Beyond Empirical Support: Structured Outlier Generation via Sinkhorn Optimal Transport

Haixiang Sun, Andrew L. Liu
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN, 47906
{sun1321, andrewliu}@purdue.edu

Abstract

Outliers are essential for evaluating and improving the robustness of machine learning systems, especially when future distributions may differ significantly from historical training data. In high-stakes applications, robustness often depends on rare cases that finite datasets fail to capture, making simple resampling or perturbation insufficient for stress scenario generation. Existing outlier synthesis methods typically rely on sparse neighborhoods, low support latent regions, or classifier boundary crossings, which can be heuristic, unstable, and tied to specific modalities or architectures. We therefore propose Sinkhorn Boundary Outlier Generation (SBOG), a structured framework for latent-space outlier generation that couples Sinkhorn optimal transport geometry with distributionally robust boundary modeling. The resulting Sinkhorn-induced support cost guides the sampler toward weakly supported boundary regions, while semantic constraints prevent uncontrolled drift from the intended context, yielding controlled deviations from the in-distribution reference measure rather than arbitrary sparse-region samples. Experiments on time series anomaly generation and image outlier synthesis show that our framework produces informative, semantically controlled outliers and improves downstream robustness evaluation across modalities, providing a foundation for stress scenario generation beyond empirical support.

1 Introduction

Outliers play an important role in building reliable machine learning systems. In many real-world applications, rare or atypical cases are often precisely the ones that determine whether a model behaves robustly under deployment. Yet such outliers are inherently difficult to study since they occur infrequently, can vary substantially in form, and therefore are poorly covered by finite training datasets. Consequently, most existing work treats outliers primarily as anomalies to be detected, monitored, and mitigated [45, 38]. Although this line of work is indispensable, it leaves open a complementary question of equal importance: how to model and generate outliers themselves. Addressing this question is important for understanding the structure and mechanisms underlying rare events. It also enables systematic construction of informative training and evaluation cases beyond standard data collection, which can strengthen model robustness, reveal failure modes before deployment, and support safer decision-making under uncertainty.

However, dedicated work on outlier generation remains relatively scarce. Existing approaches typically first learn a representation space and then identify candidate boundary samples using empirical support criteria, such as sparse nearest-neighbor neighborhoods [57], low-likelihood latent regions [13], classifier-boundary crossing rules [36], or LLM-based heuristics [1]. These candidates are then perturbed, sampled, or decoded back to the input space for synthetic outliers. Although effective in practice, such criteria often treat low support as the primary signal of outlyingness.

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

Low support alone, however, is not sufficient for useful outlier generation. A point may be sparse because it lies near an informative class boundary, but it may also be sparse because it drifts off the semantic manifold, moves toward another class, or becomes difficult for the generator to decode into a meaningful sample. Therefore, for outlier generation, the relevant objective is not merely to find low-density feature points, but to find weakly supported candidates that remain semantically anchored and generatively valid. Moreover, existing synthesis pipelines are often tailored to a specific modality or generator, with many methods designed primarily for time-series anomaly generation [10, 26, 9] or image outlier synthesis [13, 17, 66]. As a result, these modality-specific designs may not generalize well when the underlying data distribution or modality changes.

Motivated by this perspective, we use Sinkhorn optimal transport [8] as a support-aware geometry for structured outlier generation in latent space. Given an in-distribution reference measure and a ground cost induced by the learned representation, the entropic transport geometry provides a smooth way to measure how much in-distribution mass can support a candidate latent point. This support cost guides generation toward weakly supported boundary regions, while semantic constraints ensure that the generated candidates remain aligned with the intended context rather than drifting into arbitrary sparse regions. Since the construction only requires a latent reference measure and a ground cost, the same principle can be instantiated across different underlying distributions and modalities, including time series and images, without relying on modality-specific neighborhood rules.

Based on this principle, we propose Sinkhorn Boundary Outlier Generation (SBOG), a structured boundary-generation framework that combines transport-based support scoring, boundary-anchor selection and semantic feasibility. The key component of SBOG is Sinkhorn Outlier Energy (SOE), which measures the transport cost of supporting a candidate point from the in-distribution reference measure. Notably, under the empirical Euclidean case, SOE recovers a smooth kernel-support form closely related to Gaussian Kernel Density Estimation (KDE). However, here SOE is not to introduce a standalone density estimator. Rather, SBOG uses SOE as a transport-based boundary score: SOE identifies smooth boundary regions for anchor selection and candidate acceptance, while semantic constraints keep the generated outliers controlled and class-consistent. This use differs from purely statistical sparsity or density-rejection rules, which often yield unstable low-support scores.

Beyond this geometric view, the corresponding Sinkhorn Distributionally Robust Optimization (DRO) formulation provides an optimization-consistent interpretation of the boundary decisions made by SBOG [30]. The one-dimensional dual [61] exposes the same Gibbs normalizer that appears in SOE, linking high-energy candidates to deviations that are expensive to support under the transport uncertainty geometry. We use this connection to justify SBOG as a model-agnostic boundary-generation procedure: SOE selects transport-expensive candidates, while semantic constraints restrict them to meaningful class-consistent deviations. When a downstream loss is available, the same dual can be used to derive task-aware variants. As a result, SBOG provides a unified sampling framework that can be evaluated through both generation-quality diagnostics and downstream robustness performance. Our main contributions are summarized as follows:

- We formulate structured outlier generation as constrained latent boundary generation, where candidates must be weakly supported by the in-distribution reference measure while remaining semantically anchored and generatively valid. This reframes outlier synthesis beyond low-density or sparse-neighborhood search.
- We develop Sinkhorn Boundary Outlier Generation (SBOG), an algorithm that uses Sinkhorn Outlier Energy (SOE) to select boundary anchors, accept weakly supported candidates under the Sinkhorn transport geometry, enforce semantic feasibility, and decode synthetic outliers. While SOE reduces to a kernel-support score in the empirical Euclidean case, SBOG uses it as a transport-based boundary criterion rather than as a standalone density score.
- We show that the SOE boundary rule appears as Gibbs-normalization term in the Sinkhorn-DRO dual, linking SBOG's thresholding and energy-maximizing candidate selection to boundary decisions. Across time-series and image benchmarks, SBOG generates informative, semantically controlled outliers and improves downstream robustness evaluation.

2 Preliminaries

2.1 Sinkhorn Distributionally Robust Optimization

We begin by recalling the Sinkhorn Distributionally Robust Optimization framework [30], which provides the transport geometry used later to define outlyingness in latent space. Let $\mathcal{X}$ be a

measurable space, $\mu \in \mathcal{P}(\mathcal{X})$ denote a nominal distribution, $\nu \in \mathcal{P}(\mathcal{X})$ be a reference measure for admissible perturbations, and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a ground cost. For $\varepsilon > 0$, define $K_\nu(x) := \int_{\mathcal{X}} \exp\Big(-\frac{d(x,\tilde{x})}{\varepsilon}\Big)\, d\nu(\tilde{x})$, and assume $0 < K_\nu(x) < \infty$ for μ-almost every x. The associated Gibbs kernel is

$$dQ^{\nu}_{\varepsilon,x}(\tilde{x}) = \frac{\exp\big(-d(x,\tilde{x})/\varepsilon\big)}{K_\nu(x)} d\nu(\tilde{x}). \tag{1}$$

Definition 2.1 (Sinkhorn OT)**.** *For any candidate worst-case distribution $P \in \mathcal{P}(\mathcal{X})$, define the normalized Sinkhorn Optimal Transport (OT) discrepancy*

$$W^{\nu}_{\varepsilon}(\mu, P) = \inf_{\gamma \in \Gamma(\mu,P)} \big\{ \mathbb{E}_{(x,\tilde{x})\sim\gamma}\big[d(x,\tilde{x})\big] + \varepsilon \mathrm{KL}(\gamma \| \mu \otimes \nu) + \varepsilon \mathbb{E}_{x\sim\mu}\big[\log K_\nu(x)\big] \big\}, \tag{2}$$

where $\Gamma(\mu, P)$ is the set of couplings with first marginal μ and second marginal P, and $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence. Equivalently,

$$W^{\nu}_{\varepsilon}(\mu, P) = \inf_{\gamma \in \Gamma(\mu,P)} \varepsilon \mathrm{KL}(\gamma \| \mu \otimes Q^{\nu}_{\varepsilon}),$$

where $(\mu \otimes Q^{\nu}_{\varepsilon})(dx, d\tilde{x}) := \mu(dx) Q^{\nu}_{\varepsilon,x}(d\tilde{x})$. The normalization by $K_\nu(x)$ makes the minimum discrepancy over all candidate marginals equal to zero.

Definition 2.2 (Sinkhorn DRO)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be measurable and let $\rho > 0$. The Sinkhorn ambiguity set is the normalized entropic OT ball $\mathcal{U}^{\nu}_{\rho}(\mu) = \{P \in \mathcal{P}(\mathcal{X}) : W^{\nu}_{\varepsilon}(\mu, P) \le \rho\}$,and the corresponding worst-case expected value is*

$$V_\rho(f; \mu, \nu) = \sup_{P \in \mathcal{U}^{\nu}_{\rho}(\mu)} \mathbb{E}_{\tilde{x}\sim P}\big[f(\tilde{x})\big]. \tag{3}$$

By adapting the strong-duality result of [61, Theorem 1], the Sinkhorn DRO can be simplified as:

Lemma 2.3 (One-dimensional Sinkhorn-DRO dual)**.** *Assume the ambiguity set is nonempty and the exponential moments below are finite. By strong duality, $V_\rho(f; \mu, \nu)$ admits the one-dimensional dual representation $V_\rho(f; \mu, \nu) = \inf_{\lambda>0} V_D(\lambda)$, with*

$$V_D(\lambda) = \lambda\rho + \lambda\varepsilon\, \mathbb{E}_{x\sim\mu}\left[\log \mathbb{E}_{\tilde{x}\sim Q^{\nu}_{\varepsilon,x}} \exp\left(\frac{f(\tilde{x})}{\lambda\varepsilon}\right)\right]. \tag{4}$$

The kernel $Q^{\nu}_{\varepsilon,x}$ concentrates on low-cost perturbations of x, while the log-moment generating term in (4) exponentially tilts toward large values of f, yielding a softmax-type robustification governed by (ρ, ε) and the dual temperature λ.

2.2 Outlier Generation

Let $\mathcal{X}$ denote the input space, and let $D_{\text{in}} = \{(x_i, y_i)\}_{i=1}^n$ with $y_i \in \{1, \dots, C\}$ be an in-distribution dataset. Let $T : \{1, \dots, C\} \to \mathbb{R}^m$ be a fixed conditioning map, with $t_c := T(c)$. Let $h_\theta : \mathcal{X} \to \mathbb{R}^m$ be a trainable encoder and define the normalized representation $z(x; \theta) := \frac{h_\theta(x)}{\|h_\theta(x)\|_2} \in \mathcal{Z} \subseteq \mathbb{R}^m$. We write $z_i := z(x_i; \theta)$ and define the empirical In-Distriubtion (ID) embedding set $Z_n := \{z_i\}_{i=1}^n \subset \mathcal{Z}$. The first stage learns θ so that each z_i is aligned with its corresponding conditioning vector t_{y_i}. The empirical latent reference measure is $\hat{\nu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$. Outlier generation can therefore be viewed as the problem of generating new latent points relative to the geometry induced by the embedding set.

Given the learned latent space $\mathcal{Z}$ and the empirical ID set Z_n, let $\mathcal{I}(Z_n) \subseteq \mathcal{Z}$ denote the region regarded as in-distribution. An outlier sampler is a conditional distribution $S(\cdot \mid Z_n)$ over $\mathcal{Z}$ designed to place most of its mass outside $\mathcal{I}(Z_n)$. It produces latent candidates v at least approximately. Each sampled latent point v is then mapped back to the input space through a fixed conditional generator $x_{\text{ood}} \sim p(x \mid v)$, where OOD denotes out-of-distribution.. Repeating this procedure yields an auxiliary set $\mathcal{D}_{\text{ood}} = \{x_{\text{ood}}\}$.

3 SBOG: Sinkhorn Boundary Outlier Generation

3.1 Outlier Sampling under Latent Space

We instantiate the sampler $S(\cdot \mid Z_n)$ as Sinkhorn Boundary Outlier Generation (SBOG), a latent-space procedure for generating structured outliers. We work in the latent representation space $\mathcal{Z}$ and

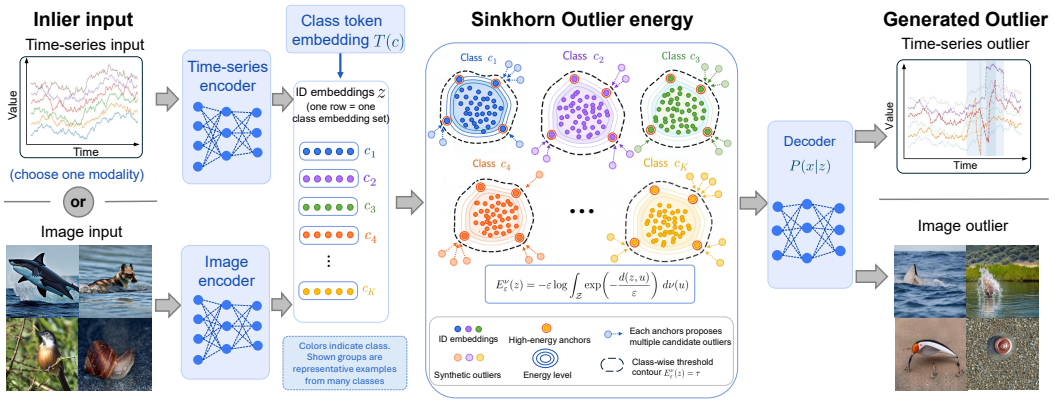

Figure 1: Pipeline of Sinkhorn Boundary Outlier Generation (SBOG).

generate synthetic outliers $v \in \mathcal{Z}$. This is reasonable because alignment-trained encoders tend to organize representations semantically: samples from the same class concentrate around their class prototype, while different classes become separated in latent space [43]. We therefore assume that the learned representation space $\mathcal{Z}$ is sufficiently aligned so that the ground cost $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ reflects local semantic geometry. Under this geometry, SBOG does not define outlyingness through a purely statistical proxy such as nearest-neighbor sparsity or classifier-boundary crossing. Instead, it measures how expensive it is to support a candidate latent point using mass from the in-distribution reference measure under an entropic transport geometry. This gives a smooth boundary-support criterion that can be instantiated across different underlying distributions and modalities.

The key support component of SBOG is a transport-based energy, which we introduce first. Let $\nu \in \mathcal{P}(\mathcal{Z})$ denote the in-distribution reference measure in the latent representation space, for a candidate latent point $z \in \mathcal{Z}$, we define its entropic Sinkhorn support as

$$K_\varepsilon^\nu(z) := \int_{\mathcal{Z}} \exp\left(-\frac{d(z,u)}{\varepsilon}\right) d\nu(u), \qquad z \in \mathcal{Z}. \tag{5}$$

This quantity aggregates the amount of in-distribution mass that can support z, where nearby mass contributes more strongly and the temperature ε controls the softness of the transport neighborhood. A large value of $K_\varepsilon^\nu(z)$ indicates that z is well supported by the reference measure, whereas a small value indicates weak support under the Sinkhorn transport geometry.

