BAD-OOD: DISCOVERING HARMFUL SYNTHETIC DIFFUSION OUTLIERS VIA CONFIDENCE CALIBRATION

Donglin Ni[1], Zhixin Feng[1], Ke Li[1], Yue Zhang[2], Yonggang Qi[1]*
[1]Beijing University of Posts and Telecommunications, China, [2]HAOHAN Data, China
{nidl,fengzhixin,like1990,qiyg}@bupt.edu.cn
zhangyue@haohandata.com

ABSTRACT

Utilizing synthetic outlier samples has shown great promise in out-of-distribution (OOD) detection. In particular, impressive results have been achieved by employing diffusion models to generate synthetic outliers in the low-density manifold. However, guiding diffusion models to generate meaningful synthetic outliers remains challenging. The synthesized samples often fall either too close to the in-distribution (ID) data (risking overlap and ambiguity) or too far (leading to visually unrealistic results). Both extremes have been shown to degrade OOD detection performance. In this work, we propose a novel OOD synthesis framework that combines a pre-trained Representation Diffusion Model (RDM) with a simple yet effective classifier calibration strategy. RDM enables global semantic embedding generation without requiring auxiliary labels or text, producing diverse yet ID-relevant outliers, thereby facilitating a more compact ID-OOD decision boundary. To ensure the utility of these samples, we calibrate a binary classifier on both ID data and synthesized OODs to assign confidence-based anomaly scores. We find that mid-confidence outliers, i.e., those balancing realism and deviation, are most informative, and using them significantly boosts detection performance. Extensive experimental results validate the superiority of our calibrated OOD sampler over several strong baselines.

1 INTRODUCTION

Out-of-distribution (OOD) detection (Liu et al., 2020; Wang et al., 2021; Sun et al., 2022; Huang et al., 2021; Behpour et al., 2023; Sharifi et al., 2024) aims to prevent neural networks from making overconfident predictions when faced with unseen data. This capability is critical for deploying trustworthy machine learning systems in open-world environments, particularly as deep models are increasingly applied to safety-critical domains such as medical diagnostics (Wei & Wang, 2023) and autonomous driving (Liu et al., 2023a).

The core challenge of OOD detection lies in identifying unknown outliers that deviate from the training data distribution. Ideally, an OOD detector functions as a binary classifier that exhibits low uncertainty for in-distribution (ID) data while flagging OOD data as uncertain. A promising recent direction for enhancing OOD detection focuses on synthesizing latent outliers (Lee et al., 2018a; Du et al., 2022; Tao et al., 2023), which can provide explicit supervision and help tighten the decision boundary around the ID manifold.

More recently, Dream-OOD (Du et al., 2023) is proposed to generate visually interpretable OOD images by sampling low-likelihood embeddings in latent space and decoding them through a pixel-space diffusion model. Specifically, it first identifies boundary ID points using KNN distances, then samples outlier embeddings from a Gaussian kernel centered at these points. However, choosing an appropriate sampling radius is non-trivial: samples drawn too far from ID data often result in visually meaningless or unrealistic images, while those too close risk overlapping with ID and confusing the detector. Dream-OOD attempts to balance this by tuning the variance of the Gaussian kernel, but this hyperparameter is difficult to optimize and often dataset- and class-dependent, making the synthesis process fragile. Moreover, the diffusion-based image generation process itself introduces inherent randomness (Ho et al., 2020; Song et al., 2022a; Song et al.), further affecting the reliability of the synthesized outliers. As a result, the above issues can lead to harmful outliers, either resembling

*Corresponding author

covariate-shifted ID or being visually implausible. Despite their importance, these concerns have been largely underexplored. This raises a key question: *Can we synthesize OOD samples that are anchored to the ID distribution – diverse enough to probe the well-covered boundary, yet reliable in quality – to better separate ID and OOD distributions?*

To achieve the goal, two critical challenges must be overcome: (i) the model should be expressive enough to synthesize diverse yet ID-relevant OOD samples to facilitate a compact decision boundary, and (ii) a robust metric is required to measure the quality of generated outliers, enabling the filtering of harmful samples. To this end, we propose a novel OOD synthesis framework that couples a representation diffusion model (RDM) (Li et al., 2024) with a classifier calibration strategy to jointly address both objectives.

Instead of relying on the conventional diffusion models (Rombach et al., 2022), which sample either in pixel space or instance-level latent space, we adopt the Representation Diffusion Model (RDM) for its ability to generate global semantic representations. This enables the synthesis of diverse outlier embeddings that remain semantically aligned with the ID data. This facilitates the construction of a more compact and robust decision boundary. Moreover, RDM requires no auxiliary class labels or textual descriptions, which are often unavailable, ambiguous, or costly to acquire in real-world settings. In our approach, a self-guided minority metric is applied in conjunction with a pre-trained RDM to generate OOD embeddings, which are subsequently decoded into pixel-space OOD images.

To tackle the second challenge, i.e., evaluating outlier quality, we introduce a simple yet effective classifier calibration strategy. Using both ID data and the full spectrum of synthesized OOD images (including unreliable ones), we train a binary classifier whose prediction probabilities of ID or OOD samples are calibrated to reflect their actual confidence (Guo et al., 2017). Crucially, we observe that the calibrated classifier can already identify harmful OOD samples: low-confidence outliers often resemble covariate-shifted ID samples, while high-confidence ones tend to be visually unrealistic or semantically irrelevant. We show that the performance is improved by retaining the OOD detector using only mid-confidence OOD samples, i.e., those that strike a balance between diversity and relevance. Furthermore, we demonstrate that our calibration strategy is model-agnostic and can be seamlessly integrated into Dream-OOD, leading to substantial gains and highlighting the generality and practicality of our approach.

Our major contributions can be summarized as follows: (i) We extend a self-guided minority metric into latent diffusion space that enables pre-trained RDM to generate diverse ID-relevant synthetic outliers without any auxiliary class labels or text descriptions, which might be difficult to obtain in practice. (ii) A classifier calibration strategy is incorporated into OOD synthesis to measure the anomaly degree of the outliers, thereby preventing harmful OOD data, thus providing more reliable supervision. We demonstrate that the calibrated classifier captures both semantic and covariate shifts by evaluating on Syn-IS (Long et al., 2024), a benchmark designed to disentangle different types of distributional shifts. (iii) Extensive experimental results validate the effectiveness of our proposed OOD calibration method, which outperforms state-of-the-art outlier synthesizers for OOD detection on several benchmarks.

2 Related Work

OOD detection with outlier exposure. Motivated by the difficulty of acquiring large, high-quality OOD datasets, various techniques have emerged to synthesize auxiliary OOD data that can learn an explicit separation boundary between ID and OOD. Early attempts like Lee et al. (2018a) adopt GAN to generate pixel-level OOD data in a visually interpretable manner. Later, VOS (Du et al., 2022) modeled ID features as a multivariate Gaussian and synthesized outliers by sampling from low-density regions. NPOS (Tao et al., 2023) relaxed the Gaussian assumption by using CLIP (Radford et al., 2021) features to form semantic clusters, and identified sparse feature regions via k-nearest neighbors. Dream-OOD (Du et al., 2023) further bridged feature-space synthesis with visual interpretability by integrating Stable Diffusion (Rombach et al., 2022), enabling image-level realization of outlier features sampled near the ID boundary. Building upon these advancements, we propose a novel *diffusion outlier guidance* strategy that not only retains visual interpretability through diffusion-based generation, but also enables more precise boundary probing without relying on predefined semantic priors. Our method improves the quality and diversity of synthetic outliers, leading to more robust and compact OOD decision boundaries.

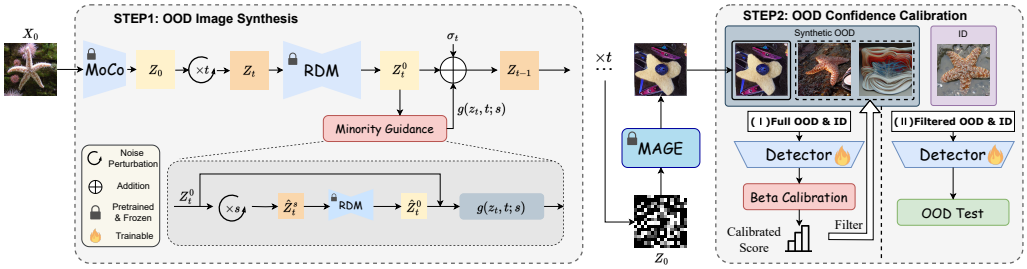


Figure 1: **Overview of our proposed method Bad-OOD.** Bad-OOD first generates auxiliary OOD samples by integrating minority guidance into the denoise process. In the second stage, the detector is initially trained using the entirety of the synthetic OOD samples in conjunction with the ID samples. Subsequent to the beta calibration of the detector, the calibrated scores are employed to filter the previously synthetic OOD samples, thereby excluding those with excessively high or low confidence levels. The refined OOD samples, with the ID samples, are then utilized to retrain the detector from scratch.

Diffusion models for OOD detection. Unlike traditional OOD detection methods, recent approaches have leveraged diffusion-based architectures to assess how well a sample aligns with the learned data distribution—typically by reconstructing the input or quantifying deviation during the generative process. Liu et al. (2023b) mask a central region of the image and apply diffusion inpainting, using the restoration quality as an OOD indicator. Gao et al. (2023) combines diffusion inversion with classifier guidance, where high reconstruction error signals a potential OOD sample. A further extension Yang et al. (2024b) adopts multi-layer representations of samples and trains a diffusion model to capture reconstruction error. Additionally, Sun et al. (2022) uses the rate-of-change from a sample toward a standard Gaussian distribution as the OOD score. In contrast to these reconstruction-based approaches, our proposed method is, to our knowledge, the first to directly synthesize feature-level outliers via a diffusion model. This enables flexible, visualizable, and controllable generation of diverse OOD data, which better supports decision boundary learning without requiring reconstruction-based inference.

