DecoGen: ZERO-SHOT CROSS-STYLE INDUSTRIAL ANOMALY GENERATION VIA EXPLICIT PHYSICAL SEMANTIC CONSTRAINT

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, industrial anomaly image generation has shown immense potential in alleviating the scarcity of real defect data. However, existing few-shot generation methods typically rely on implicit and coarse-grained visual guidance lacking real-world physical grounding, which inherently entangles anomalous features with domain-specific background textures. This representation entanglement leads to severe texture contamination and semantic misalignment, making it exceedingly difficult to transfer known defects to novel product styles that lack real anomaly images. To address these issues, we propose ***DecoGen***, a novel zero-shot cross-style anomaly transfer framework driven by **explicit physical semantic constraints**. Specifically, we construct a structured anomaly prompt bank to explicitly decompose defects into multi-dimensional, domain-invariant physical properties for flexible semantic combinations. Furthermore, a physical attribute aligner is designed to enforce rigorous disentangled alignment in the feature space, thoroughly decoupling the pure physical anomaly appearance from both the source background texture and its mask-derived spatial support. Experiments show that ***DecoGen*** yields high-fidelity and diverse cross-style anomalies. Crucially, our synthesized data significantly improves the performance of downstream anomaly detectors on unseen product styles, increasing the average image-level and pixel-level AUROC by 3.99% and 6.32%, respectively.

1 INTRODUCTION

Industrial visual inspection, encompassing anomaly detection, localization, and classification, plays a pivotal role in manufacturing quality control. However, in real-world production, genuine anomalous samples are notoriously scarce and costly to annotate. This fundamental limitation becomes especially acute in modern manufacturing scenarios where product styles change rapidly. When encountering a newly designed product (e.g., a novel fabric pattern or texture), abundant normal images are immediately available, whereas representative defects may not appear for an extended period. Consequently, mainstream unsupervised or semi-supervised anomaly detectors trained on previous styles often suffer from severe performance degradation under the resulting appearance shift, creating a critical "cold-start" problem Jeong et al. (2023); Zhang et al. (2024). To break this bottleneck, anomaly generation methods Hu et al. (2024) have emerged as a vital solution to synthesize realistic defect data for unseen styles, thereby enabling robust and high-performance supervised inspection.

Existing anomaly generation methods primarily fall into three paradigms. Early **heuristic-based methods** (Li et al., 2021; Zavrtanik et al., 2021; Schluter et al., 2022) and **GAN-based generators** (Zhang et al., 2021; Duan et al., 2023) attempt to synthesize defects but fundamentally struggle with unnatural boundaries, limited structural diversity, and a lack of explicit attribute control. Consequently, recent advancements are dominated by **diffusion-based and prompt-guided models** (e.g., AnomalyDiffusion (Hu et al., 2024), DualAnoDiff (Jin et al., 2025), SeaS (Dai et al., 2025)), which offer superior fidelity by modeling mask-conditioned embeddings or prompt alignments. However, despite their improved realism, these state-of-the-art methods **fail to achieve direct zero-shot transfer generation on novel products due to the inherent style gap**. This failure fundamentally stems from the implicit entanglement of anomalous features with source-specific background textures.

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.

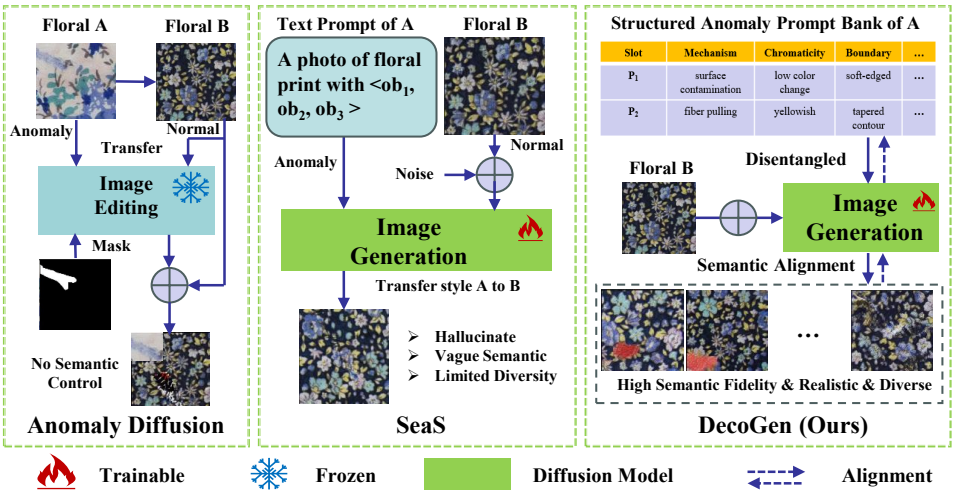


Figure 1: Comparison of cross-style anomaly transfer. Unlike rigid mask editing (left) and coarse prompts that cause texture hallucinations (middle), our ***DecoGen*** (right) achieves high-fidelity and diverse zero-shot transfer via disentangled attribute alignment.

