DIFFICULT /= EFFECTIVE: META-LEARNING OF EFFECTIVENESS HYPERPLANES VECTOR FOR ANOMALY GENERATION STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial anomaly detection increasingly relies on synthetic anomalies to overcome the scarcity of real defects. Yet most generation methods aim to make synthetic anomalies harder to detect, which is not necessarily effective for training: they may be noisy, unrealistic, or uninformative to a downstream detector. Thus, we propose Effectiveness Hyperplane Steering Anomaly Generation Framework that generates anomalies directly optimized for downstream training effectiveness. Our key idea is to learn a linear effectiveness classifier whose decision boundary defines an effectiveness hyperplane in diffusion feature space; the normal vector of this hyperplane points toward increasingly effective samples. The classifier is trained end-to-end by back-propagating real validation performance through meta-learning. Once learned, the steering vector guides a frozen anomaly generation model by direct feature addition at inference, without any detector evaluation or iterative refinement. The classifier can also score unseen samples on its own. Experiments demonstrate consistent gains across anomaly categories, yielding macro-average improvements of $6.04/9.27/11.49$ points in pixel-level AUROC/AP/AUPRO and $5.15/4.15$ points in image-level AUROC/AP on the Real-IAD dataset. The code will be public after acceptance.

1 INTRODUCTION

Industrial anomaly detection is essential for automated quality inspection, yet its performance is often limited by the scarcity of defective training samples. Since real defects are rare and expensive to collect, synthetic data generation has become an important approach for augmenting the training data of downstream detectors Li et al. (2021); Zavrtanik et al. (2021). Recent anomaly generation methods, particularly diffusion-based models, have substantially improved the realism of synthetic samples Hu et al. (2024a). However, visual realism alone does not determine their value for training: a synthetic sample may appear realistic while being redundant, overly easy, or only weakly informative to a downstream detector. This motivates a shift from generating visually plausible samples toward more effective ones that genuinely improve downstream performance.

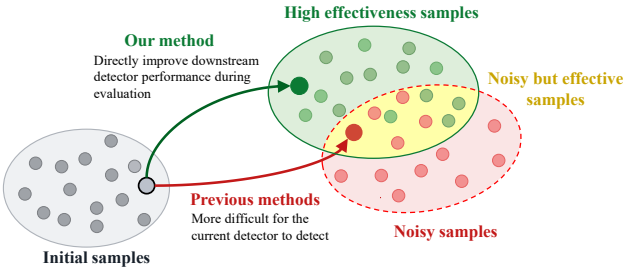

Figure 1: Difficult-to-detect does not necessarily mean high sample effectiveness. Existing methods favor synthetic samples that are more difficult for the current detector to detect, which may push generation toward noisy or uninformative regions. In contrast, our method directly targets synthetic samples that improve downstream detector performance during testing.

To improve the effectiveness of synthetic samples, existing downstream-aware methods often favor samples that are difficult for the current detector to detect Wang et al. (2024b); Yao et al. (2025a). These methods tend to follow a generate–evaluate–refine pipeline, in which the downstream detec-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tor evaluates preliminary synthetic samples and provides feedback to guide the refinement process, typically increasing the mismatch between the detector response and the ground-truth anomaly region, progressively producing samples that are more difficult to detect. However, as illustrated by the red path in Figure 1, making samples more difficult for the current detector to detect does not necessarily move them toward higher sample effectiveness. While some difficult-to-detect samples may provide useful supervision, continuing to optimize samples in this way can also drive generation toward noise, outliers, or unrealistic artifacts. Such samples provide little useful supervision and may even degrade downstream performance. This raises a more fundamental question —What kinds of synthetic anomalies can actually improve the downstream detector after training?

To this end, we propose Effectiveness Hyperplane Steering Anomaly Generation Framework, which steers generation toward synthetic samples with higher sample effectiveness. As illustrated by the green path in Figure 1, instead of continually making samples more difficult for the current detector to detect, our method guides generation toward samples that can improve downstream detector performance during evaluation. Specifically, we learn a linear effectiveness classifier from intermediate diffusion features, whose decision boundary defines an effectiveness decision hyperplane. The classifier assigns an effectiveness score to each synthetic sample, and its weight vector is normalized to obtain the effectiveness steering vector, which points toward increasing effectiveness scores. During generation, the effectiveness steering vector is injected into the corresponding intermediate diffusion features of the frozen diffusion model, steering generation toward higher-effectiveness samples without invoking the downstream detector.

To learn the linear effectiveness classifier, we adopt a meta-learning-based optimization procedure consisting of an effectiveness-weighted detector update followed by a validation-guided update of the linear effectiveness classifier Ren et al. (2018); Shu et al. (2019). First, the predicted effectiveness scores modulate the contribution of synthetic samples to the detector loss, through which the downstream detector is updated. The updated detector is then evaluated on real validation data to compute the validation objective. Since the updated detector parameters retain a differentiable dependency on the classifier parameters through the detector update, the validation gradient can be back-propagated to the linear effectiveness classifier. Consequently, the linear effectiveness classifier is optimized according to how the weighted synthetic samples affect downstream validation performance, so that samples leading to better detector performance receive higher effectiveness scores. The main contributions of the proposed method can be summarized as follows:

