ADAPTIVE ENERGY-GUIDED ROUTING FOR ROBUST OUT-OF-DISTRIBUTION DETECTION WITH SYNTHETIC IN-DISTRIBUTION DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-Distribution (OOD) detection under adversarial attacks is critical for safety-sensitive AI systems. Prevailing OOD-centric methods face a fundamental limit—the OOD space is infinite—while the in-distribution (ID) manifold is compact, suggesting a complementary strategy: enriching the model's ID representation with synthetic data from diffusion models. However, the effectiveness of synthetic ID data is tied to its generation quality: high-fidelity synthetic samples effectively improve robust OOD detection, while low-quality samples can cause severe degradation when naively incorporated. To enable robust use of synthetic ID data across varying quality levels, we propose Adaptive Energy-Guided Routing (AEGR). AEGR uses a teacher classifier's LogSumExp energy as a per-sample fidelity oracle, computing a dynamic weight γ that controls a dual-objective routing mechanism: high-quality samples ($\gamma \approx 0$) undergo standard classification, while low-quality samples ($\gamma \approx 1$) are steered toward a margin-regularized uniform loss that flattens overconfident predictions and penalizes excessive energy. We show that the explicit margin is structurally justified by a gradient zero-sum property of the uniform loss; the reject-class paradigm avoids this requirement through its implicit energy suppression. Experiments across four datasets demonstrate that AEGR consistently improves robust OOD detection, with the largest recovery over naive synthetic data usage on the most challenging dataset where generation quality is lowest, while preserving gains on simpler datasets and relaxing the margin when generation quality is sufficient.

1 INTRODUCTION

Neural networks exhibit overconfidence on Out-of-Distribution (OOD) inputs (Hendrycks & Gimpel, 2016), posing critical risks in safety-sensitive applications such as autonomous driving and medical diagnosis. Semi-supervised methods that leverage auxiliary OOD data during training (Hendrycks et al., 2018; Mohseni et al., 2020) mitigate this overconfidence to some extent, yet they remain vulnerable to *adversarial OOD attacks*—carefully crafted perturbations designed to fool the detector into accepting OOD samples as in-distribution (Sehwag et al., 2019; Hein et al., 2019). To counter these attacks, the dominant paradigm is *OOD-centric*: adversarially training on auxiliary or synthetic OOD data, as exemplified by ACET (Hein et al., 2019), ATOM (Chen et al., 2021), AT-OE (Augustin et al., 2020), and RODEO (Mirzaei et al., 2025). While effective, this paradigm faces a fundamental limitation: the OOD space is infinite and unbounded, making it impossible to exhaustively characterize all possible OOD inputs. In contrast, the in-distribution (ID) manifold is compact and of finite dimension, suggesting a *complementary* ID-centric strategy: instead of enumerating what lies *outside* the distribution, we can densely characterize what lies *inside* it. Critically, ID-centric and OOD-centric approaches address different aspects of the problem and can be combined—an ID-densified model can still benefit from OOD exposure, as the two sources of signal are orthogonal.

Under this ID-centric perspective, OOD detection quality is bounded by the completeness of the learned ID manifold representation. If a model has only observed a sparse sampling of the ID manifold, its notion of "in-distribution" is incomplete, and the boundary it draws against OOD samples will be unreliable. Synthetic ID data—generated by modern diffusion models (Karras et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2022; Rombach et al., 2022)—offers a principled way to densify the manifold. By training on diverse, high-quality synthetic ID samples, the model develops a more comprehensive understanding of the ID distribution, enabling more accurate OOD discrimination (Zhang et al., 2023). However, how synthetic ID data affects adversarially robust OOD detection—and how this effect depends on generation quality—remains unexplored.

Despite the rapid progress of diffusion models (Karras et al., 2022; Rombach et al., 2022), synthetic data inevitably deviates from the true ID distribution to some degree. On simpler datasets where generation fidelity is high (CIFAR-10/100 (Krizhevsky & Hinton, 2009)), this deviation is minimal and incorporating synthetic ID data yields clear improvements. On Tiny-ImageNet-200 (CS231N, 2015), where generation quality is substantially lower, the same strategy causes a severe degradation in robust OOD detection (Sec 4).

Addressing this quality-conditioned behavior requires a framework that can *conditionally* leverage synthetic ID data: benefiting from high-quality generations while neutralizing the harm of low-quality ones. We propose **Adaptive Energy-Guided Routing** (AEGR), built on three integrated mechanisms: (1) A *teacher-guided quality oracle*—an OOD-confidence-calibrated classifier evaluates each synthetic sample's fidelity via its LogSumExp (LSE) energy (Liu et al., 2020). (2) *Per-sample dynamic routing*—a continuous weight γ, derived from the teacher's quality assessment, controls a dual-objective routing mechanism: high-quality samples ($\gamma \approx 0$) undergo standard classification (centripetal pull toward the ID manifold), while low-quality samples ($\gamma \approx 1$) are steered toward a margin-regularized uniform loss that flattens overconfident predictions and penalizes excessive energy. (3) *Adaptive objective allocation*—the routing weight automatically reallocates each synthetic sample between classification and energy-regularized objectives according to its teacher-assessed quality, preserving the ID-augmentation benefits of high-quality samples while recovering robustness when quality degrades. We further show that the explicit energy margin is structurally justified by a gradient zero-sum property of the uniform cross-entropy loss (Sec 5.3). The framework also generalizes to the reject-class paradigm (Sec 6.4), where the same routing mechanism transfers without modification. Our core contributions are:

