanations, enhancing model explainability.

Our work not only addresses the practical limitations of current JEM frameworks but also demonstrates that adversarial training principles — typically viewed solely through the lens of robustness — can be effectively leveraged to enhance generative modeling capabilities. This approach represents a step toward developing more unified models that can perform well at both tasks without requiring separate architectures or training procedures.

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. [8] showed how generative ConvNets could be derived from discriminative ones, framing them as EBMs. Grathwohl et al. [1] 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. While focusing on scalable EBM training, Du and Mordatch [3] also demonstrated that such implicitly generative EBMs can achieve strong performance on discriminative tasks like adversarially robust classification and out-of-distribution detection. 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. [9], Jin et al. [10], and Lee et al. [11]. Flow-based models have also been explored for hybrid tasks; for instance, Residual Flows [12] 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. [1] 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, the energy $E_\theta(x, y)$ was defined 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, combining the standard cross-entropy loss for $p(y|x)$ with an EBM-based objective for $p(x)$ optimized using Stochastic Gradient Langevin Dynamics (SGLD) [13]. Through extensive experiments, Grathwohl et al. [1] demonstrated that this joint training approach allowed 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.

Learning EBMs with adversarial training Yin et al. [7] 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 high classifier probabilities (low energy) to in-distribution regions. The PGD attack used in AT to generate adversarial samples from OOD data was interpreted as a non-convergent sampler producing contrastive data, analogous to MCMC sampling in EBM training. While the resulting energy function primarily captures the data manifold rather than the exact density, their model achieves competitive image generation performance compared to explicit EBMs. Notably, this AT-based EBM learning approach was found to be significantly more stable than traditional MCMC-based EBM training and demonstrated strong performance in worst-case out-of-distribution detection, similar to methods like RATIO [14].

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. [14] 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. [14] 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. The 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.

3 Method

3.1 JEM generative modeling with adversarial training

Our approach builds upon the Joint Energy-Based Model (JEM) framework introduced by Grathwohl et al. [1], 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 using the labeled samples from $p_{\text{data}}(x, y)$. 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 of the inputs x, obtained by marginalizing y from the joint empirical distribution $p_{\text{data}}(x, y)$. This gradient has an intuitive interpretation: it decreases the energy (increases the probability) of real data samples $x \sim p_{\text{data}}(x)$ while increasing the energy (decreases the probability) of model-generated samples $x \sim p_\theta(x)$. At equilibrium, when $p_\theta(x) = p_{\text{data}}(x)$, these two terms balance out and the gradient becomes zero.

The expectation $\mathbb{E}_{x \sim p_\theta(x)}[\cdot]$ over the model distribution is approximated using MCMC methods, specifically Stochastic Gradient Langevin Dynamics (SGLD) [13]. SGLD generates samples x starting from some initial distribution $p_0(x)$ (e.g., uniform noise) and iteratively applying 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 the gradient $\nabla_x E_\theta(x_t)$ is taken with respect to the marginal energy function defined in Eq. 4.

While the JEM framework successfully integrates generative modeling into classifiers, its reliance on SGLD sampling for optimizing $\log p_\theta(x)$ introduces significant training instabilities [1, 2] and often results in poor sample quality. Our key innovation addresses these limitations by replacing the SGLD-based component with an adversarial training (AT) approach inspired by Yin et al. [7].

Specifically, we replace the standard EBM gradient (Eq. 5) with the following approximation [7]:

$$\nabla_\theta \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log p_\theta(x)] \approx \mathbb{E}_{x \sim p_{\text{data}}(x)}[-\alpha(x)\nabla_\theta E_\theta(x)] - \mathbb{E}_{x \sim p_\theta(x)}[-\beta(x)\nabla_\theta E_\theta(x)] \tag{7}$$

where $\alpha(x) = 1 - \sigma(-E_\theta(x))$ and $\beta(x) = \sigma(-E_\theta(x))$ are data-dependent scaling factors, and $\sigma(\cdot)$ is the sigmoid function. This formulation maintains the same structure as Eq. 5, but with adaptive scaling factors that modulate the gradient contributions based on the model's current energy assignments. The resulting EBM can only recover the support of $p_{\text{data}}(x)$, but in practice it is stable to train and has competitive generative modeling performance compared to standard EBMs.

In addition to the above gradient substitution, the sampling required to estimate $\mathbb{E}_{x \sim p_\theta(x)}[\cdot]$ is performed using the Projected Gradient Descent (PGD) attack [15] 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 [16] for CIFAR10 training) and performing multiple iterations of gradient ascent on the negative energy function $-E_\theta(x)$:

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

where $E_\theta(x)$ is the marginal energy function defined in Eq. 4, and η is the step size. 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 to learning EBMs doesn't have the training stability issues that plague SGLD-based methods and produces higher quality samples.

3.2 Enhancing discriminative performance through adversarial training

While our AT-based approach improves the generative capabilities of the JEM framework, 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 conditional 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 [15], generating adversarial examples through iterative gradient steps within the constraint set. The robust classification loss is then defined as:

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

This approach allows our model to maintain high classification accuracy even under adversarial perturbations, complementing the improved generative capabilities of our AT-based JEM framework.

3.3 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 dual-AT approach simultaneously enhances the model's discriminative robustness and generative capabilities, addressing the key limitations of the original JEM framework.

Our approach shares conceptual similarities with RATIO [14], 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 [7], 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.

