

ReFP-AD: Rectified Flow Preconditioning for Energy-Based Anomaly Detection

Camile Lendering, Erkut Akdag, Joaquín Figueira, and Egor Bondarev

AIMS Group, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
`c.r.lendering@tue.nl`

Abstract. Unified anomaly detection requires modeling highly heterogeneous normal data without access to anomalous samples. While foundation models like DINOv2 provide rich token representations, leveraging these spaces for explicit density estimation remains challenging. Energy-Based Models (EBMs) offer a principled formulation, but their training in high-dimensional token spaces is unstable due to anisotropy and strong cross-dimensional correlations, which degrades finite-step Markov Chain Monte Carlo (MCMC) sampling. We identify this instability as fundamentally geometric and introduce ReFP-AD (Rectified Flow Preconditioning for Anomaly Detection), which learns a geometric reparameterization that maps high-dimensional embeddings into a well-conditioned latent space via an optimal transport (OT)-coupled rectified flow. This preconditioning enables stable persistent contrastive divergence with preconditioned Stochastic Gradient Langevin Dynamics (SGLD) in full-dimensional token spaces. Anomaly scores are then derived from the learned energy landscape using gradient norms. Under a strict unified protocol on the MVTec-AD and VisA datasets, ReFP-AD achieves 98.6%/97.9% Image/Pixel AUROC on MVTec-AD and 97.3%/99.0% on VisA, outperforming prior unified EBM baselines by up to +10.8% in Image AUROC. Ablation experiments demonstrate that geometric reparameterization is critical for finite-step MCMC and accurate anomaly localization in high-dimensional token spaces. Code is available at <https://github.com/CLendering/ReFP-AD>

Keywords: Anomaly Detection · Energy-Based Models

1 Introduction

Visual anomaly detection aims to identify inputs that deviate from the distribution of normal data without access to anomalous examples during training. This setting arises across industrial inspection, autonomous systems, and scientific imaging, where models must reliably detect rare and previously unseen deviations. In the *unified* anomaly detection setting, this challenge is further amplified: a single model must generalize across diverse object categories, textures, and appearance variations. Modeling such heterogeneous, yet exclusively normal visual distributions makes it difficult to learn stable decision boundaries in high-dimensional feature spaces.

Energy-based models (EBMs) offer a principled framework for modeling complex data distributions in high-dimensional spaces. Unlike likelihood-based generative models, they assign a scalar energy $E(x)$ to each input and define an unnormalized density $p(x) \propto \exp(-E(x))$ [19], enabling flexible representation of multi-modal distributions [8] without restrictive architectural assumptions. For unsupervised anomaly detection, this formulation is compelling: anomalies are simply identified as high-energy inputs relative to the normal manifold.

In practice, EBM training requires sampling from the model distribution to approximate maximum-likelihood gradients, typically via Langevin dynamics [18] or related Markov chain Monte Carlo (MCMC) methods [33]. In high-dimensional spaces, these dynamics mix slowly and are sensitive to initialization, step size, and local curvature. As a result, practical EBM training relies on finite-step, non-equilibrium dynamics, including persistent contrastive divergence (PCD) [42]. Crucially, the stability of short-run sampling is determined not only by the energy function itself, but also by the geometry of the space in which sampling is performed.

In parallel, modern self-supervised Vision Transformers (ViTs) produce token embeddings that organize visual data into semantically structured manifolds [3, 34]. These representations have proven highly effective for anomaly detection, often with simple similarity- or distance-based scoring [37]. However, treating token embeddings as fixed feature spaces leaves explicit generative density modeling largely unexplored. While applying EBMs in token space is conceptually appealing, training with MCMC-based negative sampling remains sensitive and often unstable, especially under limited-step sampling and complex feature geometries [7, 33].

This limitation is fundamentally geometric rather than architectural. Standard Langevin dynamics corresponds to diffusion under a fixed Euclidean metric with isotropic noise [28]. In contrast, transformer token distributions are highly anisotropic and strongly correlated across dimensions, violating the underlying Euclidean assumption [35]. Under finite-step sampling, this metric mismatch yields poorly conditioned trajectories and unstable negative phases, necessitating delicate hyperparameter tuning. Consequently, prior EBMs often rely on compressing visual representations into low-dimensional bottlenecks [46], sacrificing rich token-level semantics and limiting unified modeling capacity, to stabilize Markov chains.

To address this limitation, *ReFP-AD* is proposed as a novel strategy to explicitly reshape the sampling geometry *before* energy-based training. By leveraging the rectified flow formulation [26], our method learns an optimal transport (OT)-coupled map that transforms tokens into a well-conditioned latent coordinate system. Then, persistent contrastive divergence with preconditioned SGLD [21] is performed in transported coordinates, effectively inducing a data-adaptive geometry under which the finite-step sampling becomes stable and predictable. Importantly, the energy model itself remains unconstrained; the intervention operates entirely at the level of representation geometry.

Rather than modifying the finite-step sampling procedure itself, the geometric intervention bridges the mismatch between high-dimensional semantic token

spaces (e.g., 1536-D features for DINOv2-G) and finite-step MCMC. The rectified flow is optimized purely as a preconditioner, with model selection guided by finite-step Langevin diagnostics (conditioning, residual correlation, tail stability). This enables stable high-dimensional EBM training without dimensionality reduction.

Experiments on MVTec-AD and VisA demonstrate that the proposed framework achieves strong performance among density-based anomaly detectors. To summarize, the main contributions of this work are as follows:

- A geometric diagnosis of unified token-space EBM instability, where anisotropy and correlation in foundation-token representations violate the metric assumptions of standard Langevin dynamics.
- **ReFP-AD**: a novel geometric reparameterization that maps high-dimensional vision tokens into an isotropic space conditioned for stable finite-step MCMC, enabling unconstrained EBMs in visual foundation-token spaces.
- A comprehensive unified evaluation, demonstrating strong detection and localization performance (98.6/97.9% Image/Pixel AUROC on MVTec-AD and 97.3/99.0% on VisA), including targeted ablations isolating the impact of geometric conditioning under fixed sampling budgets.