- We conduct the first study, to our knowledge, of synthetic ID data for adversarially robust OOD detection, spanning four datasets, and discover a quality-dependent effect: high-fidelity synthetic data consistently improves robust OOD detection, while low-quality samples can cause severe degradation when naively incorporated.
- We propose **Adaptive Energy-Guided Routing (AEGR)** with a teacher-guided quality oracle that routes samples between standard classification and a margin-regularized uniform loss, together with an analysis establishing the gradient zero-sum property of the uniform OE loss that explains the necessity of an explicit energy margin.
- Extensive experiments demonstrate that AEGR consistently improves robust OOD detection, with the largest recovery over naive synthetic data on the most challenging dataset where generation quality is lowest (Tiny-ImageNet-200, where AEGR recovers robust detection AUC from 78.61% naive to 95.61%), while preserving gains on simpler datasets. The same routing mechanism also generalizes to the reject-class paradigm with significant improvements (from 95.56% to 97.14%).

2 Related Work

2.1 Adversarial Robustness

Adversarial training (AT) (Madry et al., 2017) is the leading defense against adversarial examples, solving a min-max optimization over perturbation budgets, though it often reduces clean accuracy (Tsipras et al., 2019). Subsequent work pushes the robustness frontier using additional data, including unlabeled (Carmon et al., 2019) and synthetic (Wang et al., 2023; Karras et al., 2022) data; notably, synthetic data benefits AT even when distributionally shifted, because additional samples provide effective regularization against robust overfitting. Whether this tolerance to shift carries over to OOD detection—which must draw a precise ID/OOD boundary—is the question we study: our findings indicate it does not, motivating quality-conditioned strategies.

2.2 Robust Out-of-Distribution Detection

Semi-supervised methods with auxiliary OOD data (Hendrycks et al., 2018; Mohseni et al., 2020) reduce OOD overconfidence but remain vulnerable to adversarial attacks (Sehwag et al., 2019). The dominant paradigm for robust OOD detection is OOD-centric: ACET (Hein et al., 2019) adversarially trains on auxiliary OOD data with uniform pseudo-labels; ATOM (Chen et al., 2021) mines informative outliers via reject-class confidence; later methods combine ID and OOD adversarial training (Augustin et al., 2020; Chen et al., 2022); and RODEO (Mirzaei et al., 2025) / AROS (Mirzaei & Mathis, 2025) synthesize OOD data or stabilize embeddings for robustness. A common thread is OOD-centricity: improving the model's exposure to, or robustness against, OOD data. Our work *complements* these by adopting an ID-centric perspective—densifying the learned in-distribution manifold with synthetic ID data.

2.3 Synthetic Data for Out-of-Distribution Detection

Generative models are increasingly used to synthesize data for OOD detection. SIO (Zhang et al., 2023) pioneered GAN-generated ID data for clean OOD detection under static mixing, and Azizi et al. (Azizi et al., 2023) showed diffusion-generated ID data improves ImageNet classification. Recent methods synthesize OOD samples via diffusion (SONA (Yoon et al., 2025), BOOD (Liao et al., 2025), NCIS (Doorenbos et al., 2024)) or exploit diffusion reconstruction paths for OOD scoring (DiffPath (Heng et al., 2024)). A key distinction is that *all* these methods synthesize OOD data or scores, while we present what is, to our knowledge, the first study of synthetic *ID* data for the adversarially robust setting. Our work further differs in two ways: (1) we introduce *per-sample dynamic routing* via teacher-guided quality assessment, contrasting with static batch-level mixing; (2) we identify the quality-dependent regime where synthetic ID data can be harmful—a finding absent from prior literature. A detailed discussion is provided in Appendix A.

3 Preliminaries

3.1 Notation and OOD Detection

Consider a K-way classifier $f_\theta : \mathcal{X} \to \mathbb{R}^K$ with parameters θ. For an input x, the logits are $z = f_\theta(x)$. We adopt the LogSumExp (LSE) energy score (Liu et al., 2020) for OOD detection:

$$E(x; f_\theta) = \log \sum_{i=1}^{K} \exp(z_i). \tag{1}$$

Samples with higher LSE energy are considered in-distribution. Detection performance is measured by the area under the ROC curve (AUC$^{\uparrow}$) and false positive rate at 95% true positive rate (FPR-95$^{\downarrow}$). We report both clean OOD detection (Clean^{o}) and robust detection under attacks.