The complete training procedure for our combined objective (Eq. 12) can be found 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 EBMs, there are broadly two strategies for drawing samples from $p_\theta(x)$ (see similar discussion in Grathwohl et al. [1]):

1. Direct sampling from the marginal distribution by using gradient-based MCMC (like SGLD or PGD) on the marginal energy function $E_\theta(x) = -\log \sum_y \exp(f_\theta(x)[y])$. This was the approach implied in our initial formulation of PGD for EBMs in Eq. 8.
2. Ancestral sampling: first sample a label $y \sim p_{\text{data}}(y)$, then sample $x \sim p_\theta(x|y)$ by performing gradient-based MCMC on the joint energy function $E_\theta(x, y) = -f_\theta(x)[y]$.

While both approaches can, in principle, produce fair samples to estimate the necessary expectations, we found ancestral sampling to be practically superior for training stability, as training with direct sampling from the marginal distribution often leads to divergence. The stability of ancestral sampling likely due to several factors: (1) ancestral sampling provides a more focused learning signal for each class distribution, (2) it leverages the classifier's existing strong class representations, (3) it

exhibits better mode coverage and mixing properties than direct marginal sampling, and (4) it yields lower-variance gradient estimates, leading to more stable training.

In terms of test-time sample quality, we find ancestral sampling (conditional generation) yields substantially better FID than directly sampling from marginal distribution (unconditional generation) on CIFAR10 (9.07 vs. 20.57), moderately better FID on CIFAR100 (10.70 vs. 13.56), while both methods perform similarly on RestrictedImageNet with FID scores of approximately 64 (see Table 9).

Based on these findings, we adopt ancestral sampling in our implementation (Algorithm 1) for generating contrastive samples. Specifically, we first sample a label $y' \sim p_{\text{data}}(y)$, then generate a contrastive sample x_T by performing T iterations of projected gradient ascent 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 Dual-AT 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

3.4 Data augmentation decoupling

A key innovation of our approach is applying separate augmentation strategies to the discriminative and generative components. While existing joint models such as JEM typically use a single type of data augmentation for their joint objective, we identified a fundamental conflict between optimal augmentation strategies. Achieving robust classification often necessitates strong augmentations (e.g., AutoAugment [17] with Cutout [18] for CIFAR10 training), which significantly transform the input data to improve generalization against perturbations [19, 20]. However, these aggressive transformations can distort the underlying data distribution in ways detrimental to learning a generative model. For instance, applying the AutoAugment policy to the generative component in our CIFAR-10 experiments resulted in generated samples exhibiting artificial color shifts and much worse FIDs, indicating that the augmentation has distorted the true data distribution characteristics.

Our proposed training objective (Eq. 12), structured as a sum of a distinct adversarially robust classification loss ($\mathcal{L}_{\text{AT-CE}}$) and an AT-based EBM loss ($\mathcal{L}_{\text{BCE}}$), inherently enables the decoupling of augmentation strategies. This separation allows us to apply strong, robustness-enhancing augmentations (like AutoAugment + Cutout) specifically to the input samples for the $\mathcal{L}_{\text{AT-CE}}$ term, and employ much milder augmentations (e.g., only random cropping) for the samples for the generative $\mathcal{L}_{\text{BCE}}$ term, thereby preserving the data fidelity needed for learning a high-quality generative model. Our CIFAR10 experiment in Figure 2 quantitatively supports this benefit, showing improved FID scores when using milder augmentations for the generative component compared to aggressive ones.

3.5 Two-stage training

Another fundamental challenge in training joint models is the use of batch normalization (BN) [21]. While BN is highly beneficial for stabilizing standard deep network training [21], it is often found to interfere with the learning dynamics of EBMs and their sampling procedures [1, 7, 4, 22].

This incompatibility stems from a fundamental mismatch between BN's operating principles and EBM sampling dynamics. As noted by Zhao et al. [4], EBM training involves sampling steps where the distribution $p_\theta(x_t)$ continuously evolves throughout the sampling chain. The fixed statistics tracked by batch normalization (running means and variances) become progressively misaligned with these evolving distributions at different sampling steps, creating instability in the training process.

Consistent with these findings, we observed that enabling BN during joint training severely destabilized the optimization of the generative modeling term $\mathcal{L}_{\text{BCE}}$, leading to oscillating losses and failure to converge. Consequently, stable optimization of the generative component necessitates disabling BN. 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:

1. **Discriminative pre-training (with BN):** 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 focuses on leveraging the benefits of BN to achieve strong robust classification performance.
2. **Joint training (without BN):** After the robust pre-training converges, 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 [4, 23, 22], our experimental results 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 modeling, with minimal impact on the robust accuracy established in Stage 1.

4 Experiments

4.1 Training setup

We evaluate our approach on CIFAR10 [24], CIFAR100 [24], and RestrictedImageNet [25] (9 classes, 224×224 resolution).

For discriminative pre-training (Stage 1), we follow the methodology of RATIO [14]. Importantly, we enable batch normalization during this stage. For Stage 2 joint training, we utilize the best-performing model from Stage 1 and continue training with batch normalization disabled by setting the BN modules in the model to evaluation mode (while still using the BN statistics computed during Stage 1). Complete training details can be found in Appendix A.1.1.

During Stage 2 joint training, we employ separate data augmentation strategies for the two components of our objective function: for the robust discriminative training term $\mathcal{L}_{\text{AT-CE}}$, we utilize aggressive augmentations (identical to those used by RATIO), while for the generative modeling term $\mathcal{L}_{\text{BCE}}$, we apply only basic transformations to avoid distorting the underlying data distribution. Detailed augmentation specifications can be found in Appendix A.1.1.