OOD detection and distribution shift measurement. OOD score quantifies how anomalous a sample is concerning the in-distribution (ID) data. MSP (Hendrycks & Gimpel, 2018) utilizes the highest softmax output of a classifier as the OOD score. Despite its simplicity, MSP often suffers from overconfident predictions (Liu et al., 2020), limiting its reliability in estimating OOD severity. ODIN (Liang et al., 2017) improves the separation between ID and OOD samples via temperature scaling and input perturbation. Mahalanobis-based scoring (Lee et al., 2018b) estimates confidence using the distance to the nearest class-conditional Gaussian distribution in feature space. The energy-based method (Liu et al., 2020) mitigates softmax overconfidence by deriving scores directly from the logit outputs. GradNorm (Huang et al., 2021) proposes using the norm of gradients with respect to the input as a signal for OOD detection. KNN-OOD (Sun et al., 2022) utilizes local density differences based on K-nearest neighbor distances between ID and OOD features. Different from prior works, our confidence calibration strategy leverages both reliable and unreliable synthesized outliers to learn an interpretable OOD scoring function. It distinguishes harmful samples from informative ones by aligning prediction confidence with anomaly severity, enabling more effective and targeted supervision for OOD detection.

3 Outlier Synthesis with Diffusion Models

An overview of the proposed framework is shown in Figure 1, with key components detailed below.

3.1 Conditional ID Data Representation

Our ultimate goal is to generate outliers conditioned on ID data, such that help build a compact boundary around ID data. To achieve it, we can first obtain a representation of the conditional ID data as shown in Figure 1. Specifically, given any ID image $\boldsymbol{x}_0$, we utilize MoCo-v3 (Chen et al., 2021), which is pretrained using self-supervised contrastive learning, as image encoder to map the pixel-space image into the latent feature space. The obtained ID image feature $\boldsymbol{z}_0$ intuitively serves as an anchor point for producing its anomaly variants.

3.2 LATENT MINORITY SYNTHESIS USING RDM

Given the conditional representation of ID data $\boldsymbol{z}_0$, we employ representation diffusion model (RDM) (Li et al., 2024) to produce outlier embeddings. We choose RDM since it allows us to sample from a representation distribution without providing textual conditions, such as class labels or text descriptions, which might be unavailable or difficult to obtain when synthesizing anomalies. Firstly, we add noise to the conditional representation $\boldsymbol{z}_0$ of each ID image over t time steps to obtain a noisy version $\boldsymbol{z}_t$. Inspired by Um & Ye (2024), we extend the minority metric into latent space based on the denoising timestep t and the posterior mean $\hat{\boldsymbol{z}}_0$, which is obtained via the Tweedie's formula (Efron, 2011) and the pretrained diffusion model. Formally, it is defined as:

$$\mathcal{L}(\boldsymbol{z}_0; t) := \mathbb{E}_{q_{\alpha_t}(\boldsymbol{z}_t|\boldsymbol{z}_0)}[d(\boldsymbol{z}_0, \hat{\boldsymbol{z}}_0(\boldsymbol{z}_t))], \tag{1}$$

where $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)$ denotes the model's estimate of the clean latent $\boldsymbol{z}_0$ given the noisy latent $\boldsymbol{z}_t$ at timestep t, since the true $\boldsymbol{z}_0$ is only obtained at the end of the reverse process, this intermediate prediction is necessarily used for guiding OOD synthesis. $d(\cdot, \cdot)$ is a distance metric, e.g., L1 or L2, to measure the discrepancy between two features. Intuitively, a higher minority score indicates sample in a low-likelihood region farther from the anchor $\boldsymbol{z}_0$, thus being more abnormal. Therefore, we can derive the sampling towards low-likelihood areas by leveraging the gradient of the minority score. Specifically, for a noisy latent sample $\boldsymbol{z}_t$ at timestep t, we first exploit the pretrained RDM $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})$ to denoise $\boldsymbol{z}_t$ into $\boldsymbol{z}_t^0$, i.e., the predicted clean latent at timestep t. We then reintroduce s additional noise steps to obtain $\hat{\boldsymbol{z}}_t^s$, which is denoise again into $\hat{\boldsymbol{z}}_t^0$. According to the minority score mentioned above, our guided gradient $\boldsymbol{g}$ can be obtained by:

$$\boldsymbol{g}(\boldsymbol{z}_t, t; s) := \nabla_{\boldsymbol{z}_t}\mathcal{L}(\boldsymbol{z}_t^0; s) = \nabla_{\boldsymbol{z}_t}\mathbb{E}_{q_{\alpha_s}(\hat{\boldsymbol{z}}_t^s|\boldsymbol{z}_t^0)}[d(\boldsymbol{z}_t^0, \hat{\boldsymbol{z}}_t^0)))]. \tag{2}$$

Lastly, as shown in Figure 1, we follow the standard classifier-guided formulation in ADM (Dhariwal & Nichol, 2021) and use the negative gradient direction indicated by minority guidance to push samples away from dense ID regions, thereby resulting in the outlier representations that naturally occupy the periphery of the ID distribution while retaining coherent structures or semantic cues.

3.3 OUTLIER IMAGE GENERATOR

Simialr to Dream-OOD, we explicitly generate outliers in pixel space, yielding visually interpretable samples, preserving fine-grained details such as texture and color, and aligns more naturally with downstream tasks that operate directly on images. Specifically, MAGE (Li et al., 2023) is leveraged as the image generator to convert the outlier embeddings in latent space into images. During inference, a fully masked image serves as the input of generator, and MAGE reconstructs an output guided by the outlier embeddings.

3.4 ENERGY-BASED OOD REGULARIZER

In order to utilize synthetic OOD samples for the separation of ID and OOD distributions, we apply the energy-based uncertainty regularization loss following Du et al. (2022). Specifically, assume the OOD detector's output logit of the n-th ID class is $f_n(\mathbf{x}; \theta)$, then the energy function can be formulated as the LogSumExp operation among all these N logits:

$$E(\mathbf{x}; \theta) := -\log \sum_{n=1}^{N} e^{f_n(\mathbf{x};\theta)}. \tag{3}$$

Then the uncertainty loss can be defined using a binary sigmoid loss based on the energy function, and the threshold for distinguishing ID and OOD should be set to zero:

$$\mathcal{L}_{\text{uncertainty}} := \mathbb{E}_{\mathbf{v}\mathcal{OOD}}\left[-\log \frac{1}{1+\exp^{-\phi(E(\mathbf{v};\theta))}}\right] + \mathbb{E}_{\mathbf{x}\mathcal{ID}}\left[-\log \frac{\exp^{-\phi(E(\mathbf{x};\theta))}}{1+\exp^{-\phi(E(\mathbf{x};\theta))}}\right], \tag{4}$$

where ϕ denotes a nonlinear MLP. The overall training objective is to optimize the cross entropy for multi-class classification within ID data coupled with the uncertainty regularizer:

$$\mathcal{L}_{all} = \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{uncertainty}}, \tag{5}$$

where β is to balance the effect of the regularization term. Once obtained the classifier, we can measure the anomaly degree of OOD samples by the sigmoid of the energy function $E(\mathbf{x}; \theta)$. Formally, we obtain:

$$S(\mathbf{x}; \theta) := \sigma(E(\mathbf{x}; \theta)) = \frac{1}{1+e^{-E(\mathbf{x};\theta)}}. \tag{6}$$

3.5 CONFIDENCE CALIBRATION

Empirically, we find that extreme outliers (i.e., overly noisy near-OOD or semantically irrelevant far-OOD samples) are common across synthetic OOD generators, and calibration substantially changes their prevalence (see Appendix G.2). Therefore, we propose a calibration refined strategy, which leverages a calibrated detector to score samples from an auxiliary dataset and systematically selects high-quality data. Essentially, calibration aims to align the confidence output by the classifier with the actual probability, and the confidence output by a perfectly calibrated detector can reflect the degree of anomaly. Therefore, we introduce beta calibration (Kull et al., 2017), which is not only applicable to sigmoid-based output score but also capable of fitting a wide range of probability distributions. Specifically, given an uncalibrated classifier $\mathcal{C}$ and an auxiliary calibration dataset $\mathcal{D}$, beta calibration learns a map function μ_{beta} that transforms the raw scores S of samples from $\mathcal{D}$ into new scores S_c, ensuring the new scores S_c closely approximate the actual probabilities within $\mathcal{D}$. The formula for the mapping function follows:

$$\mu_{beta}(s; a, b, c) = \frac{1}{1 + 1/\left(e^{c}\frac{s^{a}}{(1-s)^{b}}\right)}, \tag{7}$$

where the parameters a, b and c are learnable with $a, b \geq 0$. Practically, we perform calibration using a hold-out validation set consisting of 10% of the ImageNet-100 samples as the ID data and ImageNet-OOD dataset for OOD (see Appendix C.2 for more implementation details). Unlike simpler methods such as temperature (Liang et al., 2017) or Platt scaling (Berta et al., 2024), beta calibration introduces parameters (a, b, c in Eq.(7)) that can better capture skewness and heavy tails, which are typical in the asymmetric, long-tailed score distributions of generated samples. Please refer to Appendix E for more analysis of adopting beta calibration. Once the detector has been calibrated, it can be used to evaluate the anomaly degree of the synthetic OOD samples, which is formulated as:

$$S_c(\mathbf{x}; \theta) = \mu_{beta}(S(\mathbf{x}; \theta); a, b, c). \tag{8}$$

4 EXPERIMENTS

In this section, we first compare the OOD detection performance of our approach with several strong competitors. Then, we provide a thorough analysis about the effectiveness and properties of our proposed confidence calibrated OOD detector.

4.1 EXPERIMENTAL SETUP

Dataset. To evaluate OOD detection performance, we follow the experimental settings of Tao et al. (2023); Du et al. (2023) and use ImageNet-100 as the ID dataset, which is carefully selected from the original ImageNet-1k dataset (Deng et al., 2009). Additionally, we adopt the same OOD test datasets as in Huang & Li (2021); Tao et al. (2023); Du et al. (2023), including subsets of iNaturalist (Horn et al., 2018), SUN (Xiao et al., 2010), Places (Zhou et al., 2018), and Textures (Cimpoi et al., 2013). We also follow the common practice of adopting CIFAR-100 as the ID data for training as well. More details are provided in the Appendix A.1.