When forced into zero-shot cross-style settings without target anomaly exemplars, their coarse-grained prompt tokens act as "texture sponges." They inadvertently absorb localized background remnants from the source domain and drag them into the target, causing severe texture contamination. This reveals that mere textual decoupling is profoundly insufficient when visual representations remain implicitly entangled.

As intuitively illustrated in Figure 1, transferring anomalies from a source domain (Floral A) to a target domain (Floral B) exposes the fundamental vulnerabilities of existing paradigms. Mask-guided methods (e.g., AnomalyDiffusion) lack semantic control and degrade into rigid image-editing processes, directly pasting the source defect images onto the target and introducing severe boundary artifacts. Conversely, text-guided approaches (e.g., SeaS) attempt to directly apply the source domain's textual prompts to anomaly generation in the target domain. However, directly applying the source domain's defect description to the target domain is highly unreasonable; it inadvertently entangles anomalous features with the source-specific background, leading to vague semantics and texture hallucinations when applied to the target domain. In summary, existing methods forcibly edit either the source defect images or semantics onto the target domain, which generates unrealistic images with limited diversity and ultimately hinders downstream defect recognition performance.

Motivated by this, our approach builds upon a crucial observation: **although product backgrounds vary drastically across domains, the underlying physical generation mechanisms of industrial defects (e.g., structural rupture or material residue) remain fundamentally consistent**. Therefore, effective cross-style anomaly generation should fundamentally aim to transfer these intrinsic generation mechanisms, rather than resorting to the superficial transfer of images or entangled semantics seen in previous methods. Based on this insight, we propose ***DecoGen***, a zero-shot cross-style generation framework driven by physical semantic constraints. First, a structured anomaly prompt bank explicitly decomposes source defects into mechanism-aware physical attributes. A subsequent dual-level alignment mechanism thoroughly isolates the pure physical defect from its source background. This ensures accurate mechanism migration, enabling high-fidelity and contamination-free anomaly synthesis on novel target styles. Our framework provides the following contributions:

- We propose ***DecoGen***, a zero-shot cross-style anomaly transfer framework driven by explicit physical semantic constraints to address the cold-start bottleneck for novel styles.
- We construct a **Structured Anomaly Prompt Bank** to decompose source defects into mechanism-aware physical attributes, extracting background-agnostic semantics for disentangled transfer.
- We design a **dual-level physical semantic alignment mechanism**. Combining category-level prototype constraints and fine-grained spatial decoupling, it isolates pure physical defects from source textures to ensure accurate mechanism migration and generation.

2 Related Work

Industrial Anomaly Synthesis. Early synthesis methods rely on hand-crafted copy-and-paste operations (Li et al., 2021; Schluter et al., 2022) or GANs (Zhang et al., 2021; Duan et al., 2023), struggling to capture complex defect distributions or suffering from unstable training. Recently, diffusion models (Hu et al., 2024; Jin et al., 2025; Dai et al., 2025; Choi et al., 2026) have significantly improved few-shot generation fidelity via mask-guided insertion or prompt alignment. However, these methods inherently assume access to real anomalies from the target product. When forced into zero-shot cross-style generation, they inevitably suffer from representation entanglement and severe texture contamination.

Diffusion Personalization and Semantic Control. Parameter-efficient personalization methods adapt pre-trained diffusion priors (Rombach et al., 2022) to specific subjects via text embeddings (Gal et al., 2023), weight fine-tuning (Ruiz et al., 2023), or multi-concept extensions (Avrahami et al., 2023; Kumari et al., 2023). Meanwhile, cross-attention manipulation (Hertz et al., 2022; Chefer et al., 2023) enables fine-grained spatial control. Yet, industrial anomalies differ from ordinary objects as their semantics are local, fine-grained, and deeply coupled to product textures. Consequently, using standard implicit tokens inevitably absorbs the source background, failing to encode transferable defects.

Few-Shot and Cross-Domain Generation. Recent cross-domain and few-shot anomaly synthesis methods attempt to transfer distributions by establishing cross-domain correspondences (Ojha et al., 2021) or aligning target features via specific feature manipulations and fine-tuning (Duan et al., 2023; Hu et al., 2024; Dai et al., 2025). However, these distribution-alignment approaches strictly require target anomalous exemplars to guide the visual transfer, rendering them ineffective in extreme cold-start scenarios where only normal images are available. In contrast, ***DecoGen*** completely circumvents this reliance by migrating style-agnostic physical generation mechanisms rather than superficial features, seamlessly integrating pure defect priors into novel styles via mask-preserving composition.

3 Method

3.1 Overview

As illustrated in Figure 2, the ***DecoGen*** training process operates through four synergistic stages. First, in **S1 (Structured Anomaly Prompt Bank Construction)**, it extracts mechanism-aware physical prototypes from source defects. Next, in **S2 (Defect Prompt Generator)**, it fuses these semantic prototypes with mask-derived spatial embeddings to create a unified condition sequence. Third, in **S3 (Model Train)**, a mask-guided latent diffusion model synthesizes the anomaly conditioned on this sequence. Finally, in **S4 (Attribute Aligner)**, it enforces explicit physical semantic constraints on the generated output, ensuring strict feature disentanglement and accurate mechanism migration.