For CIFAR experiments, we use a WideResNet-34-10 architecture following the official implementation of RATIO. We use the 80 million tiny images dataset [16] as the out-of-distribution dataset (p_{ood}) for CIFAR experiments, and for RestrictedImageNet, we employ a ResNet50 architecture with samples from the remaining ImageNet classes serving as p_{ood}. For CIFAR10 comparisons, we use RATIO's official WideResNet-34-10 checkpoint, while for CIFAR100 and RestrictedImageNet, we retrain RATIO models with their official code.

4.2 Evaluation metrics

We measure both classification robustness and generative modeling quality. For classification, we report clean accuracy and robust accuracy ($L_2, \epsilon = 0.5$) computed using AutoAttack [26], following the evaluation protocol of RATIO [14]. For generative modeling quality, we evaluate sample diversity and visual fidelity using Fréchet Inception Distance (FID) [27] and Inception Score (IS) [28]. We focus on conditional generation and the details can be found in Appendix A.2.

We use expected calibration error (ECE) [29] and AUROC as the metrics for calibration and out-of-distribution detection. 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 real 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

Table 1 summarizes the performance of our proposed model compared to existing hybrid and generative models across classification accuracy, adversarial robustness, and generative quality metrics (IS and FID). Our approach substantially improves upon the original JEM baseline in both robust classification accuracy (75.66% vs. 40.5%) and generative modeling performance (FID 9.07 vs. 38.4). Compared with RATIO, our method achieves a significantly better FID score (9.07 vs. 21.96) while incurring only a minor decrease in standard accuracy (91.86% vs. 92.23%) and robust accuracy (75.66% vs. 76.25%). On CIFAR-10 conditional generation, our model's FID surpasses several dedicated generative models such as SNGAN and BigGAN. Our model also achieves an Inception Score of 9.96, exceeding that of RATIO, JEM, and StyleGAN2.

As shown in Table 2, our model achieves an FID score of 10.70, substantially outperforming RATIO's 21.18 on CIFAR100. While there is a more noticeable trade-off in clean accuracy (65.55% vs. RATIO's 71.58%), our approach maintains comparable robust accuracy (45.97% vs. 47.74%). On RestrictedImageNet, our approach demonstrates improvements across all metrics over RATIO: our model achieves higher standard accuracy (74.52% vs. 70.37%) and robust accuracy (50.59% vs. 48.96%), while also exhibiting superior generative quality with an FID of 64.12. These results demonstrate that our method successfully combines near state-of-the-art adversarial robustness with competitive generative capabilities.

Figure 5 (Appendix) shows generated samples of our approach and RATIO. Our generated samples exhibit high visual fidelity, while some samples from the RATIO baseline show potential artifacts (e.g., saturated or unnatural colors, see examples at Row 4, Col 7; Row 4, Col 10; Row 5, Col 10; Row 7, Col 7 in Figure 5c) possibly linked to the aggressive AutoAugment policy used for model training. This visual difference highlights the benefit of our decoupled augmentation strategy, which uses milder augmentations for the generative component.

Table 1: CIFAR10 classification and generative modeling results

Hybrid Models

Method	Acc% ↑	Robust Acc% ↑	IS ↑	FID ↓
Residual Flow [12]	70.3	–	3.6	46.4
Glow [30]	67.6	–	3.92	48.9
IGEBM [3]	49.1	–	8.3	37.9
JEM [1]	92.9	40.5	8.76	38.4
RATIO [14]	92.23	76.25	9.61	21.96
Dual-AT (ours)	91.86 ± 0.03	75.66 ± 0.01	9.96 ± 0.02	9.07 ± 0.03

Generative Models

Method	IS ↑	FID ↓
Conditional		
SNGAN [23]	8.59	25.5
BigGAN [31]	9.22	14.73
StyleGAN2 [32]	9.53	6.96
StyleGAN2 ADA [33]	10.24	3.49
Unconditional		
NCSNv2 [34]	8.4	10.87
CF-EBM [4]	–	16.71
EBM-Diffusion [35]	8.3	9.58
DDPM [36]	9.46	3.17

Table 2: Classification and generative modeling results on CIFAR100 and RestrictedImageNet

CIFAR100

Method	Acc% ↑	Robust Acc% ↑	FID ↓
JEM	72.2	–	–
RATIO	71.58	47.74	21.18
Dual-AT (ours)	65.55 ± 0.62	45.97 ± 0.49	10.70 ± 0.22
LeCAM + BigGAN	–	–	11.20

RestrictedImageNet

Method	Acc% ↑	Robust Acc% ↑	FID ↓
RATIO	70.37	48.96	78.16
Standard AT	73.26	47.56	79.89
Dual-AT (ours)	74.52 ± 0.34	50.59 ± 0.73	64.12 ± 1.10

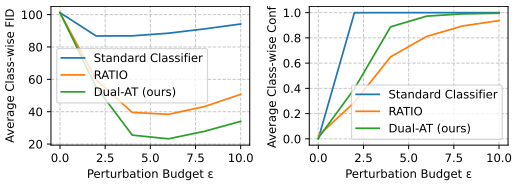


Figure 1: Counterfactual FIDs and classifier confidences under different perturbations.

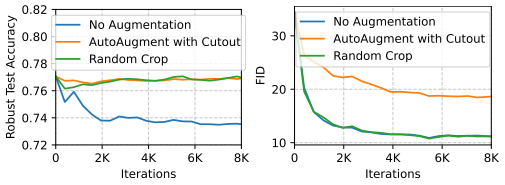


Figure 2: Training curves under different data augmentations during stage 2 joint training.