Definition 3.1 (Sinkhorn Outlier Energy)**.** *Let $\nu \in \mathcal{P}(\mathcal{Z})$ be a reference measure on the latent representation space $\mathcal{Z}$, let $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be a ground cost, and let $\varepsilon > 0$. For any $z \in \mathcal{Z}$ with $0 < K_\varepsilon^\nu(z) < \infty$, we define the Sinkhorn outlier energy of z as*

$$E_\varepsilon^\nu(z) := -\varepsilon \log K_\varepsilon^\nu(z). \tag{6}$$

Sinkhorn Outlier Energy (SOE) is the transport-support energy used inside SBOG: larger values indicate candidate points that are more expensive to support from the in-distribution reference measure under the chosen ground geometry. SBOG uses SOE to define boundary contours for anchor selection and candidate acceptance, and its relation to empirical kernel support is clarified in Lemma 3.2.

Lemma 3.2 (Entropic support cost of Sinkhorn outlier energy)**.** *For any $z \in Z$ such that $0 < K_\epsilon^\nu(z) < \infty$, given the Sinkhorn outlier energy $E_\epsilon^\nu(z)$, in particular, for the empirical reference measure $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ with distinct atoms, we have*

$$E_\epsilon^{\hat{\nu}_n}(z) = \min_{w \in \Delta_n} \left\{ \sum_{i=1}^n w_i d(z, z_i) + \epsilon \sum_{i=1}^n w_i \log(n w_i) \right\},$$

where $\Delta_n = \{w \in \mathbb{R}_+^n : \sum_{i=1}^n w_i = 1\}$ with the minimizer being $w_i^\star(z) = \frac{\exp(-d(z,z_i)/\epsilon)}{\sum_{j=1}^n \exp(-d(z,z_j)/\epsilon)}$.

Relation to kernel support. Lemma 3.2 shows that, for the empirical reference measure $\hat{\nu}n = \frac{1}{n}\sum i = 1^n \delta_{z_i}$, SOE admits a kernel-support form. Only in the special case of uniform empirical support with squared Euclidean cost does SOE reduce to a monotone transform of an unnormalized RBF-KDE score. Beyond this restricted setting, SOE is not equivalent to KDE, and our experiments

further show that this transport-based boundary criterion yields stronger outlier anchors than KDE-based alternatives. In SBOG, we use this connection to reinterpret kernel support as an entropic transport support cost: $E_\epsilon^\nu(z) = -\epsilon \log K_\epsilon^\nu(z)$ measures how costly it is, under the chosen ground geometry, for the in-distribution reference measure to assign support to z. Inside SBOG, this energy defines smooth boundary contours for anchor selection and candidate acceptance, while semantic feasibility and decoding constraints turn these contours into valid outlier candidates. The construction is modality-agnostic in latent space and, through the shared Gibbs normalizer in the Sinkhorn-DRO dual, links boundary selection to a robust downstream decision view.

Given the empirical latent reference set Z_n, SBOG first computes $\widehat{E}(z_i) := E_\varepsilon^{\widehat{\nu}_n}(z_i)$ for all embeddings $z_i \in Z_n$ and selects the top-energy embeddings as boundary anchors. Around each selected anchor, SBOG draws local latent proposals by adding isotropic Gaussian perturbations and projecting them back to $\mathcal{Z}$ when needed. A proposal v is feasible only if it crosses the SOE support-cost contour $\widehat{E}(v) \geq \tau_E$ and satisfies the semantic feasibility constraint $\cos(v, t_{y_i}) \geq \tau_{\text{sem}}$, where t_{y_i} is the class anchor associated with the source embedding z_i. Among the feasible proposals, SBOG selects the candidate with the largest SOE, the most transport-expensive point that remains semantically anchored, and then decodes it into input space via $x_{\text{ood}} \sim p(x \mid v)$. Unlike purely statistical density-rejection or hard k-NN-based methods [57, 13], SBOG does not rely on a low-support criterion alone. It uses a smooth transport-support contour for boundary selection and explicitly couples this boundary criterion with semantic anchoring and generative validity.

Algorithm 1 Sinkhorn Boundary Outlier Generation (SBOG)

Require: $D_{\text{in}} = \{(x_i, y_i)\}_{i=1}^n$, encoder h_θ, decoder $p(x \mid v)$
Require: class anchors $\{t_c\}_{c=1}^C$, semantic floor $\tau_{\text{sem}} \geq 0$
Require: $\varepsilon, \tau_E, \sigma > 0$, boundary anchors A, proposals M, samples N_{ood}, projection $\Pi_{\mathcal{Z}}$, $\widehat{E}(\cdot) := E_\varepsilon^{\widehat{\nu}_n}(\cdot)$

$Z_n \leftarrow \{z_i = z(x_i; \theta)\}_{i=1}^n$, $\widehat{\nu}_n \leftarrow \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$
let $\mathcal{I}_A$ be the indices of the top-A values of $\{\widehat{E}(z_i)\}_{i=1}^n$
$D_{\text{ood}} \leftarrow \emptyset$
for $k = 1, \ldots, N_{\text{ood}}$ **do**
  repeat
    sample $i \sim \text{Unif}(\mathcal{I}_A)$
    $v_j \leftarrow \Pi_{\mathcal{Z}}(z_i + \eta_j)$, $\eta_j \sim \mathcal{N}(0, \sigma^2 I)$, $j \in [M]$
    $\mathcal{J} \leftarrow \{j : \widehat{E}(v_j) \geq \tau_E \text{ and } \cos(v_j, t_{y_i}) \geq \tau_{\text{sem}}\}$
  until $\mathcal{J} \neq \emptyset$
  $v \leftarrow \arg\max_{j \in \mathcal{J}} \widehat{E}(v_j)$, $x_{\text{ood}} \sim p(x \mid v)$
  $D_{\text{ood}} \leftarrow D_{\text{ood}} \cup \{x_{\text{ood}}\}$
end for
return D_{ood}

The preceding sampler uses SOE as a model-agnostic boundary energy. We next show that this energy is not an isolated support heuristic, but the support-normalization term induced by the same Sinkhorn-DRO geometry. For a radius $r_{\text{DRO}} > 0$, the Sinkhorn uncertainty set around μ defines

$$\sup_{P \in \mathcal{P}(\mathcal{Z}):\ W_\varepsilon^\nu(\mu, P) \leq r_{\text{DRO}}} \mathbb{E}_{u \sim P}[f(u)], \tag{7}$$

for a latent loss or score function f. The next result explains why the SOE threshold and energy-maximizing feasible proposal in SBOG correspond to transport support-boundary acquisition.

Proposition 3.3 (Sinkhorn outlier energy in the Sinkhorn-DRO dual)**.** *Under the definitions above, for any measurable $f : \mathcal{Z} \to \mathbb{R}$ whose log-moment is finite for every $\lambda > 0$, the Sinkhorn-DRO dual objective admits the following decomposition:*

$$\begin{aligned} V_D(\lambda) &= \lambda r_{\text{DRO}} + \lambda\varepsilon \mathbb{E}_{z \sim \mu}\left[\log \mathbb{E}_{u \sim Q_{\varepsilon,z}^\nu} \exp\left(\frac{f(u)}{\lambda\varepsilon}\right)\right] \\ &= \lambda r_{\text{DRO}} + \mathbb{E}_{z \sim \mu}\left[\lambda\varepsilon \log \int_{\mathcal{Z}} \exp\left(\frac{f(u)}{\lambda\varepsilon} - \frac{d(z,u)}{\varepsilon}\right) d\nu(u) + \lambda E_\varepsilon^\nu(z)\right], \end{aligned} \tag{8}$$

where $Q_{\varepsilon,z}^\nu$ is the Gibbs kernel and $E_\varepsilon^\nu(z) = -\varepsilon \log K_\varepsilon^\nu(z)$ is the Sinkhorn outlier energy from Definition 3.1. Thus, the Sinkhorn-DRO dual objective contains the outlier energy term $\lambda \mathbb{E}_{z \sim \mu}[E_\varepsilon^\nu(z)]$.

Proposition 3.3 clarifies the role of SOE in SBOG. The decomposition shows that SOE is the support-normalization cost induced by the Sinkhorn uncertainty geometry. SBOG uses this term in a model-agnostic way: the threshold $\widehat{E}(v) \geq \tau_E$ selects candidates beyond a transport support-cost contour, and the final maximization chooses the most transport-expensive candidate among the semantically feasible proposals. Thus, the DRO connection justifies the support-boundary acquisition rule used by SBOG, rather than claiming that the submitted algorithm solves a downstream task-specific DRO problem. If a downstream model is available, one can augment the acquisition with a task loss or uncertainty score using the same Gibbs kernel, and we leave this task-aware variant as a

direct extension. Since SBOG relies on the empirical SOE threshold $\widehat{E}(v) \geq \tau_E$ for anchor selection and proposal acceptance, we next analyze the finite-sample stability of this boundary decision.

Theorem 3.4 (Finite-sample stability of SOE thresholding)**.** *For a fixed finite candidate pool $\mathcal{V} \subset \mathcal{Z}$, define $\underline{K} := \inf_{v\in\mathcal{V}} K_\varepsilon^{\bar{\nu}}(v) > 0$, $\Delta_E = \inf_{v\in\mathcal{V}} |E_\varepsilon^{\bar{\nu}}(v) - \tau_E| > 0$, and $\tau_n(\delta) = \sqrt{\log(2|\mathcal{V}|/\delta)/2n}$. If $\tau_n(\delta) \leq \underline{K}/2$, then with probability at least $1-\delta$,*

$$\sup_{v\in\mathcal{V}} \left| E_\varepsilon^{\widehat{\nu}_n}(v) - E_\varepsilon^{\bar{\nu}}(v) \right| \leq \frac{2\varepsilon}{\underline{K}} \tau_n(\delta).$$

Consequently, if $\tau_n(\delta) \leq \min\left\{\frac{\underline{K}}{2}, \frac{\underline{K}\Delta_E}{4\varepsilon}\right\}$, then with probability at least $1-\delta$, $\mathbf{1}\{E_\varepsilon^{\widehat{\nu}_n}(v) \geq \tau_E\} = \mathbf{1}\{E_\varepsilon^{\bar{\nu}}(v) \geq \tau_E\}$ for any $v \in \mathcal{V}$. In particular, it suffices that

$$n \geq \max\left\{ \frac{2}{\underline{K}^2} \log\frac{2|\mathcal{V}|}{\delta},\ \frac{8\varepsilon^2}{\underline{K}^2\Delta_E^2} \log\frac{2|\mathcal{V}|}{\delta} \right\}.$$

For a fixed candidate pool $\mathcal{V}$, the sufficient sample size scales as $O\left(\frac{\varepsilon^2}{\underline{K}^2\Delta_E^2}\log\frac{|\mathcal{V}|}{\delta}\right)$ up to the $\underline{K}^{-2}\log(|\mathcal{V}|/\delta)$ normalizer condition. Thus, the dependence on the latent geometry, the Sinkhorn temperature, and the placement of the candidate pool is captured by $\underline{K}$ and by the attainable energy margin Δ_E. In particular, if ε is too small or the candidate pool lies in extremely weakly supported regions, then $\underline{K}$ may be small and the bound can become loose. We also analyze the sample complexity of neighborhood-based methods in Appendix C, highlighting the advantages of SOE.

We next give a theorem for the consistency of the generated distribution. Write $\widehat{E}_n = E_\varepsilon^{\widehat{\nu}_n}$ and $E = E_\varepsilon^{\bar{\nu}}$. Let Q_n^{SBOG} and $Q_{n,\mathrm{SOE}}^\star$ be the class–latent output distributions of Algorithm 1 using $\widehat{E}_n$ and E, respectively, on the same observed bank. Denote their top-A anchor sets by $\widehat{I}_{A,n}$ and $I_A^\star$, and their energy thresholds by $\widehat{\tau}_n$ and $\tau_n^\star$. All remaining components and tie-breaking rules are shared. Define $e_n = \sup_{v\in\mathcal{Z}} |\widehat{E}_n(v) - E(v)|$ and $a_n = \frac{|\widehat{I}_{A,n} \triangle I_A^\star|}{2A}$,where $\triangle$ denotes symmetric difference. In one population-SOE proposal round, let $S = \{j : \cos(V_j, t_{y_I}) \geq \tau_{\mathrm{sem}}\}$ and $J^\star = \{j \in S : E(V_j) \geq \tau_n^\star\}$. Let Δ_n be the gap between the two largest energies in $J^\star$, with $\Delta_n = \infty$ when $|J^\star| < 2$. Conditional on the observed bank, define

$$b_n(t) = \Pr\left(\min_{j\in S} |E(V_j) - \tau_n^\star| \leq 2t \text{ or } \Delta_n \leq 2t \right),$$

and $p_n^\star = \Pr(J^\star \neq \varnothing)$, with $\min\varnothing = \infty$.

Theorem 3.5 (Population-SOE consistency)**.** *Suppose $|\widehat{\tau}_n - \tau_n^\star| \leq e_n$, $p_n^\star \geq p_0 > 0$, and $a_n + b_n(e_n) < p_0$. Then the empirical sampler terminates almost surely, and*

$$d_{\mathrm{TV}}\big(Q_n^{\mathrm{SBOG}}, Q_{n,\mathrm{SOE}}^\star\big) \leq \min\left\{1, \frac{2\{a_n + b_n(e_n)\}}{p_0}\right\} =: \zeta_n, \tag{9}$$

where $d_{\mathrm{TV}}(P,Q) = \sup_B |P(B) - Q(B)|$. The same bound holds after applying a shared conditional decoder. If $e_n \to 0$ and $a_n + b_n(e_n) \to 0$ in probability over the bank, and the threshold and acceptance conditions hold with probability tending to one, then $d_{\mathrm{TV}}(Q_n^{\mathrm{SBOG}}, Q_{n,\mathrm{SOE}}^\star) \to 0$ in probability.

The proof is given in Appendix D. This result concerns consistency with the population SOE criterion, not alignment with an independently defined target boundary.

3.2 Sinkhorn-DRO Alignment for Boundary-Aware Representations

SBOG relies on a latent ground cost whose local neighborhoods reflect both geometric proximity and semantic similarity. If the representation is trained only with pointwise alignment logits, the resulting distances may separate classes at training points but still give poorly calibrated boundary geometry for outlier generation. We therefore use the same Sinkhorn-DRO principle to robustify the alignment objective: instead of matching each sample only to its class prototype at a point, we require the alignment to remain stable under a local Sinkhorn transport neighborhood. Concretely, consider the standard alignment objective

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[-\log \frac{\exp\big(T(y)^\top z(x;\theta)/t\big)}{\sum_{c=1}^{C} \exp(T(c)^\top z(x;\theta)/t)} \right], \tag{10}$$

where $z(x;\theta) = h_\theta(x)/\|h_\theta(x)\|_2$ and $\{T(c)\}_{c=1}^C$ are class prototypes. Since these logits are purely pointwise, they may not adequately reflect the boundary geometry of the in-distribution, where informative outlier signals tend to concentrate.