3.2 Preliminaries

Problem Formulation. Let $\mathcal{S}_{\mathrm{src}}$ and $\mathcal{S}_{\mathrm{tgt}}$ denote the sets of source and target product styles, respectively, where $\mathcal{S}_{\mathrm{src}} \cap \mathcal{S}_{\mathrm{tgt}} = \emptyset$ to establish a strict cross-style setting. During training, we access a source dataset $\mathcal{D}_{\mathrm{src}} = \{(\mathbf{x}_i, \mathbf{M}_i, s_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^{3\times H\times W}$ is a real anomalous image, $\mathbf{M}_i \in \{0,1\}^{1\times H\times W}$ is its pixel-level mask, $s_i \in \mathcal{S}_{\mathrm{src}}$ is the background style, and y_i represents the defect's physical category. For inference, the target dataset $\mathcal{D}_{\mathrm{tgt}} = \{\mathbf{x}^j\}_{j=1}^{m}$ consists exclusively of defect-free normal images from the unseen styles $\mathcal{S}_{\mathrm{tgt}}$, with no real target anomalies available. Our goal in this zero-shot anomaly transfer setting is to synthesize a high-fidelity paired sample $(\hat{\mathbf{x}}^{\mathrm{a}}, \mathbf{M})$: given a normal reference $\mathbf{x} \in \mathcal{D}_{\mathrm{tgt}}$, a spatial mask $\mathbf{M}$, and a requested defect type y^*, the generated image $\hat{\mathbf{x}}^{\mathrm{a}}$ must accurately exhibit the physical semantics of y^* within the unmasked region while perfectly preserving the background texture of $\mathbf{x}^{\mathrm{n}}$.

Latent Diffusion Model. We build ***DecoGen*** upon a latent diffusion model (LDM) (Rombach et al., 2022). Given an input image $\mathbf{x}_0 \in \mathbb{R}^{3\times H\times W}$, a pre-trained VAE encoder $\mathcal{E}$ first compresses it

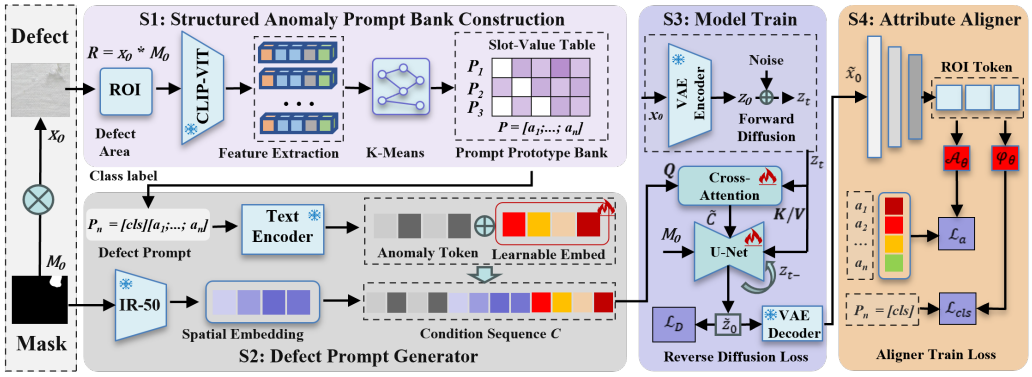


Figure 2: Overview of ***DecoGen***. A structured prompt bank provides mechanism-aware anomaly prototypes and explicit physical attribute supervision during source-domain training. At inference, prototype-aware adaptation composes transferable physical embeddings for an unseen target style, while mask-guided DDIM sampling preserves the target background, ensuring accurate and contamination-free mechanism migration.

into a low-dimensional latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. The forward diffusion process gradually perturbs $\mathbf{z}_0$ by adding Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ over T timesteps. At any timestep t, the noisy latent $\mathbf{z}_t$ is defined as:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},$$

where $\bar{\alpha}_t$ is the pre-defined cumulative noise schedule.

In the reverse denoising process, a U-Net network $\boldsymbol{\epsilon}_\theta$ is trained to iteratively predict and remove the added noise to recover the clean latent. To enable condition-driven anomaly synthesis, we inject a condition context $\mathbf{c}$ (e.g., text embeddings) into the U-Net via cross-attention mechanisms. The model is optimized using the conditional denoising objective:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right].$$

Finally, a VAE decoder maps the denoised latent $\mathbf{z}_0$ back to the pixel space.