4.3.2 Counterfactual generation, calibration, and OOD detection

Figure 1 compares counterfactual quality across different models while accounting for classifier confidence. We find our approach reaches much lower FIDs when models achieve similar confidence levels. 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 Dual-AT 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 approach generates examples that are substantially more faithful to the true visual characteristics of that class, indicating higher-quality and more plausible counterfactuals. We provide visualizations of counterfactuals in Appendix A.3.

While our model inherits JEM's calibration benefits (see Figure 3), its out-of-distribution (OOD) detection capability generally underperforms RATIO across various OOD datasets in both clean and adversarial settings, particularly for noise detection (Appendix A.4). This performance gap likely stems from our design choice to use milder augmentation strategies that favor high-quality sample generation over OOD sensitivity. Conversely, RATIO employs aggressive augmentation, which enhances the diversity of its OOD training data, leading to better generalization against novel test OOD inputs, particularly synthetic noise.

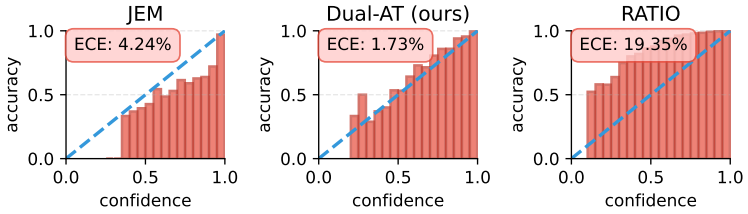
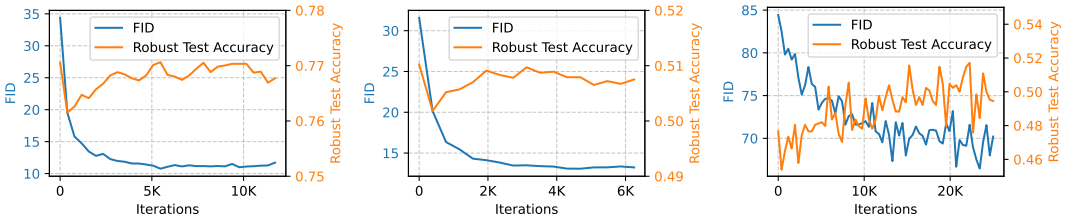


Figure 3: CIFAR10 calibration results (without temperature rescaling, see Appendix A.5 for details)

4.3.3 Training analysis



(a) CIFAR10 (b) CIFAR100 (c) RestrictedImagenet

Figure 4: Training curves (robust test accuracies are evaluated using the PGD attack and FID scores are measured using 10K generated samples)

Figure 2 illustrates CIFAR10 training curves from Stage 2 joint trianing 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 Crop, which minimally distort

the underlying data distribution p_{data}. Overall we find Random Crop provides the optimal balance between discriminative and generative performances. These findings underscore the importance of selecting different augmentation strategies for the generative and discriminative components.

Figure 4 illustrates training curves from Stage 2 joint training. The curves demonstrate a substantial improvement in FID scores while maintaining the robust test accuracy established during Stage 1. The consistent gains on datasets that vary in scale and complexity further highlight the approach's capacity to generalize across diverse image-classification tasks.

5 Conclusion

We addressed the challenge of developing models that excel simultaneously at robust classification and high-fidelity generative modeling. While Joint Energy-Based Models (JEM) [1] offer a promising foundation, they suffer from training instability and poor generation quality. Our approach integrates adversarial training principles for both components: replacing unstable SGLD-based EBM learning with an AT-based approach, while maintaining standard AT for classification robustness. Experiments across multiple datasets with increasing complexities demonstrate that our dual-AT framework significantly outperforms existing hybrid models in generative quality while maintaining competitive adversarial robustness. These improved generative capabilities also translate to higher-quality counterfactual explanations, enhancing model explainability. Future research could extend this approach to larger-scale datasets with high-capacity models, explore methods to improve OOD detection while maintaining generative quality, and investigate the theoretical connections between adversarial robustness and effective generative modeling.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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*, pages 9155–9164, 2018.

[6] 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.

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

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

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

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

[11] 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*, pages 3702–3711, 2018.

[12] 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.

[13] 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)*, pages 681–688. Citeseer, 2011.

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

[15] 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.

[16] 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.

[17] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

[18] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[19] 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.

[20] 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.

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

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

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

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

[25] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[26] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[27] 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.

[28] 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.

[29] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

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

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

[32] 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*, pages 8110–8119, 2020.

[33] 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, 2020.

[34] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[35] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125*, 2020.

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

A Technical Appendices and Supplementary Material

A.1 Model training

A.1.1 Training setup

We implement our 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 (discriminative pre-training), we train robust classifiers following RATIO's [14] methodology with batch normalization enabled. We optimize only the robust classification objective $\mathcal{L}_{\text{AT-CE}}$ using the adversarial settings detailed in Table 4. CIFAR10/100 models are trained for 300 epochs with a cosine learning rate schedule, while RestrictedImageNet training uses a step decay schedule for 75 epochs.

For Stage 2 (joint training), we initialize from the best-performing (in terms of robust test accuracy) model from Stage 1 (the EMA model for CIFAR10/100 and the standard model for RestrictedImageNet) 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 continues to use the same adversarial settings as Stage 1, while the generative component employs the parameters detailed in Table 5.

For the generative adversary used to optimize $\mathcal{L}_{\text{BCE}}$, we implement a curriculum learning strategy following Yin et al. [7] that begins with fewer PGD steps and progressively increases them when the loss value falls below predefined thresholds.