We incorporate local Sinkhorn geometry through a reference measure $\nu_x \in \mathcal{P}(\mathcal{X})$ centered around x. Let $W_\varepsilon^{\nu_x}$ denote the normalized Sinkhorn OT discrepancy with reference measure ν_x and regularization $\varepsilon > 0$. For class c, define $\ell_{\text{align}}(\tilde{x}, c; \theta) = -\frac{T(c)^\top z(\tilde{x};\theta)}{t}$ and $z(\tilde{x};\theta) = \frac{h_\theta(\tilde{x})}{\|h_\theta(\tilde{x})\|_2}$. We replace the nominal logit by the Sinkhorn-DRO robust logit

$$V_c(x;\theta) := - \sup_{q \in \mathcal{P}(\mathcal{X}): W_\varepsilon^{\nu_x}(\delta_x, q) \le \rho} \mathbb{E}_{\tilde{x}\sim q}[\ell_{\text{align}}(\tilde{x}, c; \theta)]\,,$$

where δ_x is the point mass at x, and $r_{\text{DRO}} > 0$ is the local DRO radius. By Lemma 2.3, this is equivalently given by the corresponding one-dimensional dual with $f(\tilde{x}) = \ell_{\text{align}}(\tilde{x}, c; \theta)$, namely,

$$V_c(x;\theta) = -\inf_{\lambda>0} \left\{ \lambda r_{\text{DRO}} + \lambda\varepsilon \log \mathbb{E}_{\tilde{x}\sim Q_{\varepsilon,x}^{\nu_x}} \left[\exp\left(\frac{\ell_{\text{align}}(\tilde{x}, c; \theta)}{\lambda\varepsilon} \right) \right] \right\}.$$

Using $\{V_c(x;\theta)\}_{c=1}^C$ in place of the nominal similarity logits yields

$$\mathcal{L}_{\text{rob}} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[-\log \frac{\exp(V_y(x;\theta))}{\sum_{c=1}^C \exp(V_c(x;\theta))} \right]. \tag{11}$$

Remark 3.6. *If $r_{\text{DRO}} = 0$ and $\nu_x = \delta_x$, then $Q_{\varepsilon,x}^{\nu_x} = \delta_x$ and $V_c(x;\theta) = -\ell_{\text{align}}(x, c; \theta) = T(c)^\top z(x;\theta)/t$. Therefore $\mathcal{L}_{\text{rob}}$ reduces to the standard alignment loss. More generally, when ν_x is chosen as a local neighborhood measure and $d_\mathcal{X}$ is a metric, the robust logit aggregates local evidence through a log-sum-exp tilt, recovering neighbor- or margin-based variants as limiting cases.*

This alignment module makes the learned representation geometry more compatible with SBOG by using the same local Sinkhorn neighborhoods in both robust alignment and SOE. The robust logits aggregate local perturbations through the Sinkhorn transport geometry, while SOE uses the corresponding Gibbs kernel to measure support for boundary generation.

4 Experiments

Since synthetic outliers do not have a canonical target distribution or a universally accepted fidelity metric, we evaluate them through their downstream utility, following standard practice in the outlier-synthesis literature [13, 10]. Specifically, we borrow detector-side benchmarks as a practical and task-relevant proxy: better synthesized outliers should better cover informative regions outside the in-distribution support, and therefore lead to a clearer boundary between inliers and outliers therefore improved discrimination performance [21]. All experiments are conducted on a machine equipped with 13th Gen Intel(R) Core(TM) i9-13900HX CPU (24 cores) and NVIDIA A100 GPU (40GB). The experimental details are in Appendix Section F. The code is available at `https://github.com/hSun08/SBOG`.

4.1 Time Series Generation

We evaluate SBOG on five standard multivariate time-series anomaly detection benchmarks, including SMD [53], MSL [24], SMAP, SWaT [40], and WADI [2] and baselines of COUTA [66], Ts2Vec [74], Dcdetector [69], TranAD [58], AnomalyTransformer [67], ScatterAD [70], THOC [49] and Sub-Adjacent Transformer [73]. We use F1 and AU-PR as the main evaluation metrics for each method. To avoid overestimating detection performance [28], our main tables use the stricter point-wise evaluation without Point Adjustment.

The upper part of Table 1 reports general time-series anomaly detection baselines, while CARLA [10] is the only compatible framework that explicitly uses synthetic negatives for time-series anomaly detection. We therefore use CARLA as a controlled testbed with only the negative-sample construction rule changed. Specifically, we compare random negatives, k-NN based sparse negatives, KDE-based support negatives, and SBOG negatives. It is worth noting that the robust alignment module is not used in the time-series experiments because these benchmarks do not contain semantic class anchors or prototype labels that define logits $T(c)^\top z$. Under this matched setting, CARLA-SBOG achieves

Table 1: Comparison results on SMD, MSL, SMAP, SWaT, and WADI datasets. **Bold** indicates the best result among the synthesis baselines, and † indicates the best result among all baselines.

Methods	SMD		MSL		SMAP		SWaT		WADI		Average	
	AU-PR	F1	AU-PR	F1	AU-PR	F1	AU-PR	F1	AU-PR	F1	AU-PR	F1
AnomalyTransformer [67]	0.082	0.137	0.129	0.218	0.203	0.307	0.375	0.531	0.064	0.145	0.171	0.268
DCdetector [69]	0.044	0.086	0.115	0.204	0.131	0.225	0.401	0.537	0.059	0.109	0.150	0.232
TranAD [58]	0.389	0.450	0.204	0.314	0.196	0.308	0.464†	0.684†	0.040	0.109	0.259	0.373
COUTA [66]	0.409†	0.461†	0.321	0.403	0.309	0.382	0.303	0.535	0.207	0.266	0.310	0.410
TS2Vec [74]	0.113	0.186	0.145	0.253	0.165	0.276	0.337	0.532	0.057	0.114	0.163	0.272
ScatterAD [70]	0.300	0.360	0.144	0.234	0.143	0.251	0.347	0.532	0.053	0.121	0.198	0.299
THOC [49]	0.321	0.381	0.227	0.322	0.235	0.333	0.301	0.538	0.053	0.125	0.227	0.340
Sub-Adjacent Transformer [73]	0.185	0.269	0.295	0.408	0.171	0.269	0.384	0.533	0.458†	0.571†	0.299	0.410
Synthesis baselines												
CARLA [10]	0.236	0.296	0.337	0.420	0.304	0.378	0.402	0.606	0.058	0.109	0.267	0.362
CARLA-RandomNeg	0.219	0.285	0.334	0.431	0.274	0.352	0.319	0.531	0.028	0.102	0.235	0.340
CARLA-k-NNNeg	0.371	0.434	0.250	0.350	0.390	0.451	0.402	0.604	0.045	0.109	0.291	0.390
CARLA-KDENeg	0.258	0.348	0.351	0.445	0.391	0.465	0.350	0.578	0.049	0.129	0.280	0.393
CARLA-SBOG	**0.413**	**0.471**	**0.352**†	**0.463**†	**0.449**†	**0.515**†	**0.369**	**0.625**	**0.063**	**0.133**	**0.312**†	**0.432**†

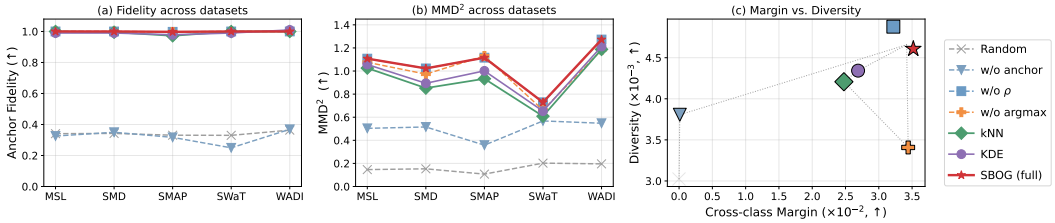

Figure 2: Component ablation of the time-series sampler.

the best average AU-PR and F1 among all CARLA variants. This indicates that the SBOG boundary rule produces more useful negatives for downstream anomaly detection.

To further examine whether the gains come from the SOE sampler itself, we ablate the main mechanisms of Algorithm 1, including high-energy anchor selection, energy-threshold acceptance, and argmax candidate selection, and additionally compare with k-NN and KDE support-based variants. All variants use the same trained encoder, so the comparison isolates the effect of the sampler and boundary scoring rule. Panels (a) and (b) show that SBOG maintains high anchor fidelity while inducing a stronger distributional shift as measured by MMD^2, whereas random sampling and removing anchor selection reduce fidelity. Panel (c) further shows that SBOG achieves a better margin–diversity trade-off than the ablated and support-score variants: it moves generated negatives farther along the boundary while retaining high intra-negative diversity. This suggests that selecting the most energetic feasible candidate improves boundary informativeness beyond generic low-support scoring.

Table 2: Sensitivity to the threshold τ_E

τ_E	Avg AUPR	Avg F1	Gap	Fill
0.50	0.279	0.387	0.225	1.000
0.70	0.295	0.409	0.224	1.000
0.90	0.310	0.430	0.217	1.000
0.95	**0.312**	**0.432**	0.126	1.000
0.98	0.305	0.422	0.109	0.999
0.99	0.277	0.384	0.079	0.945

We sweep the threshold τ_E using the same trained encoder and report Avg AUPR and F1 together with two proposal diagnostics. Gap measures the average margin $E(\text{neg}) - \tau_E$ of accepted negatives, and Fill is the fraction of anchors that yield at least one feasible proposal. Table 2 shows that performance is stable over a broad middle range of thresholds, while overly high thresholds reduce both the margin and feasible-anchor coverage. This is consistent with Theorem 3.4: reliable thresholding requires a nontrivial energy margin and a well-conditioned support normalizer.

4.2 Image Generation

For a more comprehensive evaluation, we also apply our method to image outlier generation. We follow the experimental protocol used in recent image outlier synthesis works [57, 13, 17, 36] and the standard benchmark setting in [23] for evaluation. For fair comparison, we adopt ResNet-34 [20] as the ID classifier and CIFAR-100 and ImageNet-100 [11] as the in-distribution datasets. Following prior baselines, we generate OOD samples with Stable Diffusion v1.5. On ImageNet-100, we use iNaturalist [60], SUN [65], Places [77], and Textures [7] as test datasets. On CIFAR-100, we use SVHN [41], Places365 [77], LSUN-R [72], iSUN [68], and Textures [7] as test datasets.

We compare our method against representative baselines for synthesis methods about outlier detection, including GAN [32], VOS [14], NPOS [57], DreamOOD [13], BOOD [36], FodFoM [5], and GOOD [17]. Following the standard evaluation protocol used by these baselines, we assess the quality of the generated data using FPR95, the false positive rate when the true positive rate for ID samples is 95% and AUROC, the area under the receiver operating characteristic curve. The results on ImageNet-100 and CIFAR-100 are reported in Table 3 and Table 4, respectively. For methods without official released code, we report the summarized results provided in recent benchmark papers [17, 36]. Methods marked with an asterisk (*) do not have publicly available official implementations and are therefore not included in the formal ranking. In the tables, **bold** indicates the best result and underline indicates the second-best result.

Table 3: Comparative evaluation of outlier detection performance on ImageNet-100 as the ID dataset.

Methods	iNaturalist		Places		SUN		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
GAN [32]	83.10	71.35	83.20	69.85	84.40	67.56	91.00	59.16	85.42	66.98
VOS [14]	43.00	93.77	47.60	91.77	39.40	93.17	66.10	81.42	49.02	90.03
NPOS [57]	53.84	86.52	59.66	83.50	53.54	87.99	**8.98**	**98.13**	44.00	89.04
Dream-OOD [13]	24.10	96.10	39.87	93.11	36.88	93.31	53.99	85.56	38.76	92.02
NCIS [12]	20.70	96.56	34.60	94.07	35.43	94.13	44.83	88.50	33.89	93.32
BOOD* [36]	18.33	96.74	33.33	94.08	37.92	93.52	51.88	85.41	35.37	92.44
GOOD* [17]	9.22	97.61	24.79	94.82	17.20	96.10	19.51	96.60	17.68	96.30
SBOG	**9.44**$_{\pm 0.73}$	**98.09**$_{\pm 0.10}$	14.40$_{\pm 1.01}$	97.09$_{\pm 0.16}$	13.22$_{\pm 0.92}$	97.61$_{\pm 0.11}$	24.26$_{\pm 1.06}$	94.82$_{\pm 0.23}$	15.33$_{\pm 0.46}$	96.90$_{\pm 0.08}$
SBOG (robust)	9.89$_{\pm 0.76}$	97.99$_{\pm 0.10}$	**13.24**$_{\pm 0.85}$	**97.21**$_{\pm 0.15}$	**11.81**$_{\pm 0.99}$	**97.64**$_{\pm 0.11}$	24.06$_{\pm 1.06}$	95.05$_{\pm 0.22}$	**14.75**$_{\pm 0.46}$	**96.97**$_{\pm 0.08}$

The results show that SBOG achieves near state-of-the-art outlier detection performance on both CIFAR-100 and ImageNet-100. More importantly, SOE improves over several representative statistical support-based synthesis methods, especially VOS [14] and Dream-OOD [13], which sample virtual outliers from low-likelihood regions, and NPOS [57], which is the cloest non-parametric density estimation baseline. The results support the central design of SBOG: SOE identifies informative weak-support boundary regions, while semantic anchoring controls class drift. The SBOG (robust) variant uses the Sinkhorn-DRO alignment module from Section 3.2. Its small but consistent improvement suggests that calibrating the latent representation with the same Sinkhorn geometry used by SOE can improve the boundary candidates generated by SBOG.

We first test whether the boundary component of SBOG can be replaced by generic support scores. We replace the SOE anchor rule with two statistical alternatives, k-NN sparsity and KDE-RBF scoring, while keeping the remaining proposal, semantic filtering, and decoding steps fixed. Table 5 shows that SOE gives better token alignment, diversity, and feasible yield. This indicates that the transport boundary energy selects more useful anchors than purely statistical support scores, even when the rest of the SBOG pipeline is unchanged.

Table 5: Anchor-selection ablation.

Strategy	Token cos.	Diversity	Yield
k-NN	0.490	0.635	0.277
KDE-RBF	0.464	0.626	0.272
SBOG	**0.691**	**0.835**	**0.490**

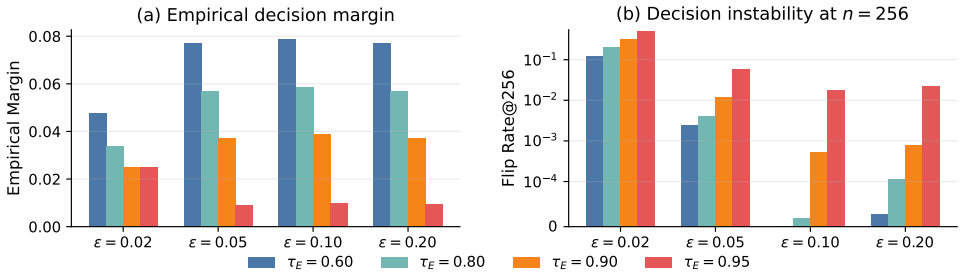

Figure 3: Finite-sample stability of the outlier-energy decision under varying ϵ and m.