2 Related Work

Visual anomaly detection in pretrained feature spaces. Most modern industrial visual anomaly detection (VAD) methods avoid explicit generative modeling and, instead, score anomalies directly in pretrained feature spaces. Retrieval and distribution-based approaches like PatchCore [37] and PaDiM [6] remain strong baselines due to their reliable localization on benchmarks including MVTec-AD [2], VisA [51], and Real-IAD [43]. Subsequent work has emphasized practical deployment constraints, including real-time inference (EfficientAD [1]) and robustness to contaminated or noisy normal training data (SoftPatch [16]). However, these approaches rely on feature-space statistics rather than explicit probabilistic density modeling.

Foundation models and unified anomaly detection. Self-supervised Vision Transformers (ViTs) produce token embeddings with rich semantic and spatial structure [3, 34], further improving anomaly detection in feature-space. In few-shot regimes, simple token-level scoring on DINOv2 already yields competitive performance, whether via patch similarity (AnomalyDINO [5]) or subspace reconstruction residuals (SubspaceAD [20]), indicating that pretrained tokens encode strong normality priors. Recent studies extend this paradigm to unified evaluation protocols spanning heterogeneous regimes, including semantic out-of-distribution (OOD) shifts and industrial defects (GeneralAD [40]). Vision-language models further expand this paradigm to category-scalable and zero-/few-shot settings through prompt alignment (WinCLIP [15], AnomalyCLIP [50], PromptAD [22]). Despite enabling open-set detection, they similarly default to discriminative scoring instead of explicit density estimation.

Density modeling in feature space. A complementary line of work models densities over intermediate representations using normalizing flows, thereby enabling tractable likelihood-based detection and localization. Methods, such as DifferNet [38], FastFlow [47], and CFLOW-AD [12], estimate feature-space likelihoods via invertible transformations, while diffusion-based approaches adopt reconstruction or denoising objectives for anomaly detection [13, 29, 49]. More recent unified diffusion/flow-matching frameworks (e.g., DTG [44]) introduce temporally adaptive guidance for multi-class anomaly detection under generative reconstruction settings. However, likelihood- and reconstruction-based models are known to assign high likelihoods or faithful reconstructions to OOD inputs [17, 30], highlighting sensitivity to representation geometry and objective design. In contrast, the proposed ReFP-AD model does not employ flows or diffusion models as generative estimators; instead a learned transport map is used as a geometric preconditioner to reshape the representation space prior to energy-based training.

Energy-based models and non-equilibrium training. Energy-based models define unnormalized densities via scalar energy functions and are trained by lowering energy on data, while raising it on negative samples drawn from the model [19]. In practice, maximum-likelihood gradients are approximated by short-run Langevin dynamics, persistent contrastive divergence (PCD), or replay buffers. Prior analyses show that finite-step, non-equilibrium sampling choices, such as step size, number of updates, and initialization strategy, can qualitatively affect the learned energy landscape [32]. EBMs have also been applied to OOD detection through energy scores derived from discriminative models (Energy-OOD [25]) and classifier-as-EBM formulations (JEM [11]). Recent unified EBM approaches, including MPDR [46], stabilize MCMC by collapsing visual representations into low-dimensional CNN feature vectors, trading representational richness for sampling stability. In contrast, the proposed ReFP-AD approach preserves full-dimensional foundation-token representations and addresses instability at the level of sampling geometry.

Transport maps and geometry-aware preconditioning. Sampling efficiency in high-dimensional correlated spaces is strongly influenced by geometry, motivating adaptive preconditioning and learned transport maps that transform targets into better-conditioned coordinates (e.g., NeuTra HMC [14], latent-space Langevin for flows [31], and EBM-flow hybrids [9]). Rectified flow and flow matching provide stable ways to learn vector-field transports [24, 26], and have recently been used directly for tabular anomaly detection via one-step contraction/deviation scoring [23]. In contrast, ReFP-AD does not use flow matching as the detector: the rectified flow only preconditions high-dimensional visual tokens for finite-step PCD, while anomaly scores come from the learned EBM.

3 Method

Figure 1 outlines our approach: standardized ViT tokens in z -space are first mapped into a well-conditioned latent u -space via OT-coupled rectified flow

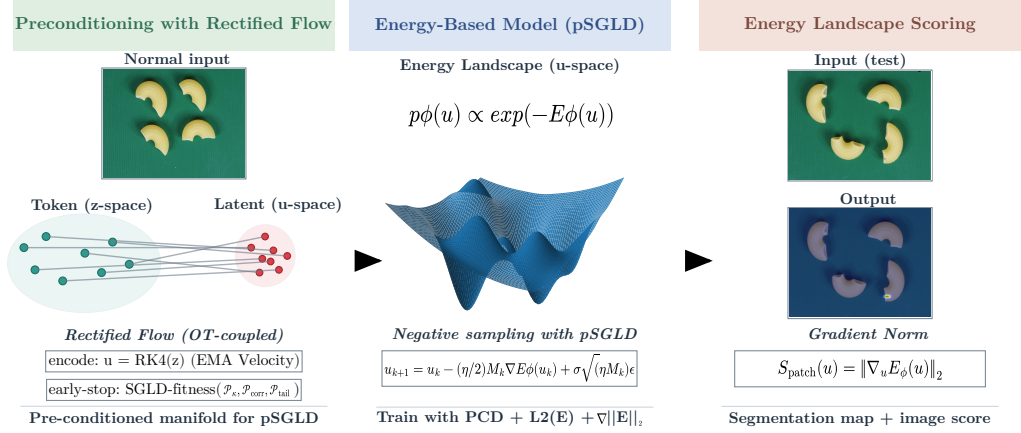


Fig. 1: Overview of **ReFP-AD**. Standardized DINOv2 tokens in z space are transported to a better-conditioned latent space u via an OT-coupled rectified flow, with checkpoint selection guided by MCMC-oriented geometric diagnostics. An unconstrained EBM $E_\phi(u)$ is trained using PCD with pSGLD, and anomalies are scored at inference by the gradient norm $\|\nabla_u E_\phi(u)\|_2$ under a fixed finite-step sampling budget.

