ANOMALYLOOP: CLOSING THE GENERATION–DETECTION LOOP FOR OPEN-SET TRAINING-FREE ANOMALY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial visual anomaly detection has long been constrained by the scarcity of real anomalous samples and the open-ended nature of anomaly types. To alleviate this issue, open-set training-free anomaly generation has gradually emerged as a promising data synthesis paradigm, enabling the synthesis of unseen defects without anomaly specific training. However, existing methods still face two key challenges: spatial control is difficult to confine to fine-grained candidate anomaly formation regions, and generation remains separate from anomaly evaluation, such that insufficiently formed anomalies are only identified after generation and cannot be corrected through detection feedback. To address these challenges, we propose **AnomalyLoop**, a two-stage framework for open-set training-free anomaly generation comprising the exploration of fine-grained anomaly formation regions and generation–detection closed-loop trajectory optimization. The key idea is to transform anomaly detection from post-generation evaluation into terminal feedback for the full diffusion trajectory, thereby closing the loop between anomaly generation and detection. In the first stage, we propose an Anomaly Specific Region Exploration (ASRE) mechanism to construct fine-grained candidate anomaly formation regions. In the second stage, we use terminal detection results as the primary feedback and propagate this feedback backward through Trajectory Optimal Control (TOC) to optimize control decisions along the full diffusion trajectory, enabling detection-driven anomaly formation. Furthermore, we jointly incorporate semantic consistency and history-aware diversity as auxiliary terminal constraints to preserve target semantics while maintaining sample diversity. AnomalyLoop achieves state-of-the-art generation quality and diversity on MVTec AD, while overall improving downstream detection on MVTec AD and VisA.

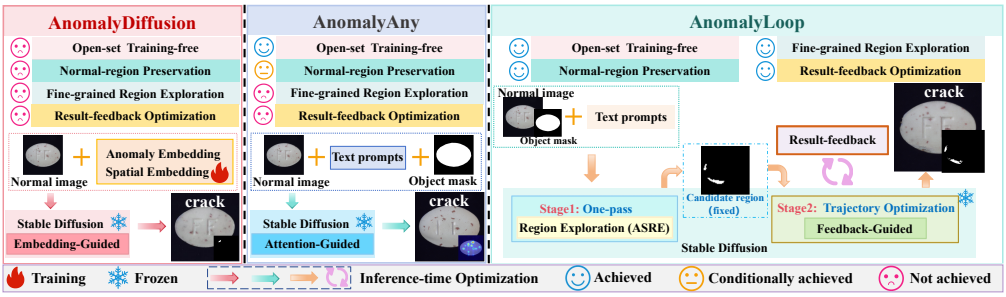

Figure 1: Comparison between representative diffusion-based anomaly generation methods. AnomalyLoop enables anomaly generation with fine-grained region exploration (one-pass denotes a single execution), normal-region preservation, and result-feedback trajectory optimization.

1 INTRODUCTION

Industrial visual anomaly detection is essential to manufacturing quality inspection, yet scarce and open-ended defects limit training coverage and model generalization. Anomaly generation mitigates this scarcity by augmenting training data. Early pseudo-anomaly methods lack realism and structural

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.

consistency (Li et al., 2021; Zavrtanik et al., 2021; Schlüter et al., 2022) , whereas GAN-based methods (Zhang et al., 2021; Duan et al., 2023) and diffusion-based methods (Jiang et al., 2026; Duan et al., 2025; Hu et al., 2024; Jin et al., 2025; Dai et al., 2025) improve generation quality and diversity but typically require real anomalies or anomaly specific training.

With the development of large-scale pretrained diffusion models, training-free anomaly generation (Xu et al., 2025; Rao et al., 2026) has emerged to regulate anomaly formation by leveraging model priors and imposing control during inference. Among these methods, open-set training-free anomaly generation (Zuo et al., 2025; Sun et al., 2025) further eliminates the dependence on real anomaly samples, synthesizing unseen anomalies solely from normal images and text descriptions. Recent methods in this line regulate anomaly location and appearance using diffusion-process signals such as cross-attention and denoising discrepancies, providing a generative paradigm with greater generalization potential in open-set industrial settings.

Despite substantial progress, existing methods still face two key limitations in candidate region exploration and anomaly formation regulation, as illustrated in Figure 1. First, object-level masks alone are insufficient to narrow the generation space to fine-grained anomaly formation regions, often leading to excessive modifications across the object. Meanwhile, anomaly-conditioned responses used for region exploration may contain components aligned with normal-conditioned responses and vary across denoising states, making stable fine-grained spatial guidance difficult. Second, diffusion-based methods regulate anomaly formation through control decisions applied at successive steps of the denoising trajectory. However, they typically assess whether anomalies are sufficiently formed only after generation, without feeding detection results back into the generation process. Consequently, insufficiently formed or inconspicuous anomalies can only be identified post hoc rather than corrected by revising these earlier trajectory-control decisions.

Based on these observations, we focus on two objectives: constructing fine-grained candidate anomaly formation regions and feeding terminal detection results back to guide anomaly formation. To this end, we propose AnomalyLoop, a two-stage framework for open-set training-free anomaly generation. In the first stage, anomaly specific region exploration suppresses response components aligned with the normal-conditioned semantic direction and aggregates the remaining evidence along the normal-conditioned diffusion trajectory, thereby constructing fine-grained candidate anomaly formation regions. In the second stage, we perform generation–detection closed-Loop trajectory optimization within the resulting candidate region. A frozen anomaly evaluator provides terminal detection feedback, which Trajectory Optimal Control (TOC) propagates backward to optimize control decisions throughout the diffusion trajectory and promote anomaly formation. To preserve target anomaly semantics and suppress cross-sample pattern repetition during detection-driven anomaly formation, we further introduce semantic consistency and history-aware diversity as complementary auxiliary constraints. Specifically, history-aware diversity compares the current anomaly change representation with its nearest historical representations and penalizes excessive similarity to previously generated local change patterns. Our contributions are summarized as follows:

- We propose AnomalyLoop, a two-stage closed-loop framework for open-set training-free anomaly generation that feeds detection feedback from the final generated result back into the generation process, allowing insufficiently formed anomalies to be corrected through full-trajectory optimization rather than merely identified after generation.
- We introduce Anomaly Specific Region Exploration to autonomously construct stable, fine-grained candidate anomaly formation regions without manually specified fine-grained anomaly regions. Building on these regions, we use detection results from the final generated image to optimize the full diffusion trajectory via TOC and promote anomaly formation, while semantic consistency and history-aware diversity jointly serve as auxiliary constraints to provide semantic and diversity regularization for trajectory optimization.
- Experimental results show that AnomalyLoop achieves overall superior performance in generation quality, diversity, and downstream anomaly detection.

2 RELATED WORK

2.1 ANOMALY GENERATION

Diffusion-based anomaly generation includes training-based, reference-guided training-free, and open-set training-free methods. Training-based methods learn from real anomalies to improve real-

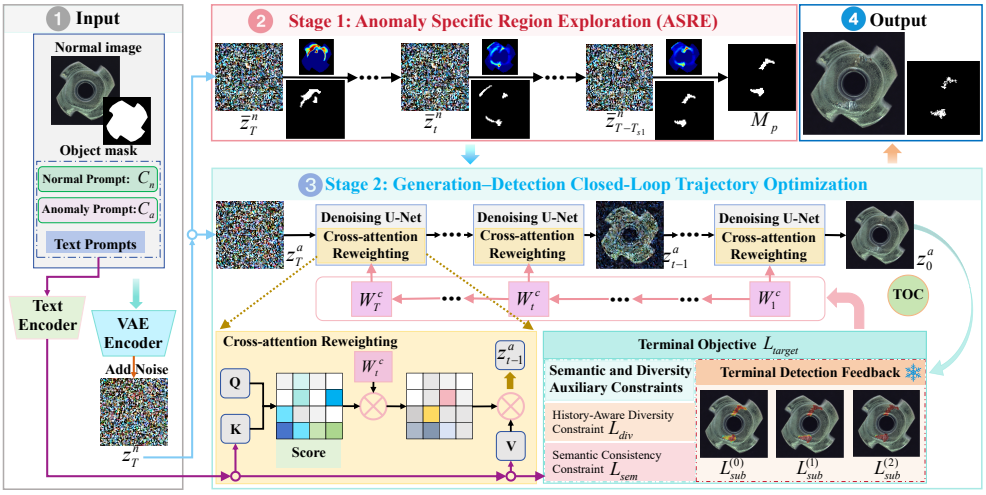

Figure 2: Overview of AnomalyLoop. Stage 1 uses Anomaly Specific Region Exploration to construct a fine-grained candidate anomaly formation region. Stage 2 optimizes anomaly formation through a generation–detection closed loop driven by terminal detection feedback, with semantic consistency and history-aware diversity as auxiliary constraints.

ism and controllability (Jin et al., 2025; Dai et al., 2025; Wang et al., 2026), but remain tied to known anomaly distributions. Reference-guided methods avoid fine-tuning by transferring appearance or feature priors from real anomalies (Xu et al., 2025; Rao et al., 2026), yet still require anomaly references. Open-set training-free methods instead generate unseen anomalies from normal images and text using diffusion signals such as attention responses and denoising discrepancies (Sun et al., 2025; Zuo et al., 2025). However, they do not feed final anomaly evaluations back into anomaly formation, whereas AnomalyLoop closes this loop.

2.2 Anomaly Detection

Normal-reference methods detect deviations from normal representations through feature matching or prompt learning (Roth et al., 2022; Li et al., 2024a), whereas zero-shot methods (Jeong et al., 2023; Zhou et al., 2024; Li et al., 2024b) exploit vision-language priors or inter-image relationships for cross-category detection. SubspaceAD (Lendering et al., 2026) builds a normal subspace from frozen pretrained features and scores departures from it. Because anomaly generation and detection are typically independent, generated samples are not explicitly optimized for detection utility. SubspaceAD requires only normal references and no anomaly specific training, making it suitable for our closed-loop objective.

2.3 Objective-Guided Diffusion and Optimal Control

Objective-guided diffusion modulates the sampling process of pretrained models at inference time using observation consistency, external energies, or reward signals (Chung et al., 2023; Yu et al., 2023; Ye et al., 2024), where the guidance typically acts on the current state or an intermediate approximation of the final result. Optimal-control methods further formulate diffusion or flow matching as a controlled dynamical system and optimize the generation trajectory through time-varying control variables (Rout et al., 2025; Domingo i Enrich et al., 2025). TOC (Chang et al., 2026) propagates terminal-objective signals via adjoint states to per-timestep controls along the full trajectory. We use it to propagate terminal detection feedback to temporal cross-attention controls, thereby closing the generation–detection loop. This enables result-driven refinement of the generation trajectory, thereby promoting sufficient anomaly formation.

3 Method

3.1 Problem Formulation and Controlled Diffusion

Problem setup and overview. Given a normal image I_n, an object mask M_{obj}, a normal prompt C_n, a target anomaly description C_a, and a candidate region area budget $r_{\text{area}} \in (0, 1]$, we aim to generate an anomalous image I_a without using real anomalous samples or updating model parameters. As shown in Figure 2, AnomalyLoop comprises two stages. Stage 1 constructs a candidate anomaly for-

mation region M_p through Anomaly Specific Region Exploration (ASRE). Stage 2 restarts the full denoising process from the same source latent and optimizes temporal cross-attention controls using terminal detection feedback. Semantic consistency and history-aware diversity are auxiliary terminal constraints in Stage 2. During candidate-region aggregation, r_{area} specifies an upper bound on the final latent-grid area of M_p. M_p denotes only a candidate formation region, not a true anomaly location or final anomaly mask.