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

Table 3: Training hyperparameters for both stages

Parameter	CIFAR10/100	RestrictedImageNet
Architecture	WideResNet-34-10	ResNet50
Optimizer	SGD with Nesterov (momentum=0.9)	SGD with Nesterov (momentum=0.9)
Weight decay	5×10^{-4}	5×10^{-4}
Batch size	128	128
EMA [20]	Yes	No
Learning rate (Stage 1)	0.1 (cosine schedule, 300 epochs)	0.1 (step decay at epochs 30, 60, 75)
Learning rate (Stage 2)	0.001 (CIFAR10), 0.01 (CIFAR100)	0.001
Batch normalization (Stage 1)	Enabled (train mode)	Enabled (train 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)

Parameter	CIFAR10/100	RestrictedImageNet
PGD steps	10	10
PGD step size	0.1	0.7
L_2 perturbation bound	0.5	3.5

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

Parameter	CIFAR10/100	RestrictedImageNet
Max PGD steps	45	18
PGD step size	0.1	0.7
L_2 perturbation bound	None (unconstrained)	None (unconstrained)
OOD data source	80M Tiny Images [16]	Non-RestrictedImageNet classes

Data augmentation As described in Section 3.4, 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 [14] for all datasets. The effects of these augmentations can be found in Appendix A.1.1.

Table 6: Data augmentation strategies for discriminative and generative components

Dataset	Component	Augmentation Strategy
CIFAR10/100	$\mathcal{L}_{\text{AT-CE}}$	AutoAugment + Cutout + Random horizontal flip
	$\mathcal{L}_{\text{BCE}}$	Random crop + Random horizontal flip
RestrictedImageNet	$\mathcal{L}_{\text{AT-CE}}$	Random crop + Random horizontal flip + Color jitter + Lighting transform
	$\mathcal{L}_{\text{BCE}}$	Random crop + Random horizontal flip

Samples produced by different augmentations on CIFAR10 (note that Autoaugment includes significant color transformations)

A.1.2 RATIO model reproduction

For CIFAR10 comparisons, we use RATIO's [14] official WideResNet-34-10 checkpoint. The ResNet50 model used in their original evaluation was not publicly available. For CIFAR100 and RestrictedImageNet, we reproduced RATIO models using their official code repository and the training configuration described in their paper.

For RestrictedImageNet specifically, the authors reported using a mixture of clean and adversarial samples during training to improve clean accuracy. Despite using their official code, we were unable to reproduce the reported clean accuracy with mixed training. Therefore, we employed standard adversarial training for reproducing their RestrictedImageNet models. Our reproduced models achieve comparable robust accuracy to those reported in the original paper, but with reduced clean accuracy.

Table 7 provides a detailed comparison between our reproduced RATIO models and the performance metrics reported in the original publication.

Table 7: Comparison between our reproduced RATIO models and originally reported results

Dataset	Model	Architecture	Clean Acc (%)	Robust Acc (%)
CIFAR10	Official checkpoint	WideResNet-34-10	92.23	76.25
	Reported in [14]	ResNet50	91.08	73.27
CIFAR100	Reproduced	WideResNet-34-10	71.58	47.74
	Reported in [14]	ResNet50	69.17	45.55
RestrictedImageNet	Reproduced (adversarial training)	ResNet50	70.37	48.96
	Reproduced (mixed training)	ResNet50	74.29	44.15
	Reported in [14] (mixed training)	ResNet50	93.94	49.22

A.2 Sample quality evaluation

We evaluate generative performance using Fréchet Inception Distance (FID) and Inception Score (IS). Following Karras et al. [33], 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.

We consider both conditional and unconditional generation approaches. For conditional generation, we generate an equal number of samples for each class. The optimal number of PGD steps for each model and dataset combination (shown in Table 8) was determined through grid search. 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:

$$x_{t+1} = x_t + \eta \frac{\nabla_x(-E_\theta(x_t, y))}{||\nabla_x(-E_\theta(x_t, y))||_2} \tag{14}$$

where T is the number of PGD steps from Table 8 and η is the corresponding step size.

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 9. As shown in the table, conditional generation consistently outperforms unconditional generation on CIFAR10 and CIFAR100, while both methods yield similar results on RestrictedImageNet. Additional samples from unconditional generation can be found in Appendix A.6.

