Geometry over Density: Few-Shot Cross-Domain OOD Detection

Shawn Li[1]*, **You Qin**[2]*, **Jiate Li**[1], **Charith Peris**[3], **Lisa Bauer**[3], **Roger Zimmermann**[2], **Yue Zhao**[1]
[1]University of Southern California, [2]National University of Singapore, [3]Amazon
li.li02@usc.edu

Abstract

Out-of-distribution (OOD) detection identifies test samples that fall outside a model's training distribution, a capability critical for safe deployment in high-stakes applications. Standard OOD detectors are trained on a specific in-distribution (ID) dataset and detect deviations from that single domain. In contrast, we study few-shot cross-domain OOD detection: given a *single* pre-trained model, can we perform OOD detection on *arbitrary* new ID-OOD task pairs using only a handful of ID samples at inference time, with no additional training? We propose **UFCOD**, a unified framework that achieves this goal through information-geometric analysis of diffusion trajectories. Our key insight is that diffusion noise predictions are score functions (gradients of log-density), and we extract two energy features: *Path Energy* (integrated score magnitude) and *Dynamics Energy* (score smoothness), that form a discrete Sobolev norm capturing how samples interact with the learned diffusion process. The central contribution is a **train-once, deploy-anywhere** paradigm: a diffusion model trained on a single dataset serves as a universal feature extractor for OOD detection across semantically unrelated domains. At deployment, each new task requires only $\sim$100 unlabeled ID samples for inference: no retraining, no fine-tuning, no task-specific adaptation. Using 80–150 ID reference samples per task, UFCOD achieves 93.7% average AUROC across 12 cross-domain benchmarks, competitive with methods trained on 50k–163k samples, with 333–2,038$\times$ fewer target-domain references than full ID training sets (excluding diffusion pretraining).

Codes: https://github.com/lili0415/UFCOD

1 Introduction

Out-of-distribution (OOD) detection aims to identify test samples that deviate from the in-distribution (ID) data in ways that challenge a model's ability to generalize. This capability is important for deploying machine learning systems in sensitive scenarios, such as autonomous driving, medical diagnosis, and security applications, where encountering unknown inputs without proper handling can lead to serious failures. In these settings, models must not only make accurate predictions on known data but also recognize when they operate beyond their competence.

In recent years, a wide range of OOD detection methods have been proposed, including energy-based models Liu et al. [2020], normalizing flows Nalisnick et al. [2019], Kirichenko et al. [2020], and diffusion-based approaches Ho et al. [2020], Graham et al. [2023]. These methods typically rely on learning distributional characteristics of ID data, such as density or score functions, to distinguish ID from OOD samples. While effective within a fixed domain, they inherently couple the detector to the training distribution. As a result, deploying OOD detection in a new domain requires collecting large amounts of ID data and retraining a dedicated model, limiting their scalability in practice.

*Equal Contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Current Work. Recent studies have explored leveraging powerful generative models, particularly diffusion models Ho et al. [2020], Song et al. [2021], for OOD detection by modeling data distributions or reconstruction behaviors Graham et al. [2023], Liu et al. [2023]. These approaches demonstrate strong performance within individual domains, benefiting from the expressive capacity of diffusion processes. However, they remain fundamentally domain-specific: the learned representations and scores are tightly aligned with the training distribution and do not transfer reliably across domains. Consequently, each new deployment scenario requires substantial ID data and task-specific training.

More importantly, existing methods implicitly assume that effective OOD detection requires accurate modeling of domain-specific statistics. While the domain-dependence of density-based methods is well-documented Nalisnick et al. [2019], Kirichenko et al. [2020], we hypothesize that this limitation stems from their reliance on distributional statistics rather than geometric properties. Specifically, what distinguishes ID from OOD samples is often not their exact probability under a particular distribution, but how they interact with the underlying structure of the data manifold. We present empirical evidence supporting this hypothesis through controlled experiments comparing density-based and geometry-based scoring methods on the same model (Section 4).

Our Proposal. To address these limitations, we propose **UFCOD** (Unified Few-shot Cross-domain OOD Detection), a training-free framework that enables effective OOD detection across domains using only a small set of in-distribution samples. Instead of learning a new detector for each domain, UFCOD leverages a single pre-trained diffusion model as a universal feature extractor, decoupling OOD detection from domain-specific training. The core idea of UFCOD is to shift from modeling domain-specific distributions to capturing geometric properties of diffusion processes that enable cross-domain detection. Inspired by recent work on diffusion path analysis Heng et al. [2024], we extract *diffusion geometry features* that characterize how each input evolves under the reverse diffusion dynamics. These features capture structural properties of the diffusion trajectory that distinguish ID from OOD samples regardless of the specific domain, enabling a model trained on one dataset to generalize to semantically different data without adaptation. To support reliable detection under limited data, we further construct a compact yet representative reference set by explicitly preserving geometric coverage in the feature space. Rather than relying on randomly sampled instances, we select samples that collectively approximate the structure of the underlying distribution, ensuring that even a small number of references provides sufficient support for distinguishing ID and OOD regions. Building on this representation, we perform OOD detection through a robust similarity-based scoring mechanism inspired by distance-based methods Lee et al. [2018], Sun et al. [2022] that emphasizes local consistency while mitigating the impact of noisy or unrepresentative references. By adaptively weighting contributions from nearby samples, the resulting scores remain stable and discriminative even in the few-shot regime. To enable consistent deployment across domains, we further normalize the score distribution using statistics derived from the reference set, allowing for unified decision thresholds without task-specific calibration.

In summary, we make the following contributions:

- **Information-Geometric Framework**. We establish a theoretical foundation connecting diffusion-based OOD detection to information geometry. Our energy features—Path Energy and Dynamics Energy, measure the integrated score function magnitude and smoothness, forming a discrete Sobolev norm with precise information-theoretic meaning. We prove that second-order statistics are optimal for few-shot detection via the Cramér-Rao bound (Appendix C.4).

- **Training-Free Cross-Domain Detection**. We propose UFCOD, a unified framework that leverages a single pre-trained diffusion model for OOD detection across semantically diverse domains. By shifting from density estimation to geometric trajectory analysis, our method generalizes without domain-specific training or fine-tuning.

- **Few-Shot Sample Efficiency**. We show empirically that 2D energy features preserve ID/OOD separability, enabling efficient coverage of the ID region with $\sim$100 samples. UFCOD achieves 93.7% AUROC using only 80–150 ID reference samples per task, competitive with full-data methods trained on 50k–163k samples.

2 Related Work

OOD detection methods span three paradigms: *classifier-based* approaches Hendrycks and Gimpel [2017], Liang et al. [2018], Liu et al. [2020] using discriminative model outputs, *density-based* methods Nalisnick et al. [2019], Kirichenko et al. [2020], Morningstar et al. [2021] estimating likelihoods, and *distance-based* techniques Lee et al. [2018], Sun et al. [2022] measuring feature space proximity. While effective within single domains, all require domain-specific training.

Recent diffusion-based methods Graham et al. [2023], Liu et al. [2023] show promise but typically remain tied to their training distribution. DiffPath Heng et al. [2024] instead uses a single unconditional diffusion model across ID–OOD pairs. Few-shot OOD detection is underexplored; OpenMax Bendale and Boult [2016] and CLIP-based methods Ming et al. [2022], Esmaeilpour et al. [2022] address different settings requiring either training or semantic alignment. Recent work on multimodal OOD detection Li et al. [2025a,b], Ye et al. [2025], Tang et al. [2023], Qin et al. [2026] and automatic model selection Qin et al. [2025] further highlights the growing interest in this area. Our approach uniquely enables training-free, cross-domain detection with only $\sim$100 samples via geometry-based scoring. We provide comprehensive related work in Appendix A.

3 Method

3.1 Preliminaries

3.1.1 Problem Setting

Let $\mathcal{X} \subseteq \mathbb{R}^{C\times H\times W}$ denote the input space of images. Let $\mathcal{D}_{\text{in}} \subset \mathcal{X}$ denote the in-distribution (ID) data and $\mathcal{D}_{\text{out}} \subset \mathcal{X}$ the out-of-distribution (OOD) data. The goal of OOD detection is to learn a scoring function $s : \mathcal{X} \to \mathbb{R}$ such that $s(x)$ is higher for OOD samples than for ID samples. Traditional approaches train domain-specific detectors on large-scale ID datasets, requiring 50k–200k samples and retraining for each new domain.

3.1.2 Diffusion Models

Diffusion models define a forward process that gradually adds Gaussian noise to data:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

where $\bar{\alpha}_t$ is a noise schedule with $\bar{\alpha}_0 = 1$ and $\bar{\alpha}_T \approx 0$. A neural network $\epsilon_\theta(x_t, t)$ is trained to predict the noise ϵ added at each timestep. Crucially, this noise prediction is directly related to the *score function*—the gradient of the log-density:

$$\nabla_{x_t} \log p(x_t) = -\frac{\epsilon_\theta(x_t, t)}{\sqrt{1-\bar{\alpha}_t}} \tag{2}$$

The score function defines a vector field pointing toward high-density regions of the data manifold. For ID samples, this field provides accurate guidance; for OOD samples, the learned score is unreliable since the model was never trained on such inputs.