Latent-space diffusion trajectory. The frozen VAE (Kingma & Welling, 2014) encoder $\mathcal{E}$ maps the normal image to $z_0^n = \mathcal{E}(I_n)$. Aligning M_{obj} with this latent spatial grid yields M_o. The frozen diffusion (Rombach et al., 2022) text encoder ψ produces the normal-text condition $e_n = \psi(C_n)$, the anomaly-text condition $e_a = \psi(C_a)$, and the unconditional baseline $e_\varnothing = \psi(C_\varnothing)$, where $C_\varnothing =$ `""` is the empty-string prompt used by Stable Diffusion's unconditional branch. $e_\varnothing$ is used only for normal-response decoupling in Stage 1, not anomaly generation. For fixed noise $\epsilon \sim \mathcal{N}(0, \mathbf{I}_z)$ and the scheduler-defined cumulative signal-retention coefficient $\bar{\alpha}_t$, the normal reference state is

$$z_t^n = \sqrt{\bar{\alpha}_t}\, z_0^n + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad t \in \Gamma = (T, T-1, \ldots, 0). \tag{1}$$

Here, $\mathbf{I}_z$ is the identity covariance in latent space, while T and 0 denote the maximum-noise state and the denoising endpoint, respectively. Reusing the same ϵ, Eq. 1 constructs the normal reference sequence $\{z_t^n\}_{t\in\Gamma}$ through the diffusion scheduler. This sequence contributes to preserving normal content outside M_p. We denote its source latent by $z_T^{\text{src}} := z_T^n$. Let $\Phi_t(z, e, W)$ denote the frozen one-step denoising transition given a latent variable, a text condition, and an attention control, and let $\bar{z}_t^n$ denote the normal-conditioned exploration state in Stage 1. Stage 1 starts from $\bar{z}_T^n = z_T^{\text{src}}$ and recursively computes $\bar{z}_{t-1}^n = M_o \odot \Phi_t(\bar{z}_t^n, e_n, \mathbf{1}_c) + (1 - M_o) \odot z_{t-1}^n$, $t \in \Gamma_{s1}$, where $\mathbf{1}_c$ is an all-one control matrix that leaves the original cross-attention unchanged. The symbols Γ_{s1} and T_{s1} denote the exploration interval and the number of exploration steps, respectively. Both stages share the normal image, text conditions, source latent, and normal reference sequence. The only newly constructed output passed from Stage 1 to Stage 2 is the frozen candidate region M_p. The intermediate responses and exploration trajectory from Stage 1 are discarded. Stage 2 restarts full denoising from the shared source latent by setting $z_T^a = z_T^{\text{src}}$.

Controlled Cross-Attention. To make anomaly generation adjustable by terminal detection feedback, let z_t^a denote the anomalous latent variable at timestep t in Stage 2. For a controlled cross-attention layer with spatial resolution $N \times N$, we define $Q_t = W_Q \phi(z_t^a)$, $K_t = W_K e_a$, and $V_t = W_V e_a$, where Q_t, K_t, and V_t are the Query, Key, and Value representations. The function ϕ denotes the spatial feature map of the current layer, and W_Q, W_K, and W_V are frozen linear projections. We introduce a time-dependent control matrix W_t^c on the pre-Softmax attention logits:

$$A_t^c = \text{Softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d}} \odot W_t^c\right) V_t, \qquad W_t^c \in \mathbb{R}^{N^2 \times N_{\text{tok}}}. \tag{2}$$

Here, A_t^c is the controlled attention output, d is the attention feature dimension, and N_{tok} is the number of text tokens. The rows and columns of W_t^c correspond to image-space positions and text tokens, respectively. Each W_t^c is initialized to $\mathbf{1}_c$. After alignment with the current layer, M_p restricts its spatial optimization range, while controls outside M_p remain at their baseline values. We denote all temporal controls by $W^c = \{W_t^c\}_{t=1}^T$.

Region-constrained anomaly generation. Cross-attention controls modulate anomaly conditioned generation, while stepwise latent fusion restricts content changes outside M_p. Stage 2 first computes $\widehat{z}_{t-1}^a = \Phi_t(z_t^a, e_a, W_t^c)$, where $\widehat{z}_{t-1}^a$ is the intermediate anomalous latent variable before regional fusion. It then restores the corresponding normal reference state outside M_p:

$$z_{t-1}^a = F_t(z_t^a, e_a, W_t^c, M_p) = M_p \odot \widehat{z}_{t-1}^a + (1 - M_p) \odot z_{t-1}^n, \qquad t = T, \ldots, 1. \tag{3}$$

The candidate region M_p is aligned with the latent grid and broadcast across feature channels. The transition F_t combines controlled denoising with region-wise latent fusion. After denoising, the frozen VAE decoder $\mathcal{D}$ outputs the anomalous image $I_a = \mathcal{D}(z_0^a)$.

3.2 Anomaly Specific Region Exploration

The object mask M_o only bounds the modifiable object region. Stage 1 first explores candidate regions for local anomaly generation within M_o. However, the single-step anomaly–normal denoising difference may still contain components aligned with the normal-conditioned response, biasing the spatial evidence subsequently aggregated across timesteps. We therefore propose ASRE, which combines normal-response decoupling with region-level temporal aggregation to construct M_p from conditional responses queried at shared normal states.

Normal-response decoupling. Stage 1 operates over $\Gamma_{s1} = \{T, T-1, \ldots, T-T_{s1}+1\}$. At each $\bar{z}_t^n$, the same frozen denoiser is queried using e_a, e_n, and $e_\varnothing$. Their differences therefore characterize instantaneous conditional denoising directions at the same state rather than discrepancies accumulated along independent trajectories. For a latent-grid location u within the support of M_o, we define

$$J_t(u) = [\epsilon_\theta(\bar{z}_t^n, t, e_a) - \epsilon_\theta(\bar{z}_t^n, t, e_n)]_u, \qquad G_t(u) = [\epsilon_\theta(\bar{z}_t^n, t, e_n) - \epsilon_\theta(\bar{z}_t^n, t, e_\varnothing)]_u,$$
$$R_t(u) = J_t(u) - \frac{\langle J_t(u), G_t(u)\rangle}{\|G_t(u)\|_2^2 + \delta} G_t(u), \qquad s_t(u) = \|R_t(u)\|_2 \left(\frac{\|R_t(u)\|_2}{\|J_t(u)\|_2 + \delta}\right)^\gamma. \tag{4}$$

Here, ϵ_θ is the U-Net noise predictor with frozen parameters θ. $J_t(u)$ is the anomaly-versus-normal conditional increment, $G_t(u)$ is the normal-versus-unconditional semantic direction, and $R_t(u)$ suppresses the component of $J_t(u)$ aligned with $G_t(u)$. The operator $[\cdot]_u$ extracts a local response vector, $\gamma > 0$ controls residual-ratio modulation, and $\delta > 0$ ensures numerical stability throughout. Thus, $s_t(u)$ provides residual spatial evidence for constructing M_p, rather than a detection score or true-anomaly localization.

Region-level temporal aggregation. With $s_t(u)$ in hand, we first normalize each response within M_o as $s'_t = \mathcal{N}_{\mathrm{rob}}(s_t; M_o)$, where $\mathcal{N}_{\mathrm{rob}}$ denotes robust quantile normalization over locations inside M_o. A temporal association operator $(\bar{s}, \Pi) = \mathcal{A}_{\mathrm{trk}}(\{s'_t\}_{t\in\Gamma_{s1}}, \{\bar{\alpha}_t\}_{t\in\Gamma_{s1}}; M_o)$ extracts high-response connected regions, associates them across adjacent states using expanded overlap or centroid proximity, and returns the noise-weighted response map $\bar{s}$ and track-persistence map Π. This association accommodates limited position and shape changes while suppressing isolated responses. The resulting temporal region evidence and candidate mask are

$$\Omega(u) = \frac{M_o(u)\bar{s}(u)\Pi(u)^{\gamma_{\mathrm{trk}}}}{\max\left(\max_{u'} M_o(u')\bar{s}(u')\Pi(u')^{\gamma_{\mathrm{trk}}}, \delta\right)}, \qquad M_p = \mathcal{B}(\Omega, \bar{s}, M_o; r_{\mathrm{area}}). \tag{5}$$

Here, u' ranges over all locations on the latent spatial grid, $M_p(u) \in \{0,1\}$, $|M_p| \le \lceil r_{\mathrm{area}}|M_o|\rceil$, $\gamma_{\mathrm{trk}} > 0$ controls the persistence contribution, and $|\cdot|$ counts grid locations. The operator $\mathcal{B}$ grows a connected region from Ω and falls back to $\bar{s}$ when track evidence is insufficient. It stops when the soft budget is reached or further evidence is insufficient, without exhausting the budget by adding distant fragments. Further details of normalization, noise weighting, region association, region growing, and fallback rules are provided in Appendix B.1.

3.3 Generation–Detection Closed-Loop Trajectory Optimization

The candidate region M_p specifies only where an anomaly may form. It does not guarantee that the final image contains an anomaly within this region that satisfies the frozen evaluator's criterion. We therefore convert the output of a frozen anomaly evaluator into terminal feedback and use TOC (Chang et al., 2026) to propagate it through the full diffusion trajectory to the temporal control variables.

Terminal detection feedback. To satisfy the open-set training-free setting, the anomaly evaluator SubspaceAD (Lendering et al., 2026) remains frozen during optimization, and its normal reference is constructed solely from the current I_n. SubspaceAD uses frozen DINOv2 (Oquab et al., 2024) to extract patch-level features within the object region and employs Principal Component Analysis (PCA) (Maćkiewicz & Ratajczak, 1993) to fit a normal subspace with mean μ_n and orthonormal basis U_n. For the terminal generated image, let $f_a(u)$ denote its feature at patch location u. Its normal-subspace residual is $r_{\mathrm{det}}(u) = \left\|(\mathbf{I}_f - U_n U_n^\top)(f_a(u) - \mu_n)\right\|_2^2$. Here, $\mathbf{I}_f$ is the identity matrix in the detector feature space. The residual $r_{\mathrm{det}}(u)$ measures the degree of deviation from the normal subspace at location u without distinguishing anomaly categories. Using a normal-reference residual scale ξ_n estimated from the patch residuals of the current I_n, we calibrate r_{det} and map it to the patch and pixel scales as $\{\tilde{r}_{\mathrm{det}}^q\}_{q\in\{\mathrm{patch,pixel}\}} = \mathcal{C}_{\mathrm{det}}(r_{\mathrm{det}}; \xi_n)$, where $\mathcal{C}_{\mathrm{det}}$ denotes score calibration and scale mapping. Aligning M_p with both detection scales yields M_p^q. We define the terminal detection objective as

$$L_{\mathrm{sub}} = \sum_{q\in\{\mathrm{patch,pixel}\}} \lambda_q \frac{\sum_{u\in\mathcal{U}_q} M_p^q(u)\,[m_q - \tilde{r}_{\mathrm{det}}^q(u)]_+^2}{\max\left(\sum_{u\in\mathcal{U}_q} M_p^q(u), \delta\right)}. \tag{6}$$

Here, $\mathcal{U}_q$ is the spatial grid at scale q, $[x]_+ = \max(x, 0)$, and $m_q > 0$ and $\lambda_q \ge 0$ are the anomaly formation margin and scale weight, respectively. This objective penalizes only locations below the

margin. It encourages anomalies within M_p to satisfy the frozen evaluator's criterion rather than increasing anomaly scores without bound.

Trajectory-control objective. Following Eq. 3 and fixing $z_T^a = z_T^{\mathrm{src}}$, the base trajectory-control problem is

$$\begin{aligned} \min_{W^c} \quad & \mathcal{J}(W^c) = L_{\mathrm{target}}\left(\mathcal{D}(z_0^a)\right) + \lambda_c \mathcal{R}_c(W^c), \\ \text{s.t.} \quad & z_{t-1}^a = F_t\left(z_t^a, e_a, W_t^c, M_p\right), \qquad t = T, \ldots, 1. \end{aligned} \tag{7}$$

Here, L_{target} is the terminal objective, $\mathcal{R}_c(W^c)$ regularizes the control magnitude and spatial variation, and $\lambda_c \geq 0$ is its weight. Optimization acts only on the components of W^c within the support defined by M_p. The text condition, M_p, normal reference sequence, and model parameters remain fixed. Under detection-only feedback, $L_{\mathrm{target}} = L_{\mathrm{sub}}$. Section 3.4 further introduces semantic and diversity terms.

Adjoint propagation and closed-loop update. Let $\ell \in \{0, \ldots, L-1\}$ index the closed-loop iterations, where L is the total number of updates. At iteration ℓ, unfolding the full trajectory from the same source latent yields states $\{z_t^{a,(\ell)}\}_{t=0}^T$, controls $W^{c,(\ell)} = \{W_t^{c,(\ell)}\}_{t=1}^T$, and terminal image $I_a^{(\ell)} = \mathcal{D}(z_0^{a,(\ell)})$. Let $g_{\mathrm{term}}^{(\ell)}$ denote the effective image-space terminal feedback direction at $I_a^{(\ell)}$. Under detection-only feedback, it reduces to $\nabla_{I_a} L_{\mathrm{sub}}|_{I_a = I_a^{(\ell)}}$. The adjoint states propagate this feedback to update the temporal controls:

$$p_0^{(\ell)} = \left(\frac{\partial \mathcal{D}}{\partial z_0^a}\right)^\top g_{\mathrm{term}}^{(\ell)}, \qquad p_t^{(\ell)} = \left(\frac{\partial F_t}{\partial z_t^a}\right)^\top p_{t-1}^{(\ell)}, \quad t = 1, \ldots, T. \tag{8}$$

$$g_t^{c,(\ell)} = \left(\frac{\partial F_t}{\partial W_t^c}\right)^\top p_{t-1}^{(\ell)} + \lambda_c \nabla_{W_t^c} \mathcal{R}_c\Big(W^{c,(\ell)}\Big), W_t^{c,(\ell+1)} \quad \leftarrow W_t^{c,(\ell)} - \eta g_t^{c,(\ell)}. \tag{9}$$

Here, $p_t^{(\ell)}$ is the adjoint variable obtained by propagating the effective terminal feedback to $z_t^{a,(\ell)}$, $g_t^{c,(\ell)}$ is the corresponding control-update gradient, and $\eta > 0$ is the update step size. All derivatives are evaluated along the current trajectory. The anomaly text condition e_a and candidate region M_p remain fixed, and the update is restricted to the spatial control range defined by M_p. The generation trajectory unfolds from T to 0, while terminal feedback propagates from 0 to T. After updating $W_t^{c,(\ell)}$ to $W_t^{c,(\ell+1)}$, the model regenerates from the same source latent using $W_t^{c,(\ell+1)}$ at the corresponding denoising steps, thereby closing the generation–detection loop. Detection-score calibration and scale mapping, control regularization, and bounded updates are detailed in Appendix B.2.