Implementation details. We implement our framework with PyTorch. MoCo-v3 ViT-L is utilized for extracting representation of ID data, then representation diffusion model (RDM) is employed to generate the synthetic embeddings of outliers, and MAGE-L is adopted for representation-conditioned image generation to decode the outlier embeddings into pixel space. During outlier embedding synthesis, we adopt the DDIM (Song et al., 2022a) sampler to predict the denoised version ($\boldsymbol{z}_t^0$) of the noisy latent of ID data $\boldsymbol{z}_t$ (we empirically set $t = 300$). To obtain the gradient of the minority metric, we further perturb $\boldsymbol{z}_t^0$ with $s = 125$ forward DDPM steps. Then given a specific minority guidance scale, we generate 1,300 OOD samples per class, resulting in a total of 130,000 synthetic outlier images. To train the OOD detector, we employ ResNet-34 as the default backbone and train it for 100 epochs using 100k calibrated mid-confidence OOD samples from scratch. We set the learning rate scheduler with 5 epochs of linear warmup followed by cosine annealing, a momentum of 0.9, a weight decay of 5e-4, an initial learning rate of 0.1, and a batch size of 160. The uncertainty regularization parameter β is set to 2.0. More details are provided in the Appendix B and C.

Competitors. We compare several strong OOD detection baseline methods, categorized into non-synthesis-based and synthesis-based methods. The majority of existing methods fall into the non-synthesis category. Logit-based methods such as MSP (Hendrycks & Gimpel, 2018), ODIN (Liang et al., 2017), GODIN (Hsu et al., 2020), and Energy (Liu et al., 2020) operate directly on the

Table 1: OOD detection results for IMAGENET-100 and CIFAR-100 as the in-distribution data. The bold and underlined numerals denote the optimal and suboptimal results among the synthesis-based methods, respectively. All the results of Bad-OOD were trained from scratch using the calibrated mid-confidence OOD samples.

Methods	ImageNet-100 as In-distribution									
	iNATURALIST		PLACES		SUN		TEXTURES		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP (Hendrycks & Gimpel, 2018)	31.80	94.98	47.10	90.84	47.60	90.86	65.80	83.34	48.08	90.01
ODIN (Liang et al., 2017)	24.40	95.92	50.30	90.20	44.90	91.55	61.00	81.37	45.15	89.76
Mahalanobis (Lee et al., 2018b)	91.60	75.16	96.70	60.87	97.40	62.23	36.50	91.43	80.55	72.42
Energy (Liu et al., 2020)	32.50	94.82	50.80	90.76	47.60	91.71	63.80	80.54	48.68	89.46
GODIN (Hsu et al., 2020)	39.90	93.94	59.70	89.20	58.70	90.65	39.90	92.71	49.55	91.628
KNN (Sun et al., 2022)	28.67	95.57	65.83	88.72	58.08	90.17	12.92	90.37	41.38	91.20
ViM (Wang et al., 2022)	75.50	87.18	88.30	81.25	88.70	81.37	15.60	96.63	67.03	86.61
ReAct (Sun et al., 2021)	22.40	96.05	45.10	92.28	37.90	93.04	59.30	85.19	41.17	91.64
DICE (Sun & Li, 2022)	37.30	92.51	53.80	87.75	45.60	89.21	50.00	83.27	46.67	88.19
GAN (Lee et al., 2018a)	83.10	71.35	83.20	69.85	84.40	67.56	91.00	59.16	85.42	66.98
VOS (Du et al., 2022)	43.00	93.77	47.60	91.77	39.40	93.17	66.10	81.42	49.02	90.03
NPOS (Tao et al., 2023)	53.84	86.52	59.66	83.50	53.54	87.99	**8.98**	**98.13**	44.00	89.04
Dream-OOD (Du et al., 2023)	26.15	**95.87**	39.90	93.59	38.65	93.09	57.40	84.12	40.52	91.67
w/ Calib.	**25.62**	95.48	37.29	93.09	36.67	92.95	54.06	86.37	38.41	91.97
Bad-OOD	26.25	95.53	**32.29**	**93.76**	**34.17**	**93.48**	53.33	86.01	**36.51**	**92.20**
ResNetv2-101										
RankFeat (Song et al., 2022b)	41.31	91.91	39.34	90.93	29.27	94.07	37.29	91.70	36.80	92.15
Bad-OOD	25.10	95.51	32.50	93.75	32.40	93.72	52.71	86.14	**35.68**	**92.28**
	CIFAR-100 as In-distribution									
	SVHN		PLACES		LSUN		TEXTURES		ISUN	
Fake it (Mirzaei et al., 2022)	85.15	77.53	77.05	76.91	56.40	76.31	70.55	80.77	67.75	84.95
Dream-OOD (Du et al., 2023)	59.00	86.86	72.45	79.98	24.55	95.33	48.10	88.04	2.15	99.21
w/ Calib.	55.74	88.45	70.62	80.32	24.08	95.09	41.05	89.99	1.24	99.22
SONA (Yoon et al., 2025)	**3.10**	**99.39**	**44.00**	**88.35**	**18.20**	**96.19**	58.90	85.20	63.10	86.17
Bad-OOD	54.90	89.16	68.75	86.04	24.54	94.22	**40.21**	**91.56**	**1.10**	**99.72**

classifier's output logits, applying post-processing techniques like temperature scaling or energy scores to differentiate between ID and OOD samples. Distance-based methods, including Mahalanobis (Lee et al., 2018b) and KNN-OOD (Sun et al., 2022), leverage statistical or geometric distances in feature space to estimate how far a sample deviates from the training distribution. Feature-based methods such as ViM (Wang et al., 2022), ReAct (Sun et al., 2021), and DICE (Sun & Li, 2022) further exploit internal network representations—e.g., by measuring residuals (ViM), clipping activations (ReAct), or sparsifying weight contributions (DICE)—to improve OOD separability and robustness. We also compare with RankFeat (Song et al., 2022b), which leverages the difference of singular value distributions between ID and OOD features for OOD detection. For synthesis-based methods, OOD generation has evolved from multivariate Gaussians to GANs (Goodfellow et al., 2014), and more recently to diffusion models (Rombach et al., 2022). Dream-OOD (Du et al., 2023) generates outlier description features and employs a pre-trained diffusion model as the image decoder. SONA (Yoon et al., 2025) introduces mutual interference among semantic regions of different categories during the generation process of the diffusion model, ultimately leading to OOD results. Fake it (Mirzaei et al., 2022) utilizes early-stopped SDEs to synthesize near-distribution OOD supplementary data.

Evaluation metrics. Three conventional metrics are employed: (i) **FPR95** - Measures the false positive rate of OOD samples when the true positive rate for ID samples is 95%, making it particularly suitable for scenarios where high recall of OOD samples is required. (ii) **AUROC** - The area under the receiver operating characteristic curve, which reflects the detector's ability to distinguish between ID and OOD samples across varying thresholds. (iii) **AUPR** - Area under the precision-recall curve which places greater emphasis on the detector's capability to correctly identify OOD samples.

4.2 OOD DETECTION RESULTS

Quantitative Results. As shown in Table 1, our method consistently outperforms baseline models across multiple key OOD detection metrics. In particular, (i) it achieves a clear improvement over Dream-OOD (Du et al., 2023), demonstrating the advantage of our RDM-based synthesis and calibration strategy. (ii) Notably, integrating our confidence calibration into Dream-OOD (w/ Calib. in brown) also yields significant performance gains, highlighting the modularity and general applicability of our approach. (iii) We observe a further performance boost when upgrading the backbone from ResNet-34 to ResNet-101, indicating that stronger classifiers can better leverage our synthesized OOD samples. (iv) Similar trends hold when switching the in-distribution dataset from ImageNet-100 to CIFAR-100, confirming the robustness and scalability of our framework across

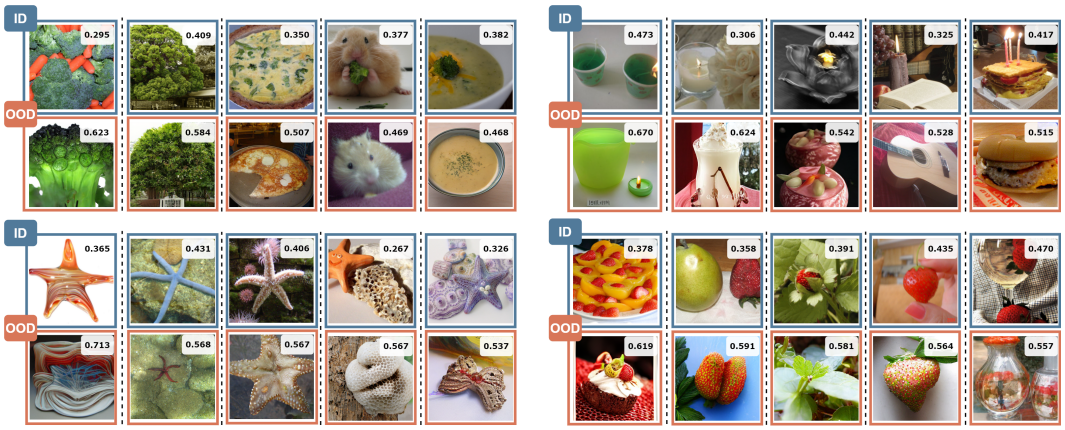


Figure 2: **ID reference samples and synthetic OOD of class broccoli, candle, starfish and strawberry.** The anomaly score is provided in the upper right of each sample.

Figure 3: **Three types of OOD samples of class broccoli, starfish and strawberry.** For OOD-II and OOD-III, the class predicted by CLIP (Radford et al., 2021) is demonstrated under each image.

varying ID complexities. Due to space limitations, we refer the reader to the Appendix for additional results on synthesis methods and architectures variants (Appendix F), generalization and robustness analysis (Appendix G), similar diffusion-based methods (Appendix J), and fine granularity OpenOOD v1.5 (Zhang et al., 2023) evaluation (Appendix K).