preconditioning, after which an unconstrained EBM is trained with PCD using pSGLD. This decouples representation geometry from density estimation: native token distributions are highly anisotropic and strongly correlated, so finite-step Langevin dynamics mixes poorly and destabilizes the negative phase. By learning a continuous-time transport map that reshapes tokens into approximately isotropic coordinates, both EBM training and gradient-norm anomaly scoring are performed entirely in the transported u -space, yielding stable MCMC behavior without architectural bottlenecks.

3.1 Unified Token Representation and Standardization

Given an input image, a frozen foundation ViT (DINOv2 ViT-G/14) produces a spatial grid of patch tokens $z_i \in \mathbb{R}^D$, where $D = 1536$. To capture both low-level structure and high-level semantics, representations from multiple deep transformer layers are aggregated via mean pooling. Specifically, for each spatial patch index i , the corresponding tokens are averaged across a selected set of transformer layers $\mathcal{L}$, excluding the CLS token.

In the unified setting, token distributions across object categories can differ substantially in scale. Without normalization, these magnitude disparities dominate the multi-class optimization landscape and obscure the underlying geometric structure. To address this, tokens are Z-score standardized per category using normal training data:

$$z \leftarrow \frac{z - \mu_c}{\sigma_c}, \quad (1)$$

where $\mu_c, \sigma_c \in \mathbb{R}^D$ are per-category feature mean and standard deviation estimated from normal training tokens, and the division is element-wise (with a small constant added to σ_c for numerical stability). This transformation constitutes a fixed affine pre-processing step. Subsequently, both the rectified flow and EBM are entirely unified, shared models that do not condition on category identity beyond this initial standardization.

3.2 Geometric Preconditioning via Optimal Transport Flow

Even after Z-score standardization, high-dimensional foundation-model token embeddings remain highly anisotropic, exhibiting strong cross-feature correlations and large covariance condition numbers. In such anisotropic spaces, MCMC literature establishes that finite-step Langevin dynamics produces inefficient, highly oscillatory “zig-zag” trajectories, where gradients dominate along narrow dimensions and mixing remains poor along elongated manifold directions [14, 21].

To eliminate this geometric limitation, a transport map is learned to reshape the token distribution into an isotropic geometry prior to EBM training. The target reference distribution is defined as an isotropic Gaussian $p_{\text{prior}}(u) = \mathcal{N}(0, \tau^2 I)$, where the temperature τ controls the scale of the latent space and the contraction–expansion balance of the transport, influencing conditioning and finite-step MCMC stability.

The transport is parameterized by a rectified flow [26], defined by the ordinary differential equation (ODE) $\dot{x}_t = v_\theta(t, x_t)$. To obtain geometrically consistent source–target pairings, data tokens z and Gaussian targets $u \sim \mathcal{N}(0, \tau^2 I)$ are coupled via an entropic optimal transport (OT)-coupling $\pi(z, u)$. Since exact OT computations are numerically unstable in high dimensions (1536-D), the coupling π is computed by stable log-domain Sinkhorn iterations [4]. Given the linear interpolation $x_t = (1 - t)z + tu$, the velocity network is optimized via the continuous-time matching objective:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{t \sim \mathcal{U}[0,1], (z,u) \sim \pi} \left[\|v_\theta(t, x_t) - (u - z)\|_2^2 \right]. \quad (2)$$

Network weights are tracked by an Exponential Moving Average (EMA) [41] to ensure stable integration. At inference, tokens are transported to u by integrating the ODE from $t=0$ to $t=1$ via a 4th-order Runge-Kutta (RK4) solver [36].

3.3 SGLD-Oriented Manifold Validation

A low flow-matching loss $\mathcal{L}_{\text{RF}}$ ensures accurate vector-field regression, but does not guarantee improved *finite-step* Langevin behavior. Because the transport is introduced as a preconditioner for short-run PCD, flow checkpoints are selected using diagnostics that directly reflect known MCMC failure modes: ill-conditioning, residual correlation, and heavy-tailed instability.

Full covariance estimation in $D=1536$ is numerically unstable; therefore, transported tokens u are projected onto a fixed random orthogonal subspace of dimension $k=128$, which preserves second-order structure in expectation

while reducing estimator variance. Let C denote the covariance of projected tokens. The following diagnostics are computed: (i) an anisotropy penalty $\mathcal{P}_\kappa = \log(\lambda_{\max}(C)/\lambda_{\min}(C))$ to reflect conditioning; (ii) a correlation penalty $\mathcal{P}_{\text{corr}}$, defined as the mean squared off-diagonal entries of the corresponding correlation matrix, to quantify residual cross-dimensional dependencies; and (iii) a tail penalty $\mathcal{P}_{\text{tail}} = \text{q}_{0.99}(\|u\|_2)/\text{median}(\|u\|_2)$ to control heavy-tailed outliers that destabilize SGLD updates.

These terms are combined into an SGLD-fitness score, represented by

$$\mathcal{F}(u) = \mathcal{P}_\kappa + \frac{1}{2}\mathcal{P}_{\text{corr}} + \frac{1}{4}\log(\mathcal{P}_{\text{tail}}), \quad (3)$$

where the coefficients are fixed globally across all experiments. The weighting reflects the relative impact of each failure mode on short-run Langevin dynamics: conditioning typically dominates mixing behavior, residual correlations have secondary influence, and heavy-tail instability affects occasional but destabilizing updates. The coefficients are selected once and kept constant; no dataset-specific tuning is performed.