3.4 Semantic–Diversity Auxiliary Constraints

Terminal detection feedback promotes deviation from the normal subspace within M_p, but this deviation may not preserve the target anomaly semantics and may repeat historical patterns. We therefore incorporate semantic consistency and history-aware diversity as auxiliary terminal constraints in the same closed-loop trajectory optimization.

Semantic Consistency. Frozen CLIP (Radford et al., 2021) evaluates state and direction relations at global and local scales. Let $\rho \in \{g, \mathrm{loc}\}$ denote the spatial scale, where g and loc correspond to global and local evaluation, respectively. Let $\kappa \in \{\mathrm{sta}, \mathrm{dir}\}$ denote the relation type, where sta and dir correspond to state and direction relations, respectively. We define $\mathcal{S}_{\mathrm{sem}} = \{(g, \mathrm{sta}), (g, \mathrm{dir}), (\mathrm{loc}, \mathrm{sta}), (\mathrm{loc}, \mathrm{dir})\}$. The four semantic losses and the resulting semantic objective are defined as

$$\left\{\mathcal{L}_\rho^\kappa\right\}_{(\rho,\kappa)\in\mathcal{S}_{\mathrm{sem}}} = \mathcal{F}_{\mathrm{sem}}\left(I_n, I_a, C_n, C_a, M_p; \tau_{\mathrm{sem}}\right), \qquad L_{\mathrm{sem}} = \sum\nolimits_{(\rho,\kappa)\in\mathcal{S}_{\mathrm{sem}}} \alpha_\rho^\kappa \mathcal{L}_\rho^\kappa. \tag{10}$$

The operator $\mathcal{F}_{\mathrm{sem}}$ computes the four CLIP cosine relations and converts their relation scores into nonnegative margin losses using the shared semantic margin $\tau_{\mathrm{sem}} > 0$. State relations measure the relative preference for anomaly semantics over normal semantics, whereas direction relations measure the alignment between visual and textual semantic changes. Global relations operate on the full image. After aligning M_p to image coordinates, the local state relation evaluates the masked image, while the local direction relation uses spatially aligned normal–generated patches within M_p. The semantic objective L_{sem} aggregates the four relation losses, where $\alpha_\rho^\kappa \geq 0$ is the corresponding loss weight. The local-direction component further weights patchwise losses according to candidate region coverage and actual image changes and is set to zero when no valid patch pair exists. Exact relation scores, margin transformations, and patch-weighting rules are provided in Appendix B.3.

History-aware diversity. The current anomaly change representation and its history-aware objective are defined as

$$h_a = \mathcal{G}_{\text{chg}}\left(I_n, I_a, M_p\right), \qquad L_{\text{div}} = \left[\tau_{\text{div}} - \frac{1}{K_h}\sum_{h_j \in \mathcal{N}_{K_h}(h_a)}\left(1 - \cos(h_a, h_j)\right)\right]_+^2. \tag{11}$$

Here, using the same frozen CLIP encoder, $\mathcal{G}_{\text{chg}}$ forms h_a by aggregating normalized differences between features extracted from spatially aligned normal and generated patches within M_p. A generation sequence denotes consecutively generated samples that share the same history bank. The history set $\mathcal{H}_a$ retains representations matched by object category, target anomaly description, and generation sequence, while prioritizing samples whose area budget r_{area} is close to that of the current sample. Let $N_h = |\mathcal{H}_a|$ denote the number of valid historical representations and $0 < K_h \leq N_h$ the number of retrieved neighbors. We denote the K_h nearest historical representations of h_a by $\mathcal{N}_{K_h}(h_a)$, where h_j denotes an individual retrieved representation. The objective penalizes their average cosine distance from h_a when it falls below $\tau_{\text{div}} > 0$ and returns zero when no valid representation or history is available. This constraint reduces repeated local change patterns without claiming coverage of the true anomaly distribution. History filtering, neighbor retrieval, and history-bank updates are detailed in Appendix B.3.

Joint terminal objective. The three feedback terms form the base joint terminal objective

$$L_{\text{target}} = L_{\text{sub}} + \lambda_{\text{sem}} L_{\text{sem}} + \lambda_{\text{div}} L_{\text{div}}. \tag{12}$$

Here, $\lambda_{\text{sem}}, \lambda_{\text{div}} \geq 0$ are the semantic and diversity weights, respectively. The detection objective L_{sub} remains the primary terminal feedback. During backpropagation, auxiliary-gradient components conflicting with the detection gradient are suppressed and their relative magnitudes are bounded, yielding $g_{\text{term}}^{(\ell)}$ used in Eqs. 8–9. All terms share the adjoint propagation and control update in section 3.3 without introducing another generation process or optimizing the text conditions, M_p, or model parameters. Complete auxiliary-gradient constraints are provided in Appendix B.3.

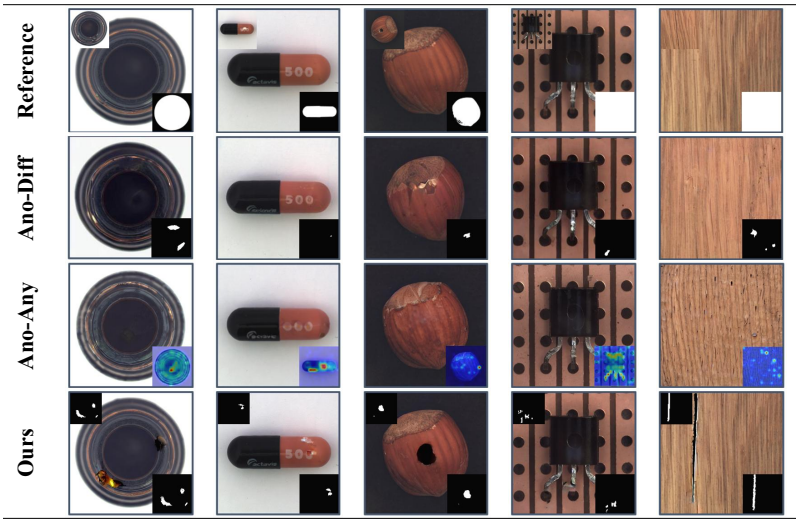

Figure 3: Qualitative comparison of anomaly generation. In the Reference row, the main image is the normal input. The upper-left inset shows a real anomaly exemplar for visual comparison only, and the lower-right inset shows the object mask. The remaining rows show results from AnomalyDiffusion, AnomalyAny, and AnomalyLoop. For AnomalyLoop, the upper-left and lower-right insets show the candidate anomaly formation region M_p and the post-generation pixel-level anomaly mask, respectively. Ano-Diff and Ano-Any denote AnomalyDiffusion and AnomalyAny.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate AnomalyLoop on two standard industrial anomaly detection benchmarks, MVTec AD (Bergmann et al., 2019) and VisA (Zou et al., 2022). MVTec AD contains 15 object and texture categories, while VisA contains 12 industrial object categories. Both provide pixel-level annotations of real anomalies across diverse object structures, materials, and defect types.

Table 1: Quantitative comparison of anomaly generation quality and diversity on MVTec AD. Best results are in bold, and second-best results are underlined. Ano-Diff and Ano-Any denote AnomalyDiffusion and AnomalyAny, respectively.

Method	Metric	bottle	cable	caps	carp	grid	hazel	leath	metal	pill	screw	tile	brush	trans	wood	zipper	Mean
Ano-Diff(2024)	IS↑	1.58	2.13	1.59	1.16	2.04	2.13	1.94	1.96	1.61	1.28	2.54	1.68	1.57	2.33	1.39	1.80
	IC-L↑	0.19	0.41	0.21	0.24	0.44	0.31	0.41	0.30	0.26	0.30	0.55	0.21	0.34	0.37	0.25	0.32
AAG(2025)	IS↑	1.54	2.33	1.87	**1.70**	**2.44**	1.98	1.79	1.94	1.87	1.30	1.56	1.97	2.00	2.51	2.41	1.95
	IC-L↑	0.17	**0.49**	**0.27**	0.31	0.48	**0.37**	0.33	**0.39**	0.30	**0.38**	0.50	**0.26**	0.36	0.38	0.27	0.35
Ano-Any(2025)	IS↑	1.73	2.06	2.16	1.10	2.31	2.55	2.26	1.82	**2.91**	1.33	2.66	1.64	1.66	1.93	2.14	2.02
	IC-L↑	0.17	0.41	0.23	0.34	0.38	0.32	0.41	0.27	0.30	0.32	0.53	0.22	0.28	0.41	0.33	0.33
Ours	IS↑	**2.50**	**2.66**	**2.17**	1.31	**2.44**	**2.57**	**2.58**	**2.03**	2.66	**1.38**	**4.27**	**2.03**	**2.51**	**2.86**	**2.63**	**2.44**
	IC-L↑	**0.32**	0.46	0.25	**0.49**	**0.50**	0.36	**0.48**	0.36	**0.35**	0.28	**0.60**	**0.26**	**0.58**	**0.47**	**0.38**	**0.41**

Evaluation Metrics. Following AnomalyAny (Sun et al., 2025), we evaluate generation quality and diversity on MVTec AD and 1-shot detection on both datasets. Generation quality and diversity are measured by Inception Score (IS) and intra-cluster pairwise LPIPS distance (IC-L), respectively. Downstream detection is evaluated by image-/pixel-level Area Under the Receiver Operating Characteristic curve (AUROC; I-AUC/P-AUC), max-F1 (I-F1/P-F1), and Per-Region Overlap (PRO). For the ablation study, we additionally report image-level Average Precision (I-AP). Further details are provided in Appendix A.2.

Implementation Details. AnomalyLoop is built on Stable Diffusion v1.5 (Rombach et al., 2022), with all diffusion model parameters frozen during inference. We set $T_{s1} = 5$ for the first stage and $T = 20$ for the second stage; r_{area} is randomly sampled within the range of possible anomaly area ratios, with the remaining implementation details provided in Appendix A.1.

4.2 Quantitative Results

Table 1 compares the quantitative generation performance of AnomalyLoop with representative published anomaly generation methods on MVTec AD, including the training-based AnomalyDiffusion (Hu et al., 2024) and the training-free methods AAG (Zuo et al., 2025) and AnomalyAny (Sun et al., 2025). AnomalyLoop achieves the best average performance on both generation quality and diversity, with IS and IC-L reaching 2.44 and 0.41, respectively. Compared with AnomalyAny, the most closely related published open-set training-free method, AnomalyLoop improves the average IS and IC-L by 20.8% and 24.2%, respectively. Overall, the results show that AnomalyLoop achieves better generation quality and sample diversity across different object and texture categories.

4.3 Qualitative Results

Figure 3 presents anomaly generation results across representative categories. Compared with AnomalyDiffusion (Hu et al., 2024) and AnomalyAny (Sun et al., 2025), which often produce weak anomalies or object-wide changes, AnomalyLoop forms clearer local defects within the explored candidate regions while better preserving the surrounding normal structures. These results support both the spatial confinement provided by ASRE and the sufficient anomaly formation promoted by detection feedback.

4.4 Downstream Detection Results

Table 2: 1-shot anomaly detection results on MVTec AD and VisA. The upper section reports baselines, and the lower section reports methods using synthetic anomalies for detector training.