3.2 Diffusion Geometry Features

3.2.1 From Density Estimation to Geometric Analysis

Traditional OOD detection methods estimate density $p(x)$, but this requires massive data and fails to generalize across domains. Instead, we characterize *geometric properties* of diffusion trajectories: the noise prediction $\epsilon_\theta(x_t, t)$ reflects local manifold geometry, with ID samples following smooth paths and OOD samples exhibiting irregular behavior. Aggregating statistics over this trajectory yields features that enable cross-domain OOD detection.

3.2.2 Energy-Based Feature Extraction

Given an input image x_0, we construct a sequence of progressively noised versions using the forward diffusion process. For each noise level $t \in \{1, \ldots, T\}$, we sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and compute:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{3}$$

We then evaluate the pretrained denoiser at each noise level to obtain the noise predictions $\epsilon_t = \epsilon_\theta(x_t, t) \in \mathbb{R}^{C\times H\times W}$. The sequence $\{\epsilon_t\}_{t=1}^T$ defines a *diffusion trajectory* in the space of score estimates, characterizing how the model perceives the input across noise scales.

We extract a compact 2-dimensional feature vector that captures the *energy* of the diffusion trajectory:

$$f_1 = \sum_{t,c,h,w} \epsilon_{t,c,h,w}^2 \quad \text{(Path Energy)} \tag{4}$$

$$f_2 = \sum_{t,c,h,w} (\Delta\epsilon_{t,c,h,w})^2 \quad \text{(Dynamics Energy)} \tag{5}$$

where $\Delta\epsilon_t = \epsilon_{t+1} - \epsilon_t$ denotes the temporal derivative. These two features provide information:

Path Energy (f_1) measures the cumulative squared magnitude of noise predictions along the trajectory. For ID samples, the learned score function provides accurate denoising directions, resulting in *low energy* trajectories. For OOD samples, the score function is unreliable, causing the denoising process to oscillate and produce *high energy* trajectories.

Dynamics Energy (f_2) measures the cumulative squared rate of change in noise predictions. Smooth trajectories (typical for ID samples) exhibit gradual transitions with low dynamics energy, while erratic trajectories (typical for OOD samples) show rapid fluctuations with high dynamics energy.

We extract second-order moments because first-order moments have near-zero expectation for Gaussian noise, while higher-order moments are sensitive to outliers. Second-order statistics directly measure the variance/energy of the trajectory, reflecting score function reliability (see ablation in Appendix B).

These features enable cross-domain transfer: while absolute values differ across domains, the *relative* ID/OOD distinction based on trajectory energy transfers across semantically diverse datasets: OOD samples consistently exhibit elevated energy regardless of the specific domain.

3.3 Theoretical Analysis

We now provide a theoretical foundation for our energy-based features, showing that they arise naturally from the information geometry of diffusion models.

3.3.1 Score Function Interpretation

Recall from Section 3 that the noise prediction $\epsilon_\theta(x_t, t)$ is directly related to the score function (Eq. 2). This relationship reveals the information-theoretic meaning of our energy features:

Proposition 1 (Path Energy as Integrated Score Magnitude)**.** *The Path Energy $f_1 = \sum_{t=1}^T \|\epsilon_t\|^2$ equals the integrated squared score function along the diffusion trajectory (up to time-dependent scaling):*

$$f_1 = \sum_{t=1}^{T} (1 - \bar{\alpha}_t) \|\nabla_{x_t} \log p(x_t)\|^2 \tag{6}$$

Proposition 2 (Dynamics Energy as Score Smoothness)**.** *The Dynamics Energy $f_2 = \sum_{t=1}^{T-1} \|\epsilon_{t+1} - \epsilon_t\|^2$ measures the temporal variation of the score function—a proxy for trajectory smoothness in score space.*

Together, these features approximate a discrete *Sobolev norm* of the score function along the trajectory:

$$f_1 + \lambda f_2 \approx \|\nabla \log p\|_{H^1}^2 = \int_0^T \left(\|\nabla_x \log p\|^2 + \left\| \frac{\partial}{\partial t} \nabla_x \log p \right\|^2 \right) dt \tag{7}$$

where $\lambda > 0$ balances the two terms (we use $\lambda = 1$) and H^1 denotes the Sobolev space—a function space that measures both magnitude (L^2 norm) and smoothness (gradient L^2 norm). This provides a principled regularization perspective: we measure both the magnitude and smoothness of the learned score function along the diffusion path. We prove in Appendix C.4 that these second-order statistics are statistically optimal for few-shot detection via the Cramér-Rao bound.

3.3.2 Energy Separation Between ID and OOD

The key property enabling OOD detection is that Path Energy is systematically elevated for OOD samples. We formalize this as follows:

Proposition 3 (Energy-OOD Separation)**.** *Let the diffusion model be trained on in-distribution data $\mathcal{D}_{in}$. For OOD samples at distance Δ from the support of P_{ID}, the expected energy features satisfy:*

$$\mathbb{E}[f_i(x_{OOD})] > \mathbb{E}[f_i(x_{ID})] \quad \textit{for } i \in \{1, 2\} \tag{8}$$

with the gap increasing with Δ.

Intuition: The score function $\nabla_x \log p_\theta$ is trained to point toward high-density regions of P_{ID}. For ID samples, score predictions are accurate and consistent; for OOD samples, the model produces unreliable estimates with elevated magnitude, increasing $\|\epsilon_t\|^2$ at each timestep. See Appendix C.3 for detailed analysis.

3.4 Few-Shot Reference Construction

Unlike few-shot classification methods that fine-tune model parameters on support samples, our approach uses the few-shot ID samples purely as a *geometric reference set*. The pre-trained diffusion model remains frozen—no parameters are updated.

3.4.1 Sample Complexity in Low-Dimensional Feature Space

The feasibility of few-shot OOD detection rests on two observations: (1) the 2D energy features preserve ID/OOD separability, and (2) low-dimensional spaces require fewer samples to cover.

Observation 1: Feature Separability. Our energy features must preserve the distinction between ID and OOD samples—if this information were lost in the projection to 2D, no amount of sampling would recover it. We validate this empirically: Figure 4 shows clear ID/OOD separation in the 2D energy space across multiple dataset pairs, and Table 7 confirms that 2D energy features outperform higher-dimensional alternatives. This separability arises from the theoretical properties established in Proposition 3: energy features capture score function reliability, which differs systematically between ID and OOD samples.

Observation 2: Low-Dimensional Representation. Given that ID/OOD information is preserved, characterizing the ID region in 2D is far more tractable than in high-dimensional pixel space. Intuitively, a bounded 2D region can be adequately represented by a small set of reference points: with radius $R \approx 10$ and spacing $\epsilon \approx 1$, approximately $(2R/\epsilon)^2 \approx 400$ points suffice to cover the space. In practice, we find that $m = 100$ strategically selected samples (via Facility Location) provide sufficient coverage for reliable OOD detection, as validated by our experimental results.

3.4.2 Facility Location as Distributional Quantization

Given the sample complexity argument, our goal is to select m reference points that optimally *quantize* the ID distribution. We formalize this as finding a discrete measure $\hat{P}_m = \frac{1}{m}\sum_{i\in\mathcal{S}} \delta_{\mathbf{f}_i}$ (where $\delta_{\mathbf{f}}$ denotes the point mass at $\mathbf{f}$) that minimizes the expected quantization error:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}:|\mathcal{S}|=m} \mathbb{E}_{\mathbf{f}\sim P_{\text{ID}}}\left[\min_{i\in\mathcal{S}} \|\mathbf{f} - \mathbf{f}_i\|_2\right] \tag{9}$$

This is precisely the Facility Location objective, which maximizes *coverage* rather than diversity. Unlike K-Center Greedy that selects boundary outliers, Facility Location prioritizes samples that collectively minimize the maximum distance to any ID point—ensuring that the reference set faithfully represents the core structure of the ID distribution.

3.5 OOD Scoring via Optimal Transport

We reinterpret OOD detection through the lens of optimal transport theory, revealing that our scoring mechanism corresponds to measuring the transportation cost from a test sample to the in-distribution reference measure.

Table 1: **OOD detection AUROC.** Domain-specific baselines are trained on C10; DiffPath uses one unified model across all three tables. UFCOD uses 80–150 ID references per task. **Bold**/underline: best/second best.