To prevent degenerate manifold collapse, we enforce structural preservation via the Spearman rank correlation ρ between pairwise distances in standardized space z and transported space u . The optimal checkpoint is selected by solving $t^* = \arg \min_t \{\mathcal{F}(u_t) \mid \rho_t \geq 0.6\}$. This threshold ensures MCMC stability is prioritized only within a feasible region of topological integrity. All coefficients and the guardrail $\rho \geq 0.6$ are fixed across all datasets to avoid over-tuning. Visualizations are provided in Suppl. C.

3.4 Unconstrained EBM with pSGLD

After transporting standardized tokens z into the latent space u , density modeling is performed with a residual MLP energy function $E_\phi(u)$ with parameters ϕ , defining the unnormalized density $p_\phi(u) \propto \exp(-E_\phi(u))$.

Training follows persistent contrastive divergence (PCD) [42]. To mitigate mode imbalance in the unified setting, the replay buffer employs vectorized stratified sampling across categories. Negative samples are updated using preconditioned Stochastic Gradient Langevin Dynamics (pSGLD) [21]:

$$u_{k+1} = u_k - \frac{\eta}{2} M_k \nabla_u E_\phi(u_k) + \sigma \sqrt{\eta M_k} \xi_k, \quad \xi_k \sim \mathcal{N}(0, I), \quad (4)$$

where u_k denotes the k -th Langevin iterate, η is the step size, σ the noise scale, ξ_k standard Gaussian noise, and M_k a positive diagonal preconditioning matrix.

The preconditioner is computed from an RMSProp-style moving average of squared gradients:

$$v_k = \beta v_{k-1} + (1 - \beta) (\nabla_u E_\phi(u_k))^2, \quad M_k = (\sqrt{v_k} + \epsilon)^{-1}, \quad (5)$$

where v_k is the running second-moment estimate, $\beta \in [0, 1)$ the decay factor, and $\epsilon > 0$ a small constant for numerical stability.

The EBM is optimized with the contrastive objective

$$\mathcal{L}_{\text{EBM}} = \mathbb{E}_{u^+}[E_\phi(u)] - \mathbb{E}_{u^-}[E_\phi(u)] + \alpha \mathbb{E}_{u^\pm}[E_\phi(u)^2] + \lambda \mathbb{E}_{u^+}[\|\nabla_u E_\phi(u)\|_2^2], \quad (6)$$

where u^+ and u^- denote positive samples from the data distribution and negative samples from the replay buffer, respectively. $\mathbb{E}$ denotes expectation over the corresponding sample sets. The coefficients α and λ control energy magnitude regularization and gradient smoothness.

3.5 Energy Landscape Scoring

During inference, normal samples concentrate near low-energy minima, whereas anomalies occupy higher-energy states. Therefore, anomaly scoring is derived directly from the learned energy landscape, by the energy gradient norm as a local deviation measure:

$$S_{\text{patch}}(u) = \|\nabla_u E_\phi(u)\|_2, \quad (7)$$

which computes the magnitude of the “restoring force” required to move a sample toward the learned normal manifold [11]. Due to the well-conditioned latent space induced by ReFP-AD, the raw energy gradient norm provides a stable and informative signal for out-of-distribution deviation, reducing reliance on explicit priors or iterative refinement heuristics.

The obtained scores are projected onto the 2D token grid, bilinearly upsampled, and smoothed with a Gaussian filter ($\sigma=4.0$). To ensure robustness to localized noise spikes, the image-level anomaly score is computed as the mean of the top 1% of the upsampled pixel scores [5]. While this gradient-based score forms our primary detection signal, a complementary ablation (Suppl. A) shows that the transported token magnitude $\|u\|_2$ already provides a strong baseline, but explicit energy modeling improves robustness on challenging categories.

4 Experiments and Results

This section evaluates *ReFP-AD* under the strict unified protocol on MVTec-AD and VisA. Main results, ablations, and qualitative localization are reported. Additional unified results on MVTec-AD 2 and Real-IAD are reported in Suppl. D.

4.1 Experimental Setup

Datasets and the Unified Evaluation Protocol. The proposed ReFP-AD method is evaluated on two industrial anomaly detection benchmarks: MVTec-AD [2] (15 categories) and VisA [51] (12 categories). A strict *unified* setting is adopted, where a *single shared* rectified flow and a *single shared* EBM are trained jointly on all normal images across categories, without per-category networks, heads, or separate training runs. Category identity is limited to fixed input standardization (Sec. 3.1) and for balanced replay-buffer initialization, reflecting industrial scenarios where category labels are available at test time.

Implementation Details. Input images are resized to 700×700 and center-cropped to 672×672 . Patch tokens (size 14×14) are extracted by a frozen DINOv2 ViT-G/14 backbone, aggregating representations from the middle seven transformer layers (layers 22 to 28) via mean pooling to obtain 1536-dimensional feature vectors. The rectified flow is parameterized by an 8-layer MLP with hidden dimension 1024. It is trained for up to 150 epochs with batch size 8192 via the Adam optimizer and a learning rate of 5×10^{-5} . The log-domain Sinkhorn OT-coupling employs an entropic regularization of 0.01, and the target Gaussian noise temperature is set to $\tau = 0.1$. To ensure integration stability, an Exponential Moving Average (EMA) with decay 0.999 is applied to the velocity network weights. Flow optimization utilizes dynamic early stopping based on the SGLD-fitness criterion with a patience of 20 epochs. At inference stage, flow integration exploits a 10-step 4th-order Runge-Kutta (RK4) solver.