Methods	MVTec AD					VisA				
	I-AUC(%)↑	I-F1(%)↑	P-AUC(%)↑	P-F1(%)↑	PRO(%)↑	I-AUC(%)↑	I-F1(%)↑	P-AUC(%)↑	P-F1(%)↑	PRO(%)↑
PatchCore(2022)	83.4±3.0	90.5±1.5	92.0±1.0	50.4±2.1	79.7±2.0	79.9±2.9	81.7±1.6	95.4±0.6	38.0±1.9	80.5±2.5
WinCLIP+(2023)	93.1±2.0	93.7±1.1	95.2±0.5	55.9±2.7	87.1±1.2	83.8±4.0	83.1±1.7	96.4±0.4	41.3±2.3	85.1±2.1
Ano-GPT(2024)	94.1±1.1	–	95.3±0.1	–	–	87.4±0.8	–	96.2±0.1	–	–
PromptAD(2024a)	94.6±1.7	–	95.9±0.5	–	–	86.9±2.3	–	96.7±0.4	–	–
AnoDiff(2024)	94.4±0.3	94.4±0.2	95.3±0.5	57.3±3.0	92.2±1.0	–	–	–	–	–
DRAEM(2021)	93.6±0.3	94.2±0.4	95.1±0.1	56.0±0.9	91.8±0.1	86.0±0.7	83.0±0.9	97.5±0.1	42.6±0.7	92.6±0.6
NSA(2022)	94.0±0.5	94.2±0.3	95.1±0.1	56.1±0.5	91.8±0.2	86.2±2.0	83.1±1.2	97.4±0.1	40.8±0.5	92.3±0.3
RealNet(2024)	92.7±0.7	93.6±0.3	95.1±0.1	56.3±1.3	91.7±0.1	86.0±1.4	82.9±1.1	97.5±0.2	41.9±1.8	**92.8±0.3**
AnoAny(2025)	94.9±0.4	94.7±0.4	95.4±0.2	57.3±0.0	91.9±0.0	89.7±0.8	85.8±0.5	97.7±0.4	**43.2±0.4**	92.5±0.1
Ours	**98.2±0.1**	**96.9±0.2**	**97.8±0.1**	**65.9±0.3**	**93.7±0.2**	**95.7±0.2**	**91.9±0.1**	**98.6±0.0**	41.3±0.1	91.9±0.1

Under the above 1-shot setting, Table 2 compares the downstream performance of AnomalyLoop with representative few-shot anomaly detection and anomaly generation methods. AnomalyLoop achieves clear advantages in image-level detection on both datasets: I-AUC/I-F1 reach 98.2%/96.9% on MVTec AD (Bergmann et al., 2019) and 95.7%/91.9% on VisA (Zou et al., 2022). For pixel-level detection, AnomalyLoop achieves the best results on P-AUC, P-F1, and PRO on MVTec AD; On

Table 3: Controlled ablation results on MVTec AD.

Variant	Stage1	L_{sub}	L_{sem} & L_{div}	IS↑	IC-L↑	I-AUC(%)↑	I-F1(%)↑	I-AP(%)↑	M-act(%)↑
ASRE	✓	–	–	1.2186	0.1863	68.24	86.52	83.00	83.65
ASRE+DF	✓	✓	–	1.2656	0.2147	71.53	86.94	84.30	88.57
Ours	✓	✓	✓	1.2631	0.2211	72.79	87.44	84.43	89.67

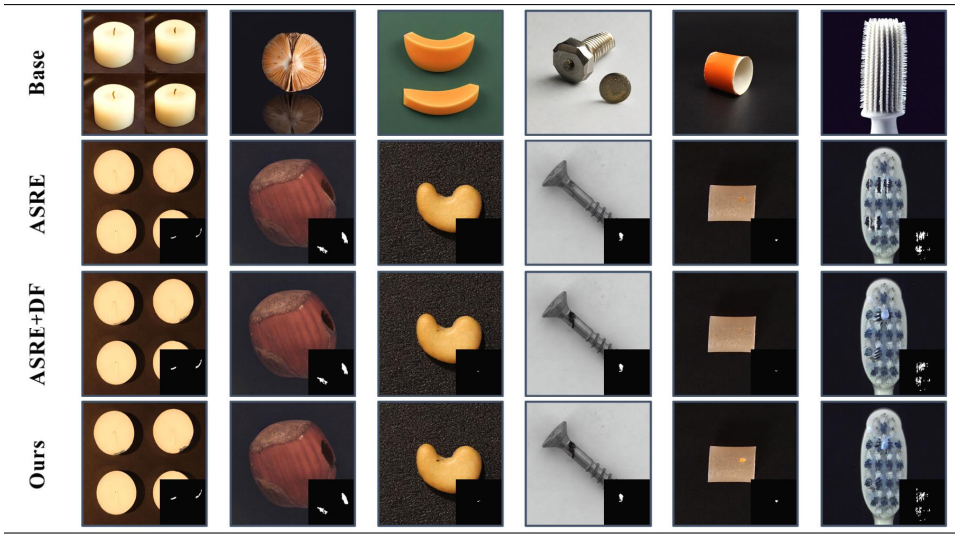

Figure 4: Qualitative ablation of AnomalyLoop. Starting from Base (SD conditioned on C_a), ASRE, ASRE+DF, and Ours progressively introduce candidate-region exploration, Detection Feedback (DF), and semantic consistency and history-aware diversity constraints.

VisA, P-AUC reaches 98.6%, while P-F1 and PRO trail the best results, indicating strong pixel-level discrimination but leaving room to improve boundaries and alignment.

4.5 Ablation Study

To analyze each component, Table 3 reports controlled ablations on MVTec AD under the 1-shot downstream detection protocol, using generated anomalies to train a lightweight U-Net detector. Results are averaged over three runs, with normal references, text conditions, candidate regions M_p, and seeds matched across variants. ASRE retains only region exploration, ASRE+DF adds closed-loop trajectory optimization driven by Detection Feedback (DF), and Ours denotes the complete AnomalyLoop. Here, M-act measures the anomaly-formation ratio within M_p, while Base denotes standard SD generation conditioned on C_a and appears only as a qualitative reference in Figure 4. ASRE+DF improves all metrics over ASRE, showing that terminal detection feedback promotes sufficient anomaly formation and improves downstream utility. Ours further improves IC-L, downstream detection, and M-act at a comparable IS, validating the joint auxiliary constraints.
Figure 4 provides consistent qualitative evidence. Base tends to induce large-scale appearance shifts, whereas ASRE confines changes to M_p but may produce weak anomalies. ASRE+DF produces clearer and more fully formed local anomalies, while Ours retains these visibility gains and further refines the anomaly appearance through the semantic-diversity auxiliary constraints.

5 Conclusion

We propose AnomalyLoop, a closed-loop trajectory optimization framework for open-set training-free anomaly generation. AnomalyLoop first constructs fine-grained candidate anomaly formation regions and then uses TOC to propagate terminal detection feedback backward through the full diffusion trajectory, thereby optimizing temporal control decisions to promote sufficient local anomaly formation. Semantic consistency and history-aware diversity further mitigate semantic drift and pattern repetition. Experimental results show that AnomalyLoop achieves leading generation quality and diversity on MVTec AD and overall improves downstream anomaly detection performance on MVTec AD and VisA. Future work may extend this framework with additional terminal evaluation signals and adaptive control over anomaly formation attributes.

AI USE STATEMENT

In this work, we used generative AI tools to generate the detailed anomaly descriptions used as C_a in our experiments. We also used generative AI tools to polish the language of the main text and appendix, assist in checking the correctness of mathematical expressions and the consistency of notation, and assist with writing and debugging selected non-core portions of the code. The research problem, core methodology, and experimental design were independently developed by the authors. All AI-assisted text and mathematical expressions were reviewed and verified by the authors, and all related code was inspected and tested by the authors. All experiments were executed and validated by the authors on local GPU hardware. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide an anonymous supplementary code package containing the implementation of AnomalyLoop for anomaly generation, together with the code for training and evaluating downstream anomaly detectors using the generated samples. The accompanying README provides instructions for running the code. The main paper and appendix further document the method, experimental setup, evaluation protocol, and necessary implementation details required to reproduce the reported results.

REFERENCES

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 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.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4183–4192, 2020.

Jinho Chang, Jaemin Kim, and Jong Chul Ye. Training-free reward-guided image editing via trajectory optimal control. In *International Conference on Learning Representations*, 2026.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Zhewei Dai, Shilei Zeng, Haotian Liu, Xurui Li, Feng Xue, and Yu Zhou. Seas: Few-shot industrial anomaly image generation with separation and sharing fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23135–23144, 2025.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pp. 475–489. Springer, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Carles Domingo i Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 53791–53846, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuanyi Duan, Wei Xu, Qinlong Wu, Guo-Sen Xie, Fang Zhao, and Caifeng Shan. Anomalycontrol: Highly-aligned anomalous image generation with controlled diffusion model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8048–8057, 2025.

Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 571–578, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1932–1940, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). 3:3, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jinwei Hu, Francesco Borsatti, Arianna Stropeni, Davide Dalle Pezze, Manuel Barusco, and Gian Antonio Susto. Mirage: Model-agnostic industrial realistic anomaly generation and evaluation for visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 7722–7731, June 2026.

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, 2024.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616, June 2023.

Yuxin Jiang, Wei Luo, Hui Zhang, Qiyu Chen, Haiming Yao, Weiming Shen, and Yunkang Cao. Anomagic: Crossmodal prompt-driven zero-shot anomaly generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5485–5493, 2026.

Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Jiafu Wu, Hao Chen, Haoxuan Wang, Wenbing Zhu, Mingmin Chi, Jun Liu, et al. 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, 2025.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Camile Lendering, Erkut Akdag, and Egor Bondarau. Subspacead: Training-free few-shot anomaly detection via subspace modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28557–28566, June 2026.

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, 2021.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. 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 (CVPR)*, pp. 16838–16848, June 2024a.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 2172–2197, 2024b.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993. doi: 10.1016/0098-3004(93)90090-R.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE, 2016.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 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, 2026.

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, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer, 2015.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, June 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free stylization using reference-based modulation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 56870–56905, 2025.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29, 2016.

Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pp. 474–489. Springer, 2022.

Jungwook Seo, Minjeong Kim, Younkwan Lee, Seungho Shin, and Sungyong Baik. Anomaly as non-conformity via training-free graph laplacian energy minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21336–21345, 2026.

Fangtao Shao, Jing Liu, Peng Wu, Zhiwei Yang, and Zhaoyang Wu. Exploiting foreground and background separation for prohibited item detection in overlapping x-ray images. *Pattern Recognition*, 2021.

Yulim So and Seokho Kang. Anostyler: Text-driven localized anomaly generation via lightweight style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15734–15742, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Fuyun Wang, Yuanzhi Wang, Xu Guo, Sujia Huang, Tong Zhang, Dan Wang, Hui Yan, Xin Liu, and Zhen Cui. Anomaly-preference image generation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Ruyi Xu, Yen-Tzu Chiu, Tai-I Chen, Oscar Chew, Yung-Yu Chuang, and Wen-Huang Cheng. Training-free industrial defect generation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24214–24223, 2025.

Minghui Yang, Jing Liu, Zhiwei Yang, and Zhaoyang Wu. Slsg: Industrial image anomaly detection with improved feature embeddings and one-class classification. *Pattern Recognition*, 156:110862, 2024a.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. *IEEE*, 2024b.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 22370–22417. Curran Associates, Inc., 2024. doi: 10.52202/079017-0704.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23174–23184, October 2023.

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, 2021.

Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, and Shijian Lu. Defect-gan: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2524–2534, January 2021.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16699–16708, 2024.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 49705–49737, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408. Springer, 2022.

Zuo Zuo, Jiahao Dong, Yanyun Qu, and Zongze Wu. Training-free anomaly generation via dual-attention enhancement in diffusion model. *arXiv preprint arXiv:2508.11550*, 2025.

A Experimental Details

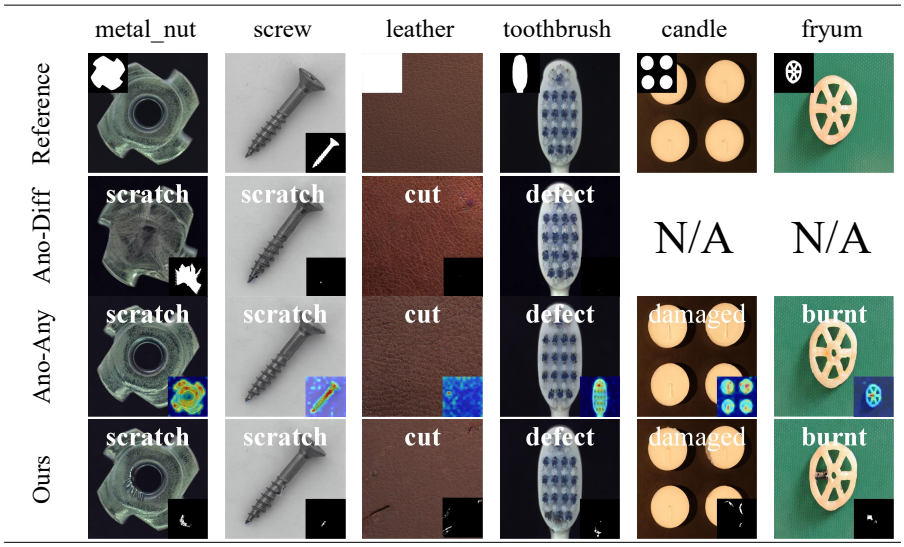

Figure 5: Additional qualitative comparisons between AnomalyLoop and existing methods on MVTec AD and VisA. The rows show the normal reference image and the results generated by AnomalyDiffusion, AnomalyAny, and AnomalyLoop, respectively. N/A indicates that the corresponding result is unavailable.