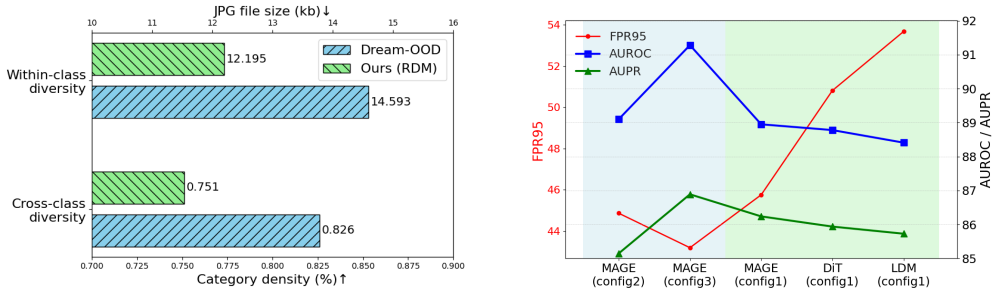
Qualitative Results. To further gain some insights about the quality of the generated outlier images, some examples are demonstrated in Figure 2. We can see from the figure that ours can generate high-resolution, semantically meaningful outlier samples similar to Dream-OOD. Uniquely, our method is capable of quantifying the abnormality of each OOD image by estimating its probability of belonging to the ID class, i.e., its classification confidence as ID data. It shows that a higher score suggests that the OOD sample closely resembles ID data, while a lower score indicates a greater degree of covariance shift, i.e., appearance anomaly. Additional visualizations of synthetic OOD samples are provided in the Appendix L.

4.3 More Analysis

Where is bad OOD? We investigate how our confidence-calibrated binary OOD classifier can locate "toxic" outliers via ablation studies using synthetic OOD samples with varying anomaly levels. Specifically, we verify whether using outliers with either too high or too low OOD scores for training will harm the performance of OOD detector. To reduce randomness, we produce synthetic outliers using multiple minority guidance scales (i.e., 0.025, 0.05, 0.075), yielding 130k synthetic outlier samples (denoted as "Mixed") as OOD candidates. After filtering with our calibrated classifier, we train an OOD detector from scratch with all ID samples ($\sim$130k). Results are presented in Table 2. We observe that: (i) Using the full set of 130k synthetic outliers without filtering ("$\times$") yields the worst performance, and randomly

Table 2: Effect with various minority scales.

Min. Scale	Calib.	FPR95↓	AUROC↑	AUPR↑
Mixed	$\times$	54.06	88.75	86.30
	Random	53.52	89.43	86.86
	High (0.3-1.0)	51.20	89.20	86.61
	Mid (0.3-0.5)	39.69	91.71	87.13
	Low (0-0.5)	45.65	90.12	86.84
0.025	Random	52.01	90.00	86.87
	✓	41.46	91.25	86.81
0.05	Random	43.20	91.28	86.89
	✓	36.51	92.20	87.32
0.075	Random	44.09	91.05	86.64
	✓	38.96	92.04	87.33



(a) (b)

Figure 4: (a) The diversity of OOD samples generated by our RDM-based method and Dream-OOD. (b) Performance using different image generators.

selecting 100k OOD samples achieves similar results. (ii) The best performance is obtained using mid-level anomaly outliers (OOD scores 0.3–0.5, denoted Mid), likely because these samples are sufficiently distinct yet still near the ID distribution, facilitating a compact decision boundary. (iii) Including near-OOD samples (0.5–1, High) significantly degrades performance (FPR95 increases from 39.69 to 51.20) due to contamination by nearly ID-identical samples; similarly, far-OOD samples provide little benefit as excessively abnormal samples do not aid in learning a compact boundary. (iv) Experiments with specific minority guidance scales further confirm that filtering with the calibrated classifier consistently improves results, demonstrating the robustness of our approach.

What are the OOD samples? To examine the content of OOD samples before and after filtering with our confidence-calibrated classifier, we use a pre-trained CLIP image encoder (Radford et al., 2021) to categorize 500 randomly selected synthetic outliers alongside their corresponding ID images. We identify three main types: (i) OOD-I: covariance shift outliers with shape/style changes but still recognizable as the original class; (ii) OOD-II: semantic shift outliers, correctly recognized by CLIP but with different class labels; (iii) OOD-III: severe semantic and appearance changes, hardly recognizable even by humans (examples in Figure 3, additional visualizations in Appendix L.2). After filtering with the optimal OOD score threshold (0.3–0.5), most removed samples belong to OOD-I and OOD-III, demonstrating that our method effectively removes near-ID and meaningless far-OOD samples (Table 3). Moreover, we further investigate the underlying mechanism for selecting mid-confidence samples and provide expanded quantitative and causal analyses to support this choice (Appendix H).

Table 3: OOD before and after calibration.

	OOD-I	OOD-II	OOD-III
Before	288 (57.6%)	45 (9.0%)	167 (33.4%)
After	233 (58.4%)	38 (9.5%)	128 (32.1%)
Reduce	55 (19.1%)	7 (15.6%)	39 (23.4%)

OOD Samples Diversity. Intuitively, compared to Dream-OOD (Du et al., 2023), our approach replaces class-conditional anchors with ID reference samples for OOD generation, removing the dependency on class labels and enabling sampling from a broader latent space. To quantify the diversity of synthesized OOD samples, we assess both cross-category and within-category diversity. Specifically, we generate OOD samples from 100 randomly selected ID categories, with 100 samples per category. For cross-category diversity, we use a pretrained ResNet-50 classifier to predict labels for all synthesized samples and compute the ratio of unique predicted categories to the total number of ID categories. For within-category diversity, we adopt the metric used in ImageNet (Deng et al., 2009), measuring the average JPEG file size of OOD samples within each category as a proxy for visual variability. As shown in Figure 4a, the RDM produces OOD samples with greater within-category diversity, which in turn encourages a more compact boundary around ID data and improves the generalization ability of the OOD detector.

Impact of Image Generators. We further investigate the how different image generators affect the quality of synthesized OOD samples and, consequently, the performance of the OOD detector. As illustrated in Figure 4b, we evaluated three image generator architectures, i.e., MAGE (Li et al., 2023), DiT (Peebles & Xie, 2023), and LDM (Rombach et al., 2022), under the MoCo-B & unconditional RDM configuration (config 1). We can see that MAGE-generated images contribute more effectively to regularizing the OOD detector. We hypothesize that compared to the diffusion-based generator, masked image modeling-based generation demonstrates stronger content understanding, which facilitates the OOD detector's learning process. Additionally, we tested other variants, including MoCo-L & class-conditional RDM (config 2) and MoCo-L & unconditional RDM (config 3), to further assess the role of encoder-decoder combinations. Moreover, an ablation that trains the detector purely in the MoCo latent space shows consistently inferior performance to pixel-space training,

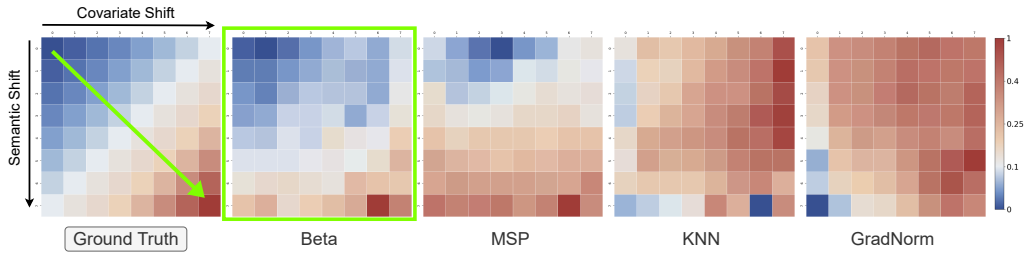


Figure 5: **Comparison of our approach and other competitors.** In each graph, the semantic shift increases from top to bottom, and the covariate shift increases from left to right. All the scores are rescaled to range between 0 and 1, and the color red indicates a higher anomaly degree.

suggesting that decoding improves the semantic fidelity, visual separability, and interpretability of synthesized OOD samples (see Appendix I).

Computation Cost Analysis. Since our method samples features at a fixed timestep and filters them with a lightweight confidence model, it is much cheaper than iterative diffusion approaches. The most time-consuming step is image generation via MAGE (single forward pass). On a single Nvidia 3090 GPU, generating 100 OOD ImageNet-100 images takes $\sim$50s, 2.9s for MoCo feature extraction, 18.2s for RDM perturbation and minority guidance, and 30.3s for latent-to-pixel generation.

4.4 Generalizability on Semantic and Covariate Shift

Here, we show that our confidence-calibrated OOD detector effectively captures both semantic and covariate shifts. We evaluate it on Syn-IS (Long et al., 2024), a benchmark specifically designed to disentangle and quantify different types of distributional shifts, which contains high-quality generated images with more diverse covariate contents.

Dataset. SynIS (Long et al., 2024), which systematically quantifies different semantic and covariance shifts relative to the ImageNet-1K dataset, is employed for evaluation. Specifically, images in SynIS are partitioned into several subsets representing eight levels of semantic shift and eight levels of covariance shift, resulting in a total of 64 distinct levels combined semantic and covariance shift for OOD images (More details in Appendix A.2).

Baselines. Competitors include: MSP (Hendrycks & Gimpel, 2018), which is a softmax-based score with the higher value indicating the higher probability of being OOD. KNN (Sun et al., 2022) uses the feature distance, with the larger distance indicating the higher probability of being OOD. GradNorm (Huang et al., 2021) adopts gradient information, with the higher back-propagated gradient indicating the higher probability of being OOD.

Evaluation Metrics. We calculate the average anomaly scores of the 64 subsets to compare the beta score employed in our method against other classical scoring metrics.

Results. As shown in Figure 5, different scoring methods exhibit distinct preferences toward distributional shifts: softmax-based scores such as MSP tend to respond primarily to semantic anomalies while showing low sensitivity to covariate shifts; in contrast, distance-based methods like KNN are more sensitive to covariate shifts but often fail to capture semantic deviations. Gradient-based approaches such as GradNorm struggle to reflect the degree of abnormality in both cases. In comparison, our confidence-calibrated detector more accurately reflects the severity of abnormalities, regardless of whether they arise from semantic or covariate shifts. This improved sensitivity to diverse distributional deviations stems from our minority-guided generation strategy, which promotes a compact and discriminative decision boundary between ID and OOD samples.