We next study the calibration of the SOE acceptance contour. Using the same trained encoder, we sweep the threshold τ_E and report the empirical margin together with flip rate. Figure 3 shows that very small ε makes the energy landscape sharp and unstable, while moderate smoothing yields larger margins and lower flip rates. This supports the use of SOE as a smooth boundary criterion inside SBOG and is consistent with the finite-sample stability condition in Theorem 3.4.

Table 6: Effect of semantic anchoring.

Metric	w/o filter	SBOG
Purity ↑	0.381	0.824
Token cos. ↑	0.329	0.602
Class margin ↑	-0.074	0.220

Table 4: Comparative evaluation of outlier detection performance on CIFAR-100 as the ID dataset.

Methods	SVHN		PLACES365		LSUN-R		iSUN		TEXTURES		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
GAN [32]	89.45	66.95	88.75	66.76	82.35	75.87	83.45	73.49	92.80	62.99	87.36	69.21
VOS [14]	78.50	73.11	84.55	75.85	59.05	85.72	72.45	82.66	75.35	80.08	73.98	79.48
NPOS [57]	11.14	97.84	79.08	71.30	56.27	82.43	51.72	85.48	35.20	92.44	46.68	85.90
DreamOOD [13]	58.75	87.01	70.85	79.94	24.25	95.23	1.10	99.73	46.60	88.82	40.31	90.15
BOOD* [36]	5.42	98.43	40.55	90.76	2.06	99.25	0.22	99.91	5.10	98.74	10.67	97.42
FodFoM [5]	33.19	94.02	**42.30**	**90.68**	28.24	95.09	33.06	94.45	35.44	93.38	34.45	93.52
GOOD* [17]	19.37	95.34	65.85	82.04	18.82	94.15	0.81	99.59	23.47	95.98	25.66	93.34
SBOG	**5.38**$_{\pm 0.64}$	98.67$_{\pm 0.10}$	47.48$_{\pm 1.70}$	89.70$_{\pm 0.34}$	3.48$_{\pm 0.52}$	**99.13**$_{\pm 0.06}$	**0.10**$_{\pm 0.07}$	**99.95**$_{\pm 0.01}$	5.74$_{\pm 0.61}$	98.69$_{\pm 0.11}$	**12.44**$_{\pm 0.39}$	**97.23**$_{\pm 0.08}$
SBOG (robust)	5.49$_{\pm 0.57}$	**98.69**$_{\pm 0.09}$	50.60$_{\pm 1.51}$	88.28$_{\pm 0.39}$	**2.96**$_{\pm 0.43}$	99.10$_{\pm 0.07}$	0.35$_{\pm 0.13}$	99.88$_{\pm 0.02}$	**5.46**$_{\pm 0.54}$	**98.78**$_{\pm 0.10}$	12.97$_{\pm 0.36}$	96.95$_{\pm 0.05}$

Finally, we isolate the role of semantic anchoring. The previous ablation shows that SOE provides stronger boundary anchors than k-NN or KDE, but weak support alone does not guarantee class consistency. We then compare an unanchored candidate pool with the token-anchored SBOG pool, using the same downstream ranking rule and evaluating the top 250 candidates. Table 6 shows that semantic anchoring improves intended-class purity, token alignment, cross-class margin, and contamination control. The t-SNE [59] in Figure 4 visualizes the same effect: anchored SBOG candidates remain near intended class boundaries, whereas unanchored candidates often mix with other class clusters or are absorbed into them. These ablations separate the two roles in SBOG: SOE improves boundary selection over purely statistical support scores, while semantic anchoring prevents the selected weak-support candidates from drifting off class.

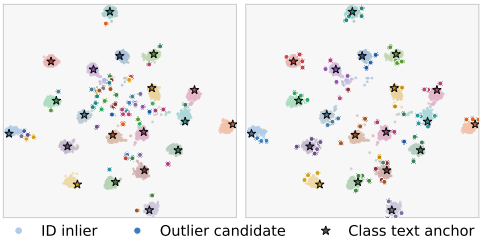

Figure 4: Visualization of anchors.

5 Conclusion

In this paper, we proposed Sinkhorn Boundary Outlier Generation (SBOG), a structured framework for latent-space outlier generation. SBOG uses Sinkhorn Outlier Energy (SOE) as a transport-support energy to identify weakly supported boundary regions, and combines it with semantic feasibility and decoding constraints to generate controlled, class-consistent outliers. We showed that SOE recovers kernel support as an empirical Euclidean special case, while its appearance in the Sinkhorn-DRO dual gives the boundary rule an optimization-consistent interpretation. Experiments on time-series anomaly generation and image outlier synthesis demonstrate that SBOG can generate informative synthetic outliers across modalities and improve downstream robustness evaluation.

SBOG depends on representation quality, transport-cost design, semantic anchoring, and generator expressiveness. Weak embeddings or limited decoders may reduce outlier fidelity and utility. Future work can develop adaptive calibration strategies for the Sinkhorn temperature and SOE threshold, jointly learn the latent representation, transport cost, and outlier generator, and extend the Sinkhorn-DRO interpretation toward downstream-aware stress generation.

References

[1] Momin Abbas, Muneeza Azmat, Raya Horesh, and Mikhail Yurochkin. Out-of-distribution detection using synthetic data generation. In *Second Conference on Language Modeling*, 2025.

[2] Chuadhry Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P. Mathur. WADI: a water distribution testbed for research in the design of secure cyber physical systems. CySWATER '17, page 25–28, New York, NY, USA, 2017. Association for Computing Machinery.

[3] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3395–3404, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Chris U. Carmona, François-Xavier Aubet, Valentin Flunkert, and Jan Gasthaus. Neural contextual anomaly detection for time series. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2843–2851. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[5] Jiankang Chen, Ling Deng, Zhiyong Gan, Wei-Shi Zheng, and Ruixuan Wang. Fodfom: Fake outlier data by foundation models creates stronger visual out-of-distribution detector. In *ACM Multimedia 2024*, 2024.

[6] Dongchan Cho, Jiho Han, Keumyeong Kang, Minsang Kim, Honggyu Ryu, and Namsoon Jung. Structured temporal causality for interpretable multivariate time series anomaly detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[9] Zahra Zamanzadeh Darban, Qizhou Wang, Geoffrey I Webb, Shirui Pan, Charu C Aggarwal, and Mahsa Salehi. Genias: Generator for instantiating anomalies in time series. *arXiv preprint arXiv:2502.08262*, 2025.

[10] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu C Aggarwal, and Mahsa Salehi. Carla: Self-supervised contrastive representation learning for time series anomaly detection. *Pattern Recognition*, 157:110874, 2025.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[13] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36:60878–60901, 2023.

[14] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

[15] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[16] Yuchen Fang, Jiandong Xie, Yan Zhao, Lu Chen, Yunjun Gao, and Kai Zheng. Temporal-frequency masked autoencoders for time series anomaly detection. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1228–1241, 2024.

[17] Xin Gao, Jiyao Liu, Guanghao Li, Yueming Lyu, Jianxiong Gao, Weichen Yu, Ningsheng Xu, Liang Wang, Caifeng Shan, Ziwei Liu, and Chenyang Si. GOOD: Training-free guided diffusion sampling for out-of-distribution detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[18] Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, and Chuan-Sheng Foo. An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2508–2517, 2021.

[19] Shantanu Ghosh, Tiankang Xie, and Mikhail Kuznetsov. Distributionally robust self-supervised learning for tabular data. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[22] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[23] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8710–8719, 2021.

[24] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 387–395, 2018.

[25] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

[26] Seungho Jeon, Kijong Koo, Daesung Moon, and Jung Taek Seo. Mutation-based multivariate time-series anomaly generation on latent space with an attention-based variational recurrent neural network for robust anomaly detection in an industrial control system. *Applied Sciences*, 14(17), 2024.

[27] Jinyong Jeong, Hyungu Kahng, and Seoung Bum Kim. Multi-expert distributionally robust optimization for out-of-distribution generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[28] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 7194–7201, 2022.

[29] Taero Kim, Subeen Park, Sungjun Lim, Yonghan Jung, Krikamol Muandet, and Kyungwoo Song. Sufficient invariant learning for distribution shift. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4958–4967, 2025.

[30] Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

[31] Tian Lan, Hao Duong Le, Jinbo Li, Wenjun He, Meng Wang, Chenghao Liu, and Chen Zhang. Towards foundation models for zero-shot time series anomaly detection: Leveraging synthetic data and relative context discrepancy. *arXiv preprint arXiv:2509.21190*, 2025.

[32] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[33] Beibu Li, Qichao Shentu, Yang Shu, Hui Zhang, Ming Li, Ning Jin, Bin Yang, and Chenjuan Guo. Crossad: Time series anomaly detection with cross-scale associations and cross-window modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[34] Chun-Guo Li. Learn hybrid prototypes for multivariate time series anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025.

[35] Hengzhuang Li and Teng Zhang. Outlier synthesis via hamiltonian monte carlo for out-of-distribution detection. In *The Thirteenth International Conference on Learning Representations*, 2025.

[36] Qilin Liao, Shuo Yang, Bo Zhao, Ping Luo, and Hengshuang Zhao. Bood: Boundary-based out-of-distribution data generation. *arXiv preprint arXiv:2508.00350*, 2025.

[37] Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

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

[39] Natalia Martinez, Fearghal O'Donncha, Wesley M. Gifford, Nianjun Zhou, Dhaval C Patel, and Roman Vaculin. Adaptive conformal anomaly detection with time series foundation models for signal monitoring. In *The Fourteenth International Conference on Learning Representations*, 2026.

[40] Aditya P. Mathur and Nils Ole Tippenhauer. SWaT: a water treatment testbed for research and training on ICS security . In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, pages 31–36, Los Alamitos, CA, USA, April 2016. IEEE Computer Society.

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

[42] Guansong Pang, Chunhua Shen, and Anton Van Den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 353–362, 2019.

[43] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[44] Qi Qian, Juhua Hu, and Hao Li. Hierarchically robust representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7336–7344, 2020.

[45] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR, 10–15 Jul 2018.

[46] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020.

[47] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[48] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[49] Lifeng Shen, Zhuocong Li, and James Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13016–13026. Curran Associates, Inc., 2020.

[50] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[51] Agnieszka Słowik and Leon Bottou. On distributionally robust optimization and data rebalancing. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1283–1297. PMLR, 28–30 Mar 2022.

[52] Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. Memto: Memory-guided transformer for multivariate time series anomaly detection. *Advances in Neural Information Processing Systems*, 36:57947–57963, 2023.

[53] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2828–2837, New York, NY, USA, 2019. Association for Computing Machinery.

[54] Haixiang Sun and Andrew Liu. Robust generalization with adaptive optimal transport priors for decision-focused learning. In *The 29th International Conference on Artificial Intelligence and Statistics*.

[55] Haixiang Sun and Andrew Liu. Diff2sp: Diffusion models for correlated scenario generation in stochastic programming. *arXiv preprint arXiv:2606.05649*, 2026.

[56] Haixiang Sun, Ye Shi, Jingya Wang, Hoang Duong Tuan, H. Vincent Poor, and Dacheng Tao. Alternating differentiation for optimization layers. In *The Eleventh International Conference on Learning Representations*, 2023.

[57] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*.

[58] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proceedings of VLDB*, 15(6):1201–1214, 2022.

[59] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[60] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[61] Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *Operations Research*, 2025.

[62] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[63] Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang, and Xiangnan He. Understanding contrastive learning via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 36:23297–23320, 2023.

[64] Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. Catch: Channel-aware multivariate time series anomaly detection via frequency patching. In *The Thirteenth International Conference on Learning Representations*.

[65] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[66] Hongzuo Xu, Yijie Wang, Songlei Jian, Qing Liao, Yongjun Wang, and Guansong Pang. Calibrated one-class classification for unsupervised time series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5723–5736, 2024.

[67] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

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

[69] Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3033–3045, 2023.

[70] Tao Yin, Shaochen Fu, Zhibin Zhang, Li Huang, Xiaohong Zhang, Yiyuan Yang, Kaixiang Yang, and Meng Yan. Scatterad: Temporal-topological scattering mechanism for time series anomaly detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

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

[73] Wenzhen Yue, Xianghua Ying, Ruohao Guo, DongDong Chen, Ji Shi, Bowei Xing, Yuqing Zhu, and Taiyan Chen. Sub-adjacent transformer: improving time series anomaly detection with reconstruction error from sub-adjacent neighborhoods. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[74] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8980–8987, 2022.

[75] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024.

[76] Fan Zhang, Sinchee Chin, Jing-Hao Xue, and Wenming Yang. ICDiffAD: Implicit conditioning diffusion model for time series anomaly detection. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

[78] Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pages 12857–12867. PMLR, 2021.

A Related works

DRO for representation Classical DRO and its Wasserstein variants learn models against worst-case shifts through ambiguity sets over distributions [48, 15, 50], later extended to local perturbations and entropic Sinkhorn formulations [61]. A closer line uses DRO to shape robust representations under subgroup or covariate shift, for example by learning hierarchically robust features [27], analyzing robust versus spurious representations [29, 56], or coupling invariant representation learning with worst-group objectives [44, 78, 54]. More recent work connects DRO to self-supervised and latent representation learning, but they are mostly concerned about interpretation [63, 55] or designing loss function for robustness in worst-case [19]. However, even in those work consider the outlier or negative points explicitly [22, 47, 25, 51], DRO is still used mainly to reweight losses or stabilize learned features thereby stabilize features and working on loss function refinement. Instead, our framework turns Sinkhorn-DRO geometry into an explicit and rigorous definition of outlyingness in latent space, yielding a training-consistent criterion for outlier identification and synthesis that is naturally more general than modality-specific generation pipelines.

Outlier in time series Most existing work on time-series outliers is still detection-oriented [49, 37, 75], focusing on learning a reliable notion of normality rather than explicitly modeling how outliers are formed. Typical approaches either learn compact normal regions, sometimes with weak anomaly supervision to calibrate the boundary [45, 42, 46, 66], or strengthen detector backbones through reconstruction-based scoring [3, 18, 58], contextual comparison [4], and representation learning with attention or contrastive objectives [67, 69]. Recent work has mostly strengthened detector backbones instead of rethinking outlier formation, by improving temporal structure modeling [62, 33], memory modules [52], hybrid prototypes [34], and exploiting masking or frequency-domain learning [16, 64]. This trend continues in newer interpretable and foundation-style detectors [6, 31, 70, 39, 76] and. By contrast, work that explicitly generates time-series outliers remains scarce. Most anomaly simulation is only auxiliary to detector training [10, 66], while dedicated synthesis methods are still rare [9, 26]. Relative to these baselines, our focus is not merely to separate normal from abnormal windows, but to define and synthesize informative outliers from a principled latent geometry, making the generated samples directly useful for further downstream evaluation and analysis.

Image outlier synthesis For image data, the literature on synthetic outliers has mostly been driven by OOD detection rather than outlier generation itself. Early training-based methods improved detectors by using curated auxiliary negatives to regularize classifiers rather than explicitly generating new outliers [21, 38]. Later work began to synthesize outliers in feature space [32], either with learned generators or with virtual sampling schemes [14, 57, 35], which improves boundary regularization but still defines candidate outliers through empirical density or sampling heuristics. More recent methods further push synthesis to the image level: [13] decodes low-likelihood latent samples with diffusion, while [71, 5, 36, 17] generate more realistic visual outliers through diffusion or foundation-model guidance. Despite the stronger visual quality, most of these methods still localize candidate outliers through heuristic proxies such as low-likelihood regions [13], sparse k-NN neighborhoods [57, 17], or boundary-crossing perturbations [36], rather than from a principled definition of outlyingness for detection. In contrast, our method is more general and not restricted to certain modality. Besides, we provide a rigorous definition of outlyingness through Sinkhorn geometric boundary, so outlier selection and synthesis are both grounded in a principled geometry rather than heuristic and purely statistic rules.