3.3 Structured Anomaly Prompt Bank Construction

A single implicit token tends to inadvertently absorb both the defect and its surrounding source texture, leading to severe representation entanglement. To extract universal physical semantics across diverse source domains, we construct an offline *Structured Anomaly Prompt Bank*. We explicitly distill these transferable priors by clustering heterogeneous source samples and leveraging an image understanding model for precise semantic description. Crucially, these generated semantic descriptions are inherently universal across multiple source domains. The construction operates through four formalized steps:

Defect Feature Extraction. For a given source sample $(\mathbf{x}_i, \mathbf{M}_i)$ in $\mathcal{D}_{\text{src}}$, we isolate the defect region of interest (ROI) and extract its visual embedding using a pre-trained CLIP-ViT encoder (Zhou et al., 2024), denoted as Φ_{CLIP}. The extracted feature $\mathbf{f}_i$ is l_2-normalized:

$$\mathbf{f}_i = \frac{\Phi_{\text{CLIP}}(\mathbf{x}_i \odot \mathbf{M}_i)}{\|\Phi_{\text{CLIP}}(\mathbf{x}_i \odot \mathbf{M}_i)\|_2} \in \mathbb{R}^{d_f}, \tag{1}$$

where $\odot$ denotes element-wise multiplication.

Mechanism-Aware Clustering and Prototype Definition. To discover underlying physical generation mechanisms, we partition the feature set $\mathcal{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_{|\mathcal{D}_{\text{src}}|}\}$ into N visually coherent clusters by K-means. The optimal cluster assignments $\mathcal{C}^* = \{C_1, \ldots, C_N\}$ and corresponding centroids $\mathcal{U}^* = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N\}$ are obtained by minimizing the intra-cluster variance:

$$\mathcal{C}^*, \mathcal{U}^* = \arg\min_{\mathcal{C},\mathcal{U}} \sum_{n=1}^{N} \sum_{\mathbf{f}_i \in C_n} \|\mathbf{f}_i - \boldsymbol{\mu}_n\|_2^2. \tag{2}$$

Each resulting cluster is then summarized and formally defined as an anomaly prototype, denoted by P_n for $n \in \{1, \ldots, N\}$.

Universal Physical Semantic Template Construction. To ensure a unified expression across diverse backgrounds, we pre-define a universal physical schema based on S shared semantic slots (i.e., attributes such as *mechanism, morphology, boundary*). While this slot architecture is completely identical across all anomaly types, the specific instantiation within these slots varies. Each attribute slot $\mathbf{a} \in S$ is associated with a pre-set discrete value vocabulary.

For a specific anomaly prototype P_n, we leverage an image understanding model to distill its physical characteristics into a dual-level semantic representation. First, at the *value level*, the model selects specific descriptive terms from the global vocabularies to explicitly define the defect. This activation is formalized as a multi-hot attribute vector $\mathbf{a}_n$:

$$\mathbf{a}_n = [a_n^{(1)}, \ldots, a_n^{(d)}]^\top \in \{0, 1\}^d, \tag{3}$$

where $a_n^{(d)} = 1$ if the d-th vocabulary term is activated for prototype P_n, and 0 otherwise. Stacking the attribute vectors of all N prototypes yields the semantic matrix $\mathcal{A}_n$, serving as explicit targets for subsequent alignment:

$$\mathcal{A}_n = [\mathbf{a}_1, \ldots, \mathbf{a}_N]^\top \in \{0, 1\}^{N \times D}. \tag{4}$$

Second, at the *slot level*, the model identifies a subset of slots as *salient axes* $\mathcal{A}_n \subseteq \{S\}$. Guided by these salient axes $\mathcal{A}_n$, a deterministic rendering function $\mathcal{R}$ synthesizes the activated discrete phrases in $\mathbf{a}_n$ into a fixed standard prompt $t_n = \mathcal{R}(\mathcal{A}_n)$.

To overcome the representation bottleneck of discrete text, we augment t_n with a sequence of learnable continuous vectors $\mathbf{v}_n \in \mathbb{R}^{m \times d_e}$ (m tokens, d_e dimension). The final structured prompt template $\mathcal{T}_n$ is thus constructed via sequence-level fusion ($\oplus$):

$$\mathcal{T}_n = \mathcal{R}(\mathcal{A}_n) \oplus \mathbf{v}_n. \tag{5}$$

For complete slot definitions and standard prompts across all prototypes, please refer to the Appendix.

3.4 Defect Prompt Generator

Building upon the structured prompt template $\mathcal{T}_n$ and its semantic placeholder '$\star$', this module synthesizes a unified multi-modal condition sequence $\mathbf{c} \in \mathbb{R}^{L \times d}$ to guide the denoising process. To bridge discrete texts with continuous visual controls, we introduce two decoupled embeddings. First, an appearance token $\mathbf{E}_a \in \mathbb{R}^{K \times d}$ is obtained by processing $\mathcal{T}_n$ through the CLIP text encoder to capture the physical semantics. Second, a spatial token $\mathbf{E}_s \in \mathbb{R}^{K \times d}$ is extracted from the binary mask $\mathbf{M}$ via a spatial encoder $\mathcal{S}_p$:

$$\mathbf{E}_s = \mathcal{S}_p(\mathbf{M}). \tag{6}$$

To explicitly map these decoupled features into a joint conditioning space, we concatenate the appearance and spatial tokens along the sequence dimension to formulate the final control vector $\mathbf{c}$:

$$\mathbf{c} = [\mathbf{E}_a; \mathbf{E}_s], \tag{7}$$

This strict separation before sequence fusion ensures that the visual texture of the physical mechanism is completely decoupled from its geometric shape, a property essential for flexible cross-style transfer.

