Scalable Energy-Based Models via Adversarial Training: Unifying Discrimination and Generation

Xuwang Yin[1] Claire Zhang[2] Julie Steele[2] Nir Shavit[2] Tony T. Wang[2]

ICLR 2026

[1]Independent Researcher

[2]MIT

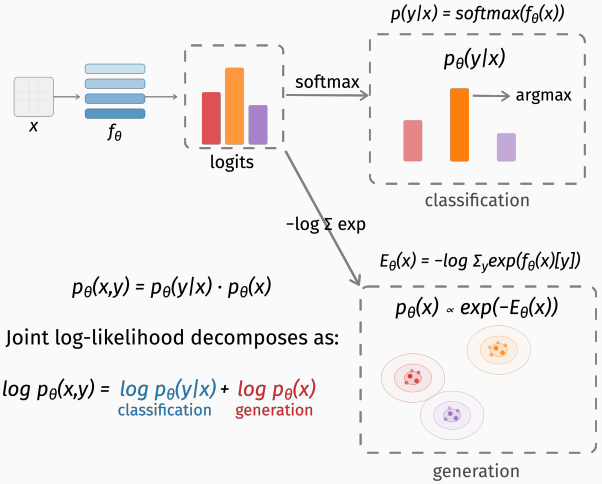
Motivation

A model that truly understands should be able to both **recognize** and **imagine**. Generation quality reflects how well a model captures the data distribution; when classification draws on this same understanding, its decisions become grounded and explainable.

Challenge: JEM provides the framework for joint modeling, but suffers from unstable training, poor generation quality, and has weak adversarial robustness.

	Classification	Generation	Robustness	Explanation
Discriminative (Robust classifier)	✓	—	✓	✓
Generative (Diffusion, GANs)	—	✓	—	—
Joint (JEM, IGEBM)	✓	✓	✓	✗
DAT (Ours)	✓	✓	✓	✓

Joint Energy-Based Model (Grathwohl et al., 2019)



Three sources of instability in JEM:

1. Standard EBM objective — energy values can grow unbounded
2. SGLD sampling — sensitive to step size and noise parameters
3. Energy function — requires explicit regularization

Fix 1: Stabilizing the EBM Gradient via BCE

The standard EBM gradient is unconstrained — $E_\theta(x)$ can grow unbounded, causing numerical instability:

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

We replace it with adaptive scaling factors that attenuate gradients at extreme energies:

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

where $\alpha(x) = 1 - \sigma(-E_\theta(x)), \quad \beta(x) = \sigma(-E_\theta(x))$. When $-E_\theta(x)$ is extreme, sigmoid saturation drives α or $\beta \to 0$, preventing overflow/underflow.

The corresponding loss whose gradient recovers this formulation is a BCE loss:

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

Fix 2: Contrastive Sample Generation via PGD

JEM uses SGLD from noise to draw samples from $p_\theta(x)$:

$$x_{t+1} = x_t - \frac{\alpha}{2}\nabla_x E_\theta(x_t) + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \alpha),\ x_0 \sim \mathcal{U}(0, 1)$$

Sensitive to step size α and noise ξ_t. We replace it with deterministic PGD, initialized from OOD data p_{ood}:

$$x_{t+1} = x_t - \eta \frac{\nabla_x E_\theta(x_t)}{\|\nabla_x E_\theta(x_t)\|_2}, \quad x_0 \sim p_{\text{ood}}$$

- OOD initialization provides structured starting points — PGD transforms them toward the data manifold.
- Random noise initialization also works in our framework, but with inferior generation quality.

Fix 3: Implicit Regularization via AT

$$\log p_\theta(x, y) = \underbrace{\log p_\theta(y|x)}_{\text{classification}} + \underbrace{\log p_\theta(x)}_{\text{generation}}$$

Standard classification: $\mathcal{L}_{\text{CE}} = -\mathbb{E}_{p_{\text{data}}}[\log p_\theta(y|x)]$

Robust classification:

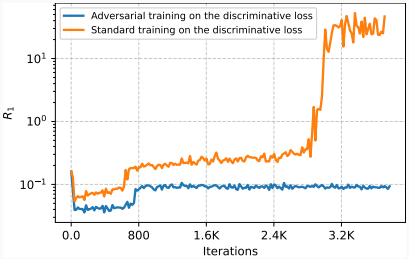
$$\mathcal{L}_{\text{AT-CE}} = \mathbb{E}_{p_{\text{data}}}\left[\max_{\|\delta\|_p \leq \epsilon} L(f(x+\delta), y)\right]$$

Roth et al. (2020): AT $\equiv$ operator norm regularization:

$$\mathcal{L}_{\text{AT-CE}} \equiv \mathcal{L}_{\text{CE}} + \lambda(\epsilon) \cdot \sigma_{\max}(J_f(x))$$

Connection to R_1: $\|\nabla_x f_y\| \leq \sigma_{\max}(J_f)$

$\Rightarrow$ AT implicitly bounds R_1 — the exact regularization EBMs need, obtained **for free**.



AT keeps R_1 bounded; standard training diverges.