3.2 OOD Training Paradigm

The primary framework is Outlier Exposure (OE) (Hendrycks et al., 2018; Augustin et al., 2020) for robust OOD detection, where auxiliary OOD samples are assigned a uniform target distribution over the K ID classes. The OE loss is the cross-entropy with this uniform target:

$$\mathcal{L}_{\text{OE}}(z) = -\frac{1}{K} \sum_{i=1}^{K} \log \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}, \tag{2}$$

which simplifies to $\text{LSE}(z) - \frac{1}{K}\sum_{i=1}^{K} z_i$ with gradient $\partial \mathcal{L}_{\text{OE}} / \partial z_i = \text{softmax}(z)_i - 1/K$ (gradient zero-sum; analyzed in Sec 5.3). Detection uses the LSE energy score of the K ID logits. An alternative is the reject-class ($K+1$) paradigm (Chen et al., 2021), where an additional output serves as a reject class for OOD data and detection uses the softmax probability of this class.

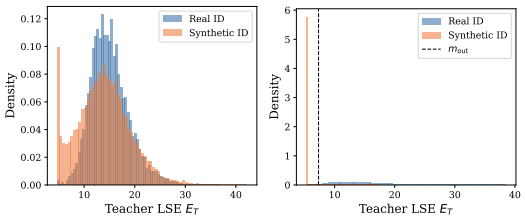
Adversarial Training. In both paradigms, all training samples (real ID, auxiliary OOD, and synthetic ID) undergo adversarial perturbation via PGD-10 (Madry et al., 2017) with $\epsilon = 8/255$, step size $2/255$. For ImageNet-200 (Deng et al., 2009), we use PGD-5 with $\epsilon = 4/255$ (step size $1/255$) following standard practice for large-scale training (Wang et al., 2023).

3.3 Datasets and Evaluation Protocol

ID Datasets. CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009) (32×32), Tiny-ImageNet-200 (64×64), and ImageNet-200 (224×224). **Auxiliary OOD Training Sets.** 300K Random Images (Hendrycks et al., 2018) for 32×32 datasets; the remaining 800 ImageNet classes for Tiny-ImageNet-200 and ImageNet-200. **OOD Test Sets.** For CIFAR-10/100: Textures/DTD (Cimpoi et al., 2014), iSUN (Xu et al., 2015), Places365 (Zhou et al., 2017), LSUN (crop/resize) (Yu et al., 2015), LSUN-resize, and SVHN (Netzer et al., 2011). For Tiny-ImageNet-200: Places365, DTD, iNaturalist, and SUN. Test sets are mixed for default evaluation.

Threat Model and Evaluation. We follow the L_∞ threat model ($\epsilon = 8/255$) and use Auto-PGD (APGD) (Croce & Hein, 2020) with 500 iterations and 5 restarts. Multi-targeted APGD (100 iterations, 1 restart) maximizes each ID class probability sequentially while minimizing the OOD detection score, denoted TAPGD^{lgt}. Clean and robust OOD detection are measured by the area under the ROC curve (AUC$^{\uparrow}$). In the results tables, *Acc* denotes standard accuracy on clean ID test data, *AA* denotes ID accuracy under AutoAttack (Croce & Hein, 2020) (L_∞, $\epsilon = 8/255$) in its default configuration, and *Clean*^{o} and *TAPGD*^{lgt} denote OOD-detection AUC on clean and adversarially perturbed OOD inputs, respectively.

4 Synthetic ID Data and Its Quality-Dependent Effects



(a) CIFAR-100 (b) Tiny-ImageNet-200

Figure 1: Teacher LSE (E_T) distributions for real vs. synthetic ID (measured, density-normalized). Synthetic samples closely match real ID on CIFAR-100 but are clearly separated from it on Tiny-ImageNet-200, falling below the routing boundary m_{out}.

Modern diffusion models (Karras et al., 2022; Rombach et al., 2022) enable high-fidelity generation of diverse ID samples. Incorporating EDM-generated synthetic ID data into adversarial OOD training improves robust OOD detection when generation quality is high: on CIFAR-100 (32×32), TAPGD AUC increases from 68.40% to 69.46%. On Tiny-ImageNet-200 (64×64), where generation quality is substantially lower—consistent with the observation that more challenging datasets yield higher FID for diffusion generators (Wang et al., 2023)—the same strategy causes a significant degradation in robust OOD detection (93.61%→78.61%). This quality-dependent behavior—synthetic data is beneficial at high fidelity but harmful at low fidelity—motivates the need for a framework that can *conditionally* leverage synthetic ID data.

The root cause of this quality-dependent behavior lies in the distribution shift between real and synthetic data, as revealed by the teacher LSE distributions (Figure 1). On CIFAR-100, the teacher's LSE values for real and synthetic ID are close (Real: 15.02 ± 3.75, Synth: 13.85 ± 5.25), indicating that EDM-generated samples closely match the real ID manifold—the distribution shift is minimal and the synthetic data behaves as an effective ID augmentation. On Tiny-ImageNet-200, synthetic samples are clearly separated from real ID (Real: 14.44 ± 4.45, Synth: 5.31 ± 0.24): the teacher assigns low confidence to the vast majority of synthetic samples, so with the percentile-based boundary m_{out} they are routed to the regularizing OOD branch rather than being forced into the ID manifold through standard OE training, where they erode the OOD detection boundary rather than reinforce it.

