

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 BAD-OOD: DISCOVERING HARMFUL SYNTHETIC DIF- FUSION OUTLIERS VIA CONFIDENCE CALIBRATION

Anonymous authors

Paper under double-blind review

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 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.

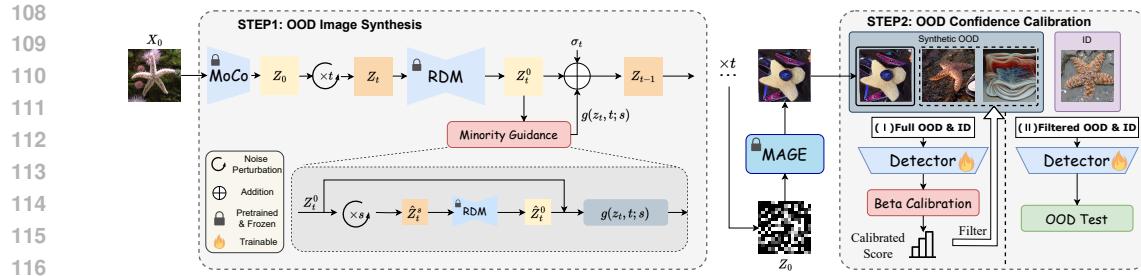


Figure 1: **Overview of our proposed method Bad-OOD.** Bad-OOD first generates auxiliary OOD samples by integrating minority guidance into the denoising process. In the second stage, the detector is initially trained using the entirety of the synthetic OOD samples in conjunction with the ID samples. 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 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 z_0 intuitively serves as an anchor point for producing its anomaly variants.

162 3.2 LATENT MINORITY SYNTHESIS USING RDM
163

164 Given the conditional representation of ID data z_0 , we employ representation diffusion model (RDM)
165 (Li et al., 2024) to produce outlier embeddings. We choose RDM since it allows us to sample
166 from a representation distribution without providing textual conditions, such as class labels or text
167 descriptions, which might be unavailable or difficult to obtain when synthesizing anomalies. Firstly,
168 we add noise to the conditional representation z_0 of each ID image over t time steps to obtain a noisy
169 version z_t . Inspired by Um & Ye (2024), we extend the minority metric into latent space based on the
170 denoising timestep t and the posterior mean \hat{z}_0 , which is obtained via the Tweedie's formula (Efron,
171 2011) and the pretrained diffusion model. Formally, it is defined as:
172

$$\mathcal{L}(z_0; t) := \mathbb{E}_{q_{\alpha_t}(z_t|z_0)}[d(z_0, \hat{z}_0(z_t))], \quad (1)$$

173 where $\hat{z}_0(z_t)$ denotes the model's estimate of the clean latent z_0 given the noisy latent z_t at timestep
174 t , since the true z_0 is only obtained at the end of the reverse process, this intermediate prediction is
175 necessarily used for guiding OOD synthesis. $d(\cdot, \cdot)$ is a distance metric, e.g., L1 or L2, to measure
176 the discrepancy between two features. Intuitively, a higher minority score indicates sample in a
177 low-likelihood region farther from the anchor z_0 , thus being more abnormal. Therefore, we can
178 derive the sampling towards low-likelihood areas by leveraging the gradient of the minority score.
179 Specifically, for a noisy latent sample z_t at timestep t , we first exploit the pretrained RDM $\epsilon_\theta(z_t, t, c)$
180 to denoise z_t into z_t^0 , i.e., the predicted clean latent at timestep t . We then reintroduce s additional
181 noise steps to obtain \hat{z}_t^s , which is denoise again into \hat{z}_t^0 . According to the minority score mentioned
182 above, our guided gradient \mathbf{g} can be obtained by:
183

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

184 Lastly, as shown in Figure 1, we follow the standard classifier-guided formulation in ADM (Dhariwal
185 & Nichol, 2021) and use the negative gradient direction indicated by minority guidance to push
186 samples away from dense ID regions, thereby resulting in the outlier representations that naturally
187 occupy the periphery of the ID distribution while retaining coherent structures or semantic cues.
188

189 3.3 OUTLIER IMAGE GENERATOR

190 Simialr to Dream-OOD, we explicitly generate outliers in pixel space, yielding visually interpretable
191 samples, preserving fine-grained details such as texture and color, and aligns more naturally with
192 downstream tasks that operate directly on images. Specifically, MAGE (Li et al., 2023) is leveraged as
193 the image generator to convert the outlier embeddings in latent space into images. During inference,
194 a fully masked image serves as the input of generator, and MAGE reconstructs an output guided by
195 the outlier embeddings.
196

197 3.4 ENERGY-BASED OOD REGULARIZER
198

199 In order to utilize synthetic OOD samples for the separation of ID and OOD distributions, we apply
200 the energy-based uncertainty regularization loss following Du et al. (2022). Specifically, assume
201 the OOD detector's output logit of the n-th ID class is $f_n(\mathbf{x}; \theta)$, then the energy function can be
202 formulated as the LogSumExp operation among all these N logits:
203

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

205 Then the uncertainty loss can be defined using a binary sigmoid loss based on the energy function,
206 and the threshold for distinguishing ID and OOD should be set to zero:
207

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

208 where ϕ denotes a nonlinear MLP. The overall training objective is to optimize the cross entropy for
209 multi-class classification within ID data coupled with the uncertainty regularizer:
210