- We introduce a sample-effectiveness perspective for synthetic sample generation in industrial anomaly detection, focusing on whether synthetic samples can improve downstream detector performance during evaluation rather than only on visual realism or whether they are difficult for the current detector to detect.
- We propose the Effectiveness Hyperplane Steering Anomaly Generation Framework, which learns a linear effectiveness classifier whose decision boundary defines an effectiveness decision hyperplane. Its weight vector is normalized to obtain the effectiveness steering vector, which guides generation toward higher-effectiveness synthetic samples.
- We develop a two-stage optimization procedure consisting of an effectiveness-weighted detector update followed by a meta-update of the linear effectiveness classifier. The dependency of the updated detector parameters on the classifier parameters allows the gradient of the meta-validation objective to be back-propagated through the detector update to the linear effectiveness classifier.
- Extensive experiments on Real-IAD dataset show consistent improvements across five macro-average detection metrics ranging from 4.15 to 11.49 points. In particular, pixel-level AUPRO and image-level AUROC improve by 11.49 and 5.15 points, respectively.

2 Related Work

2.1 Industrial Anomaly Generation

Diffusion models have recently become an important approach for generating samples for industrial anomaly detection, with substantial progress in realism, diversity, and controllability Yao et al. (2025b). Hu et al. (2024b) improves the authenticity of synthetic samples and their alignment with

anomaly masks, while Jin et al. (2025) further enhances sample diversity, realism, and mask accuracy. Recent methods such as Sun et al. (2025) and Rao et al. (2026) extend this line of research toward more flexible generation and better alignment with real defect distributions. These methods have substantially improved the visual quality and controllability of synthetic samples.

Beyond visual quality, several works further incorporate task-related signals to obtain synthetic samples that are more useful for downstream detector training. Sun et al. (2025) introduces attention-guided optimization to generate samples that are difficult for the detector to detect, while Yao et al. (2025a) incorporates feedback from downstream detection results to refine synthetic samples toward the same goal. Although these strategies explicitly incorporate downstream detector feedback, their objectives mainly encourage samples to be more difficult to detect rather than directly reflecting whether they improve downstream detector performance. Thus, being more difficult for the detector does not necessarily imply high sample effectiveness. In contrast, our method learns sample effectiveness through downstream detector updates and real validation feedback, and guides generation toward higher-effectiveness synthetic samples.

2.2 Training Data Valuation

Training samples do not contribute equally to model learning. Training data valuation therefore aims to estimate how much each sample contributes to model performance Ghorbani & Zou (2019); Sim et al. (2022). Existing methods estimate sample value in different ways. Some measure the influence of training samples on model predictions, while others evaluate their contribution to model performance or directly learn a data value function Koh & Liang (2017); Ghorbani & Zou (2019); Yoon et al. (2020). These estimated values are commonly applied to sample weighting and data selection Ghorbani & Zou (2019); Yoon et al. (2020); Sim et al. (2022).

A closely related line of work learns sample weights from downstream validation performance. Meta-reweighting methods either directly optimize sample weights through validation loss or learn a weighting function that maps training losses to sample weights Ren et al. (2018); Shu et al. (2019). Although such methods capture how training samples affect model learning, the resulting weights are primarily used for reweighting or selecting existing samples and remain tied to the current training process. In contrast, our method learns a reusable linear effectiveness classifier directly from diffusion representations using downstream detector updates and real validation feedback. This enables sample effectiveness to be estimated for unseen synthetic samples and further used to guide generation toward higher-effectiveness samples.

3 Method

3.1 Overview

The proposed method learns a linear effectiveness classifier $\mathcal{S}(\cdot;\Phi)$ from downstream detector D_θ feedback to estimate whether a synthetic sample X_i can better improve detector performance during evaluation. The learned classifier $\mathcal{S}(\cdot;\Phi)$ assigns an effectiveness score s_i to each synthetic sample X_i based on its anomaly representation $\mathbf{F}_i$ and defines an effectiveness decision hyperplane $\mathcal{S}(\mathbf{F};\Phi)=0$ in the diffusion feature space. The normalized direction $\hat{\mathbf{n}}$ of this hyperplane is termed the effectiveness steering vector and guides anomaly generation toward higher effectiveness. The same classifier $\mathcal{S}(\cdot;\Phi)$ can also be directly applied to evaluate the effectiveness of synthetic samples without invoking the downstream detector D_θ.

3.2 Linear Effectiveness Classifier Optimization

The key challenge is to learn this linear effectiveness classifier $\mathcal{S}(\cdot;\Phi)$ that estimates whether a synthetic anomaly can improve detector performance during evaluation. The training data for learning the linear effectiveness classifier contain two parts. The first part is the synthetic data set $\{(X_i, M_i)\}_{i=1}^{K}$, where X_i is a synthetic anomaly image and M_i is its corresponding anomaly-region mask. Intermediate diffusion feature maps recorded during the generation of X_i are used as inputs to the linear effectiveness classifier. The second part is the real validation data set $\{(\widehat{X}_j, \widehat{M}_j)\}_{j=1}^{R}$, providing downstream feedback to optimize the effectiveness classifier.