5 Conclusion

In this work, we introduced a novel OOD synthesis approach that integrates a classifier confidence calibration strategy with a pre-trained latent diffusion model. Unlike prior methods, our approach enables controlled sampling in the latent space without relying on auxiliary labels or text descriptions, ensuring the synthesized OOD samples remain both diverse and semantically relevant to the ID distribution. Moreover, we demonstrated that a confidence-calibrated classifier can effectively measure the anomaly degree of synthetic OOD samples, allowing for the removal of harmful outliers that could hinder OOD detection performance. Our findings suggest that utilizing mid-confidence OOD samples leads to a more compact decision boundary and improved detection robustness.

ACKNOWLEDGMENT

This work was supported by the Hainan Provincial Joint Project of Li'an International Education Innovation Pilot Zone (Grant No.624LALH008), BUPT Kunpeng&Ascend Center of Cultivation, NSFC (Grant No.61601042), and the Super Computing Platform of BUPT.

ETHICS STATEMENT

We are aware of the potential biases in the training data that may lead to the generation of misleading content. Measures have been taken using the method like probability calibration. Overall, our approach is in line with ethical standards, aiming to enhance the reliability and robustness of the model and is reliable for its intended purpose.

REPRODUCIBILITY STATEMENT

We clarify that the sources of all datasets used in the experiments have been provided in Appendix A. Moreover, the theoretical derivations and properties of the diffusion model and beta calibration have been included in the Appendix B and C, which aids in understanding the rationale for employing these methods. Lastly, the model weights and training code will be made publicly available in due course.

REFERENCES

Sima Behpour, Thang Long Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. *Advances in Neural Information Processing Systems*, 36:38206–38230, 2023.

Eugene Berta, Francis Bach, and Michael Jordan. Classifier calibration with roc-regularized isotonic regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 1972–1980. PMLR, 2024.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.02057.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild, 2013. URL https://arxiv.org/abs/1311.3618.

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, 2009. doi: 10.1109/CVPR.2009.5206848.

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

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis, 2022. URL https://arxiv.org/abs/2202.01197.

Xuefeng Du, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models, 2023. URL https://arxiv.org/abs/2309.13415.

Bradley Efron. Tweedie's formula and selection bias. *Publications of the American Statistical Association*, 106(496):1602–1614, 2011.

Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. Diffguard: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models, 2023. URL https://arxiv.org/abs/2308.07687.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018. URL https://arxiv.org/abs/1610.02136.

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

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset, 2018. URL https://arxiv.org/abs/1707.06642.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data, 2020. URL https://arxiv.org/abs/2002.11297.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space, 2021. URL https://arxiv.org/abs/2105.01879.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pp. 623–631. PMLR, 2017.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples, 2018a. URL https://arxiv.org/abs/1711.09325.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018b.

Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis, 2023. URL https://arxiv.org/abs/2211.09117.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method, 2024. URL https://arxiv.org/abs/2312.03701.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yuyuan Liu, Choubo Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation, 2023a. URL https://arxiv.org/abs/2211.14512.

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q. Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting, 2023b. URL https://arxiv.org/abs/2302.10326.

Xingming Long, Jie Zhang, Shiguang Shan, and Xilin Chen. Rethinking the evaluation of out-of-distribution detection: A sorites paradox, 2024. URL https://arxiv.org/abs/2406.09867.

Hossein Mirzaei, Mohammadreza Salehi, Sajjad Shahabi, Efstratios Gavves, Cees GM Snoek, Mohammad Sabokrou, and Mohammad Hossein Rohban. Fake it till you make it: Towards accurate near-distribution novelty detection. *arXiv preprint arXiv:2205.14297*, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *CoRR*, abs/2104.10972, 2021. URL https://arxiv.org/abs/2104.10972.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Sina Sharifi, Taha Entesari, Bardia Safaei, Vishal M Patel, and Mahyar Fazlyab. Gradient-regularized out-of-distribution detection. In *European Conference on Computer Vision*, pp. 459–478. Springer, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022a. URL https://arxiv.org/abs/2010.02502.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:17885–17898, 2022b.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection, 2022. URL https://arxiv.org/abs/2111.09805.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations, 2021. URL https://arxiv.org/abs/2111.12797.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis, 2023. URL https://arxiv.org/abs/2303.02966.

Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. In *European Conference on Computer Vision*, pp. 414–430. Springer, 2024.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching, 2022. URL https://arxiv.org/abs/2203.10807.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 34:29074–29087, 2021.

Jie Wei and Guotai Wang. Fine-grained out-of-distribution detection of medical images using combination of feature uncertainty and mahalanobis distance. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2023. doi: 10.1109/ISBI53787.2023.10230433.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

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.

William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-ood: Deciphering modern out-of-distribution detection algorithms, 2024a. URL `https://arxiv.org/abs/2310.01755`.

Ying Yang, De Cheng, Chaowei Fang, Yubiao Wang, Changzhe Jiao, Lechao Cheng, and Nannan Wang. Diffusion-based layer-wise semantic reconstruction for unsupervised out-of-distribution detection, 2024b. URL `https://arxiv.org/abs/2411.10701`.

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*, volume 39, pp. 13078–13086, 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.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 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, 2018. doi: 10.1109/TPAMI.2017.2723009.

Bad-OOD: Discovering Harmful Synthetic Diffusion Outliers via Confidence Calibration (Appendix)

A Details of Datasets

A.1 Quantitative evaluation of OOD detector performance

We follow the settings of (Tao et al., 2023; Du et al., 2023) to evaluate the performance of OOD detectors, where the classes of OOD test sets do not overlap with ID dataset. For the ImageNet-100 benchmark, we collect subsets from iNaturalist (Horn et al., 2018), SUN (Xiao et al., 2010), Places (Zhou et al., 2018), and Texture (Cimpoi et al., 2013) as large-scale OOD datasets. For the CIFAR-100 (Krizhevsky, 2009) benchmark, SVHN (Netzer et al., 2011), Places (Zhou et al., 2018), LSUN (Yu et al., 2015), iSUN (Xu et al., 2015), and Textures (Cimpoi et al., 2013) are employed as OOD datasets. We provide a brief introduction for each dataset as follows.

ImageNet-100 is created from ImageNet-1k (Deng et al., 2009) train split by randomly sampling 100 classes, and each class contains approximately 1,300 images with a resolution of 256×256.

CIFAR-100 (Krizhevsky, 2009) comprises 100 fine-grained labels, each containing 600 color images with a resolution of 32×32. Among these, 500 images per class are designated for the training set, while 100 images are allocated to the test set. We utilize its training set to assist in training the detector and employ its test set to evaluate the detector.

iNaturalist (Horn et al., 2018) contains images of natural world. It has 13 super-categories and 5,089 sub-categories covering plants, insects, birds, mammals, and so on. We use the subset that contains 110 plant classes which do not overlap with ImageNet-1k.

SUN (Scene UNderstanding) (Xiao et al., 2010) contains 899 categories that cover more than indoor, urban, and natural places with or without human beings appearing in them. We use the subset which contains 50 natural objects not in ImageNet-1k.

Places is a large scene photographs dataset (Zhou et al., 2018), which contains photos that are labeled with scene semantic categories from three macro-classes: Indoor, Nature, and Urban. The subset we use contains 50 categories that are not present in ImageNet-1k.

Textures (Describable Textures) (Cimpoi et al., 2013) contains images of textures and abstracted patterns. As no categories overlap with ImageNet-1k, we use the entire dataset as in (Huang & Li, 2021).

SVHN (Street View House Numbers) (Netzer et al., 2011) includes 10 classes and each represents a digit from 0 to 9. We use the subset which contains 10,000 samples across all classes.

LSUN (Large-scale Scene UNderstanding) (Yu et al., 2015) contains 10 scene categories, such as dining room, bedroom, chicken, outdoor church, and so on. We use the subset which contains 10,000 samples not in CIFAR-100.

iSUN (Xu et al., 2015) contains eye tracking ground-truth images derived from the SUN (Xiao et al., 2010) dataset, and we use the entire set of 8926 samples.

A.2 Additional datasets used

In addition to evaluating OOD detection performance, we also introduce the following datasets at the confidence calibration stage and the generalization assessment stage, respectively. Using the ImageNet-OOD (Yang et al., 2024a) dataset, we calibrate our OOD scores with respect to semantic shifts, thereby balancing the model's sensitivity to both semantic and covariate shifts. The detailed calibration procedure is described in Appendix C.2. Additionally, we employ the SynIS (Long et al., 2024) dataset to assess the ability of various OOD scores to capture both types of distributional shifts.

ImageNet-OOD (Yang et al., 2024a) is a dataset designed to evaluate the anomaly degree of OOD detection methods purely on semantic shifts, comprising 31,807 images from 637 categories selected directly from ImageNet-21K (Ridnik et al., 2021). It provides a standard test of model performance under semantic shifts.

SynIS (Long et al., 2024) consists of 64 subsets, each containing 5,000 images, covering 8 levels of semantic shift and 8 levels of covariate shift, simulating distribution changes by altering semantic content and visual features. This dataset provides diverse distribution shift scenarios for the complement of IS-OOD benchmark, enabling the evaluation of model generalizability under semantic and covariate variations.

B PRELIMINARIES

Diffusion models (Ho et al., 2020; Song et al., 2022a; Song et al.) are generative frameworks that learn data distributions through a gradual denoising process, which enable controlled generation by incorporating external guidance (e.g., class labels, text prompts) during training and sampling.