The unconstrained EBM is a 3-layer residual MLP with a hidden dimension of 1024. Training is performed for 15 epochs with batch size 8192 using Adam (learning rate 10^{-5} , weight decay 10^{-5} , with $\gamma = 0.4$ decay at epochs 8 and 12). The objective is stabilized with an L_2 energy penalty ($\alpha = 0.1$) and a gradient norm penalty ($\lambda = 10.0$). The PCD replay buffer contains 400,000 states and employs a 5% stratified re-initialization rate from the data manifold to mitigate multi-class mode collapse. Negative sampling executes 60 steps of preconditioned SGLD per update with step size of 10^{-3} , momentum $\beta = 0.99$, and noise standard deviation 0.1.

4.2 Unified Anomaly Detection Performance

While recent unified detectors often rely on similarity or reconstruction scoring in pretrained feature spaces, this work focuses on density-based modeling with EBMs; therefore, our primary comparisons are restricted to explicit density estimators, with a contextual comparison to recent reconstruction-based methods provided in Suppl. B. For context, the evaluated methods are grouped into general baseline approaches, normalizing-flow-based methods, and energy-based models.

The proposed ReFP-AD model achieves the strongest performance among the methods considered in Table 1, with particularly large gains over prior EBM baselines in the unified VisA setting. Notably, it outperforms the leading normalizing-flow-based model, HGAD [45], achieving 98.6% Image AUROC on MVTec-AD and 97.3% on the heterogeneous VisA dataset. These results indicate that, when the representation space is properly conditioned, explicit density estimation can surpass hierarchical likelihood formulations in a unified multi-class setting.

A substantial margin is also observed over prior EBM approaches that stabilize training through recovery-based objectives anchored to autoencoder manifolds. MPDR [46] serves as a representative baseline of this paradigm, employing a recovery-likelihood formulation in which sampling is regularized by a reconstruction-fidelity term that constrains MCMC near a low-dimensional autoencoder manifold (272-D CNN features in the unified setting). In the unified VisA reproduction with the official MPDR implementation, the default SGLD

Table 1: Unified anomaly detection and localization performance on MVTec-AD and VisA, reported as Image AUROC / Pixel AUROC (%). Baseline results are from HGAD [45] and stabilized MPDR[†] [46].

Method	MVTec-AD		VisA	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC
<i>Baseline Methods</i>				
PaDiM [6]	84.2	89.5	86.8	97.0
MKD [39]	81.9	84.9	74.2	93.9
DRAEM [48]	88.1	87.2	85.5	90.5
<i>Normalizing Flow Based Methods</i>				
FastFlow [47]	91.8	96.0	77.2	95.1
CFLOW [12]	89.0	94.0	88.0	95.9
HGAD [45]	<u>98.4</u>	97.9	<u>97.1</u>	<u>98.9</u>
<i>Energy-Based Models</i>				
EBM (Genc et al.) [10]	72.0	70.7	-	-
MPDR [†] [46]	96.0	96.7	86.5	96.5
ReFP-AD (Ours)	98.6	97.9	97.3	99.0

configuration was numerically unstable and resulted in near-random performance. To avoid underestimating the baseline, a stabilized MPDR variant (†) is reported, obtained by tuning only SGLD hyperparameters to ensure convergence, yielding 86.5% Image AUROC.

ReFP-AD improves upon the tuned MPDR baseline by +10.8% on VisA, highlighting the importance of geometric conditioning for stable high-dimensional EBM training. This improvement suggests that, in unified high-resolution settings, stability obtained through manifold-anchored recovery objectives alone is insufficient when the underlying representation geometry remains highly anisotropic. By directly conditioning the representation geometry, ReFP-AD enables stable sampling without architectural compression or reconstruction-based energy terms, while preserving the structured token information required for accurate localization and multi-class modeling.

4.3 Ablation Studies

To validate the core components of the proposed method, ablation studies are conducted on the MVTec-AD and VisA datasets. These experiments isolate the impact of geometric preconditioning, unified vs. per-category modeling, the number of SGLD sampling steps, backbone scale, and sampling preconditioning. Additional robustness analyses for DINOv3-7B features and input resolution are provided in Suppl. G and Suppl. F.

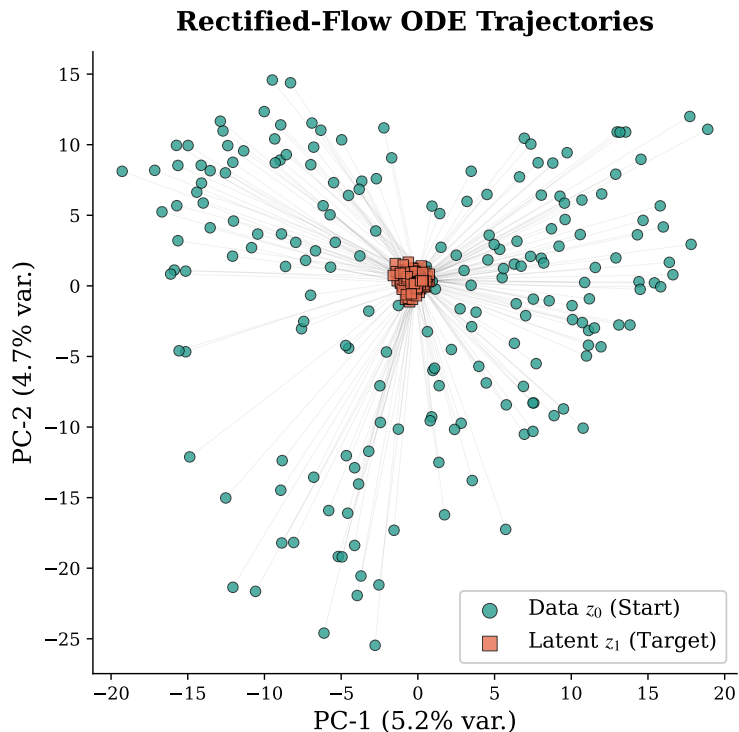


Fig. 2: PCA projection of rectified flow trajectories. Axes correspond to the first two principal components (PC-1: 5.2% variance, PC-2: 4.7% variance). The transport map transforms the correlated, heavy-tailed token distribution (z) into a well-conditioned latent space (u) with reduced anisotropy and correlations, improving finite-step Langevin dynamics.