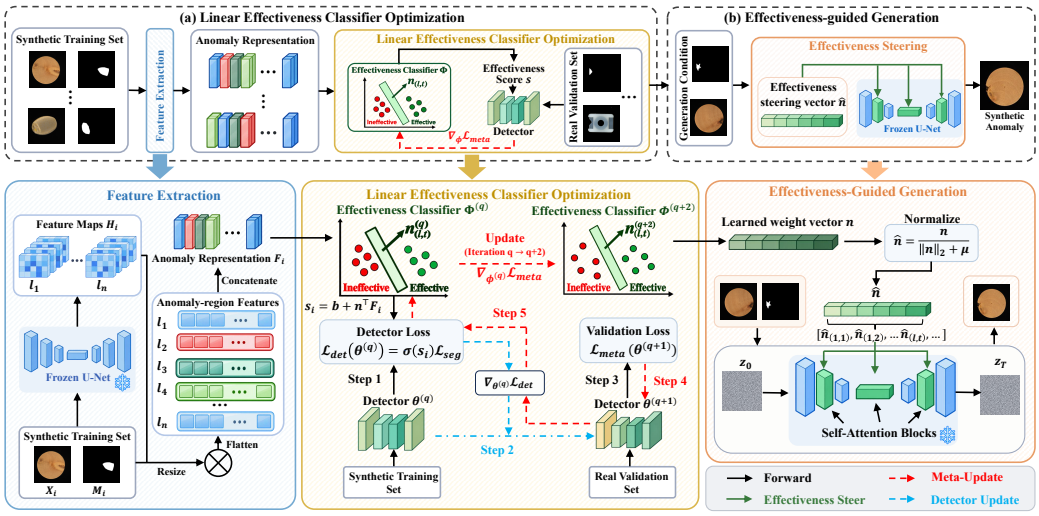

Figure 2: Overview of the proposed Effectiveness Hyperplane Steering Anomaly Generation Framework. (a) The linear effectiveness classifier is optimized through effectiveness-weighted detector updates and real-validation meta-updates. (b) The effectiveness steering vector is injected into the diffusion model to guide generation toward higher sample effectiveness.

3.2.1 Effectiveness Score Prediction

We first represent each synthetic sample using intermediate diffusion feature map $\mathbf{H}_{i,l,t}$ and map the resulting representation to an effectiveness score s_i through the proposed linear effectiveness classifier $\mathcal{S}(\cdot;\Phi)$. Feature maps from multiple diffusion layers and denoising timesteps are aggregated to capture effectiveness information across the generation process. Specifically, at the denoising timestep t, let z_t denote the noisy latent. The frozen diffusion model predicts the noise as $\widehat{\boldsymbol{\epsilon}}_t = \epsilon_\psi(z_t, t, c)$, where ψ denotes the fixed parameters of the diffusion model and c denotes its original conditioning. During the generation of X_i, we collect intermediate diffusion feature map $\mathbf{H}_{i,l,t}$ from multiple self-attention layers and denoising timesteps of the diffusion model. Let $\mathbf{H}_{i,l,t} \in \mathbb{R}^{N_l \times C_l}$ denote the feature map of X_i at layer l and timestep t, where $N_l = P_l Q_l$ is the number of spatial tokens obtained by flattening the feature map of size $P_l \times Q_l$, and C_l is the number of feature channels. We resize its anomaly mask M_i to the feature resolution of layer l and flatten it into a token-level mask $\widetilde{\mathbf{m}}_{i,l} \in \{0,1\}^{N_l}$, whose n-th element is denoted by $\widetilde{m}_{i,l,n}$. To obtain the anomaly-region feature vector $\mathbf{f}_{i,l,t}$, we aggregate the diffusion feature map $\mathbf{H}_{i,l,t}$ within the masked region, calculated as follows:

$$\mathbf{f}_{i,l,t} = \frac{\sum_{n=1}^{N_l} \widetilde{m}_{i,l,n} \cdot \mathbf{H}_{i,l,t,n}}{\sum_{n=1}^{N_l} \widetilde{m}_{i,l,n}}, \tag{1}$$

where $\mathbf{f}_{i,l,t} \in \mathbb{R}^{C_l}$ is the anomaly-region feature vector of X_i at layer l and timestep t. We concatenate the anomaly-region feature vector $\mathbf{f}_{i,l,t} \in \mathbb{R}^{C_l}$ from all selected layers and timesteps and denote the anomaly representation by $\mathbf{F}_i \in \mathbb{R}^C$, where $C = \sum_{l=1}^{L} \sum_{t=1}^{T} C_l$. For each layer–timestep feature space (l,t), we learn a vector $\mathbf{w}_{l,t} \in \mathbb{R}^{C_l}$ and a positive gate $g_{l,t} > 0$. We collect them as $\mathcal{W} = \{\mathbf{w}_{l,t} \mid 1 \le l \le L,\ 1 \le t \le T\}$ and $\mathcal{G} = \{g_{l,t} \mid 1 \le l \le L,\ 1 \le t \le T\}$, respectively. The learnable parameters of the linear effectiveness classifier are denoted by $\Phi = (\mathcal{W}, \mathcal{G}, b)$. We define the weight vector of the linear effectiveness classifier as follows:

$$\mathbf{n} = \frac{\mathrm{Concat}\,(g_{1,1}\mathbf{w}_{1,1}, \ldots, g_{L,T}\mathbf{w}_{L,T})}{\sum_{l=1}^{L} \sum_{t=1}^{T} g_{l,t}} \in \mathbb{R}^C. \tag{2}$$

The linear effectiveness classifier $\mathcal{S}(\cdot;\Phi)$ predicts the effectiveness score s_i of X_i, calculated as follows:

$$s_i = \mathcal{S}(\mathbf{F}_i;\Phi) = b + \mathbf{n}^\top \mathbf{F}_i, \tag{3}$$

where a larger s_i indicates higher sample effectiveness.