Given a data sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and a condition $\mathbf{y}$, the forward process gradually corrupts $\mathbf{x}_0$ by adding Gaussian noise over T timesteps. The noised sample $\mathbf{x}_t$ at timestep t is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{9}$$

where $\beta_t \in (0, 1)$ controls the noise schedule. The marginal distribution at timestep t can be expressed in closed form:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{10}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

The reverse process learns to iteratively denoise $\mathbf{x}_t$ while conditioning on $\mathbf{y}$, parameterized by a neural network ϵ_θ:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, \mathbf{y}), \Sigma_t \mathbf{I}\right), \tag{11}$$

where the mean μ_θ is derived from:

$$\mu_\theta(\mathbf{x}_t, t, \mathbf{y}) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t, \mathbf{y})\right). \tag{12}$$

Lastly the sampling procedure with time dependent constants σ_t follows:

$$\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t, \mathbf{y}) + \sigma_t \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}). \tag{13}$$

Logistic calibration. A logistic regression classifier is said to be *well-calibrated* if its predicted scores align with the empirical class distribution of the data (Kull et al., 2017). Formally, for a binary classifier predicts scores $s = f(\mathbf{x}) \in [0, 1]$, it's required that the proportion of positive instances among those assigned score s equals s. Mathematically, this is expressed as:

$$s = \mathbb{E}[Y \mid f(X) = s], \tag{14}$$

where X and Y denote the input sample and binary label of a randomly drawn instance from the dataset. The expectation in equation 14 can be computed empirically as:

$$\mathbb{E}[Y \mid f(X) = s] = \frac{\sum_{j=1}^{n} y_j \cdot \mathbb{I}[f(\mathbf{x}_j) = s]}{\sum_{j=1}^{n} \mathbb{I}[f(\mathbf{x}_j) = s]}, \tag{15}$$

where $\mathbb{I}[\cdot]$ is the indicator function, and n is the total number of instances.

For any fixed classifier f, there exists a unique calibration map $\mu(s) = \mathbb{E}[Y \mid f(X) = s]$ that achieves perfect calibration on the training data. However, such a map risks severe overfitting, particularly when f assigns unique scores to individual instances. In this case, the calibration map collapses to $\mu(s_i) = y_i$, resulting in overconfident 0/1 predictions that generalize poorly to unseen data. Thus, the goal of calibration methods is to learn a *generalizable* map $\mu(s)$ that preserves ordinality while aligning predictions with empirical frequencies on held-out data.

C BETA CALIBRATION

C.1 DERIVATION OF BETA CALIBRATION.

We provide a core derivation of the beta calibration method (Kull et al., 2017). The key assumption is that the classifier's output scores for each class follow the beta distributions. Therefore, this distribution function can be used to derive the confidence calibration procedure for sigmoid-based binary classification.

Class-conditional score distributions. Assume the scores s for positive ($Y = 1$) and negative ($Y = 0$) instances follow beta distributions with parameters (α_1,β_1) and (α_0,β_0), respectively:

$$p(s \mid Y = 1) = \frac{s^{\alpha_1 - 1}(1 - s)^{\beta_1 - 1}}{B(\alpha_1, \beta_1)}, \tag{16}$$

$$p(s \mid Y = 0) = \frac{s^{\alpha_0 - 1}(1 - s)^{\beta_0 - 1}}{B(\alpha_0, \beta_0)}, \tag{17}$$

where $B(\cdot, \cdot)$ is the beta function.

Likelihood ratio derivation. The likelihood ratio (LR) between the two classes is:

$$LR(s) = \frac{p(s \mid Y = 1)}{p(s \mid Y = 0)} = \frac{s^{\alpha_1 - \alpha_0}(1 - s)^{\beta_0 - \beta_1}}{B(\alpha_1, \beta_1)/B(\alpha_0, \beta_0)}. \tag{18}$$

Let $a = \alpha_1 - \alpha_0$, $b = \beta_0 - \beta_1$, and $K = B(\alpha_1, \beta_1)/B(\alpha_0, \beta_0)$. Reparameterizing $K = e^{-c}$, the LR becomes:

$$LR(s; a, b, c) = \frac{s^a}{(1 - s)^b} \cdot e^c. \tag{19}$$

Calibrated probability mapping. The calibrated probability $\mu_{beta}(s)$ is derived from the posterior odds under a uniform class prior:

$$\mu_{\text{beta}}(s; a, b, c) = \frac{1}{1 + LR(s; a, b, c)^{-1}} = \frac{1}{1 + \frac{(1-s)^b}{s^a e^c}}. \tag{20}$$

Simplifying yields the beta calibration family:

$$\mu_{\text{beta}}(s; a, b, c) = \frac{1}{1 + \frac{1}{e^c \cdot \frac{s^a}{(1-s)^b}}}. \tag{21}$$

Monotonicity constraints. To ensure the calibration map is non-decreasing, parameters a and b must satisfy $a, b \geq 0$. This preserves the ordinal relationship between classifier scores and calibrated probabilities.

C.2 DETAILED IMPLEMENTATION OF BETA CALIBRATION.

We provide a detailed description of the beta calibration method for OOD detectors in Algorithm 1. Specifically, the algorithm takes as input an uncalibrated classifier and calibration datasets, where the ID validation set is drawn from 10% of the held-out data in ImageNet-100, and the OOD validation set is constructed from the ImageNet-OOD (Yang et al., 2024a) dataset. The calibration parameters "abc" are initialized at the beginning of the algorithm. By iterating over both the ID and OOD validation sets, the algorithm computes the pre-calibration probabilities for each sample and fits the calibration parameters according to the sample labels. The final output is a classifier calibrated with beta parameters.

Algorithm 1 Beta Calibration of OOD Detector

Input: Uncalibrated classifier $\mathcal{F}_\theta$, validation ID dataset $\mathcal{D}_{in}$, validation OOD dataset $\mathcal{D}_{out}$
Output: Calibrated classifier $\mathcal{F}'_\theta$
Initialize Beta calibration parameters a_0, b_0, c_0, empty probability list $P = []$ and label list $Y = []$
foreach *mini-batch* (x_{id}, y_{id}), (x_{ood}, y_{ood}) *in* $\mathcal{D}_{in}$, $\mathcal{D}_{out}$ **do**
  Compute p=$\mathcal{F}_\theta(\text{concat}(x_{id}, x_{ood}))$
  $P \leftarrow p, \quad Y \leftarrow \text{concat}(y_{id}, y_{ood})$
end
Fit beta calibration: $a, b, c \leftarrow$ fit(P, Y)
Predict beta calibration: $\hat{P} \leftarrow$predict(P)
return calibrated model $\mathcal{F}'_\theta$ with beta calibration applied

D PARAMETER SENSITIVITY ANALYSIS.

D.1 PERTURBATION STEPS t.

The parameter analysis on perturbation steps t during OOD synthesis reveals critical insights into the trade-offs between detection performance and model robustness. Notably, $t = 300$ achieves the optimal balance, while further increasing t to 400 introduces a marginal degradation in FPR95 and ID ACC, despite a slight AUPR improvement.

Table 4: Analysis on the perturbation steps during OOD synthesis.

t	FPR95↓	AUROC↑	AUPR↑	ID ACC↑
200	45.78	90.89	86.87	87.74
300	**43.20**	**91.28**	86.87	**88.22**
400	43.52	90.84	**87.06**	87.44

D.2 REGULARIZATION WEIGHT β.

The parameter analysis on regularization weight β during OOD detector training shows critical trade-offs between OOD detection robustness and ID classification accuracy. We ablate different choices, as shown in the table below, a moderate value $\beta = 0.2$ achieves the optimal balance.

Table 5: Analysis on the regularization weight during OOD detector training.

β	FPR95↓	AUROC↑	AUPR↑	ID ACC↑
0.1	47.42	89.59	97.50	78.76
0.2	**36.51**	**92.20**	**98.08**	**87.44**
0.5	50.21	89.47	97.58	79.68
1.0	69.24	81.12	95.43	67.62

E COMPARISON WITH OTHER CALIBRATION STRATEGIES

We conducted a comparison with temperature scaling (Liang et al., 2017) and isotonic regression (Berta et al., 2024), which are two widely-used calibration techniques. As shown in Table 6, beta calibration consistently yields better AUROC and FPR@95% results across multiple benchmarks, demonstrating its superiority in aligning anomaly scores with the underlying uncertainty of generated OOD data.

Table 6: Comparisons of different calibration strategies with ImageNet-100 as in-distribution data.

Methods	iNaturalist		Places		Sun		Textures		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Beta calibration	**26.25**	**95.53**	**32.29**	**93.76**	**34.17**	**93.48**	**53.33**	86.01	**36.51**	**92.20**
Temperature scaling	38.85	92.50	51.04	88.34	55.73	85.56	53.65	89.10	49.82	90.38
Isotonic Regression	41.77	90.26	53.44	86.86	55.94	83.37	54.38	**89.31**	51.38	88.95

F Comparison of Outlier Synthesis Methods and Model Architectures

In this section, we separately evaluate the impact of outlier embedding generation methods and model architectures. It is important to note that all experiments in this part are conducted under calibrated settings. We first introduce Gaussian noise at each denoising step of RDM (Li et al., 2024) for outlier synthesis and observe that its performance is significantly inferior to that of the minority guidance approach, indicating that the low-likelihood target provided by minority guidance is more explicit. Furthermore, we assess our method on both ResNet-101 (He et al., 2016) and T2T-ViT (Yuan et al., 2021) architectures. Compared to RankFeat (Song et al., 2022b), our results show that the samples selected after calibration contain more effective boundary information than those based on singular value.

Table 7: Comparison of outlier synthesis methods and model architectures.

Outlier Synthesis	Architecture	FPR95↓	AUROC↑	AUPR↑
Add Gaussion Noise	ResNet-34	57.14	87.24	83.89
Minority Guidance	ResNet-34	36.51	92.20	87.32
RankFeat (Song et al., 2022b)	ResNet-101	36.80	92.15	-
Minority Guidance	ResNet-101	35.68	92.28	87.37
RankFeat (Song et al., 2022b)	T2T-ViT	51.58	85.60	-
Minority Guidance	T2T-ViT	50.21	89.47	84.58

G Generalization and Robustness Analysis

G.1 Generalizability on semantic and covariate shift.