B Discussion of Sinkhorn Boundary Outlier Energy

We record the variational identity underlying the Sinkhorn outlier energy in Definition 3.1. For any $v \in \mathcal{Z}$, recall that

$$K_\varepsilon^\nu(v) := \int_{\mathcal{Z}} \exp\left(-\frac{d(v,u)}{\varepsilon}\right) d\nu(u), \qquad E_\varepsilon^\nu(v) := -\varepsilon \log K_\varepsilon^\nu(v).$$

Assume $0 < K_\varepsilon^\nu(v) < \infty$. Then

$$\inf_{r \in \mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{u\sim r}[d(v,u)] + \varepsilon \mathrm{KL}(r\|\nu)\} = -\varepsilon \log K_\varepsilon^\nu(v) = E_\varepsilon^\nu(v). \tag{12}$$

Therefore, $E_\varepsilon^\nu(v)$ is exactly the optimal value of the entropic cost–KL tradeoff centered at v relative to the reference measure ν. Larger values of $E_\varepsilon^\nu(v)$ correspond to smaller kernel mass under ν,

meaning that v is less supported by the reference geometry and is therefore more likely to be an outlier under the proposed energy criterion.

Fix $v \in \mathcal{Z}$, and define the Gibbs-tilted probability measure $r^{\nu,\star}_{v,\varepsilon} \in \mathcal{P}(\mathcal{Z})$ by

$$dr^{\nu,\star}_{v,\varepsilon}(u) := \frac{\exp(-d(v,u)/\varepsilon)}{K^\nu_\varepsilon(v)} d\nu(u).$$

Then for any $r \ll \nu$,

$$\begin{aligned}
\mathrm{KL}(r\|r^{\nu,\star}_{v,\varepsilon}) &= \int_{\mathcal{Z}} \log\left(\frac{dr}{dr^{\nu,\star}_{v,\varepsilon}}\right) dr \\
&= \int_{\mathcal{Z}} \log\left(\frac{dr}{d\nu}\right) dr + \int_{\mathcal{Z}} \frac{d(v,u)}{\varepsilon} dr(u) + \log K^\nu_\varepsilon(v) \\
&= \mathrm{KL}(r\|\nu) + \frac{1}{\varepsilon}\mathbb{E}_{u\sim r}[d(v,u)] + \log K^\nu_\varepsilon(v).
\end{aligned}$$

Rearranging gives

$$\mathbb{E}_{u\sim r}[d(v,u)] + \varepsilon \mathrm{KL}(r\|\nu) = -\varepsilon \log K^\nu_\varepsilon(v) + \varepsilon \mathrm{KL}(r\|r^{\nu,\star}_{v,\varepsilon}) \geq -\varepsilon \log K^\nu_\varepsilon(v),$$

with equality if and only if $r = r^{\nu,\star}_{v,\varepsilon}$. This proves (12). Hence the definition

$$E^\nu_\varepsilon(v) = -\varepsilon \log K^\nu_\varepsilon(v)$$

is exactly the entropic variational cost associated with v relative to the reference measure ν.

Finally, the thresholded outlier set

$$O_{\nu,\varepsilon}(\tau_E) := \{v \in \mathcal{Z} : E^\nu_\varepsilon(v) \geq \tau_E\}$$

is equivalently characterized by

$$K^\nu_\varepsilon(v) \leq e^{-\tau_E/\varepsilon}.$$

Similarly, we give the proof of Lemma 3.2.

Lemma 3.2 (Entropic support cost of Sinkhorn outlier energy) *For any $z \in Z$ such that $0 < K^\nu_\epsilon(z) < \infty$, given the Sinkhorn outlier energy $E^\nu_\epsilon(z)$, in particular, for the empirical reference measure $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ with distinct atoms, we have*

$$E^{\hat{\nu}_n}_\epsilon(z) = \min_{w\in\Delta_n} \left\{ \sum_{i=1}^n w_i d(z,z_i) + \epsilon \sum_{i=1}^n w_i \log(n w_i) \right\},$$

where $\Delta_n = \{w \in \mathbb{R}^n_+ : \sum_{i=1}^n w_i = 1\}$ with the minimizer being $w^\star_i(z) = \frac{\exp(-d(z,z_i)/\epsilon)}{\sum_{j=1}^n \exp(-d(z,z_j)/\epsilon)}$.

Proof. Fix $z \in Z$ with $0 < K^\nu_\epsilon(z) < \infty$, and define

$$\frac{dr^\star_z}{d\nu}(u) = \frac{\exp(-d(z,u)/\epsilon)}{K^\nu_\epsilon(z)}.$$

This is a probability density with respect to ν. For any $r \in \mathcal{P}(Z)$ with $r \ll \nu$, using the chain rule for Radon–Nikodym derivatives gives

$$\log \frac{dr}{dr^\star_z}(u) = \log \frac{dr}{d\nu}(u) + \frac{d(z,u)}{\epsilon} + \log K^\nu_\epsilon(z).$$

Therefore, in the extended-real sense,

$$\epsilon \mathrm{KL}(r\|r^\star_z) = \epsilon \mathrm{KL}(r\|\nu) + \mathbb{E}_{u\sim r}[d(z,u)] + \epsilon \log K^\nu_\epsilon(z).$$

Rearranging yields

$$\mathbb{E}_{u\sim r}[d(z,u)] + \epsilon \mathrm{KL}(r\|\nu) = -\epsilon \log K^\nu_\epsilon(z) + \epsilon \mathrm{KL}(r\|r^\star_z).$$

Since $\mathrm{KL}(r\|r^\star_z) \geq 0$, the right-hand side is minimized by $r = r^\star_z$, and the minimum value is

$$-\epsilon \log K^\nu_\epsilon(z) = E^\nu_\epsilon(z).$$

For $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ with distinct atoms, any $r \ll \hat{\nu}_n$ can be written as $r = \sum_{i=1}^n w_i \delta_{z_i}$ for some $w \in \Delta_n$. In this case,

$$\mathrm{KL}(r\|\hat{\nu}_n) = \sum_{i=1}^n w_i \log(n w_i),$$

and substituting this expression into the variational formula gives the empirical form. The expression for $w_i^\star(z)$ follows by substituting $\nu = \hat{\nu}_n$ into the Gibbs reweighting above. □

Proposition 3.3 (Outlier energy appears in the Sinkhorn-DRO dual) *Following the previous definitions, for any measurable $f : \mathcal{Z} \to \mathbb{R}$ such that the log-moment below is finite for $\lambda > 0$, the Sinkhorn-DRO dual objective admits the following decomposition: for every $\lambda > 0$,*

$$\begin{aligned} V_D(\lambda) &= \lambda\tau_E + \lambda\varepsilon \mathbb{E}_{z\sim\mu}\left[\log \mathbb{E}_{u\sim Q^\nu_{\varepsilon,z}} \exp\left(\frac{f(u)}{\lambda\varepsilon}\right)\right] \\ &= \lambda\tau_E + \mathbb{E}_{z\sim\mu}\left[\lambda\varepsilon \log \int_{\mathcal{Z}} \exp\left(\frac{f(u)}{\lambda\varepsilon} - \frac{d(z,u)}{\varepsilon}\right) d\nu(u) + \lambda E^\nu_\varepsilon(z)\right], \end{aligned}$$

where $Q^\nu_{\varepsilon,z}$ is the Gibbs kernel and $E^\nu_\varepsilon(z) = -\varepsilon \log K^\nu_\varepsilon(z)$ is the Sinkhorn outlier energy from Definition 3.1. In particular, minimizing the Sinkhorn-DRO objective necessarily optimizes an objective that contains the outlier energy term $\lambda\mathbb{E}_{z\sim\mu}[E^\nu_\varepsilon(z)]$.

Proof. Fix $\lambda > 0$. By definition of the Gibbs kernel from Lemma 2.3,

$$dQ^\nu_{\varepsilon,v}(u) = \frac{\exp\big(-d(z,u)/\varepsilon\big)}{K^\nu_\varepsilon(z)} d\nu(u), \qquad K^\nu_\varepsilon(z) = \int_{\mathcal{Z}} \exp\big(-d(z,u)/\varepsilon\big) d\nu(u).$$

For each fixed representation $z \in \mathcal{Z}$, expand the inner expectation:

$$\begin{aligned} \mathbb{E}_{u\sim Q^\nu_{\varepsilon,z}} \exp\left(\frac{f(u)}{\lambda\varepsilon}\right) &= \int_{\mathcal{Z}} \exp\left(\frac{f(u)}{\lambda\varepsilon}\right) dQ^\nu_{\varepsilon,z}(u) \\ &= \int_{\mathcal{Z}} \exp\left(\frac{f(u)}{\lambda\varepsilon}\right) \frac{\exp\big(-d(z,u)/\varepsilon\big)}{K^\nu_\varepsilon(z)} d\nu(u) \\ &= \frac{1}{K^\nu_\varepsilon(z)} \int_{\mathcal{Z}} \exp\left(\frac{f(u)}{\lambda\varepsilon} - \frac{d(z,u)}{\varepsilon}\right) d\nu(u). \end{aligned}$$

Taking $\log$ gives

$$\log \mathbb{E}_{u\sim Q^\nu_{\varepsilon,z}} \exp\left(\frac{f(u)}{\lambda\varepsilon}\right) = \log \int_{\mathcal{Z}} \exp\left(\frac{f(u)}{\lambda\varepsilon} - \frac{d(z,u)}{\varepsilon}\right) d\nu(u) - \log K^\nu_\varepsilon(z).$$

Multiplying by $\lambda\varepsilon$ yields

$$\lambda\varepsilon \log \mathbb{E}_{u\sim Q^\nu_{\varepsilon,z}} \exp\left(\frac{f(u)}{\lambda\varepsilon}\right) = \lambda\varepsilon \log \int_{\mathcal{Z}} \exp\left(\frac{f(u)}{\lambda\varepsilon} - \frac{d(z,u)}{\varepsilon}\right) d\nu(u) - \lambda\varepsilon \log K^\nu_\varepsilon(z).$$

By Definition 3.1, $E^\nu_\varepsilon(z) = -\varepsilon \log K^\nu_\varepsilon(z)$, and hence

$$-\lambda\varepsilon \log K^\nu_\varepsilon(z) = \lambda E^\nu_\varepsilon(z).$$

Substituting this identity into the Sinkhorn-DRO dual objective from Lemma 2.3 and taking $\mathbb{E}_{v\sim\mu}[\cdot]$ proves the decomposition (8). The final statement follows because the normalization term defining $E^\nu_\varepsilon(v)$ appears explicitly in this decomposition of the robust log-moment term. □

C Comparison of Sample Complexity

We give discussion about sample complexity about neighborbood selection and our Sinkhorn outlier energy. To avoid confounding the comparison with pipeline differences, we keep all components fixed, and vary only the boundary scoring rule used in the sampling step. We first take k-NN as an example, then formalize the sample complexity by a hard neighborhood-based rule.

Theorem C.1 (Sample Complexity of Hard Neighborhood Selection)**.** *For any integer $k \geq 1$ and any $v \in \mathcal{Z}$, define the empirical k-NN radius by*

$$\widehat{r}_{n,k}(v) := \inf\left\{r > 0 : \widehat{\nu}_n(B(v,r)) \geq \frac{k}{n}\right\}, \qquad B(v,r) := \{u \in \mathcal{Z} : d(v,u) \leq r\}. \tag{13}$$

Assume that there exist constants $C > 0$, $m > 0$, and $r_0 > 0$ such that $\bar{\nu}(B(v,r)) \leq Cr^m$ for all $v \in \mathcal{V}$ and $r \in (0, r_0]$, where $\mathcal{V} \subseteq \mathcal{Z}$ is the set of candidate query points. Then for any fixed $v \in \mathcal{V}$ and any $0 < \Delta_r \leq r_0$, if $n < \frac{k}{2C\Delta_r^m}$, then $\mathbb{P}(\widehat{r}_{n,k}(v) > \Delta_r) \geq 1 - e^{-k/6}$.

Proof. Fix $v \in \mathcal{V}$ and define $N_n(v,r) := \sum_{i=1}^n \mathbf{1}\{z_i \in B(v,r)\}$. Then $N_n(v,r)$ is a binomial random variable with success probability $p_v(r) := \bar{\nu}(B(v,r))$. By assumption, for every $r \in (0, r_0]$, $p_v(r) \leq Cr^m$. Fix $0 < \Delta_r \leq r_0$. If $n < \frac{k}{2C\Delta_r^m}$, then

$$\mu_v := \mathbb{E}[N_n(v,\Delta_r)] = np_v(\Delta_r) \leq nC\Delta_r^m < \frac{k}{2}. \tag{14}$$

By the definition of $\widehat{r}_{n,k}(v)$, $\{\widehat{r}_{n,k}(v) \leq \Delta_r\} = \{N_n(v,\Delta_r) \geq k\}$. Since $\mu_v < k/2$, we have $k \geq 2\mu_v$. Applying the multiplicative Chernoff bound for the upper tail of a binomial random variable gives

$$\mathbb{P}(N_n(v,\Delta_r) \geq k) \leq \exp\left(-\frac{k-\mu_v}{3}\right) \leq e^{-k/6}. \tag{15}$$

Therefore,

$$\mathbb{P}(\widehat{r}_{n,k}(v) \leq \Delta_r) \leq e^{-k/6},$$

or equivalently,

$$\mathbb{P}(\widehat{r}_{n,k}(v) > \Delta_r) \geq 1 - e^{-k/6}.$$

This proves the claim for the fixed query point $v \in \mathcal{V}$. □

Theorem C.1 shows that, under the local mass upper bound $\bar{\nu}(B(v,r)) \leq Cr^m$, a hard k-NN rule cannot reliably resolve a neighborhood scale Δ_r unless the sample size is on the order of $k\Delta_r^{-m}$. This reflects the cost of estimating a local counting or radius event in weakly supported regions. The SOE rule is based on a different statistical object: for a fixed proposal pool, it asks whether the smooth Gibbs support cost crosses a threshold τ_E. The next theorem gives a finite-candidate stability guarantee for this threshold decision under explicit normalizer and margin conditions.

We next prove Theorem 3.4.