3.2.2 EFFECTIVENESS-WEIGHTED DETECTOR UPDATE

At optimization iteration q, we update the detector parameters $\theta^{(q)}$ while keeping the classifier parameters $\Phi^{(q)}$ fixed. The fixed linear effectiveness classifier $\mathcal{S}(\cdot;\Phi^{(q)})$ assigns an effectiveness score s_i to each synthetic sample X_i, and $\sigma(s_i)$ determines its contribution to detector training. Using the segmentation loss $\mathcal{L}_{\text{seg}}$, which consists of binary cross-entropy and Dice loss, we define the effectiveness-weighted detector loss $\mathcal{L}_{\text{det}}$ as follows:

$$\mathcal{L}_{\text{det}}(\Phi,\theta) = \frac{1}{B}\sum_{i=1}^{B} \sigma(s_i)\,\mathcal{L}_{\text{seg}}\left(D_\theta(X_i), M_i\right), \tag{4}$$

where $\sigma(\cdot)$ maps the effectiveness score s_i to a positive training weight. With $\Phi^{(q)}$ fixed in this step, the detector parameters are updated by:

$$\theta^{(q+1)}\left(\Phi^{(q)}\right) = \theta^{(q)} - \eta\nabla_\theta\mathcal{L}_{\text{det}}\left(\Phi^{(q)},\theta^{(q)}\right), \qquad \Phi^{(q+1)} \equiv \Phi^{(q)}, \tag{5}$$

where η is the detector learning rate. Although the classifier parameters $\Phi^{(q)}$ are not updated in this step, the sample weights $\sigma(s_i)$ are determined by $\Phi^{(q)}$. Therefore, the updated detector parameters $\theta^{(q+1)}(\Phi^{(q)})$ retain a differentiable dependence on the classifier parameters, which enables the subsequent meta-update.

3.2.3 META-UPDATE OF THE LINEAR EFFECTIVENESS CLASSIFIER

After the detector update, we fix $\theta^{(q+1)}$ and optimize the classifier parameters $\Phi^{(q+1)}$ using real validation feedback. The updated detector $D_{\theta^{(q+1)}(\Phi^{(q)})}$ is evaluated on the real validation data, while its differentiable dependence on $\Phi^{(q)}$ through Eq. 5 is retained. The resulting validation loss is then back-propagated through the detector update to optimize the linear effectiveness classifier $\mathcal{S}(\cdot;\Phi^{(q)})$. We refer to this loss as the meta-validation objective $\mathcal{L}_{\text{meta}}$, defined as follows:

$$\mathcal{L}_{\text{meta}}\left(\theta^{(q+1)}(\Phi^{(q)})\right) = \mathcal{L}_{\text{tve}} + \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{pix}}. \tag{6}$$

All three terms are computed on the real validation data using the updated detector. Specifically, $\mathcal{L}_{\text{tve}}$ is the Tversky loss Hashemi et al. (2019); $\mathcal{L}_{\text{img}}$ is the image-level binary cross-entropy loss based on generalized mean pooling; and $\mathcal{L}_{\text{pix}}$ is the foreground–background balanced pixel-wise binary cross-entropy loss. Since the updated detector parameters $\theta^{(q+1)}(\Phi^{(q)})$ depend on the classifier parameters $\Phi^{(q)}$ through the detector update in Eq. 5, the meta-validation objective provides feedback to the linear effectiveness classifier through this dependency, calculated as follows:

$$\nabla_{\Phi^{(q)}}\mathcal{L}_{\text{meta}} = \underbrace{\frac{\partial\mathcal{L}_{\text{meta}}}{\partial\Phi^{(q)}}}_{=0} + \left(\frac{\partial\theta^{(q+1)}}{\partial\Phi^{(q)}}\right)^{\top}\frac{\partial\mathcal{L}_{\text{meta}}}{\partial\theta^{(q+1)}}. \tag{7}$$

Since $\mathcal{L}_{\text{meta}}$ has no explicit dependence on $\Phi^{(q)}$, the direct term is zero. Combining Eq. 7 with the detector update in Eq. 5 gives:

$$\nabla_{\Phi^{(q)}}\mathcal{L}_{\text{meta}} = -\eta\nabla^2_{\Phi\theta}\mathcal{L}_{\text{det}}\left(\Phi^{(q)},\theta^{(q)}\right)\frac{\partial\mathcal{L}_{\text{meta}}}{\partial\theta^{(q+1)}}, \tag{8}$$

where $\nabla^2_{\Phi\theta}\mathcal{L}_{\text{det}}$ denotes the mixed second-order derivative of the detector loss $\mathcal{L}_{\text{det}}$ with respect to the classifier parameters Φ and detector parameters θ. The effectiveness parameters are updated as:

$$\Phi^{(q+2)} = \Phi^{(q+1)} - \beta\nabla_{\Phi^{(q)}}\mathcal{L}_{\text{meta}}, \qquad \theta^{(q+2)} \equiv \theta^{(q+1)}, \tag{9}$$

where β is the learning rate for the effectiveness parameters. Overall, the gradient of the meta-validation objective $\mathcal{L}_{\text{meta}}$ is back-propagated through the updated detector parameters $\theta^{(q+1)}(\Phi^{(q)})$ and the detector update in Eq. 5 to the effectiveness scores s, and finally to the classifier parameters $\Phi^{(q)}$. This allows the linear effectiveness classifier $\mathcal{S}(\cdot;\Phi)$ to be optimized according to the downstream validation performance induced by the weighted synthetic samples.