5 Adaptive Energy-Guided Routing

The quality-dependent nature of synthetic ID data creates a fundamental tension: it can either strengthen or degrade OOD robustness, depending on generation quality. A naive solution—simply discarding all synthetic data for complex datasets—forgoes the potential benefits of manifold densification and offers no recourse as generator quality improves. We seek a principled framework that can incorporate synthetic ID data *regardless* of its quality: amplifying its benefits when generation is faithful, and neutralizing its harm when quality degrades. This section presents Adap-

tive Energy-Guided Routing (AEGR), which achieves this through three integrated mechanisms: a teacher-guided quality oracle, per-sample dynamic routing weights, and an explicit energy margin for the OE training paradigm.

5.1 Teacher-Guided Quality Assessment

The central challenge is assessing the quality of individual synthetic samples *without* access to ground-truth quality labels. Our key insight is that a classifier trained with Outlier Exposure (OE) on real ID + auxiliary OOD data serves as an effective implicit quality oracle. Consider a classifier f_T trained with the OE objective (cross-entropy on ID data + uniform distribution loss on real OOD data), which we refer to as the *teacher*. For any input x, the LogSumExp (LSE) of its logits $z^T = f_T(x)$ measures the model's confidence that x belongs to the ID manifold:

$$E_T(x) = \log \sum_{j=1}^{K} \exp(z_j^T). \tag{3}$$

High $E_T(x)$ indicates that x lies in a region the teacher recognizes as in-distribution; low $E_T(x)$ signals that x deviates from the learned ID manifold. Notably, the OE training objective is essential here: by explicitly learning to suppress energy on OOD samples, the teacher's LSE provides a naturally calibrated quality signal—a synthetic sample with low LSE is one that the teacher cannot distinguish from known OOD data, which is precisely the signature of distributional shift. Figure 1 confirms this empirically: on CIFAR-100 (high fidelity), real and synthetic LSE are nearly identical, while on Tiny-ImageNet-200 (low fidelity), both shift to a markedly higher absolute scale with wider dispersion.

We establish a reference boundary m_{out} as the p_{out}-th percentile of E_T computed over real ID training samples:

$$m_{\text{out}} = \text{Percentile}_{p_{\text{out}}}\big(\{E_T(x) : x \in \mathcal{D}_{\text{real}}\}\big). \tag{4}$$

Synthetic samples whose teacher LSE falls below m_{out} are flagged as low-quality—they deviate from the real ID manifold more than the bottom $(100 - p_{\text{out}})\%$ of real data. Crucially, this quality assessment operates on the teacher's frozen representation and incurs no training overhead beyond a single forward pass per sample.

5.2 Dual-Track Dynamic Routing

For each synthetic sample x_{syn} with pseudo-label y (the class it was generated for), we compute a per-sample routing weight $\gamma(x_{\text{syn}}) \in (0, 1)$ that interpolates between two learning objectives:

$$\gamma(x_{\text{syn}}) = \sigma(m_{\text{out}} - E_T(x_{\text{syn}}))\,, \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function. When $E_T(x_{\text{syn}}) \gg m_{\text{out}}$ (the teacher is highly confident), $\gamma \to 0$; when $E_T(x_{\text{syn}}) \ll m_{\text{out}}$, $\gamma \to 1$.

The student model f_θ, trained with adversarial perturbations for robustness, receives a routed loss for each synthetic sample:

$$\mathcal{L}_{\text{syn}}(x_{\text{syn}}, y) = \ (1-\gamma)\mathcal{L}_{\text{CE}}(f_\theta(x'_{\text{syn}}), y) + \ \gamma\big[\,\mathcal{L}_{\text{unif}}(f_\theta(x'_{\text{syn}})) + \mathcal{L}_{\text{margin}}(f_\theta(x'_{\text{syn}}))\,\big] \tag{6}$$

where x'_{syn} denotes the adversarially perturbed synthetic sample, and the two tracks operate as follows:

Track 1 ($\gamma \approx 0$, high quality). The loss reduces to standard cross-entropy $\mathcal{L}_{\text{CE}}$. The synthetic sample is treated as a genuine ID augmentation, contributing a *centripetal* gradient that pulls it toward the correct class center. This sharpens the ID manifold and improves OOD detection.

Track 2 ($\gamma \approx 1$, low quality). The sample follows a *centrifugal* path. The uniform OE loss,

$$\mathcal{L}_{\text{unif}}(z) = \text{LSE}(z) - \frac{1}{K}\sum_{i=1}^{K} z_i, \tag{7}$$

penalizes overconfidence in any single class, pushing the logits toward a flat distribution. Simultaneously, the explicit energy margin,

$$\mathcal{L}_{\text{margin}}(z) = \text{Softplus}\big(\text{LSE}(z) - m_{\text{ceil}}\big), \text{Softplus}(x) \triangleq \log(1 + e^x). \tag{8}$$

where m_{ceil} is an energy ceiling threshold, computed as a percentile of the teacher LSE on real ID data (potentially with a different percentile p_{ceil} than the routing boundary m_{out}). This decoupling allows the ceiling to be set according to generation quality: a relaxed ceiling (high value) suffices for high-quality synthetic data where the margin is unnecessary, while a tighter ceiling is beneficial on low-quality data to prevent the model from becoming overconfident on distributionally shifted samples.