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

212 where β is to balance the effect of the regularization term. Once obtained the classifier, we can measure
213 the anomaly degree of OOD samples by the sigmoid of the energy function $E(\mathbf{x}; \theta)$. Formally, we
214 obtain:
215

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

216 3.5 CONFIDENCE CALIBRATION
217

218 Since the existence of bad samples within the synthetic OOD, it is vital to establish an effective
219 mechanism to quantify the degree of anomaly in OOD samples. Therefore, we propose a calibration
220 refined strategy, which leverages a calibrated detector to score samples from an auxiliary dataset
221 and systematically selects high-quality data. Essentially, calibration aims to align the confidence
222 output by the classifier with the actual probability, and the confidence output by a perfectly calibrated
223 detector can reflect the degree of anomaly. Therefore, we introduce beta calibration (Kull et al., 2017),
224 which is not only applicable to sigmoid-based output score but also capable of fitting a wide range of
225 probability distributions. Specifically, given an uncalibrated classifier \mathcal{C} and an auxiliary calibration
226 dataset \mathcal{D} , beta calibration learns a map function μ_{beta} that transforms the raw scores S of samples
227 from \mathcal{D} into new scores S_c , ensuring the new scores S_c closely approximate the actual probabilities
228 within \mathcal{D} . The formula for the mapping function follows:

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

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

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

241 4 EXPERIMENTS
242

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

246 4.1 EXPERIMENTAL SETUP

247 **Dataset.** To evaluate OOD detection performance, we follow the experimental settings of Tao et al.
248 (2023); Du et al. (2023) and use ImageNet-100 as the ID dataset, which is carefully selected from the
249 original ImageNet-1k dataset (Deng et al., 2009). Additionally, we adopt the same OOD test datasets
250 as in Huang & Li (2021); Tao et al. (2023); Du et al. (2023), including subsets of iNaturalist (Horn
251 et al., 2018), SUN (Xiao et al., 2010), Places (Zhou et al., 2018), and Textures (Cimpoi et al., 2013).
252 We also follow the common practice of adopting CIFAR-100 as the ID data for training as well. More
253 details are provided in the Appendix A.1.

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

267 **Competitors.** We compare several strong OOD detection baseline methods, categorized into non-
268 synthesis-based and synthesis-based methods. The majority of existing methods fall into the non-
269 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
<i>ResNetv2-101</i>										
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	
	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
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

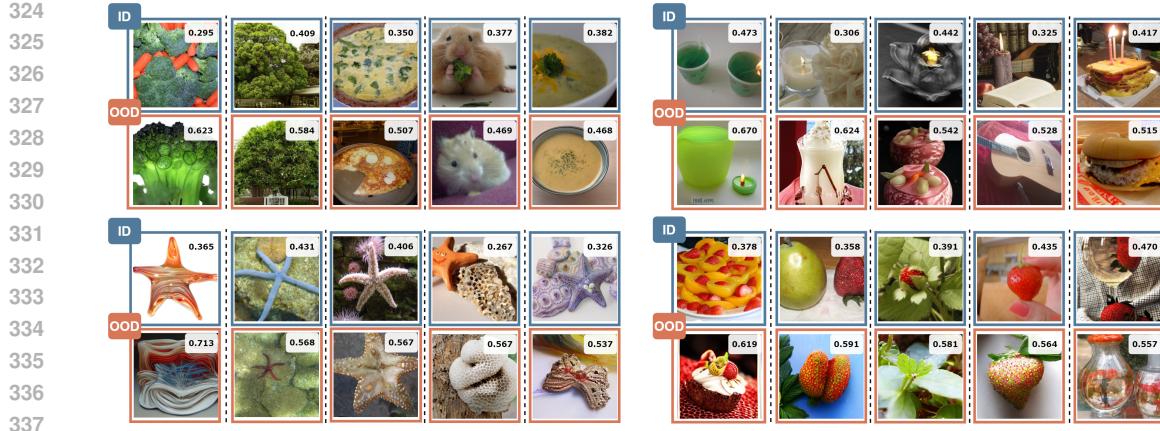


Figure 2: ID reference samples and synthetic OOD of class broccoli, candle, starfish and strawberry. The anomaly score is provided in the upper right of each sample.



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.

across 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 H), and fine granularity OpenOOD v1.5 (Zhang et al., 2023) evaluation (Appendix I).

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. We can figure 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 alongside their corresponding ID samples are provided in the Appendix J.

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 (~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 \downarrow	AUROC \uparrow	AUPR \uparrow
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
0.025	\checkmark	41.46	91.25	86.81
0.05	Random	43.20	91.28	86.89
0.05	\checkmark	36.51	92.20	87.32
0.075	Random	44.09	91.05	86.64
0.075	\checkmark	38.96	92.04	87.33