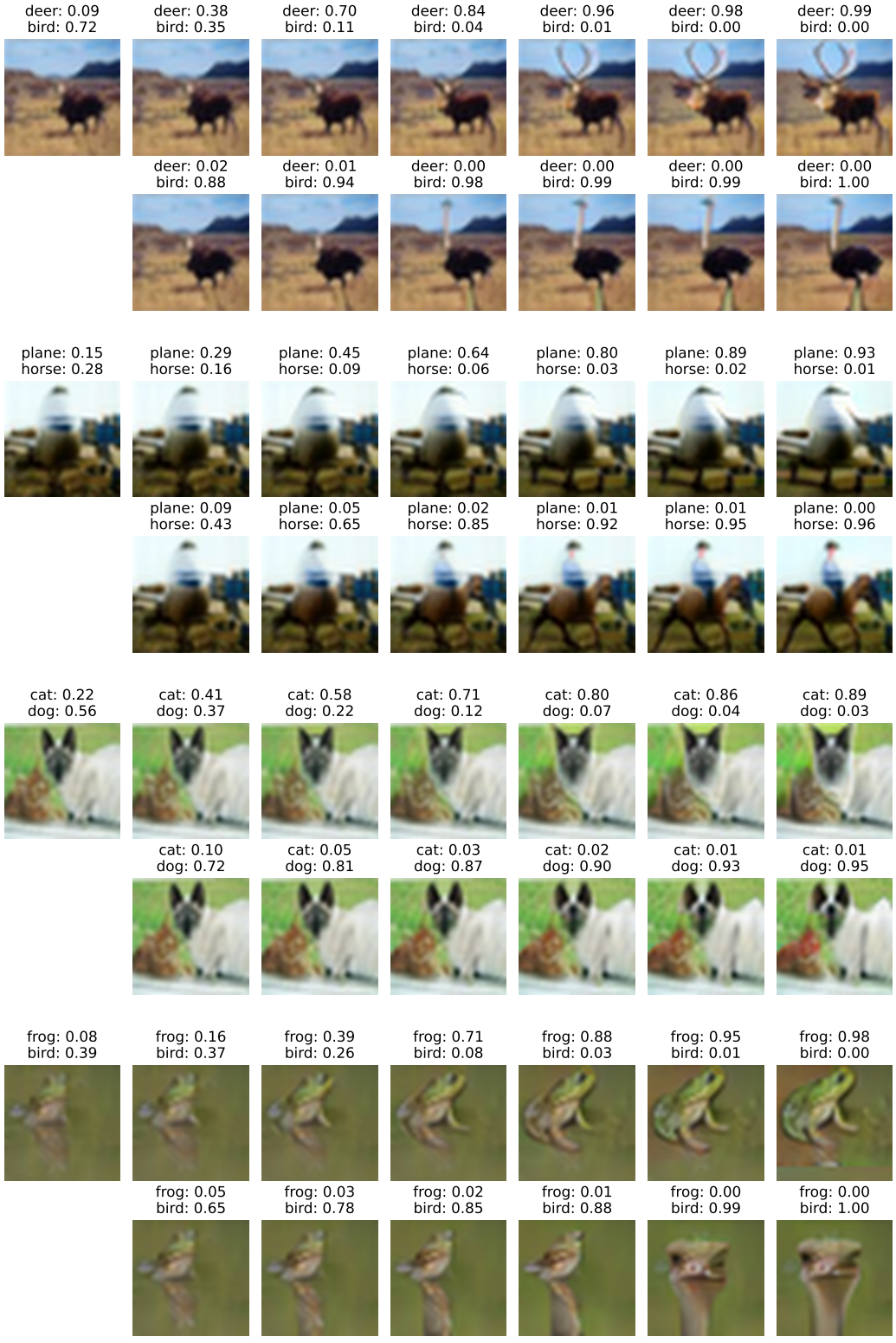
Table 8: Sample generation parameters for FID and IS evaluation

Model	Dataset	PGD Steps	Step Size	OOD Data Source
Dual-AT (Ours)	CIFAR10	33	0.2	80M Tiny Images [16]
	CIFAR100	32	0.2	80M Tiny Images [16]
	RestrictedImageNet	13	8.0	Non-RestrictedImageNet classes
RATIO	CIFAR10	31	0.2	80M Tiny Images [16]
	CIFAR100	12	0.2	80M Tiny Images [16]
	RestrictedImageNet	13	8.0	Non-RestrictedImageNet classes

Table 9: FIDs of conditional and unconditional generation of our approach

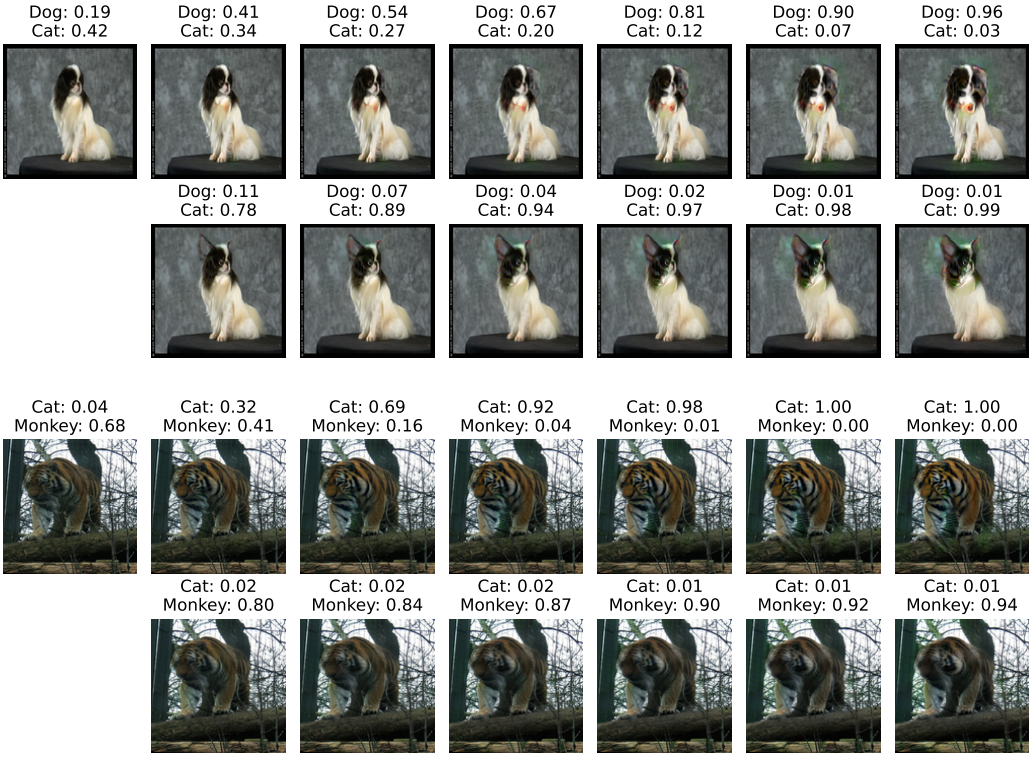
	CIFAR10	CIFAR100	RestrictedImageNet
Conditional generation	9.07 ± 0.03	10.70 ± 0.22	64.12 ± 1.10
Unconditional generation	20.57 ± 0.04	13.56 ± 0.17	63.82 ± 2.83

A.3 Counterfactual generation



CIFAR10 counterfactual examples. Perturbations limits are $0.5, 1.0, 1.5, 2.0, 2.5, 3.0$.