A.1 Anomaly Generation and Optimization Settings

All anomaly generation experiments are conducted on a single NVIDIA RTX A6000 GPU with 48 GB of memory. AnomalyLoop is implemented based on Stable Diffusion v1.5 (Rombach et al., 2022), with both the input and output image resolutions set to 512×512. All parameters of the pre-trained diffusion model remain frozen throughout generation. We use $T = 20$ denoising steps and set the classifier-free guidance scale to 7.5. Stage 1 uses $T_{s1} = 5$ exploration steps, and Stage 2 performs two rounds of closed-loop trajectory optimization. The cross-attention controls are bounded by $1 \leq W_t^c \leq 3$. The control update step size is set to $\eta = 0.2$. The outer control-regularization weight and its spatial-smoothness weight are set to $\lambda_c = 0.02$ and $\lambda_{\text{tv}} = 0.02$, respectively.
Terminal detection feedback is provided by a frozen SubspaceAD (Lendering et al., 2026) evaluator, which uses features extracted by DINOv2 (Oquab et al., 2024) with registers to construct the PCA (Maćkiewicz & Ratajczak, 1993) normal subspace. The anomaly formation margins at the patch and pixel scales are set to $m_{\text{patch}} = 10$ and $m_{\text{pixel}} = 5$, respectively. Their corresponding scale weights are set to $\lambda_{\text{patch}} = 5$ and $\lambda_{\text{pixel}} = 0.5$. The semantic auxiliary constraint uses a frozen CLIP (Radford et al., 2021) ViT-B/32 encoder. We set the outer semantic weight to $\lambda_{\text{sem}} = 5$ and the relation weights to $(\alpha_g^{\text{sta}}, \alpha_g^{\text{dir}}, \alpha_{\text{loc}}^{\text{sta}}, \alpha_{\text{loc}}^{\text{dir}}) = (1, 1, 1.2, 1.2)$, and the shared semantic margin is set to $\tau_{\text{sem}} = 0.1$. For local-direction evaluation, we use up to eight spatially aligned patch pairs of size 128×128, require a minimum candidate-region coverage of 0.05, and set $\tau_{\text{chg}} = 0.05$. For the history-aware diversity constraint, we set $\lambda_{\text{div}} = 2$, $\tau_{\text{div}} = 0.1$, $K_{\max} = 3$, and $N_{\min} = 20$. The

auxiliary-gradient cap parameters are set to $\kappa_{\text{sem}} = 0.10$ and $\kappa_{\text{div}} = 0.05$. The detector-anchored projection and saturation rules for these auxiliary gradients are detailed in Appendix B.3.
Unless otherwise specified, all generation experiments use these default configurations. IS and IC-L are computed following the evaluation protocol of AnomalyAny, with the same number of generated samples used for all compared methods.

A.2 1-Shot Downstream Anomaly Detection

To evaluate the utility of the generated anomalies for downstream detection, we follow the 1-shot anomaly detection protocol of AnomalyAny (Sun et al., 2025) and use the same AnomalyGPT-based (Gu et al., 2024) detection framework. The framework uses a CLIP (Radford et al., 2021) visual encoder to extract an image token and multi-level patch tokens, while the CLIP (Radford et al., 2021) text encoder provides embeddings for normal and anomalous states. A lightweight linear adapter maps the patch tokens into the text-aligned feature space. Vision-language predictions are computed from image/patch-to-text similarities, whereas vision-vision predictions compare the adapted test patch tokens with multi-level normal patch-token memory banks constructed from the 1-shot normal reference. The image-level and pixel-level predictions from the two branches are summed to obtain the final anomaly scores.
For each category, every 1-shot experiment uses one normal image as the reference. Based on 100 predefined anomaly descriptions and their corresponding r_{area} configurations, we generate 100 anomalous images I_a together with their pixel-level masks M_a. The anomaly masks are constructed through the post-generation readout described in Appendix A.4 and are used only as pixel-level annotations for downstream detector training.
The detector is trained following the AnomalyAny protocol for 200 epochs with a batch size of 16, using the Adam optimizer and a CosineAnnealingLR scheduler. For each experiment, we fix the normal reference image, anomaly descriptions, and r_{area} configurations. We repeat both anomaly generation and detector training using five independent random seeds. Detection performance is evaluated on the original real-anomaly test sets. We report image-level AUROC, image-level max-F1, pixel-level AUROC, pixel-level max-F1, and Per-Region Overlap, denoted as I-AUC, I-F1, P-AUC, P-F1, and PRO, respectively. All results are reported as the mean and standard deviation over the five runs.

A.3 Downstream Evaluation for the Ablation Study

To evaluate the downstream detection utility of samples generated by different components, we use a unified lightweight U-Net (Ronneberger et al., 2015) as the evaluator. The network adopts a six-level U-Net architecture with a base channel width of 32 and an input resolution of 256×256. The training objective is an equally weighted combination of Focal Loss and Dice Loss. The focusing parameter of Focal Loss (Lin et al., 2017) is set to 2, and the smoothing constant of Dice Loss (Milletari et al., 2016) is set to 1.
The evaluator is trained for 100 epochs with a batch size of 16 using the Adam optimizer. The initial learning rate is set to 1×10^{-4} and scheduled using MultiStepLR, with a multiplicative decay factor of 0.2 applied at epochs 80 and 90. For a controlled comparison, all ablation variants use the same normal reference image, text conditions, candidate anomaly formation region M_p, and random seeds. We use three independent diffusion seeds to generate the training data and train an identically configured U-Net for each seed. We report the mean I-AUC, I-F1, and I-AP over the three runs.
We further use M-act to measure the proportion of the candidate region in which visible anomalies are formed. We first downsample the pixel-level anomaly mask M_a to the spatial resolution of M_p, yielding $\widetilde{M}_a$. M-act is then defined as

$$M\text{-act} = \frac{\sum_u \widetilde{M}_a(u) M_p(u)}{\sum_u M_p(u)} \times 100\%, \tag{13}$$

where u denotes a spatial location on the candidate region grid, and M_a is the final pixel-level anomaly mask constructed as described in Appendix A.4. Thus, M-act measures the fraction of the candidate anomaly formation region that exhibits visible anomalous changes. It does not assess whether the candidate region corresponds to a real anomaly location. Table 3 reports the mean M-act over the three independent runs.

A.4 Anomaly Mask Construction

To obtain anomaly masks for synthetic-sample annotation, downstream detector training, and generation-side pixel-level analysis, we construct a final binary mask M_a after generation based on

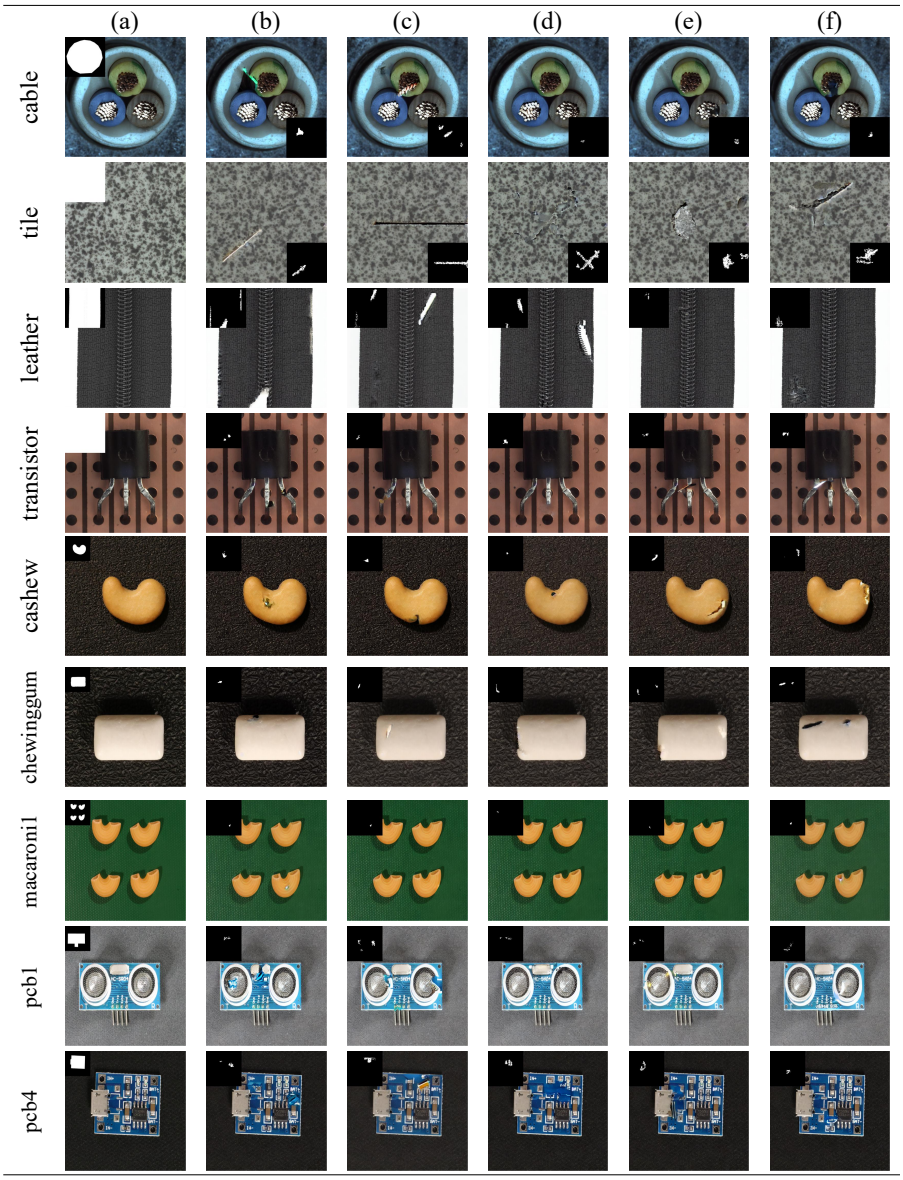

Figure 6: Additional generation results of AnomalyLoop on MVTec AD and VisA. Each row corresponds to one object category. Column (a) shows the normal reference image, and columns (b)–(f) show representative anomaly generation results.

the pixel-wise difference between the anomalous image I_a and its corresponding normal image I_n. Specifically, we first compute the pixel-wise residual between I_a and I_n. Within the local candidate region specified by M_p, we then apply a robust dual-thresholding strategy based on local residual statistics to extract the anomalous change regions. Connected-component filtering and lightweight morphological post-processing are subsequently applied to remove isolated noise and obtain the final mask M_a. All thresholds and post-processing configurations are fixed across datasets and experimental settings, with the complete configurations provided in the accompanying implementation. The resulting M_a is used only for synthetic-sample annotation, downstream detector training, and M-act computation, and is not involved in constructing M_p, computing L_{sub}, or optimizing the generation trajectory.

B Additional Method Details and Derivations

B.1 Computation Details of Anomaly Specific Region Exploration

Section 3.2 defines the spatial evidence s_t at each exploration state and presents the overall relationship between the temporal region evidence Ω and the candidate region M_p. This section details the robust normalization operator $\mathcal{N}_{\mathrm{rob}}$, the region-level temporal association operator $\mathcal{A}_{\mathrm{trk}}$, and the candidate region construction operator $\mathcal{B}$.

Robust response normalization. We first expand the robust normalization operator $\mathcal{N}_{\mathrm{rob}}$. This operator independently normalizes the spatial evidence s_t within the object region M_o at each exploration state, producing the scale-consistent response map s'_t.

Let $\mathcal{V}_t = \{s_t(u) \mid M_o(u) = 1\}$ denote the responses within the object region. We define $a_t = Q_{\rho_l}(\mathcal{V}_t)$ and $b_t = Q_{\rho_h}(\mathcal{V}_t)$ as the response values at the prescribed lower and upper quantiles, respectively, where $Q_\rho(\cdot)$ denotes the ρ-quantile. The normalized response is

$$s'_t(u) = \mathcal{N}_{\mathrm{rob}}(s_t; M_o)(u) = M_o(u)\,\mathrm{clip}_{[0,1]}\left(\frac{s_t(u) - a_t}{\max(b_t - a_t, \delta)}\right). \tag{14}$$

Here, ρ_l and ρ_h are the prescribed quantile levels, $\mathrm{clip}_{[0,1]}(\cdot)$ restricts the response to $[0, 1]$, and $\delta > 0$ is the numerical-stability constant defined in the main paper. This operation reduces scale discrepancies across exploration states and sets responses outside the object region to zero.