For real ID data, we use standard cross-entropy: $\mathcal{L}_{\text{real}} = \mathcal{L}_{\text{CE}}(f_\theta(x'), y)$. For auxiliary OOD data, we use the OE loss: $\mathcal{L}_{\text{ood}} = \mathcal{L}_{\text{unif}}(f_\theta(x'))$. The total training objective minimizes the sum over all three data sources, with adversarial perturbation applied to each.

5.3 Why OE Needs an Explicit Energy Margin

A natural question arises: why does OE-style training benefit from an explicit margin $\mathcal{L}_{\text{margin}}$? The answer lies in a structural property of the uniform CE loss. From Equation (2), the gradient with respect to each logit using the simplified form $\mathcal{L}_{\text{unif}}(z) = \text{LSE}(z) - \frac{1}{K}\sum_i z_i$ is $\partial\mathcal{L}_{\text{unif}}/\partial z_i = \text{softmax}(z)_i - \frac{1}{K}$. Critically, the sum of these gradients over all K classes vanishes:

$$\boxed{\sum_{i=1}^{K} \frac{\partial \mathcal{L}_{\text{unif}}}{\partial z_i} = 0}. \tag{9}$$

This **gradient zero-sum property** has a direct implication: $\mathcal{L}_{\text{unif}}$ is invariant to adding any constant to all logits, $\mathcal{L}_{\text{unif}}(z + c \cdot \mathbf{1}) = \mathcal{L}_{\text{unif}}(z)$. The loss only constrains the *relative shape* of the logit distribution—driving it toward uniformity across classes—but imposes no constraint on the *absolute energy scale*. As a result, the OE objective can in principle be satisfied at arbitrarily high LSE, as long as the logits are balanced across classes. For OOD detection, absolute energy magnitude is precisely what distinguishes ID from OOD; without a scale anchor, low-quality synthetic samples trained on this track may not receive a consistent downward pressure on their energy scores, leaving the energy-based OOD scoring function less reliable.

The explicit margin $\mathcal{L}_{\text{margin}}$ closes this loophole. Using the smooth Softplus formulation, its gradient,

$$\frac{\partial \mathcal{L}_{\text{margin}}}{\partial z_i} = \sigma\big(\text{LSE}(z) - m_{\text{ceil}}\big) \cdot \text{softmax}(z)_i, \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function, has a strictly positive sum ($\sum_i \partial/\partial z_i = \sigma(\text{LSE}(z) - m_{\text{ceil}}) > 0$), providing an absolute scale signal that pushes the logits downward when the energy exceeds the ceiling boundary. Together, the OE loss shapes the distribution and the margin anchors its scale—each solving half of the problem.

The reject-class exception. Under the reject-class paradigm ($K+1$ outputs with a reject class), the explicit margin $\mathcal{L}_{\text{margin}}$ is structurally unnecessary. When a low-quality synthetic sample is routed to the centrifugal track ($\gamma \approx 1$), assigning it the pseudo-label $K+1$ yields cross-entropy loss whose gradient on the K ID logits ($\partial\mathcal{L}/\partial z_i = \text{softmax}(z)_i > 0$) is always positive, naturally suppressing ID-class energy. Unlike $\mathcal{L}_{\text{unif}}$, this suppression does not vanish at any energy level—it only stops when $\text{softmax}(z)_{K+1} \to 1$, providing an implicit energy constraint that makes the explicit margin redundant. Detailed gradient analysis is provided in Appendix E.

5.4 When the Margin Engages

A critical property of AEGR is that routing and the margin play complementary roles: the routing weight γ provides *adaptive allocation*—interpolating each synthetic sample between the ID classification track and the OOD-oriented track according to the teacher's quality assessment—while the margin is the *energy anchor* that produces the large ID-OOD separation. Concretely, the margin term is only engaged where it is needed. When generation quality is high (e.g., CIFAR-100 via EDM), the teacher's LSE on synthetic samples closely matches that of real ID data; because the margin ceiling is relaxed ($m_{\text{ceil}} = \infty$) on such datasets, the margin term itself never engages regardless of the routing weights, and the routed loss acts as a uniform-regularization target that keeps synthetic data from contaminating the ID manifold.

Algorithm 1 Adaptive Energy-Guided Routing (AEGR)

Require: Student f_θ, frozen teacher f_T, real ID $\mathcal{D}_{\text{real}}$, OOD $\mathcal{D}_{\text{ood}}$, synthetic ID $\mathcal{D}_{\text{syn}}$, routing percentile p_{out}, ceiling percentile p_{ceil}
$m_{\text{out}} \leftarrow \text{Percentile}_{p_{\text{out}}}(\{E_T(x) : x \in \mathcal{D}_{\text{real}}\})$
$m_{\text{ceil}} \leftarrow \text{Percentile}_{p_{\text{ceil}}}(\{E_T(x) : x \in \mathcal{D}_{\text{real}}\})$
for each training iteration **do**
  Sample batches $\mathcal{B}_{\text{real}}, \mathcal{B}_{\text{ood}}, \mathcal{B}_{\text{syn}}$
  Compute PGD perturbations for each batch
  $\mathcal{L}_{\text{real}} \leftarrow \frac{1}{|\mathcal{B}_{\text{real}}|} \sum \mathcal{L}_{\text{CE}}(f_\theta(x'), y)$
  $\mathcal{L}_{\text{ood}} \leftarrow \frac{1}{|\mathcal{B}_{\text{ood}}|} \sum \mathcal{L}_{\text{unif}}(f_\theta(x'))$
  $\gamma \leftarrow \sigma(m_{\text{out}} - E_T(x))$ for $(x, y) \in \mathcal{B}_{\text{syn}}$
  $\mathcal{L}_{\text{syn}} \leftarrow$ Eq. (6) per-sample γ
  $\theta \leftarrow \theta - \eta \nabla_\theta (\mathcal{L}_{\text{real}} + \mathcal{L}_{\text{ood}} + \mathcal{L}_{\text{syn}})$
end for