3.5 Model Train

At training timestep t, the noisy latent $\mathbf{z}_t$, the timestep t, and the unified multi-modal condition $\mathbf{c}$ are jointly fed into the U-Net denoiser $\boldsymbol{\epsilon}_\theta$. In particular, the latent vector $\mathbf{z}_t$ is derived from the defective image in the source domain following feature extraction via a VAE. Specifically, the network feature $\mathbf{c}$ acts as the query, while the external condition $\mathbf{z}_t$ directly provides the key and value:

$$\mathrm{CrossAttn}(\mathbf{c}, \mathbf{z}_t) = \mathrm{softmax}\left(\frac{(\mathbf{c}\mathbf{W}_Q)(\mathbf{z}_t\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)(z_t\mathbf{W}_V), \tag{8}$$

where $\mathbf{W}_Q, \mathbf{W}_K$, and $\mathbf{W}_V$ are learnable projection matrices. This operation deeply fuses the physical semantics and spatial constraints of $\mathbf{c}$ into the network's generative representations.

Conditioned on these fused features, the U-Net outputs the predicted noise $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$. To ensure the anomaly is generated precisely where intended without contaminating the surrounding source texture, the denoising loss is exclusively computed within the down-sampled latent mask $\mathbf{M}_\ell \in \{0,1\}^{h\times w}$:

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}}\left[\left\|\left(\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}) - \boldsymbol{\epsilon}\right) \odot \mathbf{M}_\ell\right\|_2^2\right]. \tag{9}$$

This masked objective explicitly forces the model to synthesize the requested anomaly prototype strictly within the target region.

3.6 Attribute Aligner

The structured prompt bank successfully decouples semantics, but this decoupling is only meaningful if the generated visual regions strictly adhere to these prescribed physical constraints. To explicitly prevent semantic drift and background entanglement during generation, we propose a dual-level attribute aligner that enforces supervision at both the fine-grained attribute and global prototype levels.

Given the predicted clean defect region $\mathbf{r} = \hat{\mathbf{x}}_0 \odot \mathbf{M}$, a lightweight aligner network $f(\cdot)$, composed of a convolutional stem and a two-layer Transformer, extracts continuous attribute logits $\hat{\mathbf{a}} \in \mathbb{R}^D$ and an ℓ_2-normalized visual embedding $\boldsymbol{\phi}$.

Fine-Grained Attribute Alignment. To ensure the generated anomaly accurately reflects the specific slot-level configurations (e.g., color, morphology, structural changes), we apply a Binary Cross-Entropy (BCE) loss between the predicted logits $\hat{\mathbf{a}}$ and the ground-truth multi-hot attribute vector $\mathbf{a}$:

$$\mathcal{L}_{\mathrm{reg}} = \mathrm{BCE}\big(\hat{\mathbf{a}}, \mathbf{a}\big). \tag{10}$$

This regression term strictly bounds the localized physical properties of the generated defect.

Category-Level Prototype Alignment. While $\mathcal{L}_{\mathrm{reg}}$ controls local slots, the generated defect must also holistically match the intended physical mechanism rather than collapsing into competing anomaly types. We define a learnable semantic projector $\boldsymbol{\psi}_\theta(\cdot)$ that maps each prototype attribute vector $\mathbf{a}_n$ into a shared latent space. By computing the cosine similarity $s_n = \langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_\theta(\mathbf{a}_n)\rangle$ between the generated visual embedding $\boldsymbol{\phi}_i$ and each projected prototype, we enforce a global classification loss against the assigned target prototype label P_i:

$$\mathcal{L}_{\mathrm{cls}} = \mathrm{CE}\big(\{s_n\}_{n=1}^{N}, P_i\big). \tag{11}$$

This term effectively pushes the visual embedding $\boldsymbol{\phi}_i$ towards the correct physical mechanism and away from irrelevant prototypes.

Overall Objective. Following the dual-level design, these two complementary terms are linearly combined into a compact semantic alignment objective:

$$\mathcal{L}_{\mathrm{sem}} = \mathcal{L}_{\mathrm{reg}} + \beta\mathcal{L}_{\mathrm{cls}}. \tag{12}$$

Finally, this explicit semantic constraint is coupled with the masked diffusion loss to form the overall end-to-end training objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{D}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}}, \tag{13}$$

where we empirically set $(\lambda_{\mathrm{sem}}, \beta) = (1.0, 0.5)$. By bridging $\mathcal{L}_{\mathrm{D}}$ and $\mathcal{L}_{\mathrm{sem}}$, the framework simultaneously achieves high-fidelity textural synthesis and exact physical mechanism alignment. Without attribute supervision, the semantic branch is disabled, and Eq. 13 gracefully reduces to the default masked diffusion path.