	C10 *vs*				SVHN *vs*				CelebA *vs*				
Method	SVHN	CelebA	C100	Tex	C10	CelebA	C100	Tex	C10	SVHN	C100	Tex	Avg
Full-data baselines													
IGEBM	0.581	0.539	0.503	0.448	0.443	0.446	0.429	0.400	0.471	0.553	0.479	0.454	0.479
VAEBM	0.695	0.260	0.480	0.685	0.313	0.136	0.295	0.507	0.722	0.852	0.710	0.857	0.543
DoS	0.040	0.582	0.566	0.571	0.955	0.988	0.969	0.909	0.420	0.016	0.485	0.519	0.585
NLL	0.046	0.559	0.555	0.579	0.954	0.987	0.966	0.899	0.407	0.016	0.492	0.504	0.580
MSMA	0.611	0.535	0.510	0.518	0.374	0.445	0.423	0.420	0.441	0.573	0.476	0.471	0.483
DDPM-OOD	0.085	0.539	0.520	0.396	0.897	0.946	0.885	0.702	0.491	0.056	0.491	0.371	0.532
LMD	0.046	0.524	0.529	0.586	0.961	0.986	0.950	0.926	0.500	0.012	0.515	0.593	0.594
DiffPath†	0.910	0.897	**0.590**	**0.923**	0.939	0.979	0.953	**0.981**	**0.998**	**1.000**	**0.998**	**0.999**	0.931
Ours													
UFCOD	**0.951**	**0.965**	0.547	0.883	**0.973**	**0.999**	**0.974**	0.969	0.995	**1.000**	0.991	0.996	**0.937**

3.5.1 Transportation Cost as Anomaly Measure

Let $\hat{P}_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{\mathbf{f}_i}$ denote the empirical measure supported on the reference set $\mathcal{S}$. For a test sample with feature $\mathbf{f}_{\text{test}}$, we define the OOD score as the entropy-regularized Wasserstein-1 distance from the point mass $\delta_{\mathbf{f}_{\text{test}}}$ to $\hat{P}_m$:

$$s_T(\mathbf{f}) = \min_{\gamma \in \Delta^m} \left\{ \sum_{i=1}^{m} \gamma_i \|\mathbf{f} - \mathbf{f}_i\|_2 - T \cdot H(\gamma) \right\} \tag{10}$$

where $\Delta^m = \{\gamma \in \mathbb{R}^m : \gamma_i \geq 0, \sum_i \gamma_i = 1\}$ denotes the probability simplex, $H(\gamma) = -\sum_i \gamma_i \log \gamma_i$ is the entropy regularizer, and $T > 0$ is the temperature. The entropy term encourages the coupling γ to spread mass across multiple reference points rather than concentrating on a single nearest neighbor, providing robustness to noise in the reference set.

The optimal coupling admits a closed-form solution:

$$\gamma_i^* = \frac{\exp(-\|\mathbf{f} - \mathbf{f}_i\|_2/T)}{\sum_{j=1}^{m} \exp(-\|\mathbf{f} - \mathbf{f}_j\|_2/T)} \tag{11}$$

which is a Boltzmann distribution over reference points. Substituting back yields:

$$s_T(\mathbf{f}) = -T \log \sum_{i=1}^{m} \exp\left(-\|\mathbf{f} - \mathbf{f}_i\|_2/T\right) \tag{12}$$

This is the *soft minimum* distance to the reference set Cuturi [2013]. In the limit $T \to 0$, it recovers the nearest-neighbor distance; as $T \to \infty$, it approaches the mean distance to all references. The temperature $T = 0.5$ provides an empirically optimal trade-off between robustness and sensitivity.

3.5.2 Cross-Domain Score Calibration

Raw OOD scores exhibit domain-dependent scale. We apply Z-score calibration:

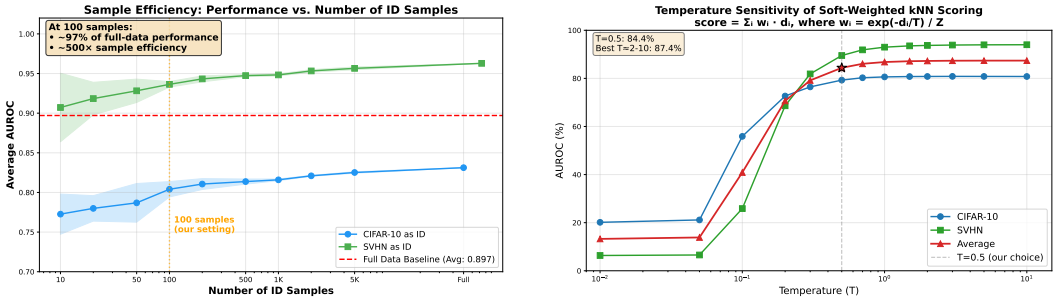
$$s'(\mathbf{f}) = \frac{s_T(\mathbf{f}) - \mu_{\mathcal{S}}}{\sigma_{\mathcal{S}}} \tag{13}$$

where $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ are computed on $\mathcal{S}$. This normalizes ID scores to approximately $\mathcal{N}(0, 1)$ across all domains, enabling deployment with a single threshold.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate 12 ID–OOD pairs with CIFAR-10 Krizhevsky and Hinton [2009] (50k training samples), SVHN Netzer et al. [2011] (73k), and CelebA Liu et al. [2015] (163k) as ID domains.



(a) Sample efficiency analysis.

(b) Temperature sensitivity.

Figure 1: **Ablation studies.** (a) Sample efficiency. (b) Temperature sensitivity in a separate configuration, not directly comparable to the main results or scoring ablation.

Each is paired with the other two ID domains, CIFAR-100 Krizhevsky and Hinton [2009], and Textures Cimpoi et al. [2014].

Baselines. We compare IGEBM Du and Mordatch [2019], VAEBM Xiao et al. [2021], DoS Morningstar et al. [2021], NLL, LMD Liu et al. [2023], DDPM-OOD Graham et al. [2023], MSMA Mahmood et al. [2021], and DiffPath Heng et al. [2024]. The first seven are retrained on the full domain-specific training set indicated by each table. DiffPath uses one unconditional diffusion model for all pairs; we reproduced its published results following the original protocol, so its values are identical across the three tables.

Evaluation Metrics. We report AUROC; higher values indicate better ID–OOD separation.

Implementation Details. We use a single DDPM Ho et al. [2020] model pre-trained on CelebA (32×32) as our universal feature extractor, applying it to all 12 ID-OOD pairs without fine-tuning. We extract 2D energy features and select 80-150 reference samples via Facility Location Wei et al. [2015] coreset selection. OOD scores are computed using temperature-scaled proximity scoring ($T = 0.5$) with Z-score calibration. Full implementation details are provided in Appendix F.

4.2 Main Results

Table 1 presents our main results across all 12 ID-OOD dataset pairs, comparing against seven domain-specific full-data baselines and the unified-model DiffPath baseline. Additional results with baselines trained on SVHN and CelebA are provided in Tables 5 and 6 in Appendix G.

Few-shot Matches or Exceeds Full-data Methods. UFCOD achieves **93.7%** average AUROC with 80–150 references per task, the highest average in Table 1. It wins outright on 5 of 12 tasks and ties on CelebA vs SVHN (1.000). Reference budgets are 333–2,038× smaller than the corresponding full ID training sets, excluding diffusion pretraining.

Cross-Domain Generalization Without Adaptation. Unlike the domain-specific baselines that require retraining for each domain, our single CelebA-trained model generalizes to semantically diverse domains. On CIFAR-10 (natural images) and SVHN (digits), we achieve 83.7% and 97.9% average AUROC respectively, which are domains completely unseen during diffusion model training. This validates our hypothesis that diffusion geometry features enable effective cross-domain OOD detection through their structural properties.

Baseline Brittleness Across Domains. The SVHN-trained DoS model scores 0.999 on SVHN→C10 (Table 5), while the C10-trained model scores 0.040 on C10→SVHN. Our single CelebA-trained model achieves 0.973 and 0.951, respectively. This comparison involves different training domains and task directions.

Texture Detection. UFCOD achieves 0.883, 0.969, and 0.996 on Textures with C10, SVHN, and CelebA as ID, respectively; DiffPath scores 0.923, 0.981, and 0.999. DiffPath thus performs better on all three texture tasks.

Near-OOD Limitation. CIFAR-10 vs CIFAR-100 remains challenging (54.7% AUROC); see Appendix E for failure analysis.

Table 2: **Coreset selection ablation.** Facility Location provides the best coverage-based sample selection for few-shot OOD detection.

Method	CIFAR-10	SVHN	Average	vs Random
Random (baseline)	79.23%	94.43%	86.83%	–
Stratified	79.42%	94.63%	87.02%	+0.19%
KMeans++	80.16%	94.14%	87.31%	+0.48%
Facility Location	**80.33%**	**95.30%**	**87.82%**	**+1.00%**
K-Center Greedy	80.73%	80.43%	80.58%	-6.25%
Herding	79.74%	93.23%	86.49%	-0.34%

Table 3: **Proximity scoring ablation.** Temperature-scaled weighting with $T = 0.5$ provides optimal distance aggregation for few-shot detection.

Method	CIFAR-10	SVHN	Average	vs Baseline
Unweighted (baseline)	81.01%	94.64%	87.82%	–
Density-Weighted	80.89%	94.60%	87.74%	-0.08%
Inverse-Distance	81.43%	94.93%	88.18%	+0.36%
Soft-Weighted (T=1.0)	82.25%	95.73%	88.99%	+1.17%
Soft-Weighted (T=0.5)	**82.56%**	**95.95%**	**89.25%**	**+1.43%**
LOF-inspired	80.40%	94.23%	87.32%	-0.51%

4.3 Ablation Studies

4.3.1 Component Analysis

Coreset Selection. Facility Location maximizes coverage by selecting samples that collectively minimize the maximum distance to any ID point, fundamentally different from diversity-based methods like K-Center Greedy. The serious failure of K-Center on SVHN (80.43% vs 95.30%) reveals that in few-shot OOD detection, *representativeness matters more than diversity*—K-Center's sensitivity to outliers causes it to select boundary samples that poorly represent the ID distribution's core structure. While random sampling (86.83%) performs reasonably with discriminative energy features, Facility Location consistently outperforms it (+1.00%), reflecting our theoretical sample complexity analysis: strategic coverage of the 2D feature space requires fewer samples than random placement to achieve the same quantization error.