3.3 EFFECTIVENESS-GUIDED GENERATION

The learned linear effectiveness classifier $\mathcal{S}(\cdot;\Phi)$ defines the effectiveness decision hyperplane $\mathcal{S}(\mathbf{F};\Phi)=0$. We normalize its weight vector as follows:

$$\widehat{\mathbf{n}} = \frac{\mathbf{n}}{\|\mathbf{n}\|_2 + \mu}, \tag{10}$$

where μ is a small constant for numerical stability. We refer to $\widehat{\mathbf{n}}$ as the effectiveness steering vector. For steering, we write $\widehat{\mathbf{n}} = \mathrm{Concat}(\widehat{\mathbf{n}}_{1,1}, \ldots, \widehat{\mathbf{n}}_{L,T})$, where $\widehat{\mathbf{n}}_{l,t} \in \mathbb{R}^{C_l}$ denotes the part associated with layer l and timestep t.

Starting from a noisy latent z_T, we follow the original reverse diffusion process. At timestep t, let $\mathbf{H}_{l,t} \in \mathbb{R}^{N_l \times C_l}$ denote the intermediate feature map from a selected self-attention layer l of the diffusion model. The anomaly mask is resized to the corresponding feature resolution and flattened into $\widetilde{\mathbf{m}}_l \in \{0,1\}^{N_l}$. We inject $\widehat{\mathbf{n}}$ into the selected intermediate feature maps as follows:

$$\mathbf{H}'_{l,t} = \mathbf{H}_{l,t} + \lambda\, \widetilde{\mathbf{m}}_l \cdot \widehat{\mathbf{n}}_{l,t}^{\top}, \tag{11}$$

where λ controls the steering strength and $\mathbf{H}'_{l,t}$ denotes the steered intermediate diffusion feature map. Since the linear effectiveness classifier $\mathcal{S}(\cdot;\Phi)$ is linear, its decision boundary $\mathcal{S}(\mathbf{F};\Phi)=0$ is a hyperplane rather than a curved surface, and its normal direction is constant throughout the diffusion feature space. Therefore, the normalized normal vector $\widehat{\mathbf{n}}$ provides a fixed, sample-independent direction toward increasing effectiveness scores. Its layer–timestep component $\widehat{\mathbf{n}}_{l,t}$ can thus be consistently injected into the corresponding intermediate diffusion feature maps to steer their representations toward higher effectiveness. The anomaly-region mask $\widetilde{\mathbf{m}}_l$ restricts this steering to the anomaly region while leaving the remaining features unchanged. The modified feature is then propagated through the remaining diffusion layers, and the subsequent latent update follows the original reverse diffusion process without modification. This operation is applied at each selected layer–timestep pair, while all diffusion model parameters remain fixed. After the final denoising step, the latent is decoded by the frozen VAE decoder to obtain the synthetic anomaly.

Moreover, such a learned linear effectiveness classifier $\mathcal{S}(\cdot;\Phi)$ can also directly evaluate an anomaly image without invoking the downstream detector D_θ. For an anomaly image X with its anomaly-region mask M, we obtain the anomaly representation $\mathbf{F}$ following the same feature extraction procedure described in Section 3.2. The effectiveness score $s = \mathcal{S}(\mathbf{F};\Phi)$ is then computed with the classifier parameters Φ fixed. The effectiveness score s estimates whether the anomaly image X can improve detector performance during evaluation, enabling direct sample scoring and selection without detector-based evaluation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We conduct all experiments on Real-IAD dataset Wang et al. (2024a) with 27 object categories with valid anomaly annotations, where anomalies are grouped into area-, line-, and point-type families according to anomaly-mask geometry. The experimental protocol comprises a synthetic set $\mathcal{D}^{syn}$ for learning the linear effectiveness classifier, together with real data in Real-IAD partitioned into a real validation set $\mathcal{D}^{val}$ and a held-out test set $\mathcal{D}^{test}$ at a ratio of 7:3. During classifier training, $\mathcal{D}^{val}$ is used to compute the meta-validation objective $\mathcal{L}_{\mathrm{meta}}$ and update the linear effectiveness classifier, whereas $\mathcal{D}^{test}$ remains held out and is used only for final performance test.

We use a reference-conditioned Stable Diffusion v1.5 inpainting model Rombach et al. (2022) as the anomaly generator and SegFormer Xie et al. (2021) as the downstream detector D_θ. Images are resized to 512×512. DDIM sampling uses $T = 50$ denoising steps, with diffusion feature maps collected from $L = 16$ self-attention layers. During linear effectiveness classifier learning, both the synthetic batch from $\mathcal{D}^{syn}$ and the real validation batch from $\mathcal{D}^{val}$ use a batch size of 64. The detector parameters θ and classifier parameters Φ are optimized with AdamW using learning rates $\eta = 1 \times 10^{-4}$ and $\beta = 2 \times 10^{-4}$, respectively. The diffusion generator remains frozen throughout classifier learning and effectiveness-steered generation.