In this section, we provide more evaluation results of the generalizability of different anomaly measurement approaches and the impact of using different OOD synthesizers. First, we comprehensively present the score performance of more softmax-based methods, distance-based methods, and gradient-based methods on the SynIS (Long et al., 2024) dataset (see Figure 6). In addition, we also assess the previously proposed NPOS (Tao et al., 2023) and Dream-OOD (Du et al., 2023) methods (see Figure 7 & 8).

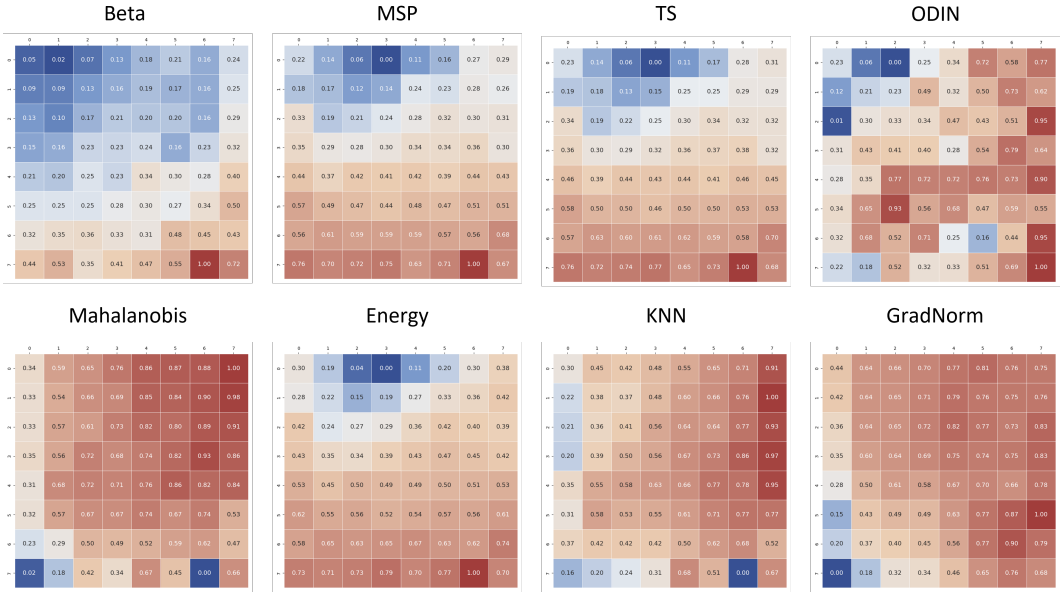


Figure 6: **Full results of beta score and other metrics.** In each graph, the semantic shift increases from top to bottom, and the covariate shift increases from left to right. All the scores are rescaled to range between 0 and 1, and the color red indicates a higher anomaly degree.

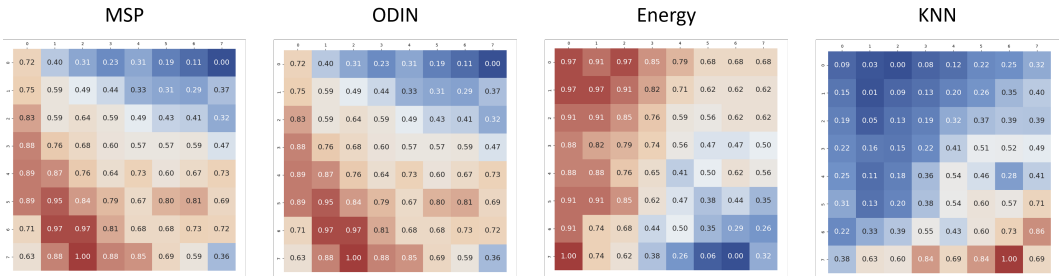


Figure 7: **Full results of SynIS using NPOS (Tao et al., 2023).** In each graph, the semantic shift increases from top to bottom, and the covariate shift increases from left to right. All the scores are rescaled to range between 0 and 1, and the color red indicates a higher anomaly degree.

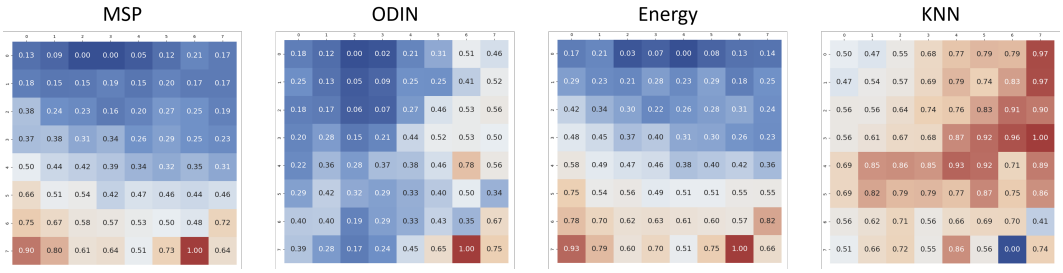


Figure 8: **Full results of SynIS using Dream-OOD (Du et al., 2023).** In each graph, the semantic shift increases from top to bottom, and the covariate shift increases from left to right. All the scores are rescaled to range between 0 and 1, and the color red indicates a higher anomaly degree.

Specifically, given an input image, the classifier generates a prediction, after which different methods compute their respective scores. We provide a detailed introduction to each method as follows.

- Maximum Softmax Probability (Hendrycks & Gimpel, 2018) method uses the largest softmax output of the classifier as the score.
- Temperature Scaling (Liang et al., 2017) method outputs the maximum softmax probability after temperature calibration.
- ODIN introduces (Liang et al., 2017) input preprocessing to enhance the separation between ID and OOD samples, and combines this with temperature scaling to produce the highest softmax score.
- Energy-based (Liu et al., 2020) method replaces the softmax probability with an energy score, addressing the issue of overconfident softmax predictions.
- Mahalanobis Distance (Lee et al., 2018b) method first learns a class-conditional Gaussian distribution on the ID data, and then uses the Mahalanobis distance between the OOD input and the nearest class Gaussian as the anomaly score.
- KNN-OOD (Sun et al., 2022) method computes the K-nearest neighbor distance between the OOD input and ID samples as the score.
- GradNorm (Huang et al., 2021) optimizes the KL divergence between the softmax output and a uniform distribution, and uses the norm of the backpropagated gradient vector as the anomaly score.

G.2 Generalizability and Necessity of Calibration Strategies.

Generating overly noisy near-OOD samples or semantically irrelevant far-OOD samples is the key challenge in synthetic OOD generation. Thus, we compare the proportion of extreme OOD samples synthesized by different methods before and after calibration. The results in Table 8 demonstrate that these undesirable outlier types are prevalent across methods, further motivating the need for a principled filtering step such as ours.

Table 8: Comparison of sample distributions before and after calibration across different OOD synthesis methods.

Method	Ours			Dream-OOD (Du et al., 2023)			SONA (Yoon et al., 2025)		
	OOD-I	OOD-II	OOD-III	OOD-I	OOD-II	OOD-III	OOD-I	OOD-II	OOD-III
Before	288 (57.6%)	45 (9.0%)	167 (33.4%)	156 (31.2%)	82 (16.4%)	262 (52.4%)	263 (52.6%)	53 (10.6%)	184 (36.8%)
After	233 (58.4%)	38 (9.5%)	128 (32.1%)	99 (29.2%)	60 (17.7%)	180 (53.1%)	193 (52.0%)	46 (12.4%)	132 (35.6%)
Reduce	**19.1%**	**15.6%**	**23.4%**	**36.5%**	**26.8%**	**31.2%**	**26.6%**	**13.2%**	**28.3%**

G.3 Robustness under intra-class and inter-class variability.

While ImageNet-100 and CIFAR-100 already exhibit a reasonable degree of intra-class variability and inter-class semantic proximity (e.g., fine-grained animal species or visually similar objects), we believe that evaluating under more challenging settings would further strengthen the validation of our framework. To this end, we have conducted additional experiments using in-distribution datasets with increased semantic ambiguity—specifically, 10 manually selected ImageNet-100 classes with high inter-class similarity, and 10 randomly selected ImageNet-R classes with higher intra-class variability. As shown in Table 9, our proposed model consistently outperforms baseline methods, indicating its robustness in these ambiguous scenarios. We believe this robustness stems from RDM's operation in the feature space of ID data, where the diversity-aware generation strategy naturally encourages the synthesis of boundary-sensitive samples, regardless of the inherent class separability.

Table 9: Generalization capacity of our RDM for outlier synthesis beyond standard benchmarks.

Settings (In-distribution Data)	**Ours**		**Dream-OOD (Du et al., 2023)**	
	FPR95 (Avg.)	AUROC (Avg.)	FPR95 (Avg.)	AUROC (Avg.)
10 Random classes from ImageNet-100 (stingray, jellyfish, Chihuahua, tiger, grasshopper, zebra, accordion, basketball, castle, lipstick)	66.85	77.41	68.59	76.93
10 Semantic overlap classes from ImageNet-100 (coyote, tabby, leopard, lion, tiger, zebra, hog, ox, impala, mink)	68.62	79.55	78.91	69.40
10 Intra-class variability classes from ImageNet-R (stingray, jellyfish, Chihuahua, tiger, grasshopper, zebra, accordion, basketball, castle, lipstick)	69.64	75.19	79.66	70.34

G.4 Robustness under class-imbalanced setting.

While our main experiments were conducted on balanced datasets, we further examined the effectiveness of the proposed strategy under class-imbalanced setting by randomly removing a varying number of samples (ranging from 0 to 1200) from each class in ImageNet-100, resulting in a new imbalanced version of the dataset. As shown in Table 10, our minority guidance strategy still yields consistent and noticeable improvements over Dream-OOD under these imbalanced conditions. This empirically suggests that our approach is inherently robust to class imbalance.

Table 10: OOD detection results using the imbalanced version of ImageNet-100 as in-distribution.

Methods	iNaturalist		Places		Sun		Textures		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Ours	**34.27**	**87.10**	**42.15**	**84.31**	**48.65**	**89.03**	49.85	**85.60**	**43.88**	**85.03**
Dream-OOD (Du et al., 2023)	50.62	82.24	50.31	80.41	59.48	76.37	**46.77**	78.04	54.30	79.26

H Mechanistic Explanation for Mid-confidence Samples

H.1 Quantitative analysis of feature-space behavior.