CIFAR100 counterfactual examples. Perturbations limits are $0.5, 1.0, 1.5, 2.0, 2.5, 3.0$.



RestrictedImageNet counterfactual examples. Perturbations limits are $5, 7, 9, 11, 13, 15$.

A.4 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.

We investigate two OOD detection functions: (1) an energy-based function $s_\theta(x) = -E_\theta(x)$, which is proportional to $\log p_\theta(x)$ up to an additive constant, and (2) 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 [14]).

To find adversarial OOD inputs for the energy-based detection function, we employ a PGD attack to maximize the negative energy:

$$x_{adv} = \underset{x' \in B(x, \epsilon_o)}{\arg\max} \; -E_\theta(x') \tag{15}$$

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

For the maximum confidence detection function, following RATIO [14], we compute adversarial OOD inputs by maximizing the cross-entropy loss against a uniform distribution:

$$x_{adv} = \underset{x' \in B(x, \epsilon_o)}{\arg\max} \; \mathcal{L}_{\text{CE}}(\theta; x', \mathbf{1}/K) \tag{16}$$

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 $\max_y p_\theta(y|x')$.

All results are computed using all the in-distribution test samples and 1024 out-distribution samples. For adversarial OOD samples, we use $\epsilon_o = 1.0$ for CIFAR10/100 datasets and $\epsilon_o = 7.0$ for RestrictedImageNet.

Table 10, 11, and 12 present the detection results for CIFAR10, CIFAR100, and RestrictedImageNet, respectively. The results reveal complementary strengths between the two detection functions: the energy-based approach ($-E_\theta(x)$) excels at uniform noise detection with near-perfect AUROC scores, while the maximum confidence variant ($\max_y p_\theta(y|x)$) performs better on natural image OOD datasets.

Since RATIO also employs $\max_y p_\theta(y|x)$ for detection, we focus our comparison with RATIO on our maximum confidence variant results. Our model generally underperforms RATIO across most OOD datasets on CIFAR10, while on CIFAR100, RATIO consistently achieves higher clean AUROC scores but our approach demonstrates superior adversarial robustness. On RestrictedImageNet, our model outperforms RATIO for most datasets except ImageNetOD and uniform noise. This performance gap with RATIO on CIFAR10 and CIFAR100 likely stems from our design choice to use milder augmentation strategies that favor high-quality sample generation over OOD sensitivity, while on RestrictedImageNet the default model (ResNet50) is a relatively small model that benefits less from aggressive augmentation.

Table 10: OOD detection performance (AUROC) with CIFAR10 as ID dataset (JEM results are from Augustin et al. [14], ImageNetOD refers to ImageNet test samples excluding those in RestrictedImageNet)

OOD Dataset	RATIO		Ours ($\max_y p_\theta(y\|x)$)		Ours ($-E_\theta(x)$)		JEM	
	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
CIFAR100	0.9157	0.7516	0.8709	0.6480	0.8484	0.6647	0.876	0.192
SVHN	0.9843	0.9130	0.9609	0.8334	0.8011	0.6046	0.893	0.073
ImageNetOD	0.9210	0.7915	0.8639	0.6582	0.8739	0.7146	–	–
Uniform noise	0.9999	0.9999	0.8922	0.8257	0.9995	0.9983	0.118	0.025

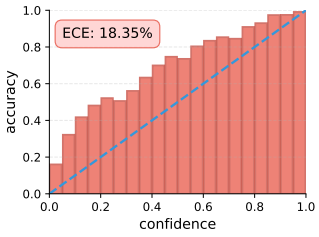
Table 11: OOD detection performance (AUROC) with CIFAR100 as ID dataset

OOD Dataset	RATIO		Ours ($\max_y p_\theta(y|x)$)		Ours ($-E_\theta(x)$)	
	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
CIFAR10	0.7320	0.3795	0.7027	0.5145	0.6689	0.4715
SVHN	0.8439	0.4356	0.8271	0.5823	0.7245	0.5392
ImageNetOD	0.7668	0.4325	0.7136	0.5211	0.7728	0.6019
Uniform noise	0.7769	0.5881	0.4024	0.2283	0.9995	0.9945

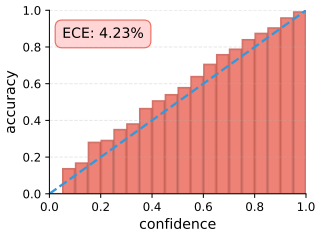
Table 12: OOD detection performance (AUROC) with RestrictedImageNet as ID dataset

OOD Dataset	RATIO		Ours ($\max_y p_\theta(y|x)$)		Ours ($-E_\theta(x)$)	
	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
CIFAR10	0.7344	0.4471	0.7989	0.4654	0.8051	0.5006
CIFAR100	0.7591	0.4908	0.8205	0.5270	0.8496	0.5871
SVHN	0.9232	0.7782	0.9415	0.7603	0.9548	0.8232
ImageNetOD	0.8348	0.5738	0.7163	0.4126	0.8593	0.6612
Uniform noise	0.8461	0.7851	0.4146	0.3080	0.9948	0.9854