Theorem 3.4 (Finite-sample stability of SOE thresholding) *For a fixed finite candidate pool $\mathcal{V} \subset \mathcal{Z}$, define $\underline{K} := \inf_{v\in\mathcal{V}} K_\varepsilon^{\bar{\nu}}(v) > 0$, $\Delta_E = \inf_{v\in\mathcal{V}} |E_\varepsilon^{\bar{\nu}}(v) - \tau_E| > 0$, and $\tau_n(\delta) = \sqrt{\log(2|\mathcal{V}|/\delta)/2n}$. If $\tau_n(\delta) \leq \underline{K}/2$, then with probability at least $1-\delta$,*

$$\sup_{v\in\mathcal{V}} \left|E_\varepsilon^{\widehat{\nu}_n}(v) - E_\varepsilon^{\bar{\nu}}(v)\right| \leq \frac{2\varepsilon}{\underline{K}}\tau_n(\delta).$$

Consequently, if $\tau_n(\delta) \leq \min\left\{\frac{\underline{K}}{2}, \frac{\underline{K}\Delta_E}{4\varepsilon}\right\}$, then with probability at least $1-\delta$, $\mathbf{1}\{E_\varepsilon^{\widehat{\nu}_n}(v) \geq \tau_E\} = \mathbf{1}\ E_\varepsilon^{\bar{\nu}}(v) \geq \tau_E\}$ for any $v \in \mathcal{V}$. In particular, it suffices that

$$n \geq \max\left\{\frac{2}{\underline{K}^2}\log\frac{2|\mathcal{V}|}{\delta},\ \frac{8\varepsilon^2}{\underline{K}^2\Delta_E^2}\log\frac{2|\mathcal{V}|}{\delta}\right\}.$$

Proof. Fix a finite candidate pool $\mathcal{V} \subset \mathcal{Z}$, and define $\psi_v(u) := \exp\left(-\frac{d(v,u)}{\varepsilon}\right)$. Since $d(v,u) \geq 0$, we have $0 < \psi_v(u) \leq 1$ for all $u \in \mathcal{Z}$. By the definition of K_ε^ν,

$$K_\varepsilon^{\bar{\nu}}(v) = \int_{\mathcal{Z}} \exp\left(-\frac{d(v,u)}{\varepsilon}\right) d\bar{\nu}(u), \qquad K_\varepsilon^{\widehat{\nu}_n}(v) = \frac{1}{n}\sum_{i=1}^n \exp\left(-\frac{d(v,z_i)}{\varepsilon}\right).$$

Therefore, for each fixed $v \in \mathcal{V}$, the quantity $K_\varepsilon^{\widehat{\nu}_n}(v)$ is the empirical mean of n i.i.d. random variables in $[0,1]$ with mean $K_\varepsilon^{\bar{\nu}}(v)$.

By Hoeffding's inequality, for any $t > 0$ and any fixed $v \in \mathcal{V}$,

$$\mathbb{P}\Big(\Big|K_\varepsilon^{\widehat{\nu}_n}(v) - K_\varepsilon^{\bar{\nu}}(v)\Big| > t\Big) \le 2e^{-2nt^2}.$$

Applying the union bound over $v \in \mathcal{V}$ gives

$$\mathbb{P}\Big(\sup_{v\in\mathcal{V}}\Big|K_\varepsilon^{\widehat{\nu}_n}(v) - K_\varepsilon^{\bar{\nu}}(v)\Big| > t\Big) \le 2|\mathcal{V}|e^{-2nt^2}.$$

Choosing $t = \tau_n(\delta)$ yields

$$\mathbb{P}\Big(\sup_{v\in\mathcal{V}}\Big|K_\varepsilon^{\widehat{\nu}_n}(v) - K_\varepsilon^{\bar{\nu}}(v)\Big| \le \tau_n(\delta)\Big) \ge 1-\delta.$$

Hence, with probability at least $1-\delta$, the event $\mathcal{E}_n := \big\{\sup_{v\in\mathcal{V}}\big|K_\varepsilon^{\widehat{\nu}_n}(v) - K_\varepsilon^{\bar{\nu}}(v)\big| \le \tau_n(\delta)\big\}$ holds. We now work on $\mathcal{E}_n$. Since $\underline{K} := \inf_{v\in\mathcal{V}} K_\varepsilon^{\bar{\nu}}(v) > 0$, the condition $\tau_n(\delta) \le \underline{K}/2$ implies

$$K_\varepsilon^{\widehat{\nu}_n}(v) \ge K_\varepsilon^{\bar{\nu}}(v) - \tau_n(\delta) \ge \underline{K} - \tau_n(\delta) \ge \frac{\underline{K}}{2}, \qquad \forall v \in \mathcal{V}.$$

Also, since $0 < \psi_v(u) \le 1$, we have $K_\varepsilon^{\bar{\nu}}(v) \le 1$ and $K_\varepsilon^{\widehat{\nu}_n}(v) \le 1$. Thus both $K_\varepsilon^{\bar{\nu}}(v)$ and $K_\varepsilon^{\widehat{\nu}_n}(v)$ lie in $[\underline{K}/2, 1]$ for every $v \in \mathcal{V}$.

Consider the function $\phi(x) := -\varepsilon \log x$ on $(0,\infty)$. Its derivative satisfies $\phi'(x) = -\varepsilon/x$, so on $[\underline{K}/2, 1]$ we have

$$|\phi'(x)| \le \frac{2\varepsilon}{\underline{K}}.$$

Hence ϕ is $(2\varepsilon/\underline{K})$-Lipschitz on $[\underline{K}/2, 1]$. Using the definition of E_ε^ν, for every $v \in \mathcal{V}$,

$$\Big|E_\varepsilon^{\widehat{\nu}_n}(v) - E_\varepsilon^{\bar{\nu}}(v)\Big| = \Big|\phi(K_\varepsilon^{\widehat{\nu}_n}(v)) - \phi(K_\varepsilon^{\bar{\nu}}(v))\Big| \le \frac{2\varepsilon}{\underline{K}}\Big|K_\varepsilon^{\widehat{\nu}_n}(v) - K_\varepsilon^{\bar{\nu}}(v)\Big|.$$

Taking the supremum over $\mathcal{V}$ and using $\mathcal{E}_n$, we obtain

$$\sup_{v\in\mathcal{V}}\Big|E_\varepsilon^{\widehat{\nu}_n}(v) - E_\varepsilon^{\bar{\nu}}(v)\Big| \le \frac{2\varepsilon}{\underline{K}}\tau_n(\delta),$$

which proves the first claim.

For the second claim, assume in addition that $\frac{2\varepsilon}{\underline{K}}\tau_n(\delta) \le \frac{\Delta_E}{2}$ for $\Delta_E := \inf_{v\in\mathcal{V}}|E_\varepsilon^{\bar{\nu}}(v) - \tau_E| > 0$. Fix any $v \in \mathcal{V}$, if $E_\varepsilon^{\bar{\nu}}(v) \ge \tau_E$, then by the definition of Δ_E, we have

$$E_\varepsilon^{\bar{\nu}}(v) - \tau_E = \big|E_\varepsilon^{\bar{\nu}}(v) - \tau_E\big| \ge \Delta_E,$$

hence $E_\varepsilon^{\bar{\nu}}(v) \ge \tau_E + \Delta_E$. Using the uniform energy bound proved above,

$$E_\varepsilon^{\widehat{\nu}_n}(v) \ge E_\varepsilon^{\bar{\nu}}(v) - \Big|E_\varepsilon^{\widehat{\nu}_n}(v) - E_\varepsilon^{\bar{\nu}}(v)\Big| \ge (\tau_E + \Delta_E) - \frac{\Delta_E}{2} = \tau_E + \frac{\Delta_E}{2} > \tau_E.$$

Therefore, $\mathbf{1}\{E_\varepsilon^{\widehat{\nu}_n}(v) \ge \tau_E\} = 1$.

If instead $E_\varepsilon^{\bar{\nu}}(v) < \tau_E$, then $\tau_E - E_\varepsilon^{\bar{\nu}}(v) = |E_\varepsilon^{\bar{\nu}}(v) - \tau_E| \ge \Delta_E$, so $E_\varepsilon^{\bar{\nu}}(v) \le \tau_E - \Delta_E$. Again using the uniform energy bound,

$$E_\varepsilon^{\widehat{\nu}_n}(v) \le E_\varepsilon^{\bar{\nu}}(v) + \Big|E_\varepsilon^{\widehat{\nu}_n}(v) - E_\varepsilon^{\bar{\nu}}(v)\Big| \le (\tau_E - \Delta_E) + \frac{\Delta_E}{2} = \tau_E - \frac{\Delta_E}{2} < \tau_E.$$

Therefore,

$$\mathbf{1}\{E_\varepsilon^{\widehat{\nu}_n}(v) \ge \tau_E\} = 0.$$

Thus, on the event $\mathcal{E}_n$,

$$\mathbf{1}\{E_\varepsilon^{\widehat{\nu}_n}(v) \ge \tau_E\} = \mathbf{1}\{E_\varepsilon^{\bar{\nu}}(v) \ge \tau_E\}, \qquad \forall v \in \mathcal{V}.$$

Since $\mathbb{P}(\mathcal{E}_n) \ge 1-\delta$, the conclusion follows with probability at least $1-\delta$.

Finally, the conditions $\tau_n(\delta) \le \frac{\underline{K}}{2}$ and $\frac{2\varepsilon}{\underline{K}}\tau_n(\delta) \le \frac{\Delta_E}{2}$ are implied by

$$\tau_n(\delta) \le \min\left\{\frac{\underline{K}}{2}, \frac{\underline{K}\Delta_E}{4\varepsilon}\right\}.$$

Rearranging these two inequalities yields the sufficient sample-size bound

$$n \ge \max\left\{\frac{2}{\underline{K}^2}\log\frac{2|\mathcal{V}|}{\delta},\ \frac{8\varepsilon^2}{\underline{K}^2\Delta_E^2}\log\frac{2|\mathcal{V}|}{\delta}\right\}.$$

This completes the proof. □

D Population-SOE consistency of the sampler

We give the definitions and proof for Theorem 3.5. Unless stated otherwise, probabilities in this subsection are conditional on the observed bank and the fixed representation. The empirical and population-SOE samplers both use Algorithm 1, including redrawing the source anchor after an unsuccessful round. The population-SOE sampler replaces the energy evaluations, not the observed anchor candidates, by their population counterparts.

Threshold calibration. Let $s_i = E(z_i)$ and $\widehat{s}_i = \widehat{E}_n(z_i)$. Write the thresholds as $\tau_n^\star = \mathcal{T}_n(s_1, \dots, s_n)$ and $\widehat{\tau}_n = \mathcal{T}_n(\widehat{s}_1, \dots, \widehat{s}_n)$, where the common calibration map satisfies

$$|\mathcal{T}_n(u) - \mathcal{T}_n(v)| \le \|u - v\|_\infty. \tag{16}$$

A common fixed threshold satisfies this condition. So does a fixed empirical quantile, computed by an order statistic or a fixed interpolation of adjacent order statistics, plus a common additive margin. Indeed, perturbing every score by at most t perturbs every order statistic, and hence such a quantile, by at most t.

Anchor-selection margin. Let $1 \le A = A_n \le n$. Let $I_A^\star$ and $\widehat{I}_{A,n}$ contain the indices of the top-A population and empirical bank energies, respectively, using the same deterministic tie-breaking rule. For $A < n$, sort the population energies as $s_{(1)} \ge \cdots \ge s_{(n)}$ and define

$$\xi_{A,n} = \frac{s_{(A)} + s_{(A+1)}}{2}, \qquad a_n(t) = \min\left\{1, \frac{1}{2A}\sum_{i=1}^{n} \mathbf{1}\{|s_i - \xi_{A,n}| \le 2t\}\right\}.$$

For $A = n$, set $a_n(t) = 0$. This quantity bounds the fraction of the uniform anchor law that can change under an energy perturbation of size t.

Proposal-selection margin. A population proposal round first draws $I \sim \mathrm{Unif}(I_A^\star)$ and then draws

$$V_j = \Pi_{\mathcal{Z}}(z_I + \eta_j), \qquad \eta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_m), \qquad j = 1, \dots, M.$$

Let $R_n^\star$ denote the joint law of $(I, V_1, \dots, V_M)$. Define the semantic and population-feasible index sets by $S = \{j : \cos(V_j, t_{y_I}) \ge \tau_{\mathrm{sem}}\}$ and $J^\star = \{j \in S : E(V_j) \ge \tau_n^\star\}$.s Let Δ_n be the difference between the largest and second-largest values of $E(V_j)$ over $j \in J^\star$. Set $\Delta_n = +\infty$ when $|J^\star| \le 1$, and use $\min \varnothing = +\infty$. Define

$$b_n(t) = R_n^\star\left(\left\{\min_{j \in S} |E(V_j) - \tau_n^\star| \le 2t\right\} \cup \{\Delta_n \le 2t\}\right), \tag{17}$$

$$p_n^\star = R_n^\star(J^\star \ne \varnothing). \tag{18}$$

The factor $2t$ in the threshold event accounts for errors in both the candidate energy and the calibrated threshold. The second event accounts for instability of the energy-maximizing feasible proposal. No semantic-margin assumption is needed because the coupled samplers use the same semantic test on the same proposals.

Theorem 3.5 (Population-SOE consistency) *Suppose $|\widehat{\tau}_n - \tau_n^\star| \le e_n$, $p_n^\star \ge p_0 > 0$, and $a_n + b_n(e_n) < p_0$. Then the empirical sampler terminates almost surely, and*

$$d_{\mathrm{TV}}\big(Q_n^{\mathrm{SBOG}}, Q_{n,\mathrm{SOE}}^\star\big) \le \min\left\{1, \frac{2\{a_n + b_n(e_n)\}}{p_0}\right\} =: \zeta_n,$$

where $d_{\mathrm{TV}}(P, Q) = \sup_B |P(B) - Q(B)|$. The same bound holds after applying a shared conditional decoder. If $e_n \to 0$ and $a_n + b_n(e_n) \to 0$ in probability over the bank, and the threshold and acceptance conditions hold with probability tending to one, then $d_{\mathrm{TV}}(Q_n^{\mathrm{SBOG}}, Q_{n,\mathrm{SOE}}^\star) \to 0$ in probability.

Proof. Fix a bank satisfying the theorem's conditions and set $\eta_n = a_n(t) + b_n(t)$. We first bound disagreement in one proposal round and then account for repeated rejection.

Anchor selection. Let $r = |\widehat{I}_{A,n} \setminus I_A^\star| = |I_A^\star \setminus \widehat{I}_{A,n}|$. The total variation distance between the two uniform anchor laws is r/A. When $r > 0$, take any $j \in \widehat{I}_{A,n} \setminus I_A^\star$ and $k \in I_A^\star \setminus \widehat{I}_{A,n}$. The empirical ranking gives $\widehat{s}_j \geq \widehat{s}_k$. Therefore,

$$s_k - s_j \leq |s_k - \widehat{s}_k| + |\widehat{s}_j - s_j| \leq 2t.$$

Since $s_k \geq s_{(A)} \geq \xi_{A,n}$ and $s_j \leq s_{(A+1)} \leq \xi_{A,n}$, both indices lie in the band $|s_i - \xi_{A,n}| \leq 2t$. All $2r$ differing indices lie in this band, so

$$d_{\mathrm{TV}}\Big(\mathrm{Unif}(\widehat{I}_{A,n}), \mathrm{Unif}(I_A^\star)\Big) = \frac{r}{A} \leq a_n(t).$$

The claim is immediate when $A = n$. Couple the two anchor draws so that they differ with probability at most $a_n(t)$. Such a coupling assigns the common probability mass to identical indices and couples the remaining mass separately.