Effect of Rectified Flow Preconditioning. The fundamental hypothesis of this work is that raw foundation-token spaces are geometrically unsuitable for finite-step Langevin dynamics. To evaluate this, the unconstrained EBM is trained directly on the standardized 1536-dimensional DINOv2 token space, without optimal transport preconditioning. Removing the geometric preconditioning yields a substantial degradation in performance, particularly on the heterogeneous VisA dataset, where the image-level AUROC drops from 97.3% down to 87.0% ($\Delta - 10.3\%$) and pixel-level AUROC decreases to 96.6% ($\Delta - 2.4\%$). On the comparatively simpler MVTec-AD benchmark, detection performance also declines significantly from 98.6% to 91.1% ($\Delta - 7.5\%$). Figure 2 illustrates how the rectified flow reshapes, decorrelates, and isotropizes the empirical token manifold. In the raw anisotropic space, finite-step Langevin dynamics mixes poorly, leading to unstable negative-phase updates and collapsed energy basins. These results demonstrate that geometric preconditioning is essential for stable EBM training in high-dimensional dense foundation-token representations.

Unified vs. Per-Category Modeling. Conventional industrial anomaly detection models are typically trained per-category, operating under a simplified, unimodal optimization landscape. Table 2 compares the proposed unified model (ReFP-AD) against its per-category variant, as well as against SimpleNet [27], a competitive per-category baseline.

Table 2: Comparison of Per-Category vs. Unified modeling. Performance is reported as Image AUROC (%). The proposed unified model remains highly competitive with dedicated per-category models, exhibiting only marginal performance degradation despite the increased complexity of modeling all categories simultaneously.

Method	Setting	MVTec-AD	VisA
SimpleNet [27]	Per-Category	99.6	96.9
ReFP-AD	Per-Category	99.2	97.7
ReFP-AD	Unified	98.6	97.3
<i>Unified Δ (vs. Per-Cat)</i>	-	-0.6	-0.4

The results indicate a marginal performance gap between the dedicated per-category networks and the unified model. Specifically, the unified model trails its per-category counterpart by only 0.6% on MVTEC-AD and 0.4% on the more complex VisA dataset. In the context of generative anomaly detection, where forcing a single model to capture 12 to 15 distinct semantic manifolds, often increasing the risk of mode imbalance and degraded density boundaries, this minimal degradation is notable.

Effect of pSGLD Sampling Steps. Energy-based models are known to be highly sensitive to the number of Langevin steps applied during the negative phase, often requiring hundreds of iterations to mix properly in high-dimensional spaces [33]. However, the proposed ReFP-AD maps the data into an isotropic, well-conditioned latent space, thereby enabling rapid mixing under preconditioned SGLD.

Table 3: Ablation on the number of preconditioned SGLD steps (K) during the EBM negative phase on the unified MVTEC-AD benchmark. The flow-preconditioned manifold enables rapid MCMC mixing, stabilizing detection performance in as few as 20 steps.

SGLD Steps (K)	10	20	40	80
Image AUROC (%)	54.7	98.2	98.5	98.6
Pixel AUROC (%)	67.8	94.1	97.9	97.9

As shown in Table 3, ablating the number of SGLD steps ($K \in \{10, 20, 40, 80\}$) on the unified MVTec-AD benchmark reveals rapid performance saturation. With only $K = 10$ steps, the Markov chains fail to mix sufficiently, resulting in a collapsed image AUROC of 54.7%. However, increasing it to $K = 20$ steps leads to a substantial improvement, with image-level performance rising to 98.2%. By $K = 40$ steps, pixel-level localization fully stabilizes (98.5% I-AUROC, 97.9% P-AUROC), while further increasing the chains to $K = 80$ steps yields negligible deviations (98.6% I-AUROC, 97.9% P-AUROC). These results demonstrate that geometric preconditioning substantially improves the sampling conditioning, so stable negative-phase updates can be obtained with relatively few finite-step pSGLD iterations.

Backbone Scalability. To analyze the robustness of the ReFP-AD model to backbone capacity, we evaluate its performance under different DINOv2 variants. DINOv2 ViT-G/14 is replaced with ViT-L/14 and ViT-B/14, while all other components are kept fixed. As shown in Table 4, performance degrades only moderately with reduced backbone size, indicating that the proposed preconditioning and EBM training remain stable across token spaces ranging from 768D–1536D.

Table 4: Ablation of backbone scale on unified Image AUROC and Pixel AUROC. The proposed ReFP-AD maintains stable performance across feature dimensions of 768D, 1024D, and 1536D.

Backbone	Dimension	MVTec-AD		VisA	
		I-AUROC	P-AUROC	I-AUROC	P-AUROC
DINOv2 ViT-B/14	768	97.7	97.6	95.7	98.8
DINOv2 ViT-L/14	1024	97.6	97.6	96.7	98.9
DINOv2 ViT-G/14	1536	98.6	97.9	97.3	99.0

Necessity of SGLD Preconditioning. Table 5 compares preconditioned SGLD (pSGLD) [21] with standard SGLD on the VisA dataset. Removing the diagonal preconditioner leads to a significant drop in image AUROC (97.3→75.3), indicating that while transport improves global conditioning, adaptive local scaling is still required for stable negative-phase mixing in the learned energy landscape. Both samplers are tuned over the same step-size/noise search range under the unified VisA setting.