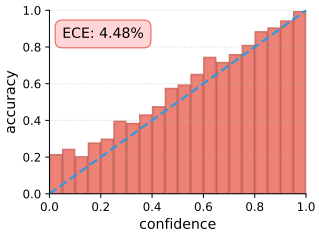
A.5 Calibration



(a) Dual-AT (Ours)

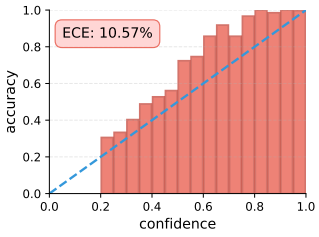


(b) Standard AT

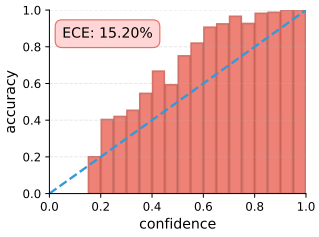


(c) RATIO

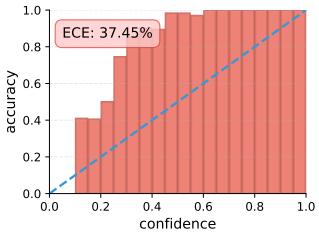
Calibration diagrams on CIFAR100 (without temperature scaling)



(a) Dual-AT (Ours)



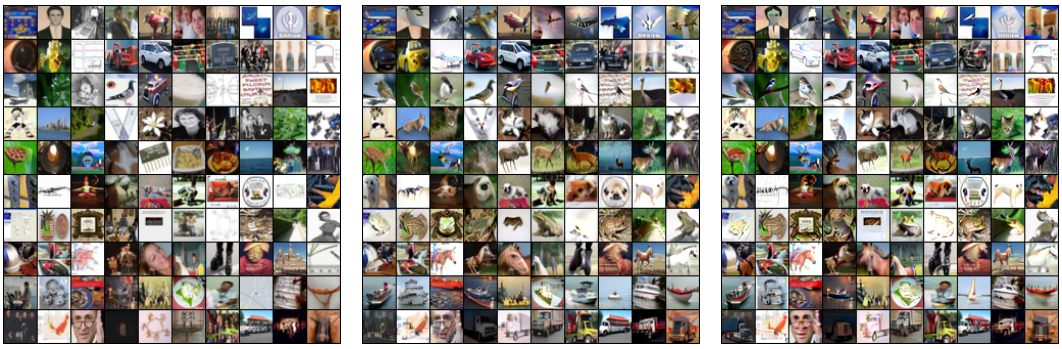
(b) Standard AT



(c) RATIO

Calibration diagrams on RestrictedImageNet (without temperature scaling)

A.6 Additional results



(a) Seed images used for producing the generated samples

(b) Uncurated class-conditional samples of our model

(c) Uncurated class-conditional samples of RATIO

Figure 5: CIFAR10 class-conditional generation results

(a) Seed images

(b) Ours (32 steps)

(c) RATIO's (32 steps)

(d) RATIO's (12 steps)

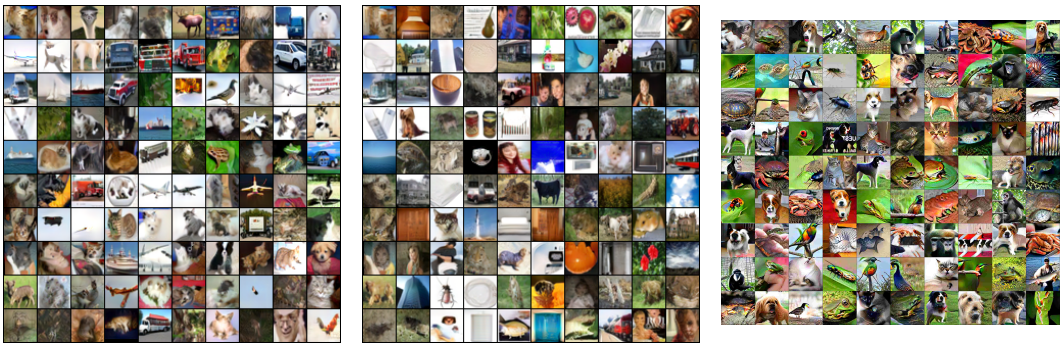
Figure 6: CIFAR100 class-conditional samples of the first 10 classes (`apple`, `aquarium_fish`, `baby`, `bear`, `beaver`, `bed`, `bee`, `beetle`, `bicycle`, `bottle`)

(a) Seed images used for producing the generated samples

(b) Uncurated class-conditional samples of our model

(c) Uncurated class-conditional samples of RATIO

Figure 7: RestrictedImageNet class-conditional generation results (classes are `dog`, `cat`, `frog`, `turtle`, `bird`, `monkey`, `fish`, `crab`, `insect`, images are in 224×224 resolution)



(a) CIFAR10 (b) CIFAR100 (c) RestrictedImageNet

Figure 8: Unconditional generation results of our model

A.7 Computational requirements

Our training was conducted across different GPU configurations. For RATIO model reproduction and Stage 1 discriminative pre-training, we used a single NVIDIA H100 GPU with 80GB memory. For Stage 2 joint training, we utilized a computational node equipped with 4 AMD Instinct MI210 GPUs (each with 64GB memory).

The total training time varies by dataset complexity. For Stage 2 joint training:

- CIFAR10: Approximately 7 hours to reach optimal FID
- CIFAR100: Approximately 10 hours to reach optimal FID
- RestrictedImageNet: Approximately 21 hours to reach optimal FID

These times are in addition to the Stage 1 pre-training, which follows RATIO's training schedule (300 epochs for CIFAR datasets and 75 epochs for RestrictedImageNet).