Region-level temporal association. We next expand the region-level temporal association operator $\mathcal{A}_{\mathrm{trk}}$. It comprises two complementary branches. The response-aggregation branch weights $\{s'_t\}_{t\in\Gamma_{s1}}$ according to the diffusion noise level and produces the temporal response strength $\bar{s}$. The region-association branch extracts high-response connected regions at each exploration state and associates them across states to obtain the track-persistence map Π. The operator $\mathcal{A}_{\mathrm{trk}}$ jointly returns these two quantities, which are subsequently combined in Eq. 5 of the main paper to form the temporal region evidence Ω.

For response aggregation, let $\nu_t = \sqrt{1 - \bar{\alpha}_t}$ denote the noise amplitude at state t. Let $\nu_{\min}$ and $\nu_{\max}$ be the minimum and maximum noise amplitudes within the exploration interval Γ_{s1}. The normalized temporal weights are

$$\widehat{\nu}_t = \frac{\nu_t - \nu_{\min}}{\nu_{\max} - \nu_{\min} + \delta}, \qquad \widetilde{\beta}_t = \beta_{\min} + (1 - \beta_{\min})\widehat{\nu}_t^{\gamma_\beta}, \qquad \beta_t = \frac{\widetilde{\beta}_t}{\sum_{\tau\in\Gamma_{s1}} \widetilde{\beta}_\tau}. \tag{15}$$

Here, $0 < \beta_{\min} \le 1$ is the lower bound on the unnormalized weight, and $\gamma_\beta \ge 0$ controls the contribution of the noise level. Consequently, $\widetilde{\beta}_t > 0$, while the normalized weights satisfy $\beta_t \ge 0$ and $\sum_{t\in\Gamma_{s1}} \beta_t = 1$.

For region association, we first extract high-response connected regions from each s'_t. Let $\mathcal{V}'_t = \{s'_t(u) \mid M_o(u) = 1\}$. We define the high-response threshold as $\tau_t = \max\{Q_{\rho_c}(\mathcal{V}'_t), \mathrm{Otsu}(\mathcal{V}'_t), \tau_{\min}\}$, where ρ_c is the candidate-response quantile, $\mathrm{Otsu}(\cdot)$ denotes Otsu thresholding, and $\tau_{\min}$ is the minimum candidate threshold. The resulting candidate region set is $\mathcal{C}_t = \mathrm{CC}(\mathbb{I}[s'_t \ge \tau_t] \odot M_o)$, where $\mathbb{I}[\cdot]$ is the indicator function and $\mathrm{CC}(\cdot)$ extracts connected regions. Candidate regions are associated in exploration order. Two regions in adjacent states are treated as consecutive observations of the same region track when their spatially dilated regions satisfy a prescribed overlap criterion or their centroids satisfy a prescribed proximity criterion. Otherwise, a new track is created. When a candidate region matches multiple existing tracks, it is assigned to the track with the strongest spatial correspondence. Stable tracks are then selected according to their noise-weighted occurrence and within-region response strength. We denote the retained set of tracks by $\mathcal{R}$. The candidate extraction, region association, and track selection criteria are fixed across all experiments.

For each track $r \in \mathcal{R}$, let $\chi_{r,t}(u) = 1$ indicate that location u lies within the region neighborhood of track r at state t, and let $\chi_{r,t}(u) = 0$ otherwise. We also set $\chi_{r,t}(u) = 0$ when track r is not observed at state t. The region-level temporal association operator returns

$$\mathcal{A}_{\mathrm{trk}}(\{s'_t\}_{t\in\Gamma_{s1}}, \{\bar{\alpha}_t\}_{t\in\Gamma_{s1}}; M_o) = (\bar{s}, \Pi), \tag{16}$$

$$\bar{s}(u) = \sum_{t\in\Gamma_{s1}} \beta_t s'_t(u), \Pi(u) = \max_{r\in\mathcal{R}} \sum_{t\in\Gamma_{s1}} \beta_t \chi_{r,t}(u). \tag{17}$$

Here, $\bar{s}(u)$ is the noise-aware response strength at location u over the exploration trajectory and is computed independently of the region tracks. The quantity $\Pi(u)$ is the maximum track-persistence

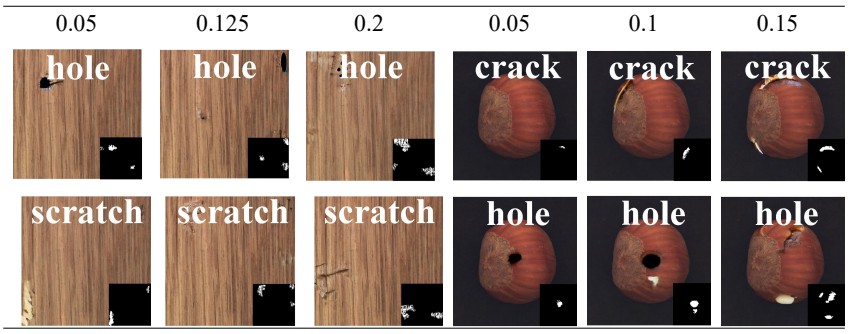

Figure 7: Effect of the candidate region area budget r_{area} on the spatial extent of anomaly formation. Each group compares results for the same object and anomaly description under different values of r_{area}.

support received by location u and is determined by cross-state region association. If no valid track is formed, namely $\mathcal{R} = \varnothing$, we set $\Pi(u) = 0$.

Region association therefore does not directly modify $\bar{s}$. Instead, it modulates the response evidence through Π. Eq. 5 of the main paper combines the two quantities as $M_o(u)\bar{s}(u)\Pi(u)^{\gamma_{\text{trk}}}$, ensuring that a strong spatial response retains a high value in Ω only when it is also supported by a persistent region track.

Candidate-region construction. We finally describe the candidate region construction operator $\mathcal{B}$. The operator primarily uses the temporal region evidence Ω from Eq. 5, while the weighted response map $\bar{s}$ provides a fallback when track evidence is insufficient. It constructs the candidate anomaly formation region M_p under the object constraint M_o and the soft area budget r_{area}.

For each retained track $r \in \mathcal{R}$, let $\mathcal{Z}_r = \{u \mid \max_{t\in\Gamma_{s1}} \chi_{r,t}(u) = 1\}$ denote its spatial coverage across exploration states, and let $u_r^\star = \arg\max_{u\in\mathcal{Z}_r} \Omega(u)$ be its region-growing seed. The retained tracks are processed according to their temporal persistence and response strength while sharing the global area budget $\lceil r_{\text{area}}|M_o|\rceil$. Let G denote the accumulated union of the grown regions. Starting from each seed, region growing iteratively adds connected locations within $\mathcal{Z}_r$ in descending order of Ω while $|G|$ remains below the global area budget. Growth from the current seed stops when the evidence at the next location falls below the prescribed stopping criterion, after which the next retained track is processed if budget remains.

If no valid track is formed, or if the retained tracks provide insufficient evidence to form a valid region, the maximum-response location in $\bar{s}$ is used as a fallback seed. Connected region growing then proceeds within M_o using $\bar{s}$ as the fallback evidence and the remaining global area budget. The resulting fallback region is included in G. The final candidate region is

$$M_p(u) = \mathcal{B}(\Omega, \bar{s}, M_o; r_{\text{area}})(u) = M_o(u)\mathbb{I}[u \in G], \qquad |M_p| \le \lceil r_{\text{area}}|M_o|\rceil\,. \tag{18}$$

Here, $|\cdot|$ denotes the number of grid locations in a region. The parameter r_{area} therefore specifies an upper bound on the candidate region area rather than an area target that must be reached. Region growing may terminate early when subsequent spatial evidence is insufficient, and it does not exhaust the budget by adding distant disconnected responses.

The resulting M_p specifies only a candidate region in which an anomaly may subsequently form. It is neither a true anomaly location nor the final pixel-level anomaly mask.

B.2 Computation Details of Terminal Detection and Trajectory Control

Section 3.3 defines the normal-subspace residual r_{det}, the terminal detection objective L_{sub}, and the updates of the adjoint states and temporal controls. This section further details the detection-score calibration operator $\mathcal{C}_{\text{det}}$, the control regularizer $\mathcal{R}_c$, and the bounded control update.

Detection-score calibration and scale mapping. We first expand the detection-score calibration operator $\mathcal{C}_{\text{det}}$. It calibrates the patch-level residuals of the terminal generated image using the normal-subspace residual distribution of the current normal image I_n, and then maps the calibrated scores to the patch and pixel scales.

Using the normal-subspace mean μ_n and orthonormal basis U_n defined in the main paper, the subspace residual at an object-region patch location v in the normal image is $r_n(v) = \|(\mathbf{I}_f - U_n U_n^\top)(f_n(v) - \mu_n)\|_2^2$, where $f_n(v)$ is the frozen DINOv2 (Oquab et al., 2024) feature of the current normal image at location v. We define

$$\xi_n = Q_{0.95}(\{r_n(v) \mid v \text{ lies within } M_o\}) \tag{19}$$

as the 95th percentile of the normal patch residuals. Detection-score calibration and scale mapping are given by $\mathcal{C}_{\mathrm{det}}(r_{\mathrm{det}};\xi_n)=\left\{\widetilde{r}_{\mathrm{det}}^{\mathrm{patch}},\widetilde{r}_{\mathrm{det}}^{\mathrm{pixel}}\right\}$,

$$\widetilde{r}_{\mathrm{det}}^{\mathrm{patch}}(u)=\frac{r_{\mathrm{det}}(u)}{\max(\xi_n,\delta)},\qquad \widetilde{r}_{\mathrm{det}}^{\mathrm{pixel}}=\mathrm{BilinearResize}\Big(\widetilde{r}_{\mathrm{det}}^{\mathrm{patch}}\Big). \tag{20}$$

Here, $\mathrm{BilinearResize}(\cdot)$ bilinearly maps the patch-level score map to the image resolution. Both scales share the same normal subspace and calibration reference ξ_n. They therefore differ only in spatial evaluation granularity rather than in the underlying anomaly criterion.

Accordingly, the candidate region M_p is mapped to the detection patch grid according to its spatial coverage, yielding M_p^{patch}, and is further aligned with the image resolution to obtain M_p^{pixel}. In Eq. 6 of the main paper, M_p^q denotes M_p^{patch} when $q=\mathrm{patch}$ and M_p^{pixel} when $q=\mathrm{pixel}$. These masks are paired with $\widetilde{r}_{\mathrm{det}}^{\mathrm{patch}}$ and $\widetilde{r}_{\mathrm{det}}^{\mathrm{pixel}}$, respectively, to form the two-scale terminal detection objective.

Control regularization. We next expand the control regularizer $\mathcal{R}_c(W^c)$ in Eq. 7 of the main paper. It constrains both the deviation of the temporal controls from the original cross-attention and abrupt local spatial variations in the control fields:

$$\mathcal{R}_c(W^c)=\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{Mean}\big((W_t^c-\mathbf{1}_c)^2\big)+\lambda_{\mathrm{tv}}\,\mathrm{TV}_{xy}(W_t^c)\right]. \tag{21}$$

Here, $\mathbf{1}_c$ is the all-one baseline control defined in the main paper, $\mathrm{Mean}(\cdot)$ averages over the elements of the control matrix, and $\mathrm{TV}_{xy}(\cdot)$ is the sum of the mean absolute differences between horizontally and vertically adjacent spatial locations in the control field of each text token. The coefficient $\lambda_{\mathrm{tv}}\geq 0$ is the spatial-smoothness weight.

The first term limits the overall deviation of W_t^c from the baseline control $\mathbf{1}_c$, while the second suppresses spatially isolated or abrupt control responses. The regularizer is applied independently at each timestep and introduces no temporal finite difference.

Bounded control updates under the candidate region constraint. We finally specify the regional constraint used by the control update in Eq. 9 of the main paper. For a controlled cross-attention layer with spatial resolution $N\times N$, the candidate region M_p is aligned with this grid to obtain the binary region $M_p^{(N)}$. The control matrix W_t^c remains defined over the complete spatial–token grid, while updates are permitted only at spatial locations within $M_p^{(N)}$.

After each iteration, the control variables are regionally constrained and bounded as

$$W_t^{c,(\ell+1)}(u,j)=\begin{cases}1, & M_p^{(N)}(u)=0,\\ \mathrm{clip}_{[1,W_{\max}^c]}\Big(W_t^{c,(\ell)}(u,j)-\eta g_t^{c,(\ell)}(u,j)\Big), & M_p^{(N)}(u)=1.\end{cases} \tag{22}$$

Here, u and j index a spatial location and a text token in the current cross-attention layer, respectively. The scalar $W_{\max}^c>1$ is the control upper bound. The quantities $\eta>0$ and $g_t^{c,(\ell)}$ are the update step size and control gradient defined in Eq. 9 of the main paper. The operator $\mathrm{clip}_{[1,W_{\max}^c]}(\cdot)$ restricts each control value to $[1,W_{\max}^c]$.