Feasibility and maximization. When the anchors agree, use the same Gaussian perturbations and projection in both samplers. Their semantic sets then agree. By (16), $|\widehat{\tau}_n - \tau_n^\star| \leq t$. Thus, for every shared proposal,

$$\Big|(\widehat{E}_n(V_j) - \widehat{\tau}_n) - (E(V_j) - \tau_n^\star)\Big| \leq 2t.$$

Outside the threshold event in (17), the feasible sets agree. If this common set contains at least two candidates and $\Delta_n > 2t$, its population maximizer $j^\star$ satisfies

$$\widehat{E}_n(V_{j^\star}) - \widehat{E}_n(V_j) \geq E(V_{j^\star}) - E(V_j) - 2t > 0 \qquad (j \neq j^\star).$$

Hence the selected candidate agrees as well. Agreement is immediate when the feasible set has at most one element.

Let W_n and $W^\star$ be the outputs of this coupled round, recording the accepted pair (y_I, V_j) or a failure symbol $\perp$. The preceding argument and a union bound give

$$\Pr(W_n \neq W^\star) \leq a_n(t) + b_n(t) = \eta_n.$$

Consequently, for every measurable set B of class–latent pairs,

$$|\mu_n(B) - \mu_n^\star(B)| \leq \eta_n, \qquad |p_n - p_n^\star| \leq \eta_n,$$

where $\mu_n(B) = \Pr(W_n \in B)$, $\mu_n^\star(B) = \Pr(W^\star \in B)$, and $p_n = \Pr(W_n \neq \perp)$.

Repeat-until-feasible sampling. Since $p_n \geq p_n^\star - \eta_n \geq p_0 - \eta_n > 0$, independent proposal rounds terminate almost surely. Summing over the number of failed rounds gives

$$Q_n^{\mathrm{SBOG}}(B) = \frac{\mu_n(B)}{p_n}, \qquad Q_{n,\mathrm{SOE}}^\star(B) = \frac{\mu_n^\star(B)}{p_n^\star}.$$

Therefore,

$$\begin{aligned}\big|Q_n^{\mathrm{SBOG}}(B) - Q_{n,\mathrm{SOE}}^\star(B)\big| &\leq \frac{|\mu_n(B) - \mu_n^\star(B)|}{p_n^\star} + \frac{\mu_n(B)}{p_n}\frac{|p_n - p_n^\star|}{p_n^\star} \\ &\leq \frac{2\eta_n}{p_n^\star} \leq \frac{2\eta_n}{p_0}.\end{aligned}$$

Taking the supremum over B and using $d_{\mathrm{TV}} \leq 1$ proves (D). The convergence statement follows on the stated events. For this statement, assign an arbitrary output law on exceptional banks where the sampler is undefined; their probability vanishes under the stated assumptions.

Decoding. Let $D(dx \mid c, v)$ be the shared conditional decoder, including $p(dx \mid v)$ as a special case. For every measurable input-space set B, the function $D(B \mid c, v)$ lies in $[0, 1]$. Integration of such a function against the difference of two probability measures is bounded by their total variation distance. Hence

$$d_{\mathrm{TV}}\big(Q_n^{\mathrm{SBOG}} D, Q_{n,\mathrm{SOE}}^\star D\big) \leq d_{\mathrm{TV}}\big(Q_n^{\mathrm{SBOG}}, Q_{n,\mathrm{SOE}}^\star\big).$$

This proves the decoded-distribution claim. □

A sufficient uniform estimation condition. We now give one setting in which the required uniform energy error vanishes. Fix the encoder before drawing an i.i.d. reference bank $z_1, \dots, z_n \sim \bar{\nu}$. Suppose $\mathcal{Z} \subset \mathbb{R}^m$ is compact, $d \geq 0$, and

$$|d(v,u) - d(w,u)| \leq L_d \|v - w\|_2 \quad \text{for all } u, v, w \in \mathcal{Z}.$$

Assume $K_0 = \inf_{v \in \mathcal{Z}} K_\varepsilon^{\bar{\nu}}(v) > 0$. Let $\mathcal{N}_{\mathcal{Z}}(h)$ be the size of a finite Euclidean h-net of $\mathcal{Z}$, and define

$$u_n(h,\delta) = \sqrt{\frac{\log\{2\mathcal{N}_{\mathcal{Z}}(h)/\delta\}}{2n}} + \frac{2L_d h}{\varepsilon}.$$

If $u_n(h,\delta) \leq K_0/2$, then, with probability at least $1-\delta$,

$$\sup_{v \in \mathcal{Z}} |\widehat{E}_n(v) - E(v)| \leq e_n(h,\delta) := \frac{2\varepsilon}{K_0} u_n(h,\delta). \tag{19}$$

To prove this claim, note that $u \mapsto \exp\{-d(v,u)/\varepsilon\}$ takes values in $[0,1]$. Hoeffding's inequality and a union bound over an h-net control the kernel-support error on that net by $\sqrt{\log\{2\mathcal{N}_{\mathcal{Z}}(h)/\delta\}/(2n)}$. The kernel is L_d/ε-Lipschitz in its first argument. Moving from any query to a net point contributes at most $2L_d h/\varepsilon$ to the difference between empirical and population supports. Thus their uniform difference is at most $u_n(h,\delta)$. Both supports are then at least $K_0/2$, where $x \mapsto -\varepsilon \log x$ is $2\varepsilon/K_0$-Lipschitz. This proves (19). Because the bound is uniform over $\mathcal{Z}$, it also applies to the data-dependent anchor locations.

For the normalized latent space, one may use $\mathcal{N}_{\mathcal{Z}}(h) \leq (1 + 2/h)^m$. Taking $h = n^{-1}$ and $\delta = n^{-2}$, with fixed m, ε, L_d, K_0, gives $e_n = O(\sqrt{\log n / n})$. Consistency also requires the selection-margin terms to vanish. For example, if, on events of probability tending to one,

$$a_n(t) + b_n(t) \leq C_{\mathrm{sel}} t^\alpha + v_n, \qquad \alpha > 0, \quad v_n \to 0, \quad p_n^\star \geq p_0 > 0,$$

for all sufficiently small t, then the theorem gives

$$d_{\mathrm{TV}}\big(Q_n^{\mathrm{SBOG}}, Q_{n,\mathrm{SOE}}^\star\big) \leq \frac{2}{p_0}\left(C_{\mathrm{sel}} e_n^\alpha + v_n\right) \to 0$$

on the corresponding events. These margin and acceptance conditions are additional assumptions, not consequences of uniform energy estimation alone. The concentration calculation above concerns an i.i.d. global reference bank; it does not by itself establish the same bound for data-dependent trimming, local-reference restrictions, or an encoder fitted on the same reference sample.

Boundary precision and coverage. The following consequence separates population-SOE consistency from alignment with an independently specified target boundary.

Corollary D.1 (Conditional boundary precision and coverage)**.** *Fix a bank for which $d_{\mathrm{TV}}(Q_n^{\mathrm{SBOG}}, Q_{n,\mathrm{SOE}}^\star) \leq \zeta_n$. For any measurable target boundary region $\mathcal{B}$,*

$$Q_n^{\mathrm{SBOG}}(\mathcal{B}) \geq Q_{n,\mathrm{SOE}}^\star(\mathcal{B}) - \zeta_n.$$

In particular, define $\mathrm{BPrec}(Q) = Q(\mathcal{B})$. If $Q_{n,\mathrm{SOE}}^\star(\mathcal{B}) = 1$, then $\mathrm{BPrec}(Q_n^{\mathrm{SBOG}}) \geq 1 - \zeta_n$.

Let $\mathcal{B}_1, \dots, \mathcal{B}_L$ be measurable target regions such that $\min_{\ell \leq L} Q_{n,\mathrm{SOE}}^\star(\mathcal{B}_\ell) \geq q_{\min} > \zeta_n$. For N independent outputs $U_1, \dots, U_N$ from the fitted empirical sampler, conditional on this bank,

$$\Pr(\textit{every } \mathcal{B}_\ell \textit{ contains at least one } U_j \mid \textit{bank}) \geq 1 - L\exp\{-N(q_{\min} - \zeta_n)\}.$$

The same statements hold for the decoded laws and input-space regions when the common decoder is applied.

Proof. The precision and per-region mass bounds follow directly from the definition of total variation. Conditional independence gives

$$\Pr(U_j \notin \mathcal{B}_\ell \text{ for all } j \mid \text{bank}) = \{1 - Q_n^{\mathrm{SBOG}}(\mathcal{B}_\ell)\}^N \leq \exp\{-N(q_{\min} - \zeta_n)\}.$$

A union bound over the L regions proves the coverage statement. The decoded version follows from the decoder contraction established in Theorem 3.5. □

The corollary is conditional on the fitted bank; outputs that share this bank are not asserted to be independent unconditionally. The coverage step uses only total variation and the assumed region masses. It does not establish that the SOE-selected regions coincide with a task-defined boundary. Also, $Q^{\star}_{n,\mathrm{SOE}}$ depends on the observed bank. Convergence to a fixed, sample-independent law $Q^{\star}_{\mathrm{SOE}}$ would additionally require convergence of these population-SOE counterparts. For example, an additional bound $d_{\mathrm{TV}}(Q^{\star}_{n,\mathrm{SOE}}, Q^{\star}_{\mathrm{SOE}}) \le \lambda_n \to 0$ would give the bound $\zeta_n + \lambda_n$ to that fixed law by the triangle inequality.

E Controlled Numerical Validation

Downstream detection performance provides only indirect evidence of generation quality. We therefore first examine whether SBOG generates boundary outliers in a controlled setting where the target boundary is known and defined independently of SOE. The population oracle uses three stable vector autoregressive (VAR) processes with perturbations Δb_c, $\Delta\Sigma_c$, and ΔA_c to their local means, innovation covariances, and transition dynamics, respectively. Using the corresponding exact population log-likelihood ratio $r_{c,k}(x)$, the target boundary band is $0 < r_{c,k}(x) \le \delta_{c,k}$, where the indices identify the process and perturbation, and $\delta_{c,k}$ determines the selected band. This oracle is distinct from the population-SOE sampler in Theorem 3.5.

We compare SBOG with RandomNeg, kNNNeg, and KDENeg. We report boundary precision (BPrec), semantic validity (SemValid), average and worst-region boundary coverage (BCov and WorstCov), sliced Wasserstein-2 distance to the oracle (SW2), and detection AUPR on held-out oracle samples. Together, these measurements assess boundary alignment, coverage, and downstream utility.

Table 7: Controlled boundary-generation evaluation against a method-independent population oracle.

Method	BPrec↑	SemValid↑	BCov↑	WorstCov↑	SW2-to-oracle↓	Oracle AUPR↑
RandomNeg	0.229±0.027	0.950±0.001	0.091±0.036	0.011±0.004	0.2349±0.0034	0.455±0.005
kNNNeg	0.298±0.036	0.977±0.014	0.033±0.005	0.002±0.002	0.1453±0.0062	0.684±0.014
KDENeg	0.305±0.043	0.972±0.011	0.070±0.008	0.005±0.003	0.1489±0.0089	0.830±0.012
SBOG	**0.629±0.032**	**0.983±0.013**	**0.236±0.006**	**0.012±0.003**	**0.1405±0.0069**	**0.896±0.052**

Table 7 shows that SBOG places more samples in the independently defined boundary band and achieves greater average boundary coverage than the competing generators, while maintaining high semantic validity. Its generated samples also support higher AUPR on held-out oracle data. Worst-region coverage nevertheless remains low, so these results do not establish uniform coverage of all boundary regions. The experiment provides direct evidence of boundary alignment in this controlled setting, complementing rather than directly verifying the population-SOE consistency result.

F Time-Series Experimental Details

F.1 Pipeline parameters

We keep the CARLA [10] pipeline fixed throughout Section 4.1 and only swap the negative-sample construction module. All five datasets (SMD, MSL, SMAP, SWaT, WADI) share the same encoder architecture and optimization schedule; the only per-dataset quantity is the input channel count C, which equals the number of variates in the benchmark.

Pretext stage (representation learning). We use a 1-D ResNet encoder with mid-channel width 4 and an MLP projection head of output dimension 4, so the latent space $\mathcal{Z}$ is a 4-D unit sphere after ℓ_2 normalization. Pretext training uses the contrastive objective with temperature $\tau_{\mathrm{ctr}} = 0.4$, batch size 50, AdamW with $\mathrm{lr} = 10^{-3}$ and weight decay 10^{-2} with cosine schedule for 30 epochs, and the augmentation pipeline of [10] with noise $\sigma_{\mathrm{aug}} = 0.01$. The pseudo-label space has $K = 10$ classes obtained by the standard CARLA neighborhood-clustering procedure.

Negative-sample construction (SOE sampler). For all five time-series benchmarks, we use the same SOE sampler hyperparameters. We set the Sinkhorn regularization to $\varepsilon = 0.02$ and define the

energy threshold as $\tau_E = \rho_q + m$, where ρ_q is the 95-th percentile of the refined ID energies and the margin is $m = 0.04$. For proposal generation, we use isotropic Gaussian noise with standard deviation $\sigma = 0.02$ and draw $M = 128$ proposals per anchor. The anchor pool contains the top-A high-energy ID points, with $A = \lceil 0.02n \rceil$ capped at 256 and lower bounded by 8 for small entities. When constructing the reference measure $\widehat{\nu}_n$, we trim the top 8% highest-energy ID points to reduce boundary contamination.

Classification stage (downstream detector). We use the same ResNet backbone, 5-NN soft-assignment, 50 epochs of Adam (lr $= 10^{-2}$, weight decay 10^{-3}) with constant schedule, batch size 50, entropy weight 2.0, and inconsistency weight 0. Anomaly contamination of the training pool is fixed at 0.99. Detection scores are reported point-wise following [28].

F.2 Quality and stability metrics

For every (sampler, entity) pair we generate negatives $\{v_i\}_{i=1}^N \subset \mathcal{Z}$ and evaluate them against the in-distribution embeddings Z_n and pseudo-labels $\{y_i\}$ on the *same* latent space the encoder produces. All metrics are computed on the 4-D unit sphere with cosine ground cost. Let $z(v)$ denote the 1-NN of v in Z_n under $d_{\cos}$, $\mathrm{src}(v_i)$ the index of the boundary anchor that proposed v_i, and $\mathcal{N}_k(v)$ its k nearest ID neighbors.

Energy gap (Table 2). $\mathrm{Gap} = \widehat{E}(v) - \tau_E$, averaged over all generated negatives. Measures by how much each accepted negative exceeds the acceptance threshold. Larger gaps mean negatives sit deeper in the high-energy outlier region, but as τ_E rises into the right tail of the ID energy distribution the gap shrinks because τ_E rises faster than $\widehat{E}(v)$ can.