3.7 Zero-Shot Target Domain Inference

To achieve zero-shot defect transfer on the unseen target domain, ***DecoGen*** leverages the offline-constructed *Structured Anomaly Prompt Bank*. For a requested defect prototype n, we first retrieve its corresponding multi-hot attribute vector $\mathbf{a}_n$ and salient axes $\mathcal{A}_n$ from the bank to synthesize the structured prompt template $\mathcal{T}_n = \mathcal{R}(\mathcal{A}_n) \oplus \mathbf{v}_n$. By processing this instantiated template through the

CLIP text encoder, we directly compute the target appearance token $\mathbf{E}_a^{\mathrm{tgt}}$, which encapsulates the pure physical mechanism and bypasses the need for target anomalous training data.

Subsequently, this semantic token is concatenated with the spatial token $\mathbf{E}_s$ (extracted from the target binary mask $\mathbf{M}$) to formulate the unified multi-modal condition $\mathbf{c}^{\mathrm{tgt}} = [\mathbf{E}_a^{\mathrm{tgt}}; \mathbf{E}_s]$. During iterative denoising on a normal target reference $\mathbf{x}^{\mathrm{n}}$, we feed $\mathbf{c}^{\mathrm{tgt}}$ into the U-Net to guide generation via cross-attention. Concurrently, we employ a spatial composition strategy at each timestep t:

$$\mathbf{z}_t \leftarrow \mathbf{z}_t \odot \mathbf{M}_\ell + \big(\mathbf{1} - \mathbf{M}_\ell\big) \odot q(\mathbf{z}_0^{\mathrm{n}}, t), \tag{14}$$

where $\mathbf{z}_0^{\mathrm{n}}$ is the clean latent encoding of the normal target image. This operation continuously anchors the unmasked region to the true background at the matched noise level, ensuring the transferred physical defect is synthesized exclusively within the mask while perfectly preserving the novel target texture.

4 Experiments

4.1 Experimental settings

Datasets. We evaluate ***DecoGen*** on a cross-style textile setting and a standard industrial benchmark. **ZJU Leaper** is the primary fabric dataset: pattern1–pattern7 provide source anomaly image–mask pairs, while pattern8–pattern15 provide only normal images during target generation (reported as eight unseen plaid and floral styles in Table 3). **MVTec AD** supplies a complementary benchmark spanning diverse industrial objects and textures.

Evaluation metrics. For generation, we report Inception Score (IS, ↑), Intra-Class LPIPS (IC-LPIPS, ↑), Kernel Inception Distance (KID, ↓), and anomaly-region IC-LPIPS(a) (↑). These metrics measure semantic quality, intra-class diversity, distributional realism, and diversity within the synthesized region, respectively. For downstream detection, we use image-level AUROC/AP/F_1-max and pixel-level AUROC/AP/F_1-max/IoU. The cross-style study summarizes image-level AUROC (I-AUROC) and pixel-level AUROC (P-AUROC).

Implementation details. We use the pretrained `text2img-large` latent diffusion backbone with five source anomalies per fabric style and 256×256 images. Architecture, optimization, and sampling details are provided in Appendix A.

4.2 Comparison in Anomaly Image Generation

Comparison methods. We compare ***DecoGen*** with AnomalyDiffusion (Hu et al., 2024), SeaS (Dai et al., 2025), MAGIC (Choi et al., 2026), and DualAnoDiff (Jin et al., 2025). All generators use the same image resolution and evaluation protocol.

Table 1: Comparison of generation quality on ZJU Leaper and MVTec AD. Bold indicates the best performance.

Methods	ZJU Leaper				MVTec AD			
	IS↑	IC-LPIPS↑	KID↓	IC-LPIPS(a)↑	IS↑	IC-LPIPS↑	KID↓	IC-LPIPS(a)↑
AnomalyDiffusion (Hu et al., 2024)	2.39	0.11	0.07	0.34	1.71	0.14	0.08	0.30
SeaS (Dai et al., 2025)	2.79	0.12	0.18	0.36	1.78	0.17	0.03	0.31
MAGIC (Choi et al., 2026)	1.73	0.20	**0.04**	0.33	1.79	0.18	0.05	0.25
DualAnoDiff (Jin et al., 2025)	2.94	0.19	0.09	0.37	1.80	0.18	0.04	0.32
Ours	**2.96**	**0.26**	0.05	**0.41**	**1.81**	**0.19**	**0.01**	**0.33**

Anomaly image generation quality. Table 1 reports generation quality. On ZJU Leaper, ***DecoGen*** achieves the best IS (2.96), IC-LPIPS (0.26), and IC-LPIPS(a) (0.41), while obtaining a competitive KID of 0.05. The diversity gains occur both globally and within the anomaly region, indicating that the model does not simply reproduce the few source defects. On MVTec AD, ***DecoGen*** ranks first on all four metrics, including a KID of 0.01. Together, these results show that structured conditioning improves diversity without sacrificing distributional realism.

Table 2: Downstream performance averaged over AnomalyCLIP, MambaAD, and VisionAD on ZJU Leaper and MVTec AD. Bold indicates the best average in each column. All metrics are reported in percent; detector-wise results are provided in Table 10.