Controls outside the candidate region therefore remain at the baseline $\mathbf{1}_c$, while those inside the candidate region are updated according to terminal feedback within the prescribed bounds. After all temporal controls have been updated, the states from the previous iteration are discarded and the full trajectory is unfolded again from the same source latent z_T^{src}. Terminal detection feedback thereby affects the next generated result through the updated temporal controls rather than by directly modifying an already generated terminal image.

B.3 Computation Details of the Semantic–Diversity Auxiliary Constraints

Section 3.4 introduces semantic consistency and history-aware diversity as auxiliary terminal constraints. This section expands the semantic-relation operator $\mathcal{F}_{\mathrm{sem}}$, the local-patch reliability weights, the anomaly change representation operator $\mathcal{G}_{\mathrm{chg}}$, and the detector-anchored auxiliary-gradient correction.

Semantic-relation operator. We first expand the semantic-relation operator $\mathcal{F}_{\mathrm{sem}}$ in Eq. 10 of the main paper. Using frozen CLIP (Radford et al., 2021), this operator evaluates the generated state and the normal-to-generated change direction at both global and local scales, and converts the four relations into nonnegative semantic losses.

Let ϕ_I and ϕ_T denote the frozen CLIP (Radford et al., 2021) image and text encoders. The global visual features of the normal and generated images are $v_n = \phi_I(I_n)$ and $v_a = \phi_I(I_a)$. The semantic features of the normal and anomaly texts are $e_n^{\text{sem}} = \text{Norm}(\phi_T(C_n))$ and $e_a^{\text{sem}} = \text{Norm}(\phi_T(C_a))$, where $\text{Norm}(\cdot)$ denotes numerically stabilized ℓ_2 normalization. The superscript sem distinguishes the CLIP text features from the diffusion text conditions e_n and e_a.

We align the candidate region with the image resolution to obtain M_p^{pixel} and define the local state feature as $v_a^{\text{loc}} = \phi_I(M_p^{\text{pixel}} \odot I_a)$. We then select up to eight spatially aligned normal–generated patch pairs according to M_p^{pixel}, and let $K_v \le 8$ denote the number selected for the current sample. Let $\mathcal{P}_k$ denote the pixel region of the k-th patch, and let $I|_{\mathcal{P}_k}$ denote the crop of image I over this region. The corresponding normal and generated visual features are $v_{n,k}^p = \phi_I(I_n|_{\mathcal{P}_k})$ and $v_{a,k}^p = \phi_I(I_a|_{\mathcal{P}_k})$, respectively. The four semantic relation scores are

$$\begin{aligned} \sigma_g^{\text{sta}} &= \cos(v_a, e_a^{\text{sem}}) - \cos(v_a, e_n^{\text{sem}}), & \sigma_g^{\text{dir}} &= \cos(v_a - v_n, e_a^{\text{sem}} - e_n^{\text{sem}}), \\ \sigma_{\text{loc}}^{\text{sta}} &= \cos(v_a^{\text{loc}}, e_a^{\text{sem}}) - \cos(v_a^{\text{loc}}, e_n^{\text{sem}}), & \sigma_{\text{loc},k}^{\text{dir}} &= \cos\Big(v_{a,k}^p - v_{n,k}^p, e_a^{\text{sem}} - e_n^{\text{sem}}\Big). \end{aligned} \tag{23}$$

Here, $\cos(\cdot,\cdot)$ denotes numerically stabilized cosine similarity. The state relations measure whether the generated visual content favors anomaly semantics over normal semantics. The direction relations measure whether the normal-to-generated visual change is aligned with the normal-to-anomaly textual change.

Local-patch reliability weights. The local direction relation should be supported only by patches that sufficiently cover the candidate region and already exhibit an actual image change. Let o_k be the candidate region coverage ratio of patch $\mathcal{P}_k$, and let $\bar{d}_k$ be the mean RGB change magnitude within the candidate portion of that patch. We define the reliability weight as

$$o_k = \frac{\sum_{x\in\mathcal{P}_k} M_p^{\text{pixel}}(x)}{|\mathcal{P}_k|}, \qquad w_k = \text{sg}\left[o_k \min\left(1, \frac{\bar{d}_k}{\tau_{\text{chg}}}\right)\right], \tag{24}$$

$$\bar{d}_k = \frac{\sum_{x\in\mathcal{P}_k} M_p^{\text{pixel}}(x)\sqrt{\frac{1}{3}\|I_a(x) - I_n(x)\|_2^2 + \delta}}{\max\Big(\sum_{x\in\mathcal{P}_k} M_p^{\text{pixel}}(x), \delta\Big)}. \tag{25}$$

Here, x is an image-pixel location, $|\mathcal{P}_k|$ is the number of pixels in the patch, $\tau_{\text{chg}} > 0$ is the saturation scale for change-based weighting, and $\text{sg}[\cdot]$ denotes stop-gradient. The quantity w_k therefore serves only as a reliability coefficient for the current patch and introduces no additional optimization direction through the weight itself.

Given the shared semantic margin $\tau_{\text{sem}} > 0$, the four semantic losses are

$$\begin{aligned} \mathcal{L}_g^{\text{sta}} &= [\tau_{\text{sem}} - \sigma_g^{\text{sta}}]_+^2, & \mathcal{L}_g^{\text{dir}} &= [\tau_{\text{sem}} - \sigma_g^{\text{dir}}]_+^2, \\ \mathcal{L}_{\text{loc}}^{\text{sta}} &= [\tau_{\text{sem}} - \sigma_{\text{loc}}^{\text{sta}}]_+^2, & \mathcal{L}_{\text{loc}}^{\text{dir}} &= \frac{\sum_{k=1}^{K_v} w_k [\tau_{\text{sem}} - \sigma_{\text{loc},k}^{\text{dir}}]_+^2}{\max\Big(\sum_{k=1}^{K_v} w_k, \delta\Big)}, \end{aligned} \tag{26}$$

where $[x]_+ = \max(x, 0)$. The semantic-relation operator in Eq. 10 of the main paper therefore returns

$$\mathcal{F}_{\text{sem}}(I_n, I_a, C_n, C_a, M_p; \tau_{\text{sem}}) = \left\{\mathcal{L}_g^{\text{sta}}, \mathcal{L}_g^{\text{dir}}, \mathcal{L}_{\text{loc}}^{\text{sta}}, \mathcal{L}_{\text{loc}}^{\text{dir}}\right\}. \tag{27}$$

When $\sum_k w_k = 0$, we set $\mathcal{L}_{\text{loc}}^{\text{dir}} = 0$. The four losses are then weighted by α_ρ^κ as defined in the main paper to obtain L_{sem}.

Anomaly change representation. We next expand the anomaly change representation operator $\mathcal{G}_{\text{chg}}$ in Eq. 11 of the main paper. The operator reuses the spatially aligned patches and reliability weights defined above to aggregate normalized normal–generated feature differences within the candidate region, rather than comparing the global appearance of the entire object.

The normalized visual change direction of the k-th patch is $d_k^v = \text{Norm}(v_{a,k}^p - v_{n,k}^p)$. The current anomaly change representation is

$$h_a = \mathcal{G}_{\text{chg}}(I_n, I_a, M_p) = \text{Norm}\left(\frac{\sum_{k=1}^{K_v} w_k d_k^v}{\max\Big(\sum_{k=1}^{K_v} w_k, \delta\Big)}\right). \tag{28}$$

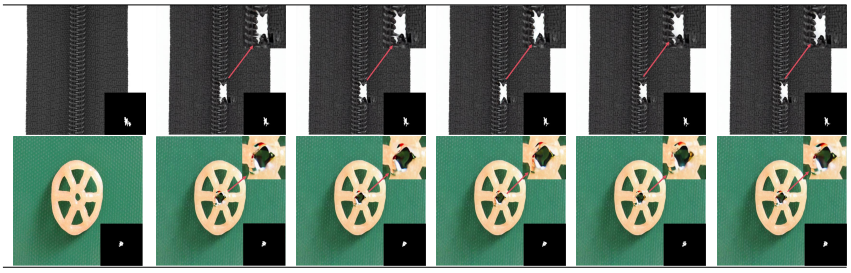

Figure 8: Qualitative results with the history-aware diversity objective L_{div}. Under otherwise identical generation conditions, the generated samples exhibit distinct local anomaly details at corresponding spatial positions within M_p.

Here, h_a is the current anomaly change direction obtained by aggregating valid local changes. When $\sum_k w_k = 0$, the current sample does not yield a valid change representation and its history-aware diversity loss is set to zero.

History-aware diversity. We provide the history-filtering and neighbor-retrieval rules used in Eq. 11. Among the completed representations matched by object category, target anomaly description, and generation sequence, we retain at most $N_{\mathrm{win}} = 50$ records whose area budgets are closest to the current r_{area}. The retained records form the area-conditioned history set $\mathcal{H}_a$. Let $N_h = |\mathcal{H}_a|$ denote the number of valid historical representations, $N_{\min} \in \mathbb{N}_+$ the minimum history size required to activate the constraint, and $K_{\max} \in \mathbb{N}_+$ the maximum number of retrieved neighbors. When $N_h \geq N_{\min}$, we set $K_h = \min(K_{\max}, N_h)$ and retrieve $\mathcal{N}_{K_h}(h_a)$ by cosine distance. We set $L_{\mathrm{div}} = 0$ when no valid change representation exists or when $N_h < N_{\min}$. The filtered history set remains fixed throughout all closed-loop iterations for the current sample, while its nearest neighbors may be retrieved again using the updated h_a at each iteration. The current representation is added to the corresponding history set only after all closed-loop iterations are complete. Intermediate generated results are not used in the history comparison for subsequent samples.

Detector-anchored auxiliary gradients. We finally expand the auxiliary-gradient restriction used to construct $g_{\mathrm{term}}^{(\ell)}$ in the main paper. This operation does not introduce a separate optimization process. Instead, it suppresses components of the auxiliary gradients that conflict with detection feedback in the terminal image space and bounds their magnitudes relative to the detection gradient. At closed-loop iteration ℓ, we define the detection gradient as $g_{\mathrm{sub}}^{(\ell)} = \nabla_{I_a} L_{\mathrm{sub}}$. For $j \in \{\mathrm{sem}, \mathrm{div}\}$, we define the auxiliary gradient as $g_j^{(\ell)} = \nabla_{I_a}(\lambda_j L_j)$, where λ_j denotes λ_{sem} or λ_{div} for the corresponding term. All gradients are evaluated at the current terminal image $I_a^{(\ell)}$.

The component of an auxiliary gradient that opposes the detection gradient is suppressed by

$$a_j^{(\ell)} = \min\left(0, \frac{\langle g_j^{(\ell)}, g_{\mathrm{sub}}^{(\ell)} \rangle}{\|g_{\mathrm{sub}}^{(\ell)}\|_2^2 + \delta}\right), \qquad \bar{g}_j^{(\ell)} = g_j^{(\ell)} - a_j^{(\ell)} g_{\mathrm{sub}}^{(\ell)}. \tag{29}$$

When the two gradients do not conflict, $a_j^{(\ell)} = 0$.

When the detection gradient is nonzero, $\kappa_j \geq 0$ specifies the maximum norm ratio of the auxiliary correction to the detection gradient. When the detection gradient is locally saturated, κ_j further limits the retained fraction of the auxiliary gradient. The conflict-corrected auxiliary gradient is scaled by

$$c_j^{(\ell)} = \begin{cases} \min\left(1, \dfrac{\kappa_j \|g_{\mathrm{sub}}^{(\ell)}\|_2}{\|\bar{g}_j^{(\ell)}\|_2 + \delta}\right), & \|g_{\mathrm{sub}}^{(\ell)}\|_2 > \delta, \\ \min(1, \kappa_j), & \|g_{\mathrm{sub}}^{(\ell)}\|_2 \leq \delta \text{ and } \|\bar{g}_j^{(\ell)}\|_2 > \delta, \\ 0, & \|\bar{g}_j^{(\ell)}\|_2 \leq \delta. \end{cases} \tag{30}$$

The final auxiliary corrections and the effective terminal feedback direction are

$$\widehat{g}_j^{(\ell)} = c_j^{(\ell)} \bar{g}_j^{(\ell)}, \qquad g_{\mathrm{term}}^{(\ell)} = g_{\mathrm{sub}}^{(\ell)} + \widehat{g}_{\mathrm{sem}}^{(\ell)} + \widehat{g}_{\mathrm{div}}^{(\ell)}. \tag{31}$$

The coefficients $a_j^{(\ell)}$ and $c_j^{(\ell)}$ are treated as constants during subsequent adjoint propagation. Semantic and diversity feedback therefore act only as bounded corrections around the detection feedback

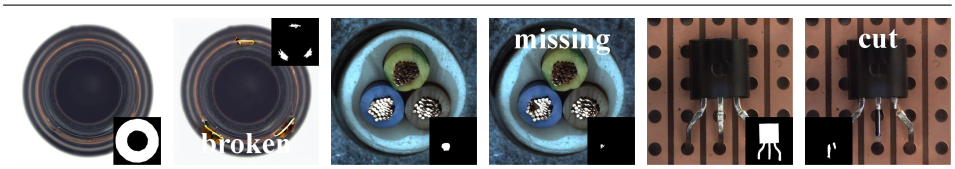

Figure 9: Anomaly generation under fine-grained spatial priors. The original object-level mask M_{obj} is replaced with an externally specified fine-grained mask to constrain the feasible region of anomaly formation.