To more rigorously characterize the behavior of synthesized OOD samples in the MoCo feature space, we perform a distance-based analysis across three confidence groups (high/mid/low). Concretely, we: (i) randomly sample 100 OOD samples from each confidence group; (ii) using their MoCo features along with the corresponding ID features to compute the L2 distance to the paired ID sample and report the group-wise averages.

As shown in Table 11, low-confidence samples exhibit small and tightly concentrated distances, indicating that they lie close to the ID manifold, while high-confidence samples show substantially larger distances and typically higher dispersion, consistent with unrealistic or semantically collapsed generations. In contrast, mid-confidence samples yield moderate and stable distances, suggesting that they populate a boundary-adjacent transition region in feature space—deviating from ID in an informative yet semantically meaningful manner. These results support our claim that selecting mid-confidence synthesized samples provides a principled way to obtain useful near-boundary OOD data, motivating our filtering strategy.

Table 11: Feature-space distances of the three confidence groups.

	High confidence	Mid confidence	Low confidence
Average feature distance	0.334	0.138	0.104

H.2 Causal explanation of why mid-confidence samples matter.

We further probe the mechanism underlying our selection strategy by conducting a controlled intervention in feature space. For each class, we compute the first principal direction of the MoCo representation using all 1,300 ID samples, and then perturb each mid-confidence OOD feature by translating it *toward* or *away* from the corresponding class-specific eigen-direction. The perturbed features are decoded back to the pixel space using MAGE, with step sizes $\epsilon \in \{0.01, 0.1\}$.

The results in the table below show a clear monotonic relationship between the perturbed samples' proximity to the decision boundary and the resulting OOD performance. This controlled manipulation demonstrates, causally, that it is indeed the feature-space position, rather than incidental correlations, that makes mid-confidence samples particularly effective for training the detector.

Table 12: Effect of feature-space perturbations of mid-confidence samples on OOD detection performance.

Perturbation direction	FPR95↓	AUROC↑	AUPR↑
Toward ID (+0.1)	53.57	88.32	85.57
Toward ID (+0.01)	41.23	91.26	86.84
Mid confidence (0)	**36.51**	**92.20**	**87.32**
Toward OOD (0.01)	40.73	91.30	86.96
Toward OOD (0.1)	44.99	90.44	86.39

I Comparison of Pixel Space and Latent Space Training for OOD Detection

To assess whether pixel-space reconstruction is essential, we conduct an ablation in which the OOD detector is trained directly in the MoCo latent space using ID features and synthesized OOD features, completely bypassing the MAGE decoding step. For a fair comparison, we disable confidence calibration so that the evaluation reflects the discriminative capacity of the latent representations themselves.

The results in the table below show that training in pixel space consistently outperforms training purely in the latent space. This indicates that MAGE decoding enhances the semantic fidelity and visual separability of the synthesized OOD samples, ultimately leading to stronger OOD detection performance. Moreover, consistent with the intuition in Dream-OOD, decoding OOD samples into pixel space provides substantially better interpretability than operating purely in the latent space.

Table 13: Comparison of OOD detection performance in pixel and latent space.

Detector type	iNaturalist		Places		Sun		Textures		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
pixel	**35.49**	**93.34**	**39.09**	**91.95**	**47.17**	**91.70**	**59.39**	**84.49**	**54.06**	**88.75**
latent	36.84	91.96	43.50	89.42	50.25	88.78	64.85	82.72	59.10	86.28

J COMPARISON WITH SIMILAR DIFFUSION-BASED METHODS

We conducted additional experiments comparing our method against SONA (Yoon et al., 2025) and Fake it till you make it (Mirzaei et al., 2022). The results are summarized in Table 14. We can see that our method achieves better performance in terms of both average FPR95 (37.10) and AUROC (92.97), compared to SONA (FPR95: 37.46 and AUROC: 91.06) and Fake it till you make it (FPR95 at 71.38 and AUROC at 79.29).

Notably, although SONA also utilizes semantic/nuisance disentanglement, our approach differs in two key aspects: (i) Instead of relying on architectural constraints for disentanglement, we employ a classifier-guided calibration that allows flexible and fine-grained control over semantic shift and harmful perturbations; (ii) Our approach is also computationally more efficient. We conducted a dedicated comparison of generation and inference times, showing that our method is 3x faster compared to SONA. This improvement is largely attributed to our use of the lightweight RDM model, which does not require additional conditioning inputs, in contrast to SONA's reliance on SD2 (Rombach et al., 2022), whose joint image-text reasoning introduces higher computational overhead.

Table 14: OOD detection results of diffusion-based methods using CIFAR-100 as in-distribution data.

Methods	SVHN		Places		LSUN		Textures		iSUN		Average		Inference Time
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	Per 100 Samples (s)
Ours	54.90	89.16	68.75	86.04	24.54	94.22	**40.21**	**91.56**	**1.10**	**99.72**	**37.10**	**92.97**	50
SONA (Yoon et al., 2025)	**3.10**	**99.39**	**44.00**	**88.35**	**18.20**	**96.19**	58.90	85.20	63.10	86.17	37.46	91.06	160
Fake it (Mirzaei et al., 2022)	85.15	77.53	77.05	76.91	56.40	76.31	70.55	80.77	67.75	84.95	71.38	79.29	-

K FINE GRANULARITY EVALUATION ON OPENOOD V1.5

OpenOOD v1.5 (Zhang et al., 2023), as a standardized benchmark in the field of OOD detection, encompasses a variety of settings with multiple ID and OOD datasets. Among them, we have selected ImageNet-200 as the ID dataset, together with near-OOD datasets (SSB-hard, NINCO), far-OOD datasets (iNaturalist, Textures, OpenImage-O), and non-semantic OOD variants (ImageNet-V2, ImageNet-C, ImageNet-R).

K.1 IMAGENET-200 BENCHMARK.

We applied both our model and SONA on ImageNet-100 to generate outliers, and then evaluated the resulting classifiers on ImageNet-200 (used as the ID dataset) together with near-OOD and far-OOD datasets. As shown in Table 15, our method consistently outperforms SONA across all categories, with more pronounced gains on far-OOD and covariate-shifted ID cases, demonstrating superior robustness and generalization.

We attribute this improvement to a fundamental difference in design principle: whereas SONA primarily enhances decision boundaries by training on semantically similar hard negatives, our approach focuses on generating diverse and realistic OOD samples in pixel space, thereby improving classifier resilience to a wider range of distribution shifts.

Table 15: Evaluation of diffusion-based methods on ImageNet-200 benchmark.

Methods	Near-OOD AUROC			Far-OOD AUROC				ID Acc
	SSB-hard	NINCO	Avg	iNaturalist	Textures	OpenImage-O	Avg	
Ours	**61.41**	**68.80**	**65.11**	**83.12**	**64.85**	**72.11**	**73.36**	**23.92**
SONA (Yoon et al., 2025)	59.31	66.77	63.04	79.61	64.09	68.44	70.71	22.28

K.2 IMAGENET-200 FULL-SPECTRUM BENCHMARK.

Additionally, we conducted the evaluation under the full-spectrum setting, which included non-semantic OOD datasets. As shown in Table 16, our method exhibits more pronounced gains over Dream-OOD, particularly on far-OOD and non-semantic OOD benchmarks in OpenOOD v1.5, underscoring our superior robustness under diverse types of distribution shifts. We attribute this improvement to our method's ability to generate OOD samples that more faithfully reflect real-world distributional variations, thereby providing more effective supervision for classifier training.

Table 16: Evaluation of diffusion-based methods on ImageNet-200 full-spectrum benchmark.

Methods	Near-OOD AUROC			Far-OOD AUROC				ID Acc (Covariate-shifted)
	SSB-hard	NINCO	Avg	iNaturalist	Textures	OpenImage-O	Avg	
Ours	**48.50**	**56.45**	**52.53**	**74.36**	**52.62**	**60.70**	**62.30**	**55.12**
Dream-OOD (Du et al., 2023)	48.12	56.01	52.06	73.65	50.88	56.80	60.44	51.63

L Additional Visual Results

L.1 Visualization of synthetic OOD.

Figure 9: **ID samples and synthetic OOD of class jeep for detector training.** The severity score is provided in the upper right corner of each figure.

L.2 Visualization of OOD samples as three types.

In addition to the three categories presented in the main text—broccoli, starfish, and strawberry, we also supplement the OOD synthesis results for the candle and jeep categories, involving three types of OOD (semantic shift, covariate shift, and the ones with severe semantic and appearance changes). It should be further noted that for OOD-II and OOD-III, our zero-shot classification results are obtained based on CLIP. Specifically, we utilize the image encoder and text encoder of CLIP (Radford et al., 2021) to compute the cosine similarity between a given image and the textual descriptions of the 1,000 categories in ImageNet-1k, then assigning the category with the highest similarity as the predicted class.

Figure 10: **OOD samples of class candle.** For OOD-II and OOD-III, the class predicted by CLIP (Radford et al., 2021) is demonstrated under each image.

Figure 11: **OOD samples of class jeep.** For OOD-II and OOD-III, the class predicted by CLIP (Radford et al., 2021) is demonstrated under each image.

M Devices and Computational Cost

We run all experiments with Python 3.8.5 and PyTorch 1.13.1, using six NVIDIA GeForce RTX 3090 GPUs.

To compare the computational performance of different methods, we conduct a comparison between Bad-OOD and other baselines, calculating the inference time of synthesizing 100 OOD samples (all experiments run on a single NVIDIA 3090 GPU). As shown in Table 17, our method is more efficient in generating OOD samples than all baselines.

Table 17: Comparison of computational cost in OOD synthesis.

Method	Ours	VOS	NPOS	Dream-OOD	GAN	SONA
Inference Time Per 100 Samples (s)	**50**	89	92	110	146	160

N The Use of Large Language Models (LLMs)

In the process of writing this paper, we only employed LLMs as tools for polishing writing and retrieving relevant knowledge (e.g., finding related work).