Augmentation	ZJU Leaper							MVTec AD						
	Image-level			Pixel-level				Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	IoU	AUROC	AP	F_1-max	AUROC	AP	F_1-max	IoU
Average	88.48	90.53	83.55	87.32	46.38	48.66	32.98	86.69	86.89	85.89	92.44	41.35	42.73	29.46
+ AnoDiff	90.26	92.41	85.64	88.69	48.85	50.68	35.05	88.51	87.74	87.56	93.36	43.35	44.37	31.34
+ SeaS	90.55	92.51	86.93	87.53	47.64	49.61	34.46	91.91	91.35	90.48	94.72	46.81	49.19	33.93
+ MAGIC	91.05	92.98	86.27	88.18	48.99	50.41	34.28	91.98	90.35	89.31	94.92	49.69	49.97	35.89
+ DualAnoDiff	90.62	92.61	86.96	86.79	48.09	49.02	33.68	91.29	90.89	89.79	94.49	**50.42**	50.30	37.40
+ Ours	**93.36**	**95.09**	**89.67**	**88.99**	**52.00**	**53.28**	**37.20**	**93.11**	**91.84**	**91.11**	**95.45**	49.83	**51.42**	**38.19**

Cross-method qualitative comparison.

Figure 3 compares two representative outputs per method on Blue Plaid. ***DecoGen*** produces localized defects while retaining the surrounding plaid structure and varying anomaly appearance. The IC-LPIPS results support perceptual diversity; realism is assessed qualitatively. Appendix **??** provides the detailed visual analysis and additional examples.

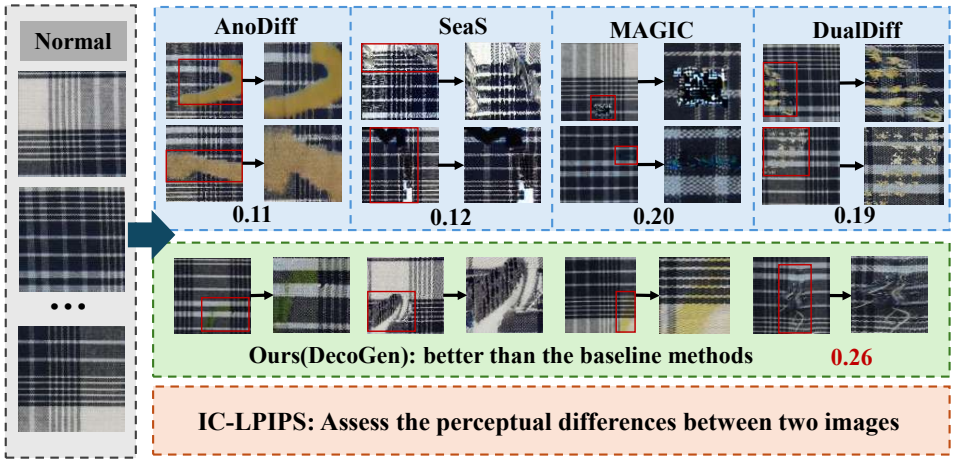


Figure 3: **Qualitative comparison on Blue Plaid(ZJU Leaper).** ***DecoGen*** yields smoother boundaries, better preservation of the surrounding weave, and broader anomaly diversity than the baselines.

Downstream anomaly detection. We next test whether the generated image-mask pairs provide useful supervision beyond visual quality. Synthetic data are combined with three detectors that use different representations: AnomalyCLIP (Zhou et al., 2024), MambaAD (He et al., 2025), and VisionAD (Wang et al., 2025). Table 2 summarizes image- and pixel-level performance averaged over the three detectors. Complete detector-wise results on both datasets are provided in Appendix C.1.

Table 2 shows that ***DecoGen*** improves all 14 averaged metrics over the unaugmented detectors and achieves the strongest augmentation average on 13 of them. The exception is MVTec AD pixel AP, where DualAnoDiff performs better. Detailed gains and detector-wise results are provided in Appendix C.1.

4.3 Cross-domain Anomaly Detection

Cross-domain anomaly detection performance. This experiment evaluates the central cross-style setting: no real anomaly from the target styles is observed during generation. We learn defect priors from the seven ZJU Leaper source styles and synthesize training pairs for eight unseen target styles. Table 3 compares each detector before and after ***DecoGen*** augmentation.

DecoGen improves all three detectors on average. For AnomalyCLIP, I-AUROC rises from 89.23 to 92.77 and P-AUROC from 88.20 to 95.44. PatchCore obtains the largest image-level gain (+6.68

Table 3: Cross-domain anomaly detection performance on our textile dataset. We compare the base unsupervised anomaly detection models against their performance when fine-tuned/augmented with synthetic defect images generated by our method (denoted as + Ours). **Bold** indicates the best performance in each column group.