4.2 Main Results

Table 1: Downstream anomaly detection on Real-IAD using synthetic samples generated by different methods. "Mean" denotes the macro average over area, line, and point anomaly families. Best results in each block are shown in **bold**.

Family	Method	Pixel-level			Image-level	
		AUROC	AP	AUPRO	AUROC	AP
Area	Baseline	87.35	46.15	81.39	86.73	87.30
	Rethinking Hard Samples	84.71	28.68	75.55	76.80	78.40
	Crucial-Diff	86.19	40.34	77.87	85.16	87.04
	Ours	**87.72**	**48.15**	**82.19**	**87.43**	**88.52**
Line	Baseline	85.92	13.61	75.59	65.68	71.39
	Rethinking Hard Samples	72.31	1.55	52.29	52.87	60.45
	Crucial-Diff	83.72	19.01	70.27	**79.47**	**81.85**
	Ours	**89.11**	**22.22**	**80.03**	73.30	78.40
Point	Baseline	78.47	2.80	55.87	58.37	71.03
	Rethinking Hard Samples	66.10	0.10	41.28	57.82	68.11
	Crucial-Diff	72.67	0.28	46.89	47.91	61.57
	Ours	**93.02**	**20.00**	**85.10**	**65.51**	**75.23**
Mean	Baseline	83.91	20.85	70.95	70.26	76.57
	Rethinking Hard Samples	74.37	10.11	56.37	62.50	68.99
	Crucial-Diff	80.86	19.88	65.01	70.85	76.82
	Ours	**89.95**	**30.12**	**82.44**	**75.41**	**80.72**

As shown in Table 1, our method consistently improves over the Baseline across area-, line-, and point-type anomalies and achieves the best macro-average results on all five metrics. Both *Rethinking Hard Samples* Wan et al. (2026) and *Crucial-Diff* Yao et al. (2025a) focus on generating harder or more challenging training samples, yet their improvements are less consistent across anomaly families. In contrast, our method directly targets sample effectiveness, leading to more consistent gains across both pixel- and image-level metrics.

The improvement is particularly pronounced for line and point anomalies at the pixel level. These anomalies occupy a small fraction of the image and are therefore more sensitive to localization errors and severe foreground–background imbalance, which makes pixel-level AP particularly challenging. Moreover, such small and localized anomaly patterns are more difficult for existing generation methods to synthesize reliably. Effectiveness steering instead guides the diffusion process toward synthetic samples that provide more useful supervision for these difficult-to-generate anomaly patterns, leading to substantially stronger pixel-level detection. The consistent gains across three anomaly families further show that the learned effectiveness steering is not restricted to a particular anomaly morphology.

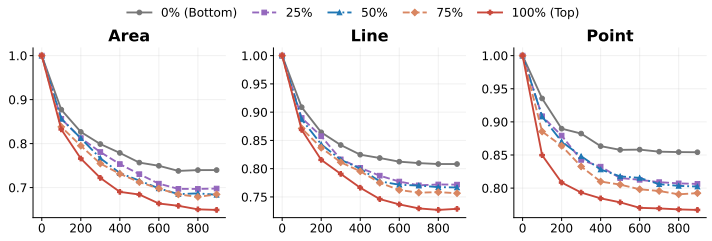

Figure 3: Training trajectories as low-effectiveness synthetic samples are progressively replaced by high-effectiveness samples.

4.3 Analysis of the Linear Effectiveness Classifier

We next examine whether the effectiveness scores learned by the linear effectiveness classifier provide a meaningful ordering of synthetic samples rather than merely distinguishing a few extreme cases. For each anomaly family, we rank the samples in $\mathcal{D}^{syn}$ by their effectiveness scores s and define the top 20% and bottom 20% as the Top and Bottom pools, respectively. Starting from a fixed-size training subset drawn entirely from the Bottom pool, we progressively replace 25%, 50%, 75%, and 100% of the samples with samples from the Top pool, while keeping the subset size and other training settings unchanged.

The training dynamics further reveal a clear overall trend as the proportion of high-effectiveness samples increases, as shown in Figure 3. The all-Bottom subset converges more slowly and maintains a relatively high loss, whereas progressively replacing Bottom-pool samples with Top-pool samples shifts the training trajectories downward. The all-Top subset consistently achieves the lowest loss across all three anomaly families. These results indicate that the classifier captures a graded ordering of sample effectiveness, with higher-ranked synthetic samples providing more useful su-

pervision for detector learning rather than simply separating samples into effective and ineffective groups.

4.4 Analysis of the Effectiveness Steering Vector

We next examine how the effectiveness steering vector changes the internal representations of the frozen diffusion model. The key question is whether injecting $\hat{\mathbf{n}}$ produces the intended shift toward higher effectiveness rather than merely perturbing the diffusion features. To isolate this effect, we construct paired generations with the same reference image, anomaly mask, and initial noise, differing only in whether effectiveness steering is applied.

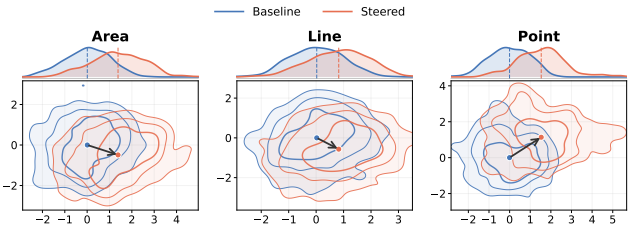

Figure 4: Paired representation shifts along the effectiveness steering direction.