4.4 Qualitative Analysis

Figure 3 (and extended results in Suppl. H) shows anomaly maps obtained via gradient-norm scoring. Operating on dense transported tokens yields sharp

Table 5: Ablation of the sampling dynamics on the unified VisA benchmark. Comparing standard SGLD with the proposed preconditioned SGLD (pSGLD). The adaptive preconditioner is essential for maintaining stable detection performance across complex, multi-class energy landscapes.

Sampling Method	I-AUROC	P-AUROC
Standard SGLD	75.3	94.3
Preconditioned SGLD (pSGLD)	97.3	99.0

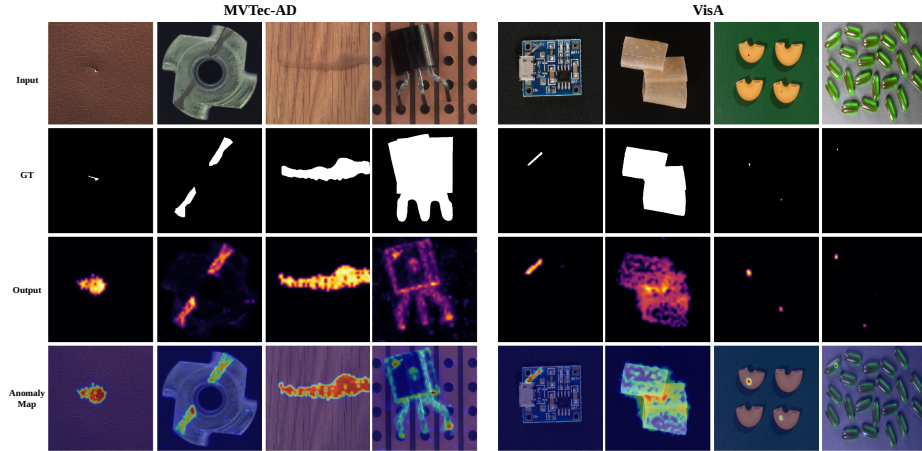


Fig. 3: Qualitative localization results on MVTec-AD and VisA datasets. From top to bottom: input image, ground-truth mask, predicted anomaly map, and overlay. The proposed model ReFP-AD produces well-localized anomaly maps that closely align with defect regions, while maintaining low scores on normal backgrounds.

localization of both subtle texture and structural defects, while suppressing background noise by concentrating normal regions near low-energy basins.

4.5 Limitations and Computational Overhead

While ReFP-AD stabilizes energy-based density modeling in high-dimensional token spaces, it increases inference cost. In contrast to feature-based detectors, such as SimpleNet [27] and PatchCore [37], which operate primarily in a feedforward manner or rely on efficient k-NN retrieval, the proposed method transports tokens with a 10-step RK4 solver and computes the gradient-norm score $S_{\text{patch}}(u) = \|\nabla_u E_\phi(u)\|_2$ via a backward pass. Although both the rectified flow and energy networks are lightweight MLPs, the combination of ODE integration and gradient evaluation leads to increased latency. Consequently, the method is better suited to offline inspection or settings where unified modeling outweighs strict real-time constraints. Latency benchmarks are reported in Suppl. E.

5 Conclusion

This work shows that the key limitation in training EBMs on modern foundation-token representations is not model capacity, but *sampling geometry*. In high-dimensional ViT token spaces, strong anisotropy and cross-dimensional correlations violate the assumptions underlying finite-step Langevin dynamics, making PCD unstable and forcing prior unified EBMs to rely on architectural bottlenecks or recovery constraints. *ReFP-AD* addresses this limitation through a geometric reparameterization: an OT-coupled rectified flow transports standardized tokens into a well-conditioned latent space in which short-run MCMC becomes stable. This enables *unconstrained* energy learning directly on dense high-dimensional DINOv2 token representations under a strict unified protocol, yielding strong detection and localization performance on the MVTec-AD and VisA datasets. In particular, ReFP-AD achieves 98.6%/97.9% Image/Pixel AUROC on MVTec-AD and 97.3%/99.0% on VisA, improving unified EBM baselines by up to +10.8% Image AUROC.

Beyond these results, ReFP-AD suggests a general principle for scaling energy-based modeling to foundation representations: *first condition the representation geometry, then learn the energy*. Future work includes distilling the preconditioned energy landscape into an efficient feed-forward scorer to bridge the gap between stable unified modeling and latency-critical applications.

Acknowledgement

This work is supported by the ADVISOR ITEA 241007 project.

References

1. Batzner, K., Heckler, L., König, R.: Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 128–138 (2024) [3](#)
2. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9592–9600 (2019) [3](#), [8](#)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) [2](#), [3](#)
4. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* **26** (2013) [6](#)
5. Damm, S., Laszkiewicz, M., Lederer, J., Fischer, A.: Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1319–1329. IEEE (2025) [3](#), [8](#)
6. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: International conference on pattern recognition. pp. 475–489. Springer (2021) [3](#), [10](#)