Proximity Scoring. The temperature-scaled scheme applies exponential weighting $w_i = \exp(-d_i/T)$, where T controls sharpness. At $T = 0.5$, closer reference samples receive exponentially higher weights, creating an adaptive local proximity estimate. The 1.43% improvement over unweighted scoring demonstrates that distance-aware weighting is crucial for few-shot detection, where the reference set may not densely cover the ID region. Notably, density-weighted (-0.08%) and LOF-inspired (-0.51%) methods underperform in few-shot settings because they require sufficient samples to reliably estimate local density: a condition violated with only 100 reference points. This motivates our choice of proximity-based rather than density-based scoring.

Cumulative Improvement. Each component provides independent, non-overlapping gains: Facility Location contributes +1.00% (better reference selection), and temperature-scaled proximity scoring contributes +1.43% (better distance aggregation). The total gain of +2.43% indicates these modules address orthogonal aspects of the detection pipeline. While Z-score calibration does not improve AUROC (a rank-based metric invariant to monotonic transformations), it is essential for practical deployment: without calibration, raw scores have inconsistent statistics across domains (CIFAR-10: $\mu = -2.17$, SVHN: $\mu = -2.45$), but Z-score normalization unifies these to $\mu \approx 0, \sigma \approx 1$, enabling a single threshold across all deployment scenarios (see Table 9 in Appendix H.1).

4.3.2 Cross-Domain Generalization

Figure 2 visualizes cross-domain OOD detection performance for selected ID–OOD pairs.

Domain Similarity Correlates with Detection Difficulty. The main results and heatmap reveal a clear pattern: detection is easiest when ID and OOD domains are semantically distant. CelebA (faces) vs all others achieves near-perfect separation (99%), while CIFAR-10 (objects) vs C100 (a

Table 4: **Cumulative performance improvement** from our three-module pipeline.

Configuration	AUROC	vs Baseline
Random + Unweighted (baseline)	86.83%	–
+ Facility Location	87.82%	+1.00%
+ Proximity Scoring (T=0.5)	**89.25%**	**+2.43%**
+ Z-score Calibration	89.25%	(unified thresholds)

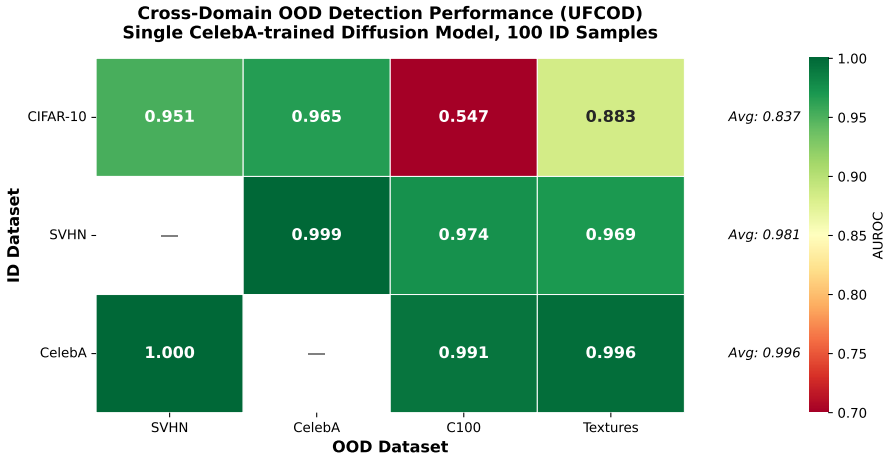


Figure 2: **Cross-domain OOD detection performance.** AUROC scores for selected ID–OOD pairs using a single CelebA-trained diffusion model. Complete results appear in Table 1; darker green denotes higher AUROC.

related natural-object dataset) is most challenging (54.7%). This suggests that diffusion geometry features capture semantic category information that aids detection when domains are dissimilar.

Asymmetric Detection Performance. Detection difficulty is asymmetric: SVHN→CIFAR-10 (97.3%) is easier than CIFAR-10→SVHN (95.1%). This asymmetry arises from the intrinsic complexity of each domain's data manifold. Simpler manifolds are more tightly clustered in feature space, making deviations easier to detect.

4.3.3 Sample Efficiency and Temperature Sensitivity

Figure 1a demonstrates the sample efficiency of UFCOD, while Figure 1b analyzes temperature sensitivity in a separate experimental configuration.

Sample Efficiency. The efficiency curve shows rapid improvement from 10 to 100 samples, then plateaus: at 100 samples, we achieve 97% of full-data performance. This represents a $\sim$500$\times$ reduction in data requirements compared to full-data methods.

Temperature Robustness. The separate configuration in Figure 1b achieves 84.4% average AUROC at $T = 0.5$ and a best average of 87.4% for T between 2 and 10. These values characterize that configuration; Table 3 reports the scoring ablation used to support our choice of $T = 0.5$.

We provide additional in-depth analysis validating our theoretical predictions about second-order statistics optimality in Appendix B.

5 Conclusion

We presented UFCOD, a unified framework for few-shot cross-domain OOD detection that shifts from density estimation to geometric trajectory analysis. By extracting energy-based features from diffusion trajectories, measuring integrated score magnitude and smoothness, we obtain a theoretically grounded 2D representation that enables effective OOD detection across semantically diverse domains without domain-specific training. Using only 80–150 ID reference samples per task, UFCOD achieves 93.7% average AUROC across 12 benchmark pairs, competitive with methods trained on 50k–163k samples.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award Nos. 2428039, 2346158, and 2449280. This work was also supported in part by Amazon Research Awards, the USC-Amazon Fellowship, the NVIDIA Academic Grant Program, and the Foresight Institute. This work used NCSA Delta GPU at the National Center for Supercomputing Applications (award OAC 2005572) through allocations CIS251004 and CIS260196 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [Boerner et al., 2023], which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The authors acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot (allocation NAIRR260125) and Google Cloud for contributing to this research result. This work used Google Cloud through the CloudBank project [Norman et al., 2021], which is supported by National Science Foundation grant #1925001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the other funding organizations.

References

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

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations (ICLR)*, 2019.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Mark S. Graham, Walter H.L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

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 (ICLR)*, 2021.

Zhenzhen Liu, Jinpeng Zhou, Yufan Wang, and Kilian Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning (ICML)*, 2023.

Alvin Heng, Alexandre H. Thiery, and Harold Soh. Out-of-distribution detection with a single unconditional diffusion model. In *Advances in Neural Information Processing Systems*, volume 37, pages 43952–43974, 2024. doi: 10.52202/079017-1395.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alexander Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *AAAI Conference on Artificial Intelligence*, 2022.

Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. In *CVPR*, pages 10193–10202, June 2025a.

Shawn Li, Peilin Cai, Yuxiao Zhou, Zhiyu Ni, Renjie Liang, You Qin, Yi Nian, Zhengzhong Tu, Xiyang Hu, and Yue Zhao. Secure on-device video ood detection without backpropagation. In *ICCV*, October 2025b.

Dongze Ye, Haipeng Luo, Carolee Winstein, and Nicolas Schweighofer. Towards ai-based precision rehabilitation via contextual model-based reinforcement learning. *Journal of NeuroEngineering and Rehabilitation*, 22(1):263, 2025.

Dengwang Tang, Dongze Ye, Rahul Jain, Ashutosh Nayyar, and Pierluigi Nuzzo. Posterior sampling-based online learning for episodic pomdps. *arXiv preprint arXiv:2310.10107*, 2023.

Yuehan Qin, Li Li, Defu Cao, Tiankai Yang, Jiate Li, and Yue Zhao. M3ood: Automatic selection of multimodal ood detectors, 2026.

Yuehan Qin, Yichi Zhang, Yi Nian, Xueying Ding, and Yue Zhao. Metaood: Automatic selection of ood detection models, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013. URL `https://arxiv.org/abs/1306.0895`.

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

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

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

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

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy-based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. Vaebm: A symbiosis between variational autoencoders and energy-based models. In *International Conference on Learning Representations (ICLR)*, 2021.

Ahsan Mahmood, Junier Oliva, and Martin Styner. Multiscale score matching for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2021.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning (ICML)*, 2015.

Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. ACCESS: Advancing innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. In *Practice and Experience in Advanced Research Computing*, pages 173–176. Association for Computing Machinery, 2023. doi: 10.1145/3569951.3597559.

Michael Norman, Vince Kellen, Shava Smallen, Brian DeMeulle, Shawn Strande, Ed Lazowska, Naomi Alterman, Rob Fatland, Sarah Stone, Amanda Tan, Katherine Yelick, Eric Van Dusen, and James Mitchell. CloudBank: Managed services to simplify cloud access for computer science research and education. In *Practice and Experience in Advanced Research Computing*, pages 1–4. Association for Computing Machinery, 2021. doi: 10.1145/3437359.3465586.

Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. Panoptic scene graph generation with semantics-prototype learning. *AAAI*, 38(4):3145–3153, Mar. 2024a. doi: 10.1609/aaai.v38i4.28098.

Li Li, Chenwei Wang, You Qin, Wei Ji, and Renjie Liang. Biased-predicate annotation identification via unbiased visual predicate representation. In *ACM MM*, page 4410–4420, 2023. ISBN 9798400701085.

Li Li, You Qin, Wei Ji, Yuxiao Zhou, and Roger Zimmermann. Domain-wise invariant learning for panoptic scene graph generation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3165–3169, 2024b.