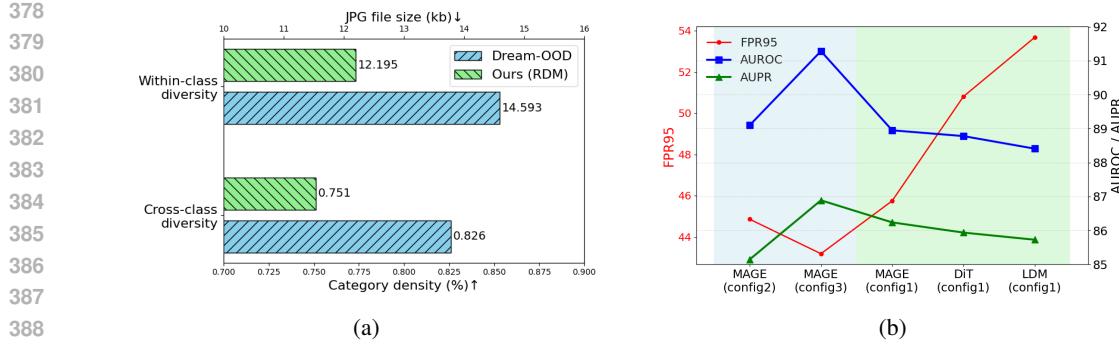


Figure 4: (a) The diversity of OOD samples generated by our RDM-based method and Dream-OOD. (b) Performance using different image generators.

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 J.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.

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—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.

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%)

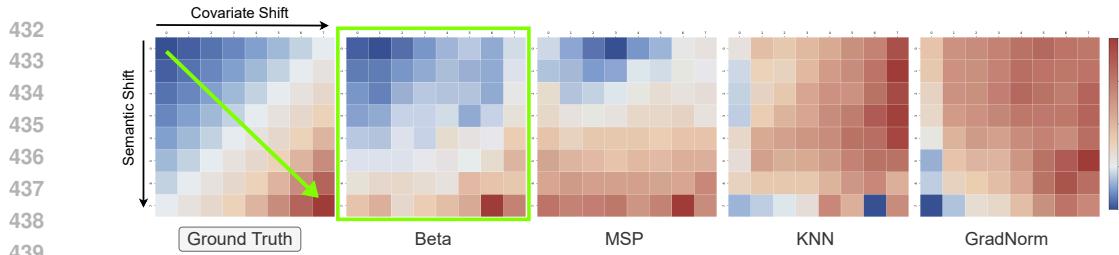


Figure 5: **Comparison of our approach and other competitors.** In each graph, the semantic shift increases from top to bottom, and the covariate shift increases from left to right. All the scores are rescaled to range between 0 and 1, and the color red indicates a higher anomaly degree.

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.

4.3.1 GENERALIZABILITY ON SEMANTIC AND COVARIATE SHIFT

In this section, we demonstrate that our confidence-calibrated OOD detector effectively captures both semantic and covariate shifts. To this end, 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 are provided in the Appendix A.2).

Competitors. We compare our approach with several other methods, including: 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.

486 ETHICS STATEMENT
487488 We are aware of the potential biases in the training data that may lead to the generation of misleading
489 content. Measures have been taken using the method like probability calibration. Overall, our
490 approach is in line with ethical standards, aiming to enhance the reliability and robustness of the
491 model and is reliable for its intended purpose.
492493 REPRODUCIBILITY STATEMENT
494495 We clarify that the sources of all datasets used in the experiments have been provided in Appendix A.
496 Moreover, the theoretical derivations and properties of the diffusion model and beta calibration have
497 been included in the Appendix B and C, which aids in understanding the rationale for employing
498 these methods. Lastly, the model weights and training code will be made publicly available in due
499 course.
500501 REFERENCES
502503 Sima Behpour, Thang Long Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: A simple
504 yet efficient out-of-distribution detection with orthogonal projection of gradients. *Advances in
505 Neural Information Processing Systems*, 36:38206–38230, 2023.506 Eugene Berta, Francis Bach, and Michael Jordan. Classifier calibration with roc-regularized isotonic
507 regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 1972–1980.
508 PMLR, 2024.
509510 Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision
511 transformers, 2021. URL <https://arxiv.org/abs/2104.02057>.512 Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. De-
513 scribing textures in the wild, 2013. URL <https://arxiv.org/abs/1311.3618>.
514515 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hier-
516 archical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*,
517 pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.518 Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances
519 in neural information processing systems*, 34:8780–8794, 2021.
520521 Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual
522 outlier synthesis, 2022. URL <https://arxiv.org/abs/2202.01197>.523 Xuefeng Du, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination
524 with diffusion models, 2023. URL <https://arxiv.org/abs/2309.13415>.
525526 Bradley Efron. Tweedie's formula and selection bias. *Publications of the American Statistical
527 Association*, 106(496):1602–1614, 2011.528 Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. Diffguard: Semantic mismatch-
529 guided out-of-distribution detection using pre-trained diffusion models, 2023. URL <https://arxiv.org/abs/2308.07687>.
531532 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,
533 Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL <https://arxiv.org/abs/1406.2661>.
534535 Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural
536 networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.
537538 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual net-
539 works. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands,
October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016.