Table 2: Quantitative analysis of paired effectiveness-coordinate shifts induced by steering.

Family	Mean shift $\Delta \uparrow$	95% CI	Positive pairs	Paired corr. $r \uparrow$
Point	$+0.896$	$[0.853, 0.939]$	100%	0.894
Line	$+0.558$	$[0.531, 0.586]$	100%	0.972
Area	$+0.751$	$[0.726, 0.775]$	100%	0.967

For each pair, we project the anomaly representation onto the effectiveness steering direction $\hat{\mathbf{n}}$ and examine the remaining variation orthogonal to this direction. Figure 4 reveals a clear and consistent pattern across area-, line-, and point-type anomalies. After steering, the representations move toward larger projections along $\hat{\mathbf{n}}$, while substantial variation is retained in the orthogonal space. Thus, steering does not force different samples toward a single feature state, but instead shifts their representations in a common effectiveness-increasing direction while largely preserving sample-specific variation.

This paired movement is further quantified in Table 2. For each paired sample, we measure the change in its projection onto $\hat{\mathbf{n}}$. The mean change is reported as Δ, where a positive value indicates movement toward increasing effectiveness scores. We additionally report the 95% Confidence Interval (CI) of Δ, the percentage of pairs with a positive shift, and the paired correlation r between the projection values before and after steering. The mean shift Δ is positive for all three anomaly families, all confidence intervals lie strictly above zero, and every paired sample exhibits a positive shift. Meanwhile, the high paired correlations indicate that the relative relationships among samples along the steering direction are largely preserved. These results quantitatively confirm that the learned steering direction induces a consistent effectiveness-increasing shift without collapsing sample-specific structure.

4.5 Ablation Studies

Table 3: Effect of different loss components in the meta-validation objective.

Feedback	Pixel-level			Image-level	
	AUROC	AP	AUPRO	AUROC	AP
$\mathcal{L}_{\text{tve}}$	88.02	30.41	79.37	75.18	80.62
$\mathcal{L}_{\text{tve}} + \mathcal{L}_{\text{img}}$	88.31	**30.57**	79.85	74.73	80.50
$\mathcal{L}_{\text{tve}} + \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{pix}}$	**89.95**	30.12	**82.44**	**75.41**	**80.72**

Table 4: Effect of synthetic samples on downstream detector performance.

Training	Pixel-level			Image-level	
	AUROC	AP	AUPRO	AUROC	AP
Real only	82.07	24.75	67.66	69.20	75.98
Real + Baseline	82.15	24.72	67.30	69.54	75.75
Real + Ours	**86.20**	**29.28**	**76.06**	**72.09**	**80.13**

Steering direction. We first examine whether the improvement comes from the learned direction itself or merely from perturbing the diffusion features, as shown in Table 5. Reversing the effectiveness steering vector consistently degrades downstream performance, while a random direction performs better than negative steering but remains clearly inferior to $+\hat{\mathbf{n}}$. This ordering is especially pronounced for localized line and point anomalies, whose small spatial support makes generation more sensitive to an incorrect feature displacement. These results support the geometric interpretation of the learned hyperplane: the positive normal direction carries meaningful effectiveness information, rather than acting as a generic feature perturbation.

Effectiveness-learning strategy. We compare four strategies for learning sample effectiveness, as reported in Table 6. *Offline Score Regression* first obtains continuous effectiveness scores from an image-level scorer and fits them from diffusion features, while *Binary Reweighting* and *Pairwise Ranking* learn from online downstream feedback using binary weights and pairwise preferences, respectively. In Pairwise Ranking, sample pairs are ordered according to their relative effects on the real validation loss rather than by regressing absolute effectiveness scores. Interestingly, Offline Score Regression performs substantially worse despite using explicit continuous supervision. We attribute this mainly to the additional distillation step. The image-level scores must be recovered from selected diffusion features by a linear regressor, which may lose effectiveness information that is not well represented in this feature space. In contrast, our method directly optimizes the classifier in the diffusion feature space according to the real validation response from detector updates, avoiding this supervision mismatch and achieving the strongest overall balance across metrics.

Meta-validation objective. To examine how the meta-validation objective $\mathcal{L}_{\mathrm{meta}}$ affects effectiveness learning, we compare three variants while keeping all other settings unchanged. As shown in Table 3, these variants use Tversky loss alone, Tversky loss with the image-level loss, and the full objective further incorporating the pixel-level loss. The Tversky term focuses on foreground overlap, while the image-level term additionally captures global anomaly recognition. The pixel-level term further provides fine-grained localization information. Overall, the full objective achieves the strongest balance across pixel- and image-level metrics, although the reduced variants perform slightly better on individual metrics such as pixel AP. These results show that combining complementary loss components provides a more complete objective for learning sample effectiveness.

Table 5: Effect of steering direction on downstream detector performance.