This is not a failure mode but a *designed property*: the margin's influence is confined to the samples routed to the OOD-oriented track, whose proportion is controlled by the routing percentile p_{out}. When quality drops (Tiny-ImageNet-200), the tightened ceiling ($p_{\text{ceil}} = 99.5$) activates the margin on the routed samples, and the framework transitions to its full routed form. The routing behavior is robust to the exact choice of routing percentile p_{out} within a reasonable range (ablation in Table 2); the ceiling percentile p_{ceil} is then set per dataset according to its generation quality—a single percentile rather than a full re-optimization of the training pipeline. Per-dataset default configurations are listed in Appendix B.

5.5 End-to-End Algorithm

Algorithm 1 summarizes the complete training procedure. For each training iteration, we sample mini-batches from three data sources: real ID, auxiliary OOD, and synthetic ID. Adversarial perturbations are computed via PGD for each data source independently. The real ID loss follows standard cross-entropy, the OOD loss follows uniform OE, and the synthetic ID loss follows the routed formulation of Equation (6). The teacher f_T is pre-trained once and frozen; its only role is to compute $E_T(x_{\text{syn}})$ for routing weight computation, which adds negligible overhead (a single forward pass per synthetic batch).

6 Experiments

6.1 Experimental Setup

All experiments use WRN-40-4 (Zagoruyko & Komodakis, 2016) for CIFAR and Tiny-ImageNet-200, and RAResNet-50 (Singh et al., 2023) for ImageNet-200. Synthetic ID data is generated with EDM (Karras et al., 2022) for CIFAR-10, CIFAR-100, and Tiny-ImageNet-200, and from the public ImageNet-1k-SD-1.4 dataset (Sariyildiz et al., 2023) (Stable Diffusion v1.4 (Rombach et al., 2022)) for ImageNet-200. The OE paradigm (Section 3.2) is our primary framework; reject-class results are reported separately. Adversarial training uses PGD-10 with $\epsilon = 8/255$ ($4/255$ for ImageNet-200). Detection uses the LSE energy score (Liu et al., 2020); MSP-based results and training overhead are in Appendices F and I.

Baselines. We compare against AT (Madry et al., 2017), OE (Hendrycks et al., 2018), ACET (Hein et al., 2019), ATOM (Chen et al., 2021), RODEO^U (Mirzaei et al., 2025), and AT-OE (adversarial training + uniform OE). RODEO^U operates in the reject-class paradigm; we adapt it to OE with a uniform-style OOD loss for comparison, using the LSE energy score. We ablate two configurations of our method: **AT-OE+SynthID** (naive synthetic ID data, no routing; "SynthID" abbreviates *synthetic in-distribution data*, unrelated to (Azizi et al., 2023)) and **AEGR** (full routing with margin). For each of CIFAR-10, CIFAR-100, and Tiny-ImageNet-200, EDM generates 1M samples (Karras et al., 2022), used in full (Appendix H.2 sweeps the effect of retaining only a top-LSE subset). The

Table 1: Main results under the OE paradigm. Acc and AA are ID accuracy (%); Clean^{o} and TAPGD^{lgt} are OOD-detection AUC (%). Gray rows de-emphasize reference and reject-class methods (AT, ACET, ATOM, RODEO^{U}) that are not the central comparison.

Dataset	Method	Acc	AA	Clean^{o}	TAPGD^{lgt}
CIFAR-10	AT	86.74	46.6	88.12	70.65
	ACET	81.6	0.0	91.9	38.3
	ATOM	82.0	13.5	91.0	41.4
	RODEO^{U}	85.99	51.37	95.37	83.67
	AT-OE	85.86	51.39	95.55	86.35
	AT-OE+SynthID	89.31	55.84	96.82	89.28
	AEGR (ours)	89.24	55.90	**97.14**	**89.36**
CIFAR-100	AT	61.21	25.59	68.70	46.30
	ACET	78.87	0.0	77.91	1.29
	ATOM	78.80	0.0	78.1	1.10
	RODEO^{U}	60.23	28.17	80.76	61.12
	AT-OE	60.44	28.66	85.91	68.40
	AT-OE+SynthID	66.02	31.62	87.97	69.46
	AEGR (ours)	66.09	31.71	**88.27**	**70.15**
Tiny-200	AT-OE	46.02	18.09	97.49	93.61
	AT-OE+SynthID	52.54	21.78	90.06	78.61
	AEGR (ours)	46.35	18.31	**99.21**	**95.61**
ImageNet-200	AT-OE	75.46	52.18	**76.41**	64.09
	AT-OE+SynthID	75.35	52.25	75.73	63.28
	AEGR (ours)	75.58	52.29	75.78	**64.88**

teacher f_T is an OE-calibrated WRN-40-4 trained on the same dataset (Section 5.1; sensitivity in Appendix H.3). Full training details are in Appendix B.