540 Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution
 541 examples in neural networks, 2018. URL <https://arxiv.org/abs/1610.02136>.
 542

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

545 Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam,
 546 Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset,
 547 2018. URL <https://arxiv.org/abs/1707.06642>.
 548

549 Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-
 550 distribution image without learning from out-of-distribution data, 2020. URL <https://arxiv.org/abs/2002.11297>.
 551

552 Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic
 553 space, 2021. URL <https://arxiv.org/abs/2105.01879>.
 554

555 Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional
 556 shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.
 557

558 Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
 559

560 Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily
 561 implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence
 562 and statistics*, pp. 623–631. PMLR, 2017.
 563

563 Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers
 564 for detecting out-of-distribution samples, 2018a. URL <https://arxiv.org/abs/1711.09325>.
 565

566 Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting
 567 out-of-distribution samples and adversarial attacks. *Advances in neural information processing
 568 systems*, 31, 2018b.
 569

570 Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan.
 571 Mage: Masked generative encoder to unify representation learning and image synthesis, 2023.
 572 URL <https://arxiv.org/abs/2211.09117>.
 573

574 Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised
 575 representation generation method, 2024. URL <https://arxiv.org/abs/2312.03701>.
 576

576 Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution
 577 image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
 578

579 Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection.
 580 *Advances in neural information processing systems*, 33:21464–21475, 2020.
 581

581 Yuyuan Liu, Choubi Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, and Gus-
 582 tavo Carneiro. Residual pattern learning for pixel-wise out-of-distribution detection in semantic
 583 segmentation, 2023a. URL <https://arxiv.org/abs/2211.14512>.
 584

585 Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q. Weinberger. Unsupervised out-of-
 586 distribution detection with diffusion inpainting, 2023b. URL <https://arxiv.org/abs/2302.10326>.
 587

588 Xingming Long, Jie Zhang, Shiguang Shan, and Xilin Chen. Rethinking the evaluation of out-
 589 of-distribution detection: A sorites paradox, 2024. URL <https://arxiv.org/abs/2406.09867>.
 590

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

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

597

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

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

603

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

606

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

609

610 Sina Sharifi, Taha Entesari, Bardia Safaei, Vishal M Patel, and Mahyar Fazlyab. Gradient-regularized
 611 out-of-distribution detection. In *European Conference on Computer Vision*, pp. 459–478. Springer,
 612 2024.

613

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

616

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

620

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

623

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

626

627 Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations,
 628 2021. URL <https://arxiv.org/abs/2111.12797>.

629

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

632

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

635

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

638

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

641

642 Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks
 643 know what they don’t know? *Advances in Neural Information Processing Systems*, 34:29074–
 644 29087, 2021.

645

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

649

650 Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database:
 651 Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference*
 652 *on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.
 653 5539970.

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

651

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

654

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

658

659 Suhee Yoon, Sanghyu Yoon, Ye Seul Sim, Sungik Choi, Kyungeun Lee, Hye-Seung Cho, Hankook
 660 Lee, and Woohyung Lim. Diffusion-based semantic outlier generation via nuisance awareness for
 661 out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*,
 662 volume 39, pp. 13078–13086, 2025.

663

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

667

668 Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi
 669 Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on
 670 imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp.
 671 558–567, 2021.

672

673 Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou
 674 Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-
 675 distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

676

677 Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10
 678 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine*
 679 *Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706 **Bad-OOD: Discovering Harmful Diffusion Outliers via**

707 **Confidence Calibration for OOD Detection (Appendix)**

708 A DETAILS OF DATASETS

709 A.1 QUANTITATIVE EVALUATION OF OOD DETECTOR PERFORMANCE

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

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

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

723 **iNaturalist** (Horn et al., 2018) contains images of natural world. It has 13 super-categories and 5,089
 724 sub-categories covering plants, insects, birds, mammals, and so on. We use the subset that contains
 725 110 plant classes which do not overlap with ImageNet-1k.

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

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

733 **Textures** (Describable Textures) (Cimpoi et al., 2013) contains images of textures and abstracted
 734 patterns. As no categories overlap with ImageNet-1k, we use the entire dataset as in (Huang & Li,
 735 2021).

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

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

742 **iSUN** (Xu et al., 2015) contains eye tracking ground-truth images derived from the SUN (Xiao et al.,
 743 2010) dataset, and we use the entire set of 8926 samples.

745 A.2 ADDITIONAL DATASETS USED

747 In addition to evaluating OOD detection performance, we also introduce the following datasets at
 748 the confidence calibration stage and the generalization assessment stage, respectively. Using the
 749 ImageNet-OOD (Yang et al., 2024a) dataset, we calibrate our OOD scores with respect to semantic
 750 shifts, thereby balancing the model’s sensitivity to both semantic and covariate shifts. The detailed
 751 calibration procedure is described in Appendix C.2. Additionally, we employ the SynIS (Long et al.,
 752 2024) dataset to assess the ability of various OOD scores to capture both types of distributional shifts.

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

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

762 B PRELIMINARIES

764 **Diffusion models** (Ho et al., 2020; Song et al., 2022a; Song et al.) are generative frameworks that
 765 learn data distributions through a gradual denoising process, which enable controlled generation by
 766 incorporating external guidance (e.g., class labels, text prompts) during training and sampling.
 767

768 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
 769 adding Gaussian noise over T timesteps. The noised sample \mathbf{x}_t at timestep t is defined as:

$$770 \quad 771 \quad q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N} \left(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I} \right), \quad (9)$$

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

$$774 \quad 775 \quad q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N} \left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I} \right), \quad (10)$$

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

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

$$779 \quad 780 \quad p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{y}) = \mathcal{N} \left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, \mathbf{y}), \Sigma_t \mathbf{I} \right), \quad (11)$$

781 where the mean μ_θ is derived from:

$$782 \quad 783 \quad \mu_\theta(\mathbf{x}_t, t, \mathbf{y}) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t, \mathbf{y}) \right). \quad (12)$$