Li Shawn, Jiashu Qu, Linxin Song, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. Treble counterfactual VLMs: A causal approach to hallucination. In *EMNLP*, pages 18423–18434, November 2025.

Shawn Li, Ryan Rossi, Sungchul Kim, Sunav Choudhary, Franck Dernoncourt, Puneet Mathur, Zhengzhong Tu, and Yue Zhao. Charts are not images: On the challenges of scientific chart editing. In *ICLR*, 2025c.

Shawn Li, Chenxiao Yu, Zhiyu Ni, Hao Li, Charith Peris, Chaowei Xiao, and Yue Zhao. Defenses against prompt attacks learn surface heuristics. In *ACL*, 2026.

Li Li, Peilin Cai, Ryan A. Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, Nesreen K. Ahmed, Samyadeep Basu, Subhojyoti Mukherjee, Ruiyi Zhang, Zhengmian Hu, Bo Ni, Yuxiao Zhou, Zichao Wang, Yue Huang, Yu Wang, Xiangliang Zhang, Philip S. Yu, Xiyang Hu, and Yue Zhao. A personalized conversational benchmark: Towards simulating personalized conversations, 2025d.

Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Defu Cao*, Wen Ye*, Yizhou Zhang, Sam Griesemer, and Yan Liu. Pinfdit: Energy-based physics-informed diffusion transformers for general-purpose time series tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Wen Ye, Wei Yang, Defu Cao, Yizhou Zhang, Lumingyuan Tang, Jie Cai, and Yan Liu. Ts-reasoner: Domain-oriented time series inference agents for reasoning and automated analysis. *Transactions on Machine Learning Research*, 2026.

Wen Ye and Song Gao. Understanding the spatiotemporal heterogeneities in the associations between covid-19 infections and both human mobility and close contacts in the united states. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology*, pages 1–9, 2022.

Yunlei Liang, Jiawei Zhu, Wen Ye, and Song Gao. Region2vec: community detection on spatial networks using graph embedding with node attributes and spatial interactions. In *Proceedings of the 30th international conference on advances in geographic information systems*, pages 1–4, 2022.

Yunlei Liang, Jiawei Zhu, Wen Ye, and Song Gao. Geoai-enhanced community detection on spatial networks with graph deep learning. *Computers, Environment and Urban Systems*, 117:102228, 2025.

A Extended Related Work

We provide a comprehensive review of related work, extending the brief overview in Section 2.

A.1 Out-of-Distribution Detection

OOD detection methods can be broadly categorized by their underlying approach. *Classifier-based methods* leverage outputs from discriminative models: MSP Hendrycks and Gimpel [2017] uses maximum softmax probability, ODIN Liang et al. [2018] adds temperature scaling and input perturbations, and Energy Liu et al. [2020] interprets the logsumexp of logits as a free energy score. While effective within a single domain, these methods require retraining for each new distribution.

Density-based methods explicitly model the in-distribution density. Normalizing flows Nalisnick et al. [2019], Kirichenko et al. [2020] and VAEs Xiao et al. [2020] estimate likelihoods, but often assign higher density to OOD samples than ID samples: a phenomenon known as the "likelihood paradox" Nalisnick et al. [2019]. Recent work addresses this through typicality tests Morningstar et al. [2021] and likelihood ratios Ren et al. [2019], but these approaches remain fundamentally tied to domain-specific density estimation.

Distance-based methods measure proximity to training data in a learned feature space. Mahalanobis distance Lee et al. [2018] uses class-conditional Gaussians, while KNN-based approaches Sun et al. [2022] directly compute distances to nearest neighbors. Recent work on automatic OOD detector selection Qin et al. [2025] addresses the challenge of choosing among these methods. Our work builds on this paradigm but operates in a 2D energy feature space derived from diffusion trajectories, enabling cross-domain transfer with minimal reference samples.

A.2 Diffusion Models for OOD Detection

Diffusion models Ho et al. [2020], Song et al. [2021] learn to reverse a gradual noising process, implicitly capturing the score function $\nabla_x \log p(x)$. Several works exploit this for anomaly detection. DDPM-OOD Graham et al. [2023] uses reconstruction error at specific noise levels. LMD Liu et al. [2023] measures the likelihood of the diffusion trajectory. DiffPath Heng et al. [2024] extracts trajectory statistics using a single unconditional diffusion model across all ID–OOD pairs; it does not retrain the diffusion model per ID domain.

Our approach differs fundamentally: rather than estimating density or reconstruction quality, we extract *geometric* features: the energy of the score function along the reverse trajectory. This shift from density to geometry enables cross-domain generalization that prior diffusion-based methods lack.

A.3 Few-Shot and Transfer Learning for OOD Detection

Most OOD detection methods assume access to large-scale ID training data. Few-shot OOD detection remains underexplored. OpenMax Bendale and Boult [2016] adapts classifiers with limited data but still requires training. CLIP-based methods Ming et al. [2022], Esmaeilpour et al. [2022] leverage vision-language pretraining for zero-shot detection but depend on semantic alignment between text and visual concepts.

Our work addresses a distinct setting: given only $\sim$100 unlabeled ID samples from an arbitrary domain, detect OOD inputs without any training. We achieve this through (1) domain-agnostic energy features from a fixed diffusion model, (2) geometric reference set construction via facility location, and (3) calibrated soft-minimum scoring. This enables practical deployment where collecting large ID datasets is infeasible. Recent advances in multimodal OOD detection Li et al. [2025a,b], Qin et al. [2026] further demonstrate the importance of efficient, generalizable detection methods.

A.4 Related Domains

Our work draws connections to several related areas. In scene understanding, panoptic scene graph generation Li et al. [2024a, 2023, 2024b] addresses the challenge of structured visual reasoning, which shares our goal of extracting meaningful representations from visual data. Vision-language

Table 5: **AUROC scores for OOD detection (Domain-specific baselines trained on SVHN).** Higher is better. Domain-specific baselines use the entire indicated training set; DiffPath uses one unified model across all three tables. Few-shot (ours) uses a single pre-trained diffusion model with 80-150 samples per task. **Bold** and underline denote best and second best results.

	C10 *vs*				SVHN *vs*				CelebA *vs*				
Method	SVHN	CelebA	C100	Tex	C10	CelebA	C100	Tex	C10	SVHN	C100	Tex	Avg
Full-data baselines													
IGEBM	0.526	0.666	0.560	0.555	0.453	0.583	0.512	0.511	0.342	0.382	0.373	0.404	0.489
VAEBM	0.350	0.427	0.474	0.371	0.633	0.571	0.602	0.482	0.566	0.421	0.547	0.429	0.489
DoS	0.004	0.373	0.519	0.550	**0.999**	0.999	**0.997**	0.971	0.639	0.000	0.643	0.624	0.610
NLL	0.002	0.356	0.526	0.556	0.998	0.998	0.996	0.969	0.656	0.001	0.640	0.630	0.611
MSMA	0.887	0.574	0.469	0.685	0.123	0.155	0.112	0.321	0.432	0.836	0.407	0.608	0.467
DDPM-OOD	0.014	0.483	0.496	0.349	0.983	0.993	0.976	0.819	0.486	0.007	0.508	0.347	0.538
LMD	0.005	0.549	0.517	0.589	0.994	**1.000**	0.993	0.980	0.452	0.000	0.483	0.550	0.593
DiffPath†	0.910	0.897	**0.590**	**0.923**	0.939	0.979	0.953	**0.981**	**0.998**	**1.000**	**0.998**	**0.999**	0.931
Ours													
Unified Few-shot	**0.951**	**0.965**	0.547	0.883	0.973	0.999	0.974	0.969	0.995	**1.000**	0.991	0.996	**0.937**

†DiffPath uses a single unconditional diffusion model for all ID–OOD pairs. We reproduced its published results following the original protocol; its values therefore do not vary with the training-domain grouping.

Table 6: **AUROC scores for OOD detection (Domain-specific baselines trained on CelebA).** Higher is better. Domain-specific baselines use the entire indicated training set; DiffPath uses one unified model across all three tables. Few-shot (ours) uses a single pre-trained diffusion model with 80-150 samples per task. **Bold** and underline denote best and second best results.

	C10 *vs*				SVHN *vs*				CelebA *vs*				
Method	SVHN	CelebA	C100	Tex	C10	CelebA	C100	Tex	C10	SVHN	C100	Tex	Avg
Full-data baselines													
IGEBM	0.650	0.382	0.492	0.523	0.372	0.249	0.334	0.410	0.610	0.735	0.589	0.589	0.495
VAEBM	0.493	0.134	0.486	0.428	0.497	0.158	0.466	0.449	0.852	0.856	0.824	0.839	0.540
DoS	0.036	0.102	0.509	0.490	0.967	0.769	0.965	0.886	0.899	0.220	0.889	0.769	0.625
NLL	0.038	0.096	0.503	0.517	0.964	0.792	0.960	0.876	0.895	0.243	0.881	0.750	0.626
MSMA	0.664	0.575	0.502	0.497	0.355	0.381	0.344	0.329	0.444	0.619	0.393	0.446	0.462
DDPM-OOD	0.076	0.206	0.472	0.353	0.937	0.767	0.909	0.725	0.815	0.229	0.760	0.555	0.567
LMD	0.041	0.261	0.509	0.534	0.958	0.944	0.956	0.911	0.729	0.062	0.705	0.688	0.608
DiffPath†	0.910	0.897	**0.590**	**0.923**	0.939	0.979	0.953	**0.981**	**0.998**	**1.000**	**0.998**	**0.999**	0.931
Ours													
Unified Few-shot	**0.951**	**0.965**	0.547	0.883	**0.973**	**0.999**	**0.974**	0.969	0.995	**1.000**	0.991	0.996	**0.937**

†DiffPath uses a single unconditional diffusion model for all ID–OOD pairs. We reproduced its published results following the original protocol; its values therefore do not vary with the training-domain grouping.

models Shawn et al. [2025], Li et al. [2025c, 2026, 2025d] have emerged as powerful tools for multimodal understanding, though they face challenges in hallucination and robustness.