6.2 Main Results

Table 1 presents the main comparison. ACET and ATOM operate in the reject-class ($K+1$) paradigm and are evaluated with the energy score for completeness (less well-suited to their design); RODEO^{U}, which generates synthetic OOD data via diffusion, is reported on the CIFAR datasets where its synthetic OOD data is public—its release lacks such data for Tiny-ImageNet-200 and ImageNet-200. RODEO^{U} underperforms the real-OOD-based AT-OE baseline on both CIFAR datasets, so it is not used in later ablations. On CIFAR-10 and CIFAR-100, where EDM achieves high fidelity, AT-OE+SynthID substantially improves robust OOD detection (TAPGD AUC from 86.35% to 89.28% on CIFAR-10; from 68.40% to 69.46% on CIFAR-100), and AEGR further improves both robust and clean detection. On Tiny-ImageNet-200, where generation quality is lower, naive synthetic data severely degrades robustness (93.61%→78.61%); AEGR not only recovers the loss but exceeds the baseline, reaching 95.61% TAPGD AUC (+17.0 over naive SynthID, +2.0 over the no-synth baseline). On ImageNet-200, synthetic data slightly degrades robustness (64.09%→63.28%), and AEGR recovers to 64.88%[1] Across all four datasets, AEGR improves over both the AT-OE baseline and naive synthetic data usage, with the gains most pronounced where naive synthetic data is most harmful. Full per-dataset results and FPR-95 metrics are in Appendix G.

6.3 Ablation Studies

A natural question is whether a simpler teacher-guided filtering strategy—discarding low-LSE synthetic samples before training—could substitute for the per-sample routing in AEGR. We provide a direct comparison in Appendix H.2, which also examines combining filtering with AEGR.

Ablation 1: γ Routing Threshold. Table 2 ablates p_{out} on Tiny-ImageNet-200, where m_{out} is the p_{out}-th percentile of teacher LSE (Eq. 5). All routing-only configurations substantially improve over naive SynthID (78.61), generally improving at higher percentiles; the best result is at $p_{\text{out}} = 99.5$ (95.61), with $p_{\text{out}} = 99.99$ yielding a similar 95.59, indicating robustness to the exact choice of p_{out} as long as the most divergent samples are routed to the OOD branch. The default is $p_{\text{out}} = 99.99$

[1]ImageNet-200 uses the public ImageNet-1k-SD-1.4 dataset (Sariyildiz et al., 2023), whose Stable Diffusion v1.4 generator differs from the EDM used for the other datasets, so gain magnitudes are not directly comparable across generators.

Table 2: γ routing threshold on Tiny-ImageNet-200. All values are TAPGD^{lgt} AUC (↑).

SynthID	+γ (p=95)	+γ (p=97)	+γ (p=99)	+γ (p=99.5)	+γ (p=99.99)
78.61	94.66	94.54	94.89	**95.61**	95.59

for CIFAR-10/100 and ImageNet-200, and $p_{\text{out}} = 99.5$ for Tiny-ImageNet-200 (a higher percentile routes more samples to the OOD track; per-dataset details in Appendix B).

Ablation 2: Margin Effect. Table 3 compares an unrestricted ceiling ($m_{\text{ceil}} = \infty$, margin never activates) against a tight one ($p_{\text{ceil}} = 99.5$). On CIFAR-100, where generation quality is high, the margin gives no significant gain (70.15 vs. 70.04); on Tiny-ImageNet-200, where naive SynthID severely degrades robustness (93.61%→78.61%), tightening the ceiling improves performance from 94.22% to 95.61%. This contrast shows the margin is necessary when synthetic data quality cannot be guaranteed—the setting that matters most for real-world deployment.

Table 3: Margin effect across datasets. CIFAR-100 uses an unrestricted ceiling; Tiny-ImageNet-200 uses $p_{\text{ceil}} = 99.5$. All values are TAPGD^{lgt} AUC (↑).

Ceiling setting	CIFAR-100	Tiny-200
relaxed ($m_{\text{ceil}} = \infty$)	70.15	94.22
tight ($p_{\text{ceil}} = 99.5$)	70.04	95.61

Ablation 3: Mixing Strategy. We reproduce static batch mixing (fixed ratio $\alpha = 0.8$ of synthetic ID per batch, same EDM data) under the OE adversarial training framework (Zhang et al., 2023). Table 4 compares it against naive SynthID and AEGR. On CIFAR-100, static mixing slightly degrades the AT-OE baseline (68.40%→67.71%), while naive SynthID improves (69.46%)—even on high-quality data, blindly mixing a fixed proportion can dilute the ID signal. On Tiny-ImageNet-200, both static mixing and naive SynthID substantially underperform AEGR (95.61%), whose per-sample γ routing avoids the dilution that static mixing causes when low-quality samples are forced uniformly into the classification track.

Table 4: Comparison of static mixing vs. adaptive routing (same EDM data).