7. Du, Y., Li, S., Tenenbaum, J., Mordatch, I.: Improved contrastive divergence training of energy based models. arXiv preprint arXiv:2012.01316 (2020) [2](#)
8. Du, Y., Mordatch, I.: Implicit generation and modeling with energy based models. *Advances in neural information processing systems* **32** (2019) [2](#)
9. Gao, R., Nijkamp, E., Kingma, D.P., Xu, Z., Dai, A.M., Wu, Y.N.: Flow contrastive estimation of energy-based models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7518–7528 (2020) [4](#)
10. Genc, E.U., Ahuja, N., Ndiour, I.J., Tickoo, O.: Energy-based anomaly detection and localization. arXiv preprint arXiv:2105.03270 (2021) [10](#)
11. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. arXiv preprint arXiv:1912.03263 (2019) [4](#), [8](#)
12. Gudovskiy, D., Ishizaka, S., Kozuka, K.: Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 98–107 (2022) [4](#), [10](#)
13. He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., Xie, L.: A diffusion-based framework for multi-class anomaly detection. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 38, pp. 8472–8480 (2024) [4](#)
14. Hoffman, M., Sountsov, P., Dillon, J.V., Langmore, I., Tran, D., Vasudevan, S.: Neutra-lizing bad geometry in hamiltonian monte carlo using neural transport. arXiv preprint arXiv:1903.03704 (2019) [4](#), [6](#)
15. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 19606–19616 (2023) [3](#)
16. Jiang, X., Liu, J., Wang, J., Nie, Q., Wu, K., Liu, Y., Wang, C., Zheng, F.: Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems* **35**, 15433–15445 (2022) [3](#)
17. Kirichenko, P., Izmailov, P., Wilson, A.G.: Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems* **33**, 20578–20589 (2020) [4](#)
18. Langevin, P.: On the theory of brownian motion. *CR Acad Sci (Paris)* **146**, 530 (1908) [2](#)
19. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F., et al.: A tutorial on energy-based learning. *Predicting structured data* **1**(0) (2006) [2](#), [4](#)
20. Lendering, C., Akdag, E., Bondarau, E.: Subspacead: Training-free few-shot anomaly detection via subspace modeling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 28557–28566 (2026) [3](#)
21. Li, C., Chen, C., Carlson, D., Carin, L.: Preconditioned stochastic gradient langevin dynamics for deep neural networks. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 30 (2016) [2](#), [6](#), [7](#), [13](#)
22. Li, X., Zhang, Z., Tan, X., Chen, C., Qu, Y., Xie, Y., Ma, L.: Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16838–16848 (2024) [3](#)
23. Li, Z., Huang, Q., Zhu, Y., Yang, L., Mohammadi Amiri, M., van Stein, N., van Leeuwen, M.: Scalable, explainable and provably robust anomaly detection with one-step flow matching. *Advances in Neural Information Processing Systems* **38**, 87834–87899 (2026) [4](#)

24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) [4](#)
25. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. *Advances in neural information processing systems* **33**, 21464–21475 (2020) [4](#)
26. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022) [2](#), [4](#), [6](#)
27. Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 20402–20411 (2023) [12](#), [14](#)
28. Ma, Y.A., Chen, T., Fox, E.: A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems* **28** (2015) [2](#)
29. Mousakhan, A., Brox, T., Tayyub, J.: Anomaly detection with conditioned denoising diffusion models. In: *DAGM German Conference on Pattern Recognition*. pp. 181–195. Springer (2024) [4](#)
30. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don’t know? arXiv preprint arXiv:1810.09136 (2018) [4](#)
31. Nijkamp, E., Gao, R., Sountsov, P., Vasudevan, S., Pang, B., Zhu, S.C., Wu, Y.N.: Mcmc should mix: Learning energy-based model with neural transport latent space mcmc. arXiv preprint arXiv:2006.06897 (2020) [4](#)
32. Nijkamp, E., Hill, M., Han, T., Zhu, S.C., Wu, Y.N.: On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 34, pp. 5272–5280 (2020) [4](#)
33. Nijkamp, E., Hill, M., Zhu, S.C., Wu, Y.N.: Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems* **32** (2019) [2](#), [12](#)
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) [2](#), [3](#)
35. Park, N., Kim, S.: How do vision transformers work? In: *International Conference on Learning Representations* (2022), <https://openreview.net/forum?id=D78Go4hVcx0> [2](#)
36. Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**(4), 341–363 (December 1996). <https://doi.org/10.2307/3318418>, <https://projecteuclid.org/journals/bernoulli/volume-2/issue-4/Exponential-convergence-of-Langevin-distributions-and-their-discrete-approximations/10.2307/3318418.full> [6](#)
37. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14318–14328 (2022) [2](#), [3](#), [14](#)
38. Rudolph, M., Wandt, B., Rosenhahn, B.: Same same but differnet: Semi-supervised defect detection with normalizing flows. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 1907–1916 (2021) [4](#)
39. Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M.H., Rabiee, H.R.: Multiresolution knowledge distillation for anomaly detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14902–14912 (2021) [10](#)
40. Sträter, L.P., Salehi, M., Gavves, E., Snoek, C.G., Asano, Y.M.: Generalad: Anomaly detection across domains by attending to distorted features. In: *European conference on computer vision*. pp. 448–465. Springer (2024) [3](#)

41. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* **30** (2017) 6
42. Tieleman, T.: Training restricted boltzmann machines using approximations to the likelihood gradient. In: *Proceedings of the 25th international conference on Machine learning*. pp. 1064–1071 (2008) 2, 7
43. Wang, C., Zhu, W., Gao, B.B., Gan, Z., Zhang, J., Gu, Z., Qian, S., Chen, M., Ma, L.: Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22883–22892 (2024) 3
44. Wang, X., Chai, L., Chen, J.: Debiasing trace guidance: Top-down trace distillation and bottom-up velocity alignment for unsupervised anomaly detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 22989–22998 (2025) 4
45. Yao, X., Li, R., Qian, Z., Wang, L., Zhang, C.: Hierarchical gaussian mixture normalizing flow modeling for unified anomaly detection. In: *European Conference on Computer Vision*. pp. 92–108. Springer (2024) 9, 10
46. Yoon, S., Jin, Y.U., Noh, Y.K., Park, F.: Energy-based models for anomaly detection: A manifold diffusion recovery approach. *Advances in Neural Information Processing Systems* **36**, 49445–49466 (2023) 2, 4, 9, 10
47. Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., Wu, L.: Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677* (2021) 4, 10
48. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 8330–8339 (2021) 10
49. Zhang, H., Wang, Z., Zeng, D., Wu, Z., Jiang, Y.G.: Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE transactions on pattern analysis and machine intelligence* (2025) 4
50. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961* (2023) 3
51. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: *European conference on computer vision*. pp. 392–408. Springer (2022) 3, 8