Family	Direction	Pixel-level			Image-level	
		AUROC	AP	AUPRO	AUROC	AP
Area	$-\widehat{\mathbf{n}}$	87.11	44.57	81.07	85.37	86.11
	Random	87.46	46.15	81.30	86.61	87.21
	$+\widehat{\mathbf{n}}$ (Ours)	**87.72**	**48.15**	**82.19**	**87.43**	**88.52**
Line	$-\widehat{\mathbf{n}}$	82.36	8.99	65.60	67.32	70.44
	Random	87.40	19.07	77.93	71.12	76.47
	$+\widehat{\mathbf{n}}$ (Ours)	**89.11**	**22.22**	**80.03**	**73.30**	**78.40**
Point	$-\widehat{\mathbf{n}}$	63.34	0.44	26.95	47.07	63.51
	Random	80.62	5.01	56.82	47.82	64.09
	$+\widehat{\mathbf{n}}$ (Ours)	**93.02**	**20.00**	**85.10**	**65.51**	**75.23**
Mean	$-\widehat{\mathbf{n}}$	77.60	18.00	57.87	66.59	73.35
	Random	85.16	23.41	72.02	68.52	75.92
	$+\widehat{\mathbf{n}}$ (Ours)	**89.95**	**30.12**	**82.44**	**75.41**	**80.72**

Contribution of synthetic samples. We further examine whether the synthetic samples generated by our method provide useful additional supervision when combined with real training data. We conduct a separate real-data augmentation experiment comparing real data only, real data with Baseline samples, and real data with samples generated by Ours, while keeping the real training data, synthetic-data budget, and generation conditions unchanged. As shown in Table 4, adding Baseline samples brings little benefit and slightly degrades some metrics, whereas samples generated by Ours consistently improve both pixel- and image-level detection. These results show that effectiveness-steered synthetic samples provide complementary supervision beyond the available real data rather than simply increasing the amount of training data.

Table 6: Comparison of different strategies for learning sample effectiveness.

Learning	Pixel-level			Image-level	
	AUROC	AP	AUPRO	AUROC	AP
Offline Score Reg.	84.69	23.73	71.13	70.14	76.93
Binary Reweighting	89.22	29.83	81.19	75.14	80.48
Pairwise Ranking	89.02	29.50	80.90	**76.23**	**80.80**
Ours	**89.95**	**30.12**	**82.44**	75.41	80.72

5 Conclusion

We propose the effectiveness hyperplane steering anomaly generation framework to generate synthetic samples with higher sample effectiveness. A linear effectiveness classifier is learned from downstream detector updates and real validation data, and its normalized weight vector is used to steer a frozen diffusion model without invoking the downstream detector during generation. Experiments on Real-IAD demonstrate consistent improvements across area-, line-, and point-type anomalies. Future work will explore richer effectiveness structures beyond a single linear decision hyperplane and more adaptive steering that can vary across denoising timesteps. It is also worth investigating stage-dependent effectiveness, since the usefulness of synthetic samples may change as the downstream detector evolves during training.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, improving writing clarity, LaTeX formatting, and limited assistance with code and figure preparation. All AI-assisted content was reviewed and verified by the authors, and all methodological decisions, experiments, results, and conclusions were independently checked by the authors. We take full responsibility for the final content of this work, including any text, code, or artifacts produced with the aid of generative AI.

REFERENCES

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 09–15 Jun 2019.

Seyed Raein Hashemi, Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Sanjay P. Prabhu, Simon K. Warfield, and Ali Gholipour. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735, 2019. ISSN 2169-3536. doi: 10.1109/access.2018.2886371.

Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. AnomalyDiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024a.

Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. AnomalyDiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8526–8534, 2024b. doi: 10.1609/aaai.v38i8.28696.

Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Jiafu Wu, Hao Chen, Haoxuan Wang, Wenbing Zhu, Mingmin Chi, Jun Liu, and Yabiao Wang. Dual-interrelated diffusion model for few-shot anomaly image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30420–30429, June 2025.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 06–11 Aug 2017.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9664–9674, June 2021.

Haoxiang Rao, Zhao Wang, Chenyang Si, Yan Lyu, Yuanyi Duan, Fang Zhao, and Caifeng Shan. One-to-More: High-fidelity training-free anomaly generation with attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28492–28501, June 2026.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343, 10–15 Jul 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, June 2022.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-Weight-Net: Learning an explicit mapping for sample weighting. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. Data valuation in machine learning: "ingredients", strategies, and open challenges. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 5607–5614. International Joint Conferences on Artificial Intelligence Organization, 7 2022.

Han Sun, Yunkang Cao, Hao Dong, and Olga Fink. Unseen visual anomaly generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25508–25517, June 2025.

Zhibin Wan, Zhiqiang Gao, Mingjie Sun, Yang Yang, Cao Min, Hongliang He, and Guohong Fu. Rethinking hard training sample generation for medical image segmentation. *Pattern Recognition*, 172:112533, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.112533.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. 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*, 2024a.

Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, and Kai Zhang. DetDiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7246–7255, June 2024b.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021.

Siyue Yao, Mingjie Sun, Eng Gee Lim, Ran Yi, Baojiang Zhong, and Moncef Gabbouj. Crucial-Diff: A unified diffusion model for crucial image and annotation synthesis in data-scarce scenarios. *IEEE Transactions on Image Processing*, 34:7024–7036, 2025a. doi: 10.1109/TIP.2025.3625380.

Siyue Yao, Haotian Xu, Mingjie Sun, Siyue Yu, Jimin Xiao, and Eng Gee Lim. DriftRemover: Hybrid energy optimizations for anomaly images synthesis and segmentation. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2251–2259. International Joint Conferences on Artificial Intelligence Organization, 8 2025b. doi: 10.24963/ijcai.2025/251. Main Track.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10842–10851, 13–18 Jul 2020.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, October 2021.