786 Lastly the sampling procedure with time dependent constants σ_t follows:

$$787 \quad 788 \quad \mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t, \mathbf{y}) + \sigma_t \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}). \quad (13)$$

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

$$794 \quad 795 \quad s = \mathbb{E}[Y | f(X) = s], \quad (14)$$

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

$$800 \quad 801 \quad \mathbb{E}[Y | f(X) = s] = \frac{\sum_{j=1}^n y_j \cdot \mathbb{I}[f(\mathbf{x}_j) = s]}{\sum_{j=1}^n \mathbb{I}[f(\mathbf{x}_j) = s]}, \quad (15)$$

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

805 For any fixed classifier f , there exists a unique calibration map $\mu(s) = \mathbb{E}[Y | f(X) = s]$ that
 806 achieves perfect calibration on the training data. However, such a map risks severe overfitting,
 807 particularly when f assigns unique scores to individual instances. In this case, the calibration map
 808 collapses to $\mu(s_i) = y_i$, resulting in overconfident 0/1 predictions that generalize poorly to unseen
 809 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.

810 C BETA CALIBRATION
811812 C.1 DERIVATION OF BETA CALIBRATION.
813814 We provide a core derivation of the beta calibration method (Kull et al., 2017). The key assumption
815 is that the classifier’s output scores for each class follow the beta distributions. Therefore, this
816 distribution function can be used to derive the confidence calibration procedure for sigmoid-based
817 binary classification.
818819 **Class-conditional score distributions.** Assume the scores s for positive ($Y = 1$) and negative
820 ($Y = 0$) instances follow beta distributions with parameters (α_1, β_1) and (α_0, β_0) , respectively:
821

822
$$p(s | Y = 1) = \frac{s^{\alpha_1-1}(1-s)^{\beta_1-1}}{B(\alpha_1, \beta_1)}, \quad (16)$$

823

824
$$p(s | Y = 0) = \frac{s^{\alpha_0-1}(1-s)^{\beta_0-1}}{B(\alpha_0, \beta_0)}, \quad (17)$$

825

826 where $B(\cdot, \cdot)$ is the beta function.
827828 **Likelihood ratio derivation.** The likelihood ratio (LR) between the two classes is:
829

830
$$LR(s) = \frac{p(s | Y = 1)}{p(s | Y = 0)} = \frac{s^{\alpha_1-\alpha_0}(1-s)^{\beta_0-\beta_1}}{B(\alpha_1, \beta_1)/B(\alpha_0, \beta_0)}. \quad (18)$$

831

832 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
833 LR becomes:
834

835
$$LR(s; a, b, c) = \frac{s^a}{(1-s)^b} \cdot e^c. \quad (19)$$

836

837 **Calibrated probability mapping.** The calibrated probability $\mu_{\text{beta}}(s)$ is derived from the posterior
838 odds under a uniform class prior:
839

840
$$\mu_{\text{beta}}(s; a, b, c) = \frac{1}{1 + LR(s; a, b, c)^{-1}} = \frac{1}{1 + \frac{(1-s)^b}{s^a e^c}}. \quad (20)$$

841

842 Simplifying yields the beta calibration family:
843

844
$$\mu_{\text{beta}}(s; a, b, c) = \frac{1}{1 + \frac{1}{e^c \cdot \frac{s^a}{(1-s)^b}}}. \quad (21)$$

845

846 **Monotonicity constraints.** To ensure the calibration map is non-decreasing, parameters a and b
847 must satisfy $a, b \geq 0$. This preserves the ordinal relationship between classifier scores and calibrated
848 probabilities.
849850 C.2 DETAILED IMPLEMENTATION OF BETA CALIBRATION.
851852 We provide a detailed description of the beta calibration method for OOD detectors in Algorithm 1.
853 Specifically, the algorithm takes as input an uncalibrated classifier and calibration datasets, where the
854 ID validation set is drawn from 10% of the held-out data in ImageNet-100, and the OOD validation
855 set is constructed from the ImageNet-OOD (Yang et al., 2024a) dataset. The calibration parameters
856 “abc” are initialized at the beginning of the algorithm. By iterating over both the ID and OOD
857 validation sets, the algorithm computes the pre-calibration probabilities for each sample and fits the
858 calibration parameters according to the sample labels. The final output is a classifier calibrated with
859 beta parameters.
860

864	Algorithm 1 Beta Calibration of OOD Detector
865	Input: Uncalibrated classifier \mathcal{F}_θ , validation ID dataset \mathcal{D}_{in} , validation OOD dataset \mathcal{D}_{out}
866	Output: Calibrated classifier \mathcal{F}'_θ
867	Initialize Beta calibration parameters a_0, b_0, c_0 , empty probability list $P = []$ and label list $Y = []$
868	foreach mini-batch $(x_{id}, y_{id}), (x_{ood}, y_{ood})$ in $\mathcal{D}_{in}, \mathcal{D}_{out}$ do
869	Compute $p = \mathcal{F}_\theta(\text{concat}(x_{id}, x_{ood}))$
870	$P \leftarrow p, Y \leftarrow \text{concat}(y_{id}, y_{ood})$
871	end
872	Fit beta calibration: $a, b, c \leftarrow \text{fit}(P, Y)$
873	Predict beta calibration: $\hat{P} \leftarrow \text{predict}(P)$
874	return calibrated model \mathcal{F}'_θ with beta calibration applied
875	
876	
877	
878	

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.