Table 4: Per-class 1-shot downstream anomaly detection results on MVTec AD. All metrics are reported as percentages. Each entry gives the mean and standard deviation over five random seeds.

Category	I-AUC	I-F1	P-AUC	P-F1	PRO
bottle	100.0 ± 0.0	99.8 ± 0.3	98.4 ± 0.1	77.9 ± 0.4	94.7 ± 0.2
cable	92.3 ± 0.8	89.1 ± 0.9	97.3 ± 0.2	67.3 ± 1.0	91.3 ± 0.7
capsule	99.3 ± 0.2	98.7 ± 0.2	98.5 ± 0.1	58.1 ± 0.5	94.9 ± 0.1
carpet	100.0 ± 0.0	100.0 ± 0.0	99.5 ± 0.0	69.0 ± 0.3	97.5 ± 0.1
grid	99.6 ± 0.2	99.1 ± 0.0	98.6 ± 0.1	54.9 ± 0.5	95.3 ± 0.3
hazelnut	95.4 ± 1.6	91.4 ± 2.1	93.2 ± 0.5	71.9 ± 0.2	86.6 ± 0.3
leather	100.0 ± 0.0	100.0 ± 0.0	99.5 ± 0.0	64.8 ± 0.8	98.4 ± 0.1
metal_nut	100.0 ± 0.0	100.0 ± 0.0	95.8 ± 0.4	78.5 ± 0.3	94.0 ± 0.4
pill	99.0 ± 0.3	98.3 ± 0.2	98.2 ± 0.1	68.7 ± 0.3	97.5 ± 0.1
screw	93.4 ± 0.7	94.8 ± 0.7	98.9 ± 0.2	48.9 ± 3.5	94.3 ± 0.8
tile	100.0 ± 0.0	100.0 ± 0.0	98.3 ± 0.1	78.9 ± 0.1	94.5 ± 0.1
toothbrush	99.2 ± 0.6	97.7 ± 1.4	99.1 ± 0.0	53.7 ± 0.8	92.5 ± 0.2
transistor	95.7 ± 0.5	88.1 ± 2.3	94.6 ± 0.9	55.1 ± 1.6	82.4 ± 2.2
wood	99.6 ± 0.1	98.2 ± 0.3	97.5 ± 0.0	65.6 ± 0.2	94.3 ± 0.2
zipper	98.8 ± 0.1	98.1 ± 0.2	99.0 ± 0.0	73.3 ± 0.4	96.5 ± 0.1
Average	$\mathbf{98.2 \pm 0.1}$	$\mathbf{96.9 \pm 0.2}$	$\mathbf{97.8 \pm 0.1}$	$\mathbf{65.9 \pm 0.3}$	$\mathbf{93.7 \pm 0.2}$

and update the temporal controls through the same adjoint propagation in Eqs. (8)–(9) of the main paper. This restriction does not imply that L_{sub} must decrease monotonically after regenerating the full trajectory.

C Additional Qualitative Results on MVTec AD and VisA

C.1 Additional Comparisons with Existing Methods

Figure 5 provides additional comparisons across different object categories and anomaly descriptions. Some results from AnomalyDiffusion and AnomalyAny exhibit weak anomaly changes, indistinct local structures, or unintended large-scale modifications to normal content. In contrast, AnomalyLoop generally produces clearer local anomalies within the candidate regions while better preserving the surrounding normal content.

C.2 Additional Generation Results of AnomalyLoop

Figure 6 further presents AnomalyLoop results across different structures, materials, and anomaly conditions, providing additional evidence of its applicability to diverse object categories and anomaly semantics.

D Per-Class 1-Shot Downstream Detection Results

Tables 4 and 5 further report the per-class downstream 1-shot anomaly detection results obtained with the generated samples. Image-level metrics and P-AUC remain high for most categories on both datasets, indicating that the overall performance is not driven solely by a small subset of categories. In contrast, P-F1 varies more noticeably across VisA categories, indicating category-dependent differences in pixel-level segmentation quality.

Table 5: Per-class 1-shot downstream anomaly detection results on VisA. All metrics are reported as percentages. Each entry gives the mean and standard deviation over five random seeds.

Category	I-AUC	I-F1	P-AUC	P-F1	PRO
candle	97.9 ± 0.1	93.5 ± 0.5	99.2 ± 0.0	39.9 ± 0.6	95.8 ± 0.3
capsules	92.7 ± 0.2	90.1 ± 0.8	99.1 ± 0.1	40.9 ± 1.0	90.1 ± 0.9
cashew	98.3 ± 0.1	95.7 ± 0.5	98.7 ± 0.2	41.2 ± 1.2	94.3 ± 1.2
chewinggum	99.8 ± 0.0	98.7 ± 0.4	99.5 ± 0.1	67.8 ± 0.5	91.1 ± 0.3
fryum	97.5 ± 0.1	94.5 ± 0.2	96.8 ± 0.2	43.9 ± 0.1	89.2 ± 0.2
macaroni1	96.2 ± 0.2	90.6 ± 0.9	99.2 ± 0.0	28.2 ± 0.3	95.1 ± 0.6
macaroni2	87.1 ± 2.9	80.5 ± 2.4	98.4 ± 0.1	22.5 ± 0.8	89.8 ± 1.8
pcb1	93.6 ± 0.7	89.8 ± 0.8	99.4 ± 0.2	45.7 ± 1.5	93.9 ± 1.4
pcb2	93.9 ± 0.7	86.8 ± 1.2	98.5 ± 0.1	41.1 ± 0.6	88.9 ± 0.9
pcb3	91.6 ± 0.7	84.5 ± 1.0	98.2 ± 0.1	37.9 ± 2.0	90.4 ± 1.0
pcb4	99.5 ± 0.2	98.3 ± 0.6	98.1 ± 0.4	38.6 ± 2.6	88.0 ± 2.0
pipe_fryum	99.8 ± 0.0	99.1 ± 0.4	98.3 ± 0.0	48.3 ± 0.3	95.8 ± 0.2
Average	$\mathbf{95.7 \pm 0.2}$	$\mathbf{91.9 \pm 0.1}$	$\mathbf{98.6 \pm 0.0}$	$\mathbf{41.3 \pm 0.1}$	$\mathbf{91.9 \pm 0.1}$

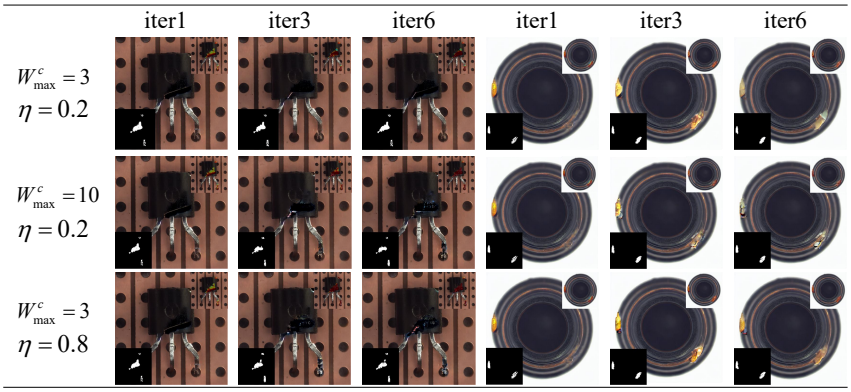

Figure 10: Effects of the number of closed-loop iterations and the control strength. The columns show results after different numbers of closed-loop iterations, and the groups correspond to different update step sizes η and control upper bounds $W^c_{\max}$. All other generation conditions are held fixed.

E Additional Analyses

E.1 Effect of the Candidate-Region Area Budget

As r_{area} increases, the spatial extent of the generated anomalies generally expands in the examples shown in Figure 7. Because r_{area} is a soft area budget that need not be exhausted, the final anomaly extent can still adapt to the object structure and anomaly semantics. The parameter therefore provides flexible control over the scale of anomaly formation rather than prescribing the exact final anomaly area.

E.2 Qualitative Behavior of History-Aware Diversity

As shown in Figure 8, under otherwise identical generation conditions, the generated samples exhibit different local details at corresponding spatial positions within M_p while preserving the target semantics and spatial constraints. These qualitative results are consistent with the intended role of the history-aware diversity objective in discouraging repeated local change patterns.

E.3 Generation under Fine-Grained Spatial Priors

As shown in Figure 9, we replace the original object-level mask M_{obj} with a finer spatial prior while keeping the remaining AnomalyLoop pipeline unchanged. The finer M_{obj} narrows the feasible region explored in Stage 1, leading to more localized anomaly formation and more visually plausible and natural generated results. These results demonstrate that AnomalyLoop can effectively incorporate fine-grained spatial priors when available.

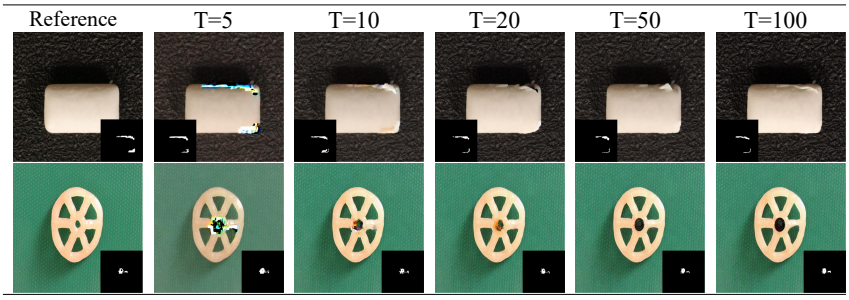

Figure 11: Effect of the number of denoising steps T. All other generation conditions are held fixed.

E.4 Effects of Closed-Loop Iterations and Control Strength

As shown in Figure 10, increasing the number of closed-loop iterations makes the generated anomalies appear more natural. Moderately increasing η or $W^c_{\max}$ strengthens anomaly formation within the candidate region and improves its visibility, whereas excessive values may introduce local distortions or unnatural artifacts.

E.5 Effect of the Number of Denoising Steps

As shown in Figure 11, small values of T tend to produce visible distortion or blur, whereas the generated results become comparatively clear when $T \geq 20$. Increasing the number of steps to $T = 50$ or $T = 100$ yields only limited additional visual changes. These observations indicate that $T = 20$ provides a favorable balance between generation quality and inference efficiency under our setting.

F Anomaly Prompt Construction and Examples

As shown in Figure 12, a predefined prompt template and GPT expand each object name–anomaly keyword pair into a detailed description used as the anomaly text condition C_a. The description is generated once before synthesis and remains fixed throughout diffusion sampling and closed-loop trajectory optimization. GPT is not invoked during the subsequent generation process.

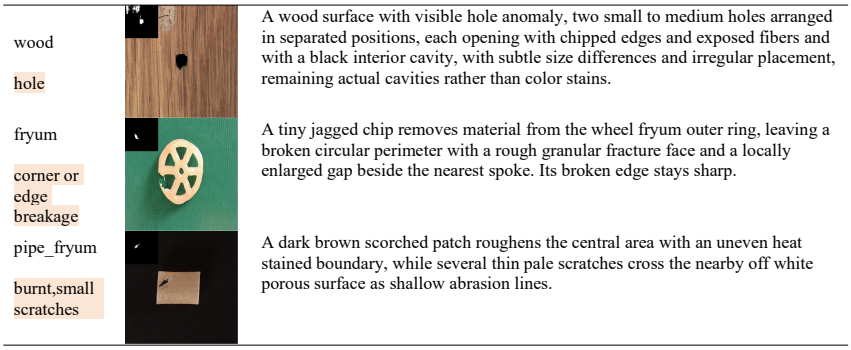

Figure 12: Examples of GPT-assisted anomaly-text construction. Each row shows an object name and anomaly keyword, the resulting condition C_a, and the generated image.