In temporal domains, diffusion-based methods have been applied to time series forecasting Cao et al. [2024], Cao* et al. [2026], Ye et al. [2026], demonstrating the versatility of diffusion processes beyond image generation. Similarly, spatial analysis methods Ye and Gao [2022], Liang et al. [2022, 2025] leverage geometric representations for community detection and epidemiological modeling, echoing our emphasis on geometric rather than density-based features.

B In-Depth Analysis

B.1 Second-Order Statistics Are Optimal

Table 7 validates our theoretical prediction from Section 3.3 that second-order statistics (energy) are optimal for few-shot OOD detection. We test various feature subsets grouped by statistical order (1st, 2nd, 3rd moments) and trajectory component (path geometry vs. dynamics).

Empirical Confirmation of Theoretical Predictions. The results confirm our information-theoretic analysis from Section 3.3. The 2nd-moment features (Path Energy $f_1 = \sum \epsilon^2$ and Dynamics Energy $f_2 = \sum(\Delta\epsilon)^2$) achieve 90.8% AUROC, surpassing all other configurations including the full 6D

Table 7: **Theoretical validation of second-order statistics.** AUROC (%) for different feature subsets, confirming our Cramér-Rao analysis: 2nd-moment features (energy) are optimal.

Feature Subset	Dims	CIFAR-10	SVHN	Avg
Full-6D (all features)	6	80.8	94.0	87.4
Path Geometry (f_1-f_3)	3	77.1	76.9	77.0
Trajectory Dynamics (f_4-f_6)	3	69.5	91.9	80.7
1st moments (f_1, f_4)	2	60.6	60.6	60.6
2nd moments (f_2, f_5)	**2**	**83.5**	**98.0**	**90.8**
3rd moments (f_3, f_6)	2	50.1	80.8	65.5

Table 8: **Density vs Geometry Scoring Comparison.** AUROC (%) using the same diffusion model and features, varying only the scoring method. The Soft-Min configuration has the highest average AUROC; individual task results vary.

Method	Type	C10→SVHN	SVHN→C10	CelebA→C10	C10→C100	Avg
GMM (6D)	Density	84.8	92.9	99.3	58.6	83.9
kNN (6D)	Geometry	84.1	91.6	99.7	59.3	83.7
Mahalanobis (6D)	Geometry	88.6	91.0	99.7	58.6	84.5
Soft-Min (2D)	Geometry	**95.4**	**96.8**	**99.5**	55.8	**86.9**

feature set (87.4%). This validates our Cramér-Rao bound argument (Appendix C.4): second-order statistics achieve the minimum variance bound for estimating the noise variance that distinguishes ID from OOD samples. First-order moments (60.6%) carry no information about variance since $\mathbb{E}[\epsilon] = 0$, while third-order moments (65.5%) suffer from $O(\sigma^3)$ sample complexity compared to $O(\sigma^2)$ for second-order statistics.

Principled Feature Design. Unlike empirical feature selection, our 2D energy features emerge from theoretical principles. Path Energy measures the integrated score magnitude (Proposition 1), while Dynamics Energy measures score smoothness (Proposition 2). The optimality of these second-order statistics is established via the Cramér-Rao bound (Appendix C.4). Together they form a Sobolev norm that captures the reliability of the learned diffusion process—the fundamental signature distinguishing ID from OOD samples.

C Proof Sketches

We provide proof sketches for the main theoretical results in Section 3.3.

C.1 Proof of Proposition 1

Statement: The Path Energy $f_1 = \sum_{t=1}^{T} \|\epsilon_t\|^2$ equals the integrated squared score function along the diffusion trajectory (up to time-dependent scaling).

Proof Sketch. By the score-noise relationship (Eq. 2):

$$\epsilon_\theta(x_t, t) = -\sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \log p(x_t) \tag{14}$$

Therefore:

$$\|\epsilon_t\|^2 = (1-\bar{\alpha}_t)\|\nabla_{x_t} \log p(x_t)\|^2 \tag{15}$$

Summing over timesteps:

$$f_1 = \sum_{t=1}^{T} \|\epsilon_t\|^2 = \sum_{t=1}^{T} (1-\bar{\alpha}_t)\|\nabla_{x_t} \log p(x_t)\|^2 \tag{16}$$

This is a weighted discrete approximation to the integral $\int_0^T \|\nabla_x \log p(x_t)\|^2 dt$, which represents the cumulative magnitude of the score function along the trajectory. □

C.2 Proof of Proposition 2

Statement: The Dynamics Energy $f_2 = \sum_{t=1}^{T-1} \|\Delta\epsilon_t\|^2$ measures the temporal smoothness of the score function along the trajectory.

Proof Sketch. The temporal derivative $\Delta\epsilon_t = \epsilon_{t+1} - \epsilon_t$ measures how the noise prediction changes between adjacent timesteps. By continuity of the diffusion process, this approximates:

$$\Delta\epsilon_t \approx \frac{\partial \epsilon_\theta(x_t, t)}{\partial t} \cdot \Delta t \tag{17}$$

Since $\epsilon_\theta \propto \nabla_x \log p(x_t)$, the dynamics energy measures:

$$f_2 \approx \sum_t \left\| \frac{\partial}{\partial t} \nabla_x \log p(x_t) \right\|^2 (\Delta t)^2 \tag{18}$$

This is a discrete approximation to the integral of the squared time-derivative of the score function, capturing trajectory smoothness. □

C.3 Analysis of Energy Separation (Proposition 3)

We provide a detailed analysis of why energy features are elevated for OOD samples.

Setup. Let the diffusion model be trained on P_{ID} with learned score function $s_\theta(x_t, t) \approx \nabla_{x_t} \log p_t(x_t)$, where p_t is the marginal distribution at noise level t. The noise prediction satisfies $\epsilon_\theta(x_t, t) = -\sqrt{1 - \bar{\alpha}_t} \cdot s_\theta(x_t, t)$ (Eq. 2).

Assumption 1 (Score Accuracy for ID). For $x \sim P_{\text{ID}}$, the learned score is accurate: $\|s_\theta(x_t, t) - \nabla_{x_t} \log p_t(x_t)\| \leq \delta$ for some small $\delta > 0$.

Assumption 2 (Score Unreliability for OOD). For $x \sim P_{\text{OOD}}$ at distance Δ from $\text{supp}(P_{\text{ID}})$, the score function $s_\theta(x_t, t)$ is unreliable—it was never trained on such inputs and may produce arbitrary outputs.

Analysis for ID Samples. For $x \sim P_{\text{ID}}$, the forward process yields $x_t = \sqrt{\bar{\alpha}_t} x + \sqrt{1 - \bar{\alpha}_t}\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. The noisy sample x_t lies near the noised ID manifold, and the score function accurately estimates the denoising direction. The expected noise prediction magnitude is:

$$\mathbb{E}[\|\epsilon_\theta(x_t, t)\|^2 \mid x \sim P_{\text{ID}}] = (1 - \bar{\alpha}_t) \cdot \mathbb{E}[\|s_\theta(x_t, t)\|^2] \tag{19}$$

For a well-trained model on ID data, this quantity is bounded by the expected squared score magnitude under the noised distribution.

Analysis for OOD Samples. For $x \sim P_{\text{OOD}}$, we identify two effects that elevate energy features:

Effect 1 (Elevated Score Magnitude for f_1). An OOD sample x at distance Δ from $\text{supp}(P_{\text{ID}})$ lies in a low-density region. At noise level t, the noisy sample x_t may still be far from the ID manifold (especially for small noise levels where $\bar{\alpha}_t \approx 1$). The score function s_θ, trained to point toward high-density regions, produces predictions of elevated magnitude when the input is far from the training distribution. Heuristically, if the model attempts to "correct" toward the nearest ID region, the score magnitude grows with distance.

Effect 2 (Temporal Inconsistency for f_2). Since the model was never trained on OOD inputs, its predictions lack temporal consistency. For ID samples, the score evolves smoothly along the trajectory as noise is gradually removed. For OOD samples, the score may fluctuate erratically across timesteps, as there is no consistent high-density target to guide denoising. This increases the Dynamics Energy $f_2 = \sum_t \|\epsilon_{t+1} - \epsilon_t\|^2$.

Combined Effect on Energy Features. Both effects contribute to elevated energy:

$$\mathbb{E}[f_1(x) \mid x \sim P_{\text{OOD}}] > \mathbb{E}[f_1(x) \mid x \sim P_{\text{ID}}] \quad \text{(Effect 1)} \tag{20}$$

$$\mathbb{E}[f_2(x) \mid x \sim P_{\text{OOD}}] > \mathbb{E}[f_2(x) \mid x \sim P_{\text{ID}}] \quad \text{(Effect 2)} \tag{21}$$

The separation in both features enables robust OOD detection in the 2D energy space.