918 Table 6: Comparisons of different calibration strategies with ImageNet-100 as in-distribution data.
919

920 Methods	iNaturalist		Places		Sun		Textures		Average				
	921 FPR95 AUROC		FPR95 AUROC		FPR95 AUROC		FPR95 AUROC		FPR95 AUROC				
	Beta calibration	923 Temperature scaling	924 Isotonic Regression	925 38.85	926 92.50	927 51.04	928 88.34	929 55.73	930 85.56	931 53.65	932 89.10	933 49.82	934 90.38

925
926
927
928
F COMPARISON OF OUTLIER SYNTHESIS METHODS AND MODEL
929 ARCHITECTURES
930931
932
933
In this section, we separately evaluate the impact of outlier embedding generation methods and model
934 architectures. It is important to note that all experiments in this part are conducted under calibrated
935 settings. We first introduce Gaussian noise at each denoising step of RDM (Li et al., 2024) for outlier
936 synthesis and observe that its performance is significantly inferior to that of the minority guidance
937 approach, indicating that the low-likelihood target provided by minority guidance is more explicit.
938 Furthermore, we assess our method on both ResNet-101 (He et al., 2016) and T2T-ViT (Yuan et al.,
939 2021) architectures. Compared to RankFeat (Song et al., 2022b), our results show that the samples
940 selected after calibration contain more effective boundary information than those based on singular
941 value.
942
943944 Table 7: Comparison of outlier synthesis methods and model architectures.
945

946 Outlier Synthesis	947 Architecture	948 FPR95↓ AUROC↑ AUPR↑		
		949 57.14	950 87.24	951 83.89
Add Gaussian Noise	ResNet-34	36.51	92.20	87.32
Minority Guidance	ResNet-34	36.80	92.15	-
RankFeat (Song et al., 2022b)	ResNet-101	35.68	92.28	87.37
Minority Guidance	ResNet-101	51.58	85.60	-
RankFeat (Song et al., 2022b)	T2T-ViT	50.21	89.47	84.58
Minority Guidance	T2T-ViT			

952
953
954
955
956
957
958
959
960
G GENERALIZATION AND ROBUSTNESS ANALYSIS
961962
963
964
G.1 GENERALIZABILITY ON SEMANTIC AND COVARIATE SHIFT.
965966
967
In this section, we provide more evaluation results of the generalizability of different anomaly
968 measurement approaches and the impact of using different OOD synthesizers. First, we comprehensively
969 present the score performance of more softmax-based methods, distance-based methods, and gradient-
970 based methods on the SynIS (Long et al., 2024) dataset (see Figure 6). In addition, we also assess the
971 previously proposed NPOS (Tao et al., 2023) and Dream-OOD (Du et al., 2023) methods (see Figure
972 7 & 8).

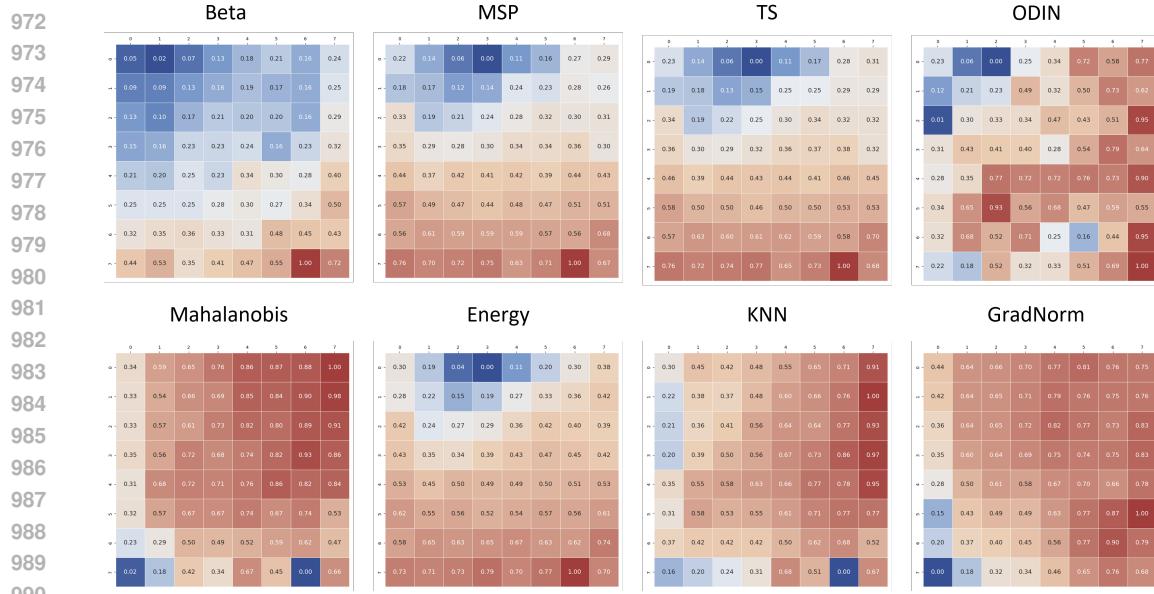


Figure 6: **Full results of beta score and other metrics.** In each graph, the semantic shift increases from top to bottom, and the covariate shift increases from left to right. All the scores are rescaled to range between 0 and 1, and the color red indicates a higher anomaly degree.