Dual AT for Joint Modeling

$$\log p_\theta(x, y) = \underbrace{\log p_\theta(y|x)}_{\text{classification}} + \underbrace{\log p_\theta(x)}_{\text{generation}}$$

Combined objective:

$$\mathcal{L} = \underbrace{\mathbb{E}_{p_{\text{data}}}\left[\max_{\|\delta\| \le \epsilon} L(f(x+\delta), y)\right]}_{\mathcal{L}_{\text{AT-CE}}\text{: models } p(y|x)} + \underbrace{-\mathbb{E}_{p_{\text{data}}}[\log \sigma(-E_\theta(x))] - \mathbb{E}_{p_\theta}[\log(1-\sigma(-E_\theta(x)))]}_{\mathcal{L}_{\text{BCE}}\text{: models } p(x)}$$

Two-stage training:

Stage 1: Adversarial Training

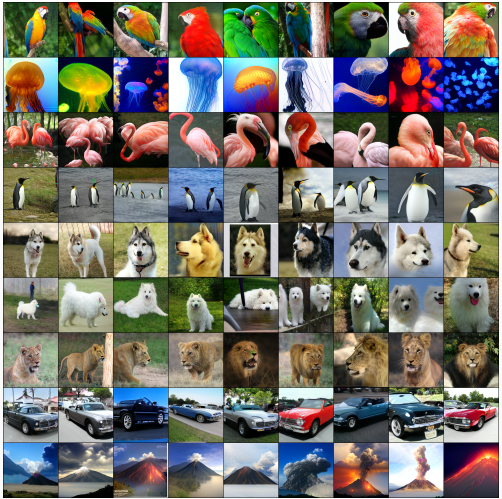
- Standard AT $\rightarrow$ robust classifier
- Reuse existing checkpoints

Stage 2: Joint Training

- Add $\mathcal{L}_{\text{BCE}}$ to $\mathcal{L}_{\text{AT-CE}}$
- Freeze BN stats (when using BatchNorm)

Generalizes across architectures with BatchNorm (ResNet) and LayerNorm (ConvNeXt).

Results: ImageNet 256×256 Generation



Curated generation results from ConvNeXt-L.

- First EBM hybrid to scale to ImageNet 256.
- ∼**5**× **faster** sample generation than LDM.

Method	Params	FID↓	Steps
BigGAN-deep	340M	6.95	1
ADM-G	608M	4.59	250
LDM-4-G	400M	3.60	250
VAR-d16	310M	3.30	10
DAT	**198M**	**3.29**	**36**

Results: Classification & Robustness

CIFAR-10

Method	Acc%	Robust Acc%†	FID↓
AT Methods			
RATIO	92.23	76.25	21.96
Std AT	92.43	75.73	28.41
Joint Models			
JEM	92.9	40.5	38.4
SADA-JEM	95.5	31.93	9.41
DAT	90.72	**74.65**	**7.57**

† AutoAttack, ℓ_2, ϵ=0.5.

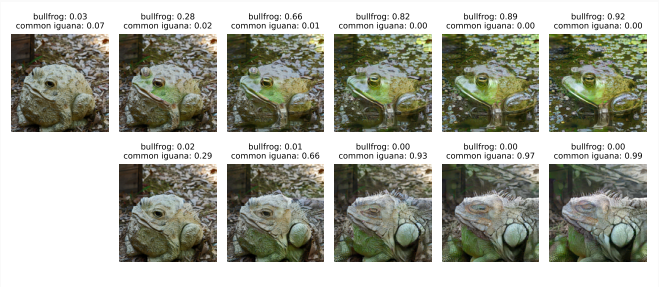
ImageNet 256×256

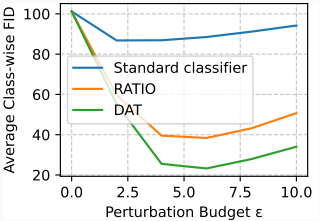
Method	Acc%	Robust Acc%‡	FID↓
AT Methods			
Std AT*	78.25	33.38	44.46
Joint Models			
EGC	78.90	13.56	6.05
DAT	75.78	**56.40**	**3.29**

* Trained with ℓ_∞ perturbations. ‡ AutoAttack, ℓ_2, ϵ=3.0.

DAT uniquely combines strong adversarial robustness with strong generation quality.

Self-Explanation via Counterfactuals

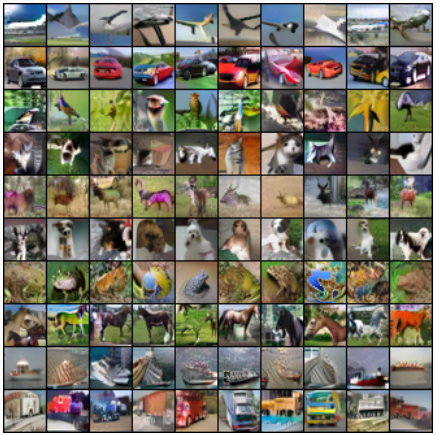




Class-wise FID of counterfactuals.

PGD from a real image toward a target class produces **semantically meaningful** transformations — classification confidence is grounded in generative understanding.

Generation from Pure Noise — No Auxiliary Data



Samples from pure noise init (CIFAR-10).

Training with OOD init requires an auxiliary dataset and leaves background residue in generated samples.

DAT can also be trained with **pure random noise** init:

Init strategy	IS↑	FID↓
OOD (80M Tiny Images)	9.86	7.57
Uniform noise	9.00	13.72

Summary

Three fixes for unstable EBM training:

1. BCE objective — bounds energy gradients
2. PGD sampling — removes SGLD sensitivity
3. Adversarial training — provides implicit regularization for free

Result: A single model that uniquely combines strong adversarial robustness with strong generation quality, and is intrinsically explainable — joint modeling does not need to compromise on either dimension.

Paper

`openreview.net/forum?id=I9iai932rK`

Code

`github.com/xuwangyin/DAT`