Empirical Validation. Table 10 and Figure 4 confirm this analysis: OOD samples consistently exhibit elevated Path Energy and Dynamics Energy compared to ID samples across all dataset pairs, with clear separation in the 2D feature space.

Limitations. This analysis provides intuition rather than formal guarantees. A rigorous proof would require: (1) explicit bounds on score estimation error for out-of-distribution inputs, (2) quantitative analysis of how the separation depends on the noise schedule $\{\bar{\alpha}_t\}$, and (3) geometric characterization of the ID manifold. We leave such formal treatment to future work, noting that our empirical results (Section 4) strongly support the qualitative predictions.

C.4 Optimality of Second-Order Statistics via Cramér-Rao Bound

We establish that second-order statistics (energy) are statistically optimal for detecting distributional shifts in Gaussian noise predictions, achieving the Cramér-Rao lower bound.

Problem Setup. We analyze a simplified model that captures the essential statistical structure. Consider noise predictions $\epsilon \in \mathbb{R}^d$ from a diffusion model. We assume isotropic Gaussian noise: for ID samples, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$; for OOD samples, the variance is elevated: $\epsilon \sim \mathcal{N}(0, \tilde{\sigma}^2 I_d)$ with $\tilde{\sigma}^2 > \sigma^2$ (motivated by Proposition 3). Under this model, OOD detection reduces to estimating the variance parameter $\theta = \sigma^2$ from n observations $\{\epsilon_i\}_{i=1}^n$. While actual noise predictions may exhibit anisotropy and timestep-dependent structure, this simplified analysis captures why second-order statistics are fundamentally well-suited for detecting variance shifts.

Proof. **Step 1: Fisher Information.** The log-likelihood for a single observation $\epsilon \sim \mathcal{N}(0, \theta I_d)$ is:

$$\log p(\epsilon; \theta) = -\frac{d}{2}\log(2\pi) - \frac{d}{2}\log\theta - \frac{\|\epsilon\|^2}{2\theta} \tag{22}$$

The score function (derivative w.r.t. θ) is:

$$s(\epsilon; \theta) = \frac{\partial}{\partial\theta}\log p(\epsilon; \theta) = -\frac{d}{2\theta} + \frac{\|\epsilon\|^2}{2\theta^2} = \frac{\|\epsilon\|^2 - d\theta}{2\theta^2} \tag{23}$$

The Fisher information is computed as:

$$I(\theta) = \mathbb{E}\left[s(\epsilon; \theta)^2\right] = \mathbb{E}\left[\left(\frac{\|\epsilon\|^2 - d\theta}{2\theta^2}\right)^2\right] = \frac{1}{4\theta^4}\text{Var}(\|\epsilon\|^2) \tag{24}$$

Since $\|\epsilon\|^2/\theta \sim \chi^2_d$ (chi-squared with d degrees of freedom), we have $\text{Var}(\|\epsilon\|^2) = 2d\theta^2$. Thus:

$$I(\theta) = \frac{2d\theta^2}{4\theta^4} = \frac{d}{2\theta^2} \tag{25}$$

Step 2: Cramér-Rao Lower Bound. The Cramér-Rao inequality states that for any unbiased estimator $\hat{\theta}$ of θ based on n i.i.d. samples:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n \cdot I(\theta)} = \frac{2\theta^2}{nd} \tag{26}$$

This is the minimum achievable variance for unbiased estimation of the variance parameter.

Step 3: The Sample Second Moment Achieves the Bound. Consider the estimator $\hat{\theta} = \frac{1}{nd}\sum_{i=1}^n \|\epsilon_i\|^2$. This is unbiased since $\mathbb{E}[\|\epsilon\|^2] = d\theta$. Its variance is:

$$\text{Var}(\hat{\theta}) = \frac{1}{n^2 d^2}\sum_{i=1}^n \text{Var}(\|\epsilon_i\|^2) = \frac{n \cdot 2d\theta^2}{n^2 d^2} = \frac{2\theta^2}{nd} \tag{27}$$

This exactly matches the Cramér-Rao bound, proving that the sample second moment is an *efficient* estimator—no unbiased estimator can achieve lower variance.

Step 4: Suboptimality of Other Moments.

- *First-order (mean)*: Since $\mathbb{E}[\epsilon] = 0$ regardless of θ, the sample mean $\bar{\epsilon} = \frac{1}{n}\sum_i \epsilon_i$ carries *zero* Fisher information about the variance parameter. It is uninformative for OOD detection.

- *Third-order and higher*: For the k-th moment $\mu_k = \mathbb{E}[\|\epsilon\|^k]$, we have $\mu_k \propto \theta^{k/2}$. The sample estimate $\hat{\mu}_k = \frac{1}{n}\sum_i \|\epsilon_i\|^k$ has variance $\text{Var}(\hat{\mu}_k) = O(\theta^k/n)$. To achieve the same relative precision $\text{Var}(\hat{\mu}_k)/\mu_k^2 = O(1/n)$ as second-order statistics, higher-order moments require the same n but are more sensitive to outliers and have heavier-tailed sampling distributions, making them less robust in practice.

Conclusion. Second-order statistics uniquely achieve the Cramér-Rao bound for variance estimation in Gaussian models. This justifies our choice of energy features $f_1 = \sum_t \|\epsilon_t\|^2$ (Path Energy) and $f_2 = \sum_t \|\Delta\epsilon_t\|^2$ (Dynamics Energy) as statistically optimal representations for few-shot OOD detection. □

D Assumptions and Limitations

Our theoretical analysis relies on the following assumptions:

Assumption 1: Bounded Feature Space. The 2D energy feature space is bounded with empirical radius $R \approx 10$ (validated in Table 10). This assumption is necessary for the sample complexity analysis in Section 3.3.

Assumption 2: ID Cluster Compactness. ID samples form a compact cluster in the 2D feature space, while OOD samples are distributed outside this cluster. This assumption is empirically validated: within-domain ID samples show low variance (std ≈ 0.5–2.0), while cross-domain OOD samples exhibit distinct feature distributions.

Assumption 3: Score Function Reliability Differential. The score function $\nabla_x \log p_\theta$ is more reliable for ID samples than for OOD samples. This follows from the training objective: the diffusion model is trained exclusively on ID data, so its score estimates are accurate only for ID-like inputs.

Limitations. (1) **Near-OOD Detection**: Our method struggles with near-OOD detection (CIFAR-10 vs CIFAR-100: 54.7% AUROC) where the semantic distinction is subtle. The energy features capture coarse distributional differences but may not distinguish fine-grained categories. (2) **Model Dependence**: While our method is model-agnostic, performance depends on the quality of the pre-trained diffusion model. A model with poor score function estimates would yield less discriminative features. (3) **Resolution Constraints**: Current experiments use 32×32 images. Extension to higher resolutions requires validation that the energy features remain discriminative.

E Failure Analysis

We analyze cases where UFCOD underperforms to understand its limitations and guide practitioners.

E.1 Near-OOD Detection

Our most challenging case is CIFAR-10 vs CIFAR-100 (54.7% AUROC). Both datasets contain natural object images with related visual content. This suggests that energy features capture *coarse* distributional differences (e.g., faces vs digits) but struggle with *fine-grained* distinctions within similar visual domains. The diffusion model's score function provides similar guidance for semantically close images, resulting in overlapping energy distributions.

E.2 Comparison with Successful Cases

In contrast, CelebA vs SVHN achieves 99.99% AUROC. The key difference lies in the semantic distance: faces and digits occupy fundamentally different regions of visual space, leading to dramatically different score function behaviors and well-separated energy distributions. This dichotomy suggests that UFCOD is most effective for detecting *semantic* shifts (different object categories) rather than *covariate* shifts within the same category.