Configuration	CIFAR-100	Tiny-200
AT-OE (baseline)	68.40	93.61
SIO (static mixing)	67.71	92.28
AT-OE+SynthID	69.46	78.61
AEGR	**70.15**	**95.61**

6.4 Generalization to the Reject-Class Paradigm

Table 5 validates AEGR under the reject-class (K+1) paradigm. RODEO^r is listed on CIFAR-100, where its published setup and synthetic OOD data are available; its release does not cover a Tiny-ImageNet-200 inlier setup, so no row is listed there. On CIFAR-100, naive synthetic ID data improves over the AT-Rej baseline (67.81%→69.34%) and AEGR achieves further gains (69.57%). On Tiny-ImageNet-200, AT-Rej+SynthID degrades robustness (95.56%→70.30%), while AEGR recovers to 97.14%, exceeding the no-synthetic AT-Rej baseline. Under the K+1 paradigm, the reject output naturally serves as an energy sink, eliminating the need for an explicit margin—consistent with the analysis in Sec 5.3.

Table 5: Results under the reject-class ($K+1$) paradigm.

Dataset	Method	Acc	AA	Clean^o	TAPGD^{lgt}
CIFAR-100	RODEO^r	60.71	27.68	85.12	66.93
	AT-Rej	60.57	27.71	84.70	67.81
	AT-Rej+SynthID	65.93	31.59	86.63	69.34
	AEGR (rej)	66.07	31.66	**87.17**	**69.57**
Tiny-200	AT-Rej	48.39	18.13	97.72	95.56
	AT-Rej+SynthID	50.52	18.41	97.10	70.30
	AEGR (rej)	48.74	18.38	**99.08**	**97.14**

7 Conclusion

Synthetic ID data is not universally beneficial for robust OOD detection: high-fidelity samples help, while low-quality ones can cause severe degradation when naively incorporated. AEGR resolves this through two complementary mechanisms—*per-sample adaptive allocation*, in which the teacher-guided routing weight interpolates each synthetic sample between the ID and OOD-oriented tracks according to its assessed quality, and an *explicit energy margin* that anchors the energy scale the uniform OE loss structurally lacks. Our measured energy distributions confirm that the margin, rather than routing alone, drives the improvement, and the same mechanism generalizes to the reject-class paradigm. The teacher and diffusion generator add one-time costs, but both are standard for synthetic-data methods. We hope this encourages broader investigation of quality-conditioned strategies for robust OOD detection.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) for editing and polishing the manuscript text and for assisting with LaTeX formatting. We have not used generative AI tools for the conception of research ideas, the design of experiments, or the analysis of results, all of which were performed by the authors. We have reviewed all AI-assisted text and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide complete experimental settings and hyperparameter configurations in the appendix, including per-dataset default configurations for the routing and margin-ceiling percentiles. The anonymized source code and scripts used for all experiments are provided as supplementary material.

ETHICS STATEMENT

This work addresses technical challenges in robust out-of-distribution detection and does not involve human subjects, sensitive data, or potentially harmful applications. We do not foresee negative societal consequences specific to this work beyond those generally associated with machine learning systems.

REFERENCES

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

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves ImageNet classification. *Transactions on Machine Learning Research (TMLR)*, 2023.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, pp. 11192–11203, 2019.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. *In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. In *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, 2022.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 2206–2216, 2020.

Stanford CS231N. Tiny ImageNet challenge (cs231n). `https://www.kaggle.com/c/tiny-imagenet`, 2015.

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.

Lars Doorenbos, Raphael Sznitman, and Pablo Márquez-Neila. Non-linear outlier synthesis for out-of-distribution detection. *arXiv preprint arXiv:2411.13619*, 2024.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

Alvin Heng, Alexandre H. Thiery, and Harold Soh. Out-of-distribution detection with a single unconditional diffusion model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

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

Qilin Liao, Shuo Yang, Bo Zhao, Ping Luo, and Hengshuang Zhao. Boundary-based out-of-distribution data generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

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.

Hossein Mirzaei and Mackenzie W. Mathis. Adversarially robust out-of-distribution detection through Lyapunov-stabilized embeddings. In *International Conference on Learning Representations (ICLR)*, 2025.

Hossein Mirzaei, Mohammad Jafari, Hamid Reza Dehbashi, Ali Ansari, Sepehr Ghobadi, Masoud Hadi, Arshia Soltani Moakhar, Mohammad Azizmalayeri, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Rodeo: Robust outlier detection via exposing adaptive out-of-distribution samples. *arXiv preprint arXiv:2501.16971*, 2025.

Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5216–5223, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Vikash Sehwag, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 105–116, 2019.

Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36:13931–13955, 2023.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pp. 36246–36263. PMLR, 2023.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Suhee Yoon, Sanghyu Yoon, Ye Seul Sim, Sungik Choi, Kyungeun Lee, Hye-Seung Cho, Hankook Lee, and Woohyung Lim. Diffusion-based semantic outlier generation via nuisance awareness for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Ryan Luley, Yiran Chen, and Hai Li. Synthetic in-distribution data benefits out-of-distribution detection. *arXiv preprint arXiv:2303.14531*, 2023.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.