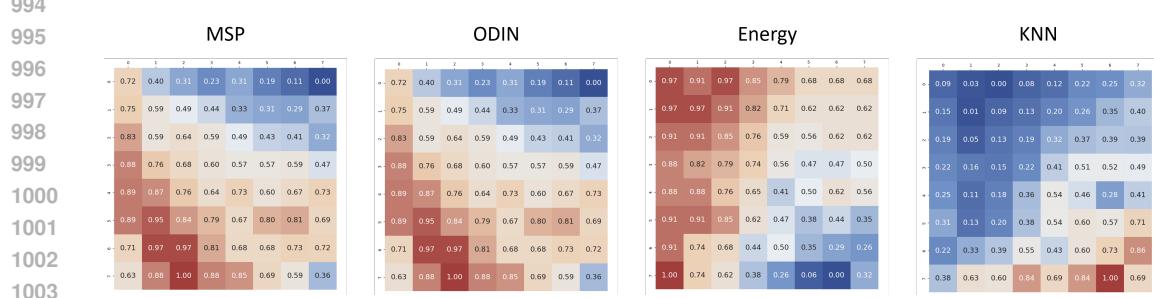


Figure 7: **Full results of SynIS using NPOS (Tao et al., 2023).** In each graph, the semantic shift increases from top to bottom, and the covariate shift increases from left to right. All the scores are rescaled to range between 0 and 1, and the color red indicates a higher anomaly degree.

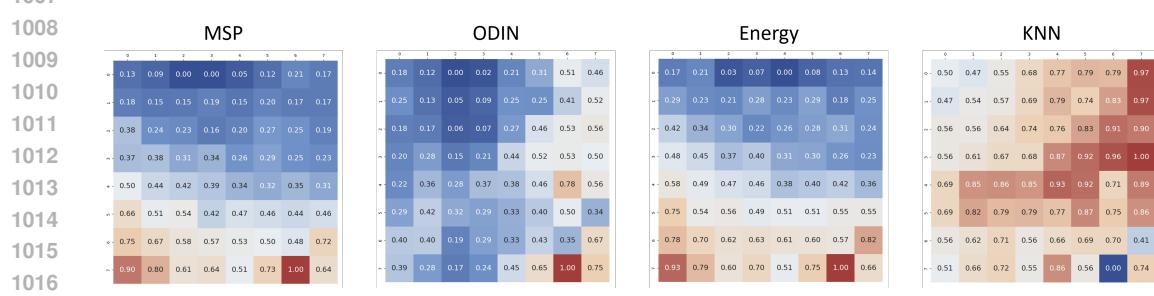


Figure 8: **Full results of SynIS using Dream-OOD (Du et al., 2023).** In each graph, the semantic shift increases from top to bottom, and the covariate shift increases from left to right. All the scores are rescaled to range between 0 and 1, and the color red indicates a higher anomaly degree.

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.

1044 G.2 ROBUSTNESS UNDER INTRA-CLASS AND INTER-CLASS VARIABILITY.

1045 While ImageNet-100 and CIFAR-100 already exhibit a reasonable degree of intra-class variability
 1046 and inter-class semantic proximity (e.g., fine-grained animal species or visually similar objects), we
 1047 believe that evaluating under more challenging settings would further strengthen the validation of our
 1048 framework. To this end, we have conducted additional experiments using in-distribution datasets with
 1049 increased semantic ambiguity—specifically, 10 manually selected ImageNet-100 classes with high
 1050 inter-class similarity, and 10 randomly selected ImageNet-R classes with higher intra-class variability.
 1051 As shown in Table 8, our proposed model consistently outperforms baseline methods, indicating its
 1052 robustness in these ambiguous scenarios. We believe this robustness stems from RDM’s operation in
 1053 the feature space of ID data, where the diversity-aware generation strategy naturally encourages the
 1054 synthesis of boundary-sensitive samples, regardless of the inherent class separability.

1055 **Table 8:** Generalization capacity of our RDM for outlier synthesis beyond standard benchmarks.

1058 Settings (In-distribution Data)	Ours		Dream-OOD	
	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

1064 G.3 ROBUSTNESS UNDER CLASS-IMBALANCED SETTING.

1065 While our main experiments were conducted on balanced datasets, we further examined the effectiveness
 1066 of the proposed strategy under class-imbalanced setting by randomly removing a varying
 1067 number of samples (ranging from 0 to 1200) from each class in ImageNet-100, resulting in a new
 1068 imbalanced version of the dataset. As shown in Table 9, our minority guidance strategy still yields
 1069 consistent and noticeable improvements over Dream-OOD under these imbalanced conditions. This
 1070 empirically suggests that our approach is inherently robust to class imbalance.

1071 **Table 9:** OOD detection results using the imbalanced version of ImageNet-100 as in-distribution.

1076 Methods	iNaturalist		Places		Sun		Textures		Average	
	FPR95	AUROC								
1077 Ours	34.27	87.10	42.15	84.31	48.65	89.03	49.85	85.60	43.88	85.03
1078 Dream-OOD	50.62	82.24	50.31	80.41	59.48	76.37	46.77	78.04	54.30	79.26

1080 H COMPARISON WITH SIMILAR DIFFUSION-BASED METHODS 1081

1082 We conducted additional experiments comparing our method against SONA (Yoon et al., 2025) and
1083 Fake it till you make it (Mirzaei et al., 2022). The results are summarized in Table 10. We can see
1084 that our method achieves better performance in terms of both average FPR95 (37.10) and AUROC
1085 (92.97), compared to SONA (FPR95: 37.46 and AUROC: 91.06) and Fake it till you make it (FPR95
1086 at 71.38 and AUROC at 79.29).