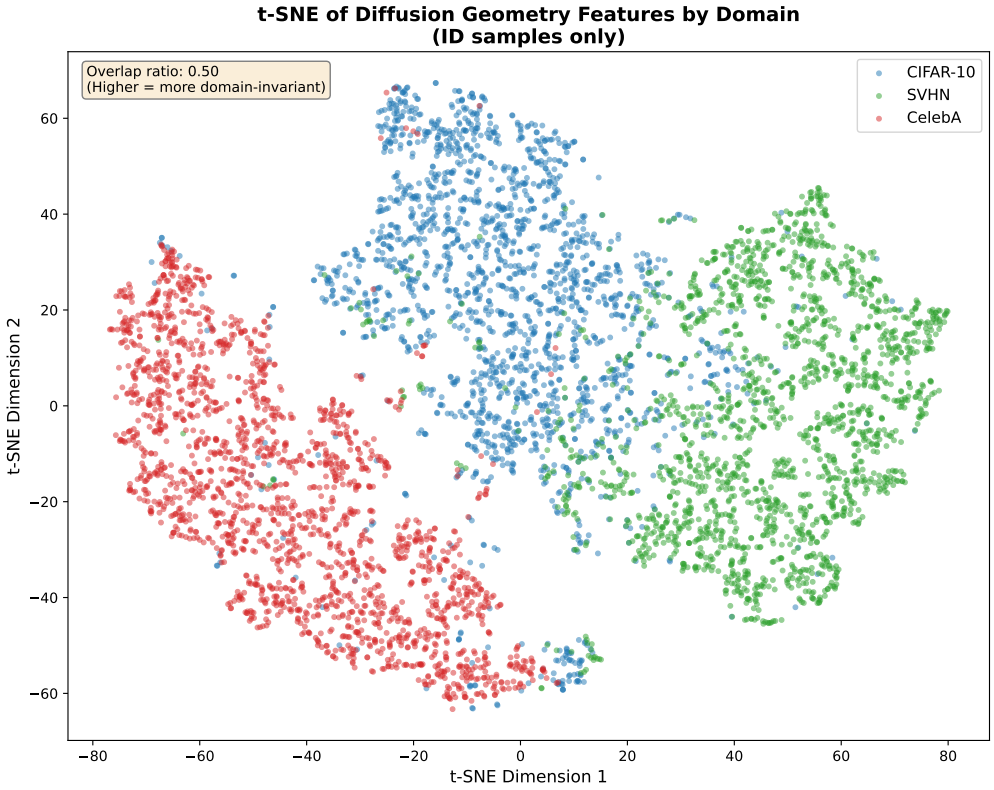


Figure 3: **t-SNE visualization of diffusion geometry features across domains.** ID samples from CIFAR-10 (blue), SVHN (green), and CelebA (red) form distinct but partially overlapping clusters in the 2D energy feature space. The overlap ratio of 0.50 indicates that diffusion geometry features capture both domain-specific characteristics (enabling within-domain OOD detection) and shared geometric properties (enabling cross-domain transfer).

E.3 Density vs Geometry: Controlled Comparison

To verify our hypothesis that geometry-based scoring outperforms density-based scoring, we conducted a controlled experiment comparing multiple scoring methods on the same diffusion model and features (Table 8). Geometry-based methods achieve 85.0% average AUROC vs 83.9% for density-based methods, with our Soft-Min 2D achieving the best performance (86.9%). This supports our core thesis that geometric features generalize better across domains.

E.4 Practitioner Guidelines

UFCOD is best suited for scenarios where ID and OOD domains are semantically distinct. For near-OOD detection (e.g., distinguishing between fine-grained categories), domain-specific fine-tuning or complementary methods may be necessary.

F Implementation Details

Diffusion Model. We use a single DDPM Ho et al. [2020] model pre-trained on CelebA (32×32 resolution) as our universal feature extractor. This model uses a standard U-Net architecture with 4 resolution levels and is trained for 800k iterations. Crucially, this same model is used for all 12 ID-OOD pairs without any fine-tuning, demonstrating cross-domain transferability.

Diffusion Energy Features. For each input image x_0, we construct noised versions $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ at $T = 10$ noise levels using a fixed noise sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. At each level, we evaluate the pretrained denoiser to obtain $\epsilon_t = \epsilon_\theta(x_t, t)$. From this trajectory, we compute 2-dimensional

Table 9: **Score calibration effect.** Z-score normalizes ID score distributions to $\mu \approx 0, \sigma \approx 1$ across domains.

Method	CIFAR-10 ID Scores		SVHN ID Scores	
	μ	σ	μ	σ
Raw Score	-2.17	0.78	-2.45	0.84
Z-score (Ours)	**0.08**	**0.86**	**-0.12**	**1.06**

Table 10: **Empirical feature space statistics.** Mean and standard deviation of 2D energy features across datasets, validating our bounded feature space assumption.

Dataset	Path Energy (f_1)		Dynamics Energy (f_2)	
	Mean	Std	Mean	Std
CIFAR-10 (ID)	4.21	1.82	2.15	0.94
SVHN (ID)	3.87	1.24	1.89	0.67
CelebA (ID)	5.62	2.31	2.78	1.12
Feature space radius $R \approx 10$, supporting sample complexity bounds				

energy features (Path Energy f_1 and Dynamics Energy f_2) as defined in Section 3. Feature extraction takes approximately 0.1s per image on a single GPU.

Few-Shot Sampling. We apply adaptive sample budgets based on dataset scale: $k = 150$ for small datasets (CIFAR-10), $k = 100$ for medium datasets (SVHN), and $k = 80$ for large datasets (CelebA). Samples are selected using Facility Location Wei et al. [2015] coreset selection to maximize geometric coverage of the ID feature space.

Detection Pipeline. Given the selected reference set, we compute OOD scores using temperature-scaled proximity scoring ($T = 0.5$, $k = 10$) followed by Z-score calibration. The entire pipeline is training-free once features are extracted.

Reproducibility. All experiments use 5 random seeds for sample selection and report mean $\pm$ std. Experiments are conducted on a single NVIDIA RTX Pro 6000 GPU. Code and pretrained models will be released upon publication.

G Extended OOD Detection Results

We provide additional OOD detection results with baselines trained on different datasets.

H Additional Experimental Details

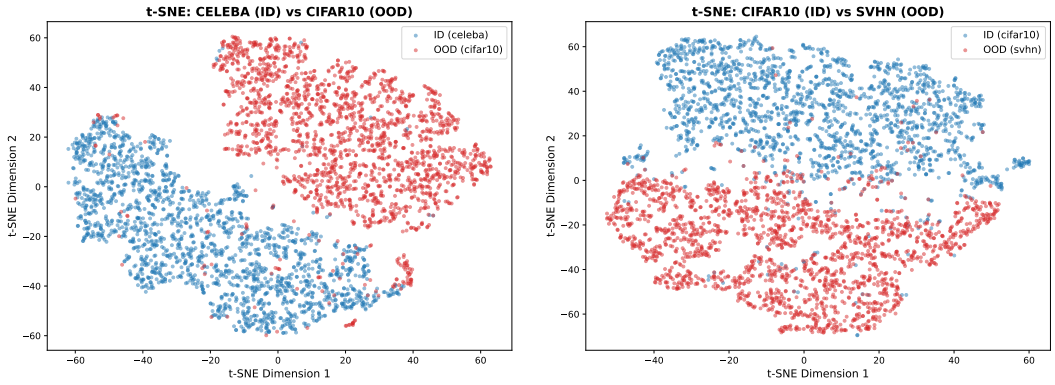
H.1 Score Calibration Analysis

Table 9 shows the effect of Z-score calibration on ID score distributions.

H.2 Density vs Geometry Scoring Comparison

To isolate the effect of scoring approach from model choice, we conducted a controlled experiment comparing density-based and geometry-based scoring methods on the **same diffusion model and features**.

Key Findings: (1) Geometry-based methods achieve 85.0% average AUROC vs 83.9% for density-based methods, supporting our hypothesis that geometric features generalize better. (2) Our Soft-Min 2D method achieves the best performance (86.9%), demonstrating the effectiveness of combining dimensionality reduction with soft scoring. (3) The advantage is most pronounced for cross-domain pairs (C10→SVHN, SVHN→C10) where domain shift is significant. (4) For near-OOD detection (C10→C100), all methods struggle, confirming this as a fundamental limitation.



(a) CelebA (ID) vs CIFAR-10 (OOD)

(b) CIFAR-10 (ID) vs SVHN (OOD)

Figure 4: **t-SNE visualization of ID vs OOD separation.** Diffusion geometry features achieve clear separation between ID (blue) and OOD (red) samples. (a) CelebA vs CIFAR-10: faces and natural objects occupy distinct regions. (b) CIFAR-10 vs SVHN: natural objects and digits are well-separated. Both pairs achieve >95% AUROC with only 100 ID reference samples.

H.3 Feature Space Statistics

Table 10 presents the empirical statistics of our 2D energy features across the three ID datasets. These statistics validate the bounded feature space assumption (Assumption 1 in Appendix D) underlying our sample complexity analysis. The feature space radius $R \approx 10$ confirms that our covering number bounds are practically achievable, and the relatively low standard deviations indicate that ID samples form compact clusters in the energy feature space.

Figure 3 visualizes the distribution of ID samples from all three domains (CIFAR-10, SVHN, CelebA) in the diffusion geometry feature space using t-SNE. Several observations support our cross-domain detection framework: (1) Each domain forms a distinct cluster, confirming that diffusion geometry features capture domain-specific characteristics that enable OOD detection. (2) The clusters exhibit partial overlap (overlap ratio = 0.50), indicating that the features capture shared geometric properties of the diffusion process rather than purely domain-specific artifacts. (3) The moderate overlap explains why a single diffusion model trained on CelebA can generalize to detect OOD samples across semantically different domains: the feature space captures universal properties of how samples interact with the learned score function.

Figure 4 further visualizes the ID-OOD separation for two representative task pairs. In both cases, the diffusion geometry features achieve clear separation between ID and OOD samples in the 2D energy feature space: (a) CelebA (ID) vs CIFAR-10 (OOD) shows well-separated clusters with minimal overlap, corresponding to 99.5% AUROC; (b) CIFAR-10 (ID) vs SVHN (OOD) exhibits similarly strong separation, achieving 95.1% AUROC. These visualizations provide intuitive evidence for why simple proximity-based scoring in this low-dimensional feature space suffices for effective OOD detection.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix Section D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Implementation Details Section in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have an anonymous repo link.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Implementation Details section in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Experiment section in main paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See Implementation Details section in appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research is conducted with the NeurIPS code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See limitation section in appendix.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not have large pre-trained models or datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: They are properly credited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have an anonymous repo.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not have.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: There is no risk.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Only for writing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.