Feasible fill ratio (Table 2). Fraction of ID anchors with at least one of M proposals satisfying $\widehat{E}(v_j) \geq \tau_E$,

$$\mathrm{Fill} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\Big\{\exists j \in [M] :\ \widehat{E}(v_j^{(i)}) \geq \tau_E\Big\}. \tag{20}$$

MMD^2 **(RBF, median heuristic, Figure 2).** Squared maximum mean discrepancy between the negative and ID samples,

$$\mathrm{MMD}^2(\widehat{P}_v, \widehat{\nu}_n) = \frac{1}{N^2}\sum_{i,j} k(v_i, v_j) + \frac{1}{n^2}\sum_{i,j} k(z_i, z_j) - \frac{2}{Nn}\sum_{i,j} k(v_i, z_j), \tag{21}$$

with Gaussian kernel $k(a, b) = \exp(-\gamma\|a - b\|_2^2)$ and bandwidth $\gamma = 1/(2\sigma_{\mathrm{med}}^2)$, where σ_{med} is the median pairwise Euclidean distance on a random 1024-point subset of the union $\widehat{P}_v \cup \widehat{\nu}_n$. Both samples are subsampled to 2048 points before evaluation. *Higher is better*: large MMD^2 means the negatives form a distribution that is detectable as different from ID.

Sliced Wasserstein-2 distance (Figure 2). The 2-Wasserstein distance averaged over 1000 random unit-norm 1-D projections,

$$\widehat{\mathrm{SW}}_2(\widehat{P}_v, \widehat{\nu}_n) = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L} W_2^2\Big(\theta_\ell^\top \widehat{P}_v, \theta_\ell^\top \widehat{\nu}_n\Big)}, \quad L = 1000, \tag{22}$$

with W_2 on each line computed in $O(n \log n)$ via sorting.

Intra-negative diversity (Figure 2). Mean pairwise cosine distance among generated negatives,

$$\mathrm{Diversity} = \frac{1}{N(N-1)}\sum_{i \neq j} \big(1 - v_i^\top v_j\big). \tag{23}$$

Anchor fidelity (intended-class purity, Figure 2). Fraction of negatives whose 1-NN in Z_n shares the pseudo-label of the boundary anchor that proposed them,

$$\mathrm{Fidelity} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big\{y_{z(v_i)} \ = \ y_{\mathrm{src}(v_i)}\big\}. \tag{24}$$

Contamination (Table 7). $\text{Contam.} = 1 - \text{Fidelity}$.

Cross-class margin (Table 7). Let μ_c be the (re-normalized) mean of Z_n within pseudo-class c. For each negative v_i with source class $c_i^\star = y_{\text{src}(v_i)}$,

$$\text{Margin}(v_i) = \cos(v_i, \mu_{c_i^\star}) - \max_{c \neq c_i^\star} \cos(v_i, \mu_c), \qquad \text{Margin} = \frac{1}{N} \sum_i \text{Margin}(v_i).$$

Table 8: Component ablation of the TS outlier-synthesis sampler.

Variant	Fidelity ↑	Contam. ↓	Margin ↑	MMD^2 ↑	Diversity ↑
Random	0.339 ± 0.014	0.661 ± 0.014	-0.000 ± 0.002	0.140 ± 0.043	0.051 ± 0.074
w/o top-A anchor	0.329 ± 0.025	0.671 ± 0.025	0.000 ± 0.002	0.467 ± 0.121	0.223 ± 0.062
w/o ρ threshold	0.988 ± 0.004	0.012 ± 0.004	0.034 ± 0.038	1.076 ± 0.227	0.182 ± 0.026
w/o argmax select	0.989 ± 0.004	0.011 ± 0.004	0.034 ± 0.039	1.049 ± 0.247	0.099 ± 0.022
k-NN	0.990 ± 0.031	0.010 ± 0.031	0.025 ± 0.030	0.934 ± 0.301	0.195 ± 0.027
KDE	0.992 ± 0.028	0.008 ± 0.028	0.028 ± 0.032	0.978 ± 0.265	0.198 ± 0.024
SBOG (ours)	0.999 ± 0.004	0.001 ± 0.004	0.082 ± 0.038	1.156 ± 0.227	0.202 ± 0.026

G Image Experimental Details

G.1 Configuration

In-distribution datasets and classifiers. We follow the experimental protocol of recent image outlier-synthesis works [57, 13, 17, 36] and the benchmark setting of [23]. The ID classifier is ResNet-34 [20], trained on CIFAR-100 and ImageNet-100 [11] with the standard cross-entropy objective. The latent representation $z(x;\theta) \in \mathbb{R}^{768}$ is the row-normalised penultimate feature.

Class text anchors. Class anchors $\{t_c\}_{c=1}^C$ are the unit-normalised text-token embeddings produced by the text encoder of Stable Diffusion v1.5 on each class name.

ID feature bank. For each class c we cache 1000 ID features, giving an ID tensor $Z_n \in \mathbb{R}^{C \times 1000 \times 768}$. All energies, anchors and candidates are computed in this normalised cosine space.

Parameters in outlier energy.

- Smoothing parameter $\varepsilon = 0.05$.
- Threshold τ_E calibrated per class as the 0.95-quantile of $\widehat{E}(z_i)$ over Z_n, plus a margin m: $\tau_E = \rho + m$ with $m = 0$ as default.
- Gaussian perturbation scale $\sigma = 0.012$.
- Each anchor proposes $M = 256$ candidates per round, up to $\texttt{max_rounds} = 10$ rounds, until the per-class accepted set reaches $N_{\text{ood}} = 10000$.
- Local-energy reference = the 128 nearest ID neighbours of each anchor; seed-cosine floor 0.7.

Decoder. Accepted latents are decoded back to image space by Stable Diffusion v1.5 conditioned on the class token of the source anchor; one image per accepted latent. The decoded set $\mathcal{D}_{\text{ood}}$ is used as auxiliary outlier exposure during detector calibration.

G.2 Definition of Diagnostic Metrics

This subsection collects the precise definitions of every metric that appears in the diagnostic tables and figures of Section 4.2 (Tables 4, 6, 8 and Figure 3). Throughout, $Z_n = \{z_i\}_{i=1}^n \subset \mathcal{Z}$ denotes the ID feature bank for the class under consideration, $t_c \in \mathcal{Z}$ its class text anchor, $\widehat{E}(\cdot) = E_\varepsilon^{\widehat{\nu}_n}(\cdot)$ the empirical Sinkhorn outlier energy of Definition 3.1 with smoothing ε, and $\tau_E = \rho + m$ the energy threshold of Algorithm 1. Per-class quantities are reported as mean $\pm$ standard deviation across the 100 ImageNet-100 classes.

Empirical decision margin Δ (Figure 4). For each candidate v in the accepted candidate pool we compute $|\widehat{E}(v) - \tau_E|$, and report

$$\Delta \;=\; \mathrm{median}\big\{|\widehat{E}(v) - \tau_E| : v \in \mathcal{V}\big\}. \tag{25}$$

A larger Δ means accepted/rejected candidates sit further from the threshold, so each individual decision is more confident. This is the quantity controlled by Theorem 3.4: the sample complexity scales as $\varepsilon^2/(\underline{K}^2\Delta_E^2)$, so Δ acts as the operationally observable proxy for Δ_E.

Flip Rate@256 (Figure 4). For each class we draw a calibration subset S of size $n = 256$ uniformly without replacement from Z_n, recompute the empirical energy $\widehat{E}^{(S)}$ from S alone, and re-derive a class-specific threshold $\tau_E^{(S)}$ by the same 0.95-quantile rule. The flip rate is

$$\mathrm{FlipRate@256} \;=\; \mathbb{E}_S\left[\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbf{1}\Big\{\mathbf{1}\{\widehat{E}^{(S)}(v) \geq \tau_E^{(S)}\} \;\neq\; \mathbf{1}\{\widehat{E}(v) \geq \tau_E\}\Big\}\right], \tag{26}$$

averaged over 30 random calibration subsets. Smaller is better: flip rate $\to 0$ means small calibration sets reach the same accept/reject decisions as the full ID bank, the regime predicted by Theorem 3.4 when the margin Δ_E is well-conditioned.

Token cosine (Table 4). Mean cosine similarity to the class text token,

$$\overline{\cos}_{\mathrm{tok}} \;=\; \frac{1}{N}\sum_{j=1}^{N}\langle v_j, t_c\rangle. \tag{27}$$

Diversity (Table 4). One minus the mean off-diagonal cosine of the candidate set,

$$\mathrm{Diversity} \;=\; 1 \;-\; \frac{1}{N(N-1)}\sum_{i\neq j}\langle v_i, v_j\rangle. \tag{28}$$

A larger value means candidates spread more uniformly along the boundary rather than collapsing to a few directions.

Yield (Table 4). The product of the on-class pass rate and the mean Sinkhorn outlier energy of the passing subset:

$$\mathrm{Yield} \;=\; \mathrm{Pass}\cdot\overline{E}_{\mathrm{pass}},$$

where, for each anchor, 32 Gaussian perturbations of scale $\sigma = 0.012$ are drawn, $\mathrm{Pass} = \Pr[\cos(v, t_c) \geq \tau_{\mathrm{sem}}]$ is the fraction that survives the semantic filter, and $\overline{E}_{\mathrm{pass}}$ is the mean Sinkhorn outlier energy of those passing perturbations. Yield captures both axes simultaneously: a strategy must produce on-class proposals *and* those proposals must be high-energy.

Intended-class purity (Table 6). Fraction of candidates whose nearest ID neighbour over *all* C classes lies in the intended class c:

$$\mathrm{Purity}_c \;=\; \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\left\{\arg\max_{i\in\bigcup_{c'} Z_n^{c'}}\langle v_j, z_i\rangle \;\in\; Z_n^c\right\}, \tag{29}$$

where $Z_n^{c'}$ is the ID bank of class c' (we use 150 samples per class for this nearest-neighbour bank). Higher = the candidate stays inside the correct class manifold.

Margin vs. other classes (Table 6).

$$\mathrm{Margin} \;=\; \frac{1}{N}\sum_{j=1}^{N}\Big[\langle v_j, t_c\rangle - \max_{c'\neq c}\langle v_j, t_{c'}\rangle\Big]. \tag{30}$$

Contamination (Table 6). Fraction of candidates whose *nearest class token* is not the intended class:

$$\mathrm{Contamination} \;=\; \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\Big\{\arg\max_{c'}\langle v_j, t_{c'}\rangle \neq c\Big\}. \tag{31}$$

Fidelity. Mean cosine similarity between each generated negative and its source anchor's pseudo-class prototype $\mu_{y(v_j)}$, averaged over $\mathcal{N}$. High fidelity means generated negatives are still recognisably class-conditional rather than generic noise.

Contamination. Per-negative defined as $\mathrm{Contam}(v_j) = \mathbf{1}\{\arg\max_{c'} \langle v_j, \mu_{c'} \rangle \neq y(v_j)\}$, averaged over $\mathcal{N}$. Lower is better; together with Fidelity it satisfies $\mathrm{Fidelity} + \mathrm{Contam} \approx 1$ in the matched-prototype regime.

Margin. $\mathrm{Margin}(v_j) = \langle v_j, \mu_{y(v_j)} \rangle - \max_{c' \neq y(v_j)} \langle v_j, \mu_{c'} \rangle$, averaged over $\mathcal{N}$. Positive margin = negatives sit closer to their pseudo-class than to any other.

MMD^2. Squared maximum-mean-discrepancy between the generated-negative distribution $\widehat{P}_{\mathcal{N}}$ and the empirical positive distribution $\widehat{P}_{Z_n}$,

$$\mathrm{MMD}^2(\widehat{P}_{\mathcal{N}}, \widehat{P}_{Z_n}) \;=\; \mathbb{E}\big[k(v, v')\big] + \mathbb{E}\big[k(z, z')\big] - 2\mathbb{E}\big[k(v, z)\big],$$

with a Gaussian RBF kernel $k(\cdot, \cdot)$ at the median heuristic bandwidth. Larger = generated negatives induce a stronger distributional shift away from positives, which is desirable for contrastive negative mining.

Intra-negative diversity. $1 - \frac{1}{N(N-1)} \sum_{i \neq j} \langle v_i, v_j \rangle$, the same formula as "Diversity" in Table 4 above, applied to the time-series generated-negative set. Higher = negatives cover a wider range of boundary directions instead of clustering on a single axis.

The last result shows the factors that influence the quality of the generation. Consider the two smoothing parameter ϵ in the outlier energy $E(v) = -\epsilon \log \int \exp\big(-d(v, u)/\epsilon\big) d\nu(u)$, and the energy margin m in the acceptance rule $E(v) \geq \rho + m$. The four panels in Figure 5 respectively summarize their influence on energy separation (Energy Gap and Energy Strength), semantic preservation $\cos(v, t_c)$ (Semantic Alignment), and sampling feasibility (Feasible Fill Ratio).

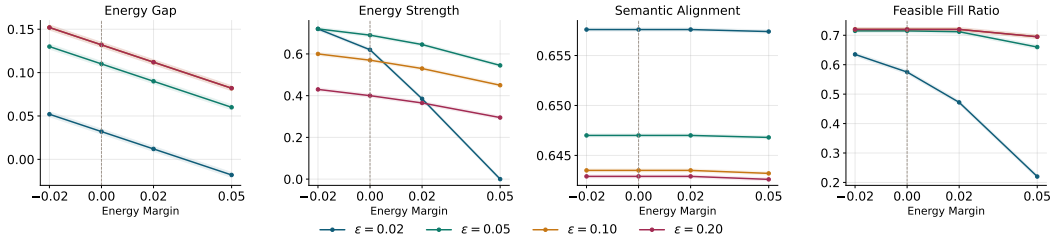

Figure 5: Generation quality validation

As m increases, the acceptance condition becomes stricter, which consistently reduces the energy advantage and the number of feasible synthesized outliers; this effect is especially severe for very small ϵ, where the energy landscape becomes overly sharp and the feasible fill ratio drops rapidly. By contrast, semantic alignment remains nearly unchanged across settings, suggesting that the dominant trade-off is not semantic drift but the balance between outlier strength and synthesis stability. Overall, the results indicate that a moderate ϵ together with a small margin provides the most reliable setting, preserving strong outlierness while maintaining stable and feasible synthesis.

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 abstract and introduction accurately summarize SBOG, SOE, the Sinkhorn-DRO connection, and the experiments on time-series and image outlier generation.

   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: The paper discusses limitations in the conclusion, especially about the precision of learning latent features.

   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: The theoretical assumptions are stated with the main lemmas, proposition, theorem, and the corresponding complete proofs are included in the 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: The paper provides the main algorithm and detailed experimental settings in the appendices, including datasets, architectures, training hyperparameters, evaluation metrics, and compute resources.

   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 uploaded the code into supplementary materials, and will release the full version if the paper got accepted.

   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: The paper provides the main algorithm and detailed experimental settings in the appendices, including datasets, architectures, training hyperparameters, evaluation metrics, and compute resources.

   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: We have reported the variance for our results in Table 3 and 4.

   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: All our experiments are conducted on a machine equipped with 13th Gen Intel(R) Core(TM) i9-13900HX CPU (24 cores) and NVIDIA A100 GPU (40GB).

   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 work uses standard public benchmarks and synthetic outlier generation for robustness evaluation, with no identified deviation from 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: [N/A]

    Justification: This is a pure machine learning paper and does not any incoporate social impacts.

    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: The paper does not release pre-trained models, generative systems, or scraped 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: The Wadi and SwaT dataset need to be obtained from `https://www.sutd.edu.sg/itrust/`.

    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: [N/A]

    Justification: The paper does not appear to release a new dataset, model, or code asset requiring separate documentation.

    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: The paper does not involve crowdsourcing experiments or research with human subjects.

    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: The paper does not involve human-subject research.

    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: LLM is only used for polishing the draft.

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.