1087 Notably, although SONA also utilizes semantic/nuisance disentanglement, our approach differs in
1088 two key aspects: (i) Instead of relying on architectural constraints for disentanglement, we employ
1089 a classifier-guided calibration that allows flexible and fine-grained control over semantic shift and
1090 harmful perturbations; (ii) Our approach is also computationally more efficient. We conducted
1091 a dedicated comparison of generation and inference times, showing that our method is 3x faster
1092 compared to SONA. This improvement is largely attributed to our use of the lightweight RDM
1093 model, which does not require additional conditioning inputs, in contrast to SONA’s reliance on
1094 SD2 (Rombach et al., 2022), whose joint image-text reasoning introduces higher computational
1095 overhead.

1096 Table 10: OOD detection results of diffusion-based methods using CIFAR-100 as in-distribution data.
1097

Methods	SVHN		Places		LSUN		Textures		iSUN		Average		Inference Time Per 100 Samples (s)
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	
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	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	85.15	77.53	77.05	76.91	56.40	76.31	70.55	80.77	67.75	84.95	71.38	79.29	-

1104 I FINE GRANULARITY EVALUATION ON OPENOOD v1.5 1105

1106 OpenOOD v1.5 (Zhang et al., 2023), as a standardized benchmark in the field of OOD detection,
1107 encompasses a variety of settings with multiple ID and OOD datasets. Among them, we have selected
1108 ImageNet-200 as the ID dataset, together with near-OOD datasets (SSB-hard, NINCO), far-OOD
1109 datasets (iNaturalist, Textures, OpenImage-O), and non-semantic OOD variants (ImageNet-V2,
1110 ImageNet-C, ImageNet-R).

1112 I.1 IMAGENET-200 BENCHMARK.

1113 We applied both our model and SONA on ImageNet-100 to generate outliers, and then evaluated the
1114 resulting classifiers on ImageNet-200 (used as the ID dataset) together with near-OOD and far-OOD
1115 datasets. As shown in Table 11, our method consistently outperforms SONA across all categories,
1116 with more pronounced gains on far-OOD and covariate-shifted ID cases, demonstrating superior
1117 robustness and generalization.

1118 We attribute this improvement to a fundamental difference in design principle: whereas SONA
1119 primarily enhances decision boundaries by training on semantically similar hard negatives, our
1120 approach focuses on generating diverse and realistic OOD samples in pixel space, thereby improving
1121 classifier resilience to a wider range of distribution shifts.

1123 Table 11: Evaluation of diffusion-based methods on ImageNet-200 benchmark.
1124

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	59.31	66.77	63.04	79.61	64.09	68.44	70.71	22.28	

1131 I.2 IMAGENET-200 FULL-SPECTRUM BENCHMARK.

1132 Additionally, we conducted the evaluation under the full-spectrum setting, which included non-
1133 semantic OOD datasets. As shown in Table 12, our method exhibits more pronounced gains over

1134 Dream-OOD, particularly on far-OOD and non-semantic OOD benchmarks in OpenOOD v1.5,
 1135 underscoring our superior robustness under diverse types of distribution shifts. We attribute this
 1136 improvement to our method’s ability to generate OOD samples that more faithfully reflect real-world
 1137 distributional variations, thereby providing more effective supervision for classifier training.
 1138
 1139

1140 Table 12: Evaluation of diffusion-based methods on ImageNet-200 full-spectrum benchmark.
 1141

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

J ADDITIONAL VISUAL RESULTS

J.1 VISUALIZATION OF SYNTHETIC OOD.

In addition to the four categories presented in the main text—broccoli, candle, starfish, and strawberry, we also include ID images of the jeep category and their corresponding OOD results. Furthermore, we provide the anomaly score for each image using our proposed method.



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.

J.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.



K 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.

1242 To compare the computational performance of different methods, we conduct a comparison between
 1243 Bad-OOD and other baselines, calculating the inference time of synthesizing 100 OOD samples (all
 1244 experiments run on a single NVIDIA 3090 GPU). As shown in Table 13, our method is more efficient
 1245 in generating OOD samples than all baselines.

1246
1247 **Table 13: Comparison of computational cost in OOD synthesis.**

1249 Method	1249 Ours	1249 VOS	1249 NPOS	1249 Dream-OOD	1249 GAN	1249 SONA
1250 Inference Time Per 100 Samples (s)	1250 50	1250 89	1250 92	1250 110	1250 146	1250 160

1252
1253

L THE USE OF LARGE LANGUAGE MODELS (LLMs)

1254 In the process of writing this paper, we only employed LLMs as tools for polishing writing and
 1255 retrieving relevant knowledge (e.g., finding related work).

1256
 1257
 1258
 1259
 1260
 1261
 1262
 1263
 1264
 1265
 1266
 1267
 1268
 1269
 1270
 1271
 1272
 1273
 1274
 1275
 1276
 1277
 1278
 1279
 1280
 1281
 1282
 1283
 1284
 1285
 1286
 1287
 1288
 1289
 1290
 1291
 1292
 1293
 1294
 1295