Target Style (Unseen)	AnomalyCLIP (Zhou et al., 2024)				PatchCore (Roth et al., 2022)				WinCLIP (Jeong et al., 2023)			
	Base		+ Ours (*DecoGen*)		Base		+ Ours (*DecoGen*)		Base		+ Ours (*DecoGen*)	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC
Twill Plaid	90.45	82.11	**92.49**	**92.50**	62.98	69.95	**68.82**	**81.85**	53.15	59.71	**55.37**	**64.30**
Blue Plaid	94.29	94.92	**96.11**	**98.90**	74.62	76.65	**76.51**	**79.19**	86.12	75.63	**87.61**	**79.89**
Brown Plaid	91.03	93.38	**94.19**	**96.26**	59.86	66.08	**61.48**	**76.56**	84.12	79.88	**86.24**	**84.38**
Gray Plaid	89.58	86.20	**94.15**	**96.19**	50.68	76.19	**56.74**	**79.01**	72.70	73.30	**74.43**	**73.93**
Red Plaid	89.89	82.06	**95.26**	**95.28**	61.79	68.45	**73.35**	**68.74**	79.29	71.95	**81.28**	**74.96**
Floral Print1	95.51	96.39	**95.90**	**99.55**	60.92	70.85	**67.84**	**75.93**	81.77	79.67	**86.65**	**86.09**
Floral Print2	83.69	88.56	**92.83**	**95.35**	49.06	57.88	**53.46**	**73.51**	72.10	51.61	**72.89**	**56.51**
Floral Print3	79.38	82.01	**81.23**	**89.51**	53.56	57.91	**68.66**	**69.85**	69.09	71.80	**67.81**	**76.46**
Average	89.23	88.20	**92.77**	**95.44**	59.18	68.00	**65.86**	**75.58**	74.79	70.44	**76.53**	**74.57**

Table 4: Component ablation on eight unseen ZJU Leaper styles. Detection-style metrics measure generated image–mask consistency and are reported in percent. Variant definitions are provided in Appendix C.4. Bold indicates the best value before rounding.

Variant	Generation		Image-level			Pixel-level			
	IS	IC-L	AUROC	AP	F_1-max	AUROC	AP	F_1-max	IoU
w/o Attribute Alignment	2.074	0.680	96.49	97.14	94.24	95.01	83.56	76.49	62.73
w/o Prototype Adaptation	2.073	0.690	97.27	97.74	95.59	95.57	**86.55**	79.87	66.87
w/o Structured Prompt Bank	2.048	0.677	96.49	97.15	94.36	95.15	84.06	77.01	63.43
Full *DecoGen*	**2.103**	**0.696**	**97.58**	**98.05**	**96.30**	**95.57**	86.54	**79.93**	**66.94**

points) together with a 7.58-point pixel-level gain. WinCLIP also benefits, increasing P-AUROC from 70.44 to 74.57. Improvements are largest on difficult styles: for Floral Print2, PatchCore P-AUROC increases from 57.88 to 73.51, while AnomalyCLIP I-AUROC increases from 83.69 to 92.83. Averaged across the three detectors, ***DecoGen*** improves I-AUROC by 3.99 points and P-AUROC by 6.32 points. This result directly supports target-anomaly-free transfer rather than adaptation from hidden target defects.

4.4 Ablation Study

Protocol. We isolate the contribution of attribute alignment, prototype-aware adaptation, and the structured prompt bank. All variants use the same training schedule and component-study setting $(\mathrm{lr}, \lambda_{\mathrm{gen}}, s) = (0.01, 0.05, 7.5)$, and each run generates 200 paired samples for each of the eight unseen target styles. The reported scores measure generated image–mask consistency rather than real-image detector performance. The scoring procedure and aggregation are detailed in Appendix B.2.

Component contribution. Table 4 shows that Full DecoGen ranks first on eight of the nine metrics; variant definitions appear in Appendix C.4. Attribute alignment contributes most to localization: removing it reduces pixel AP, F_1-max, and IoU by 2.98, 3.43, and 4.21 points. Removing the structured prompt bank lowers pixel AP by 2.49 and IoU by 3.52 points. Prototype-aware adaptation mainly affects diversity and image-level separability: its removal leaves pixel AP essentially unchanged ($+0.01$) but degrades IS, IC-L, every image-level metric, and pixel F_1-max/IoU. The full model thus yields the most balanced performance; parameter sensitivity is reported in Appendix C.5.

5 Conclusion

We presented ***DecoGen***, a semantic-guided diffusion framework that transfers defect priors to unseen product styles without target anomaly supervision. By decoupling defect attributes and adapting prototype tokens, ***DecoGen*** establishes semantic prototype transfer as a practical paradigm for cold-start anomaly detection. Looking ahead, future work will explore automated prototype construction, multi-seed robustness, and joint adaptation across broader product domains.

REFERENCES

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia Conference Papers*, pp. 1–12, 2023.

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.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4):1–10, 2023.

JaeHyuck Choi, MinJun Kim, and Je Hyeong Hong. MAGIC: Few-shot mask-guided anomaly inpainting with prompt perturbation, spatially adaptive guidance, and context awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026. doi: 10.48550/arXiv.2507.02314. URL https://arxiv.org/abs/2507.02314.

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. *arXiv preprint arXiv:2410.14987*, 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*, 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. In *International Conference on Learning Representations*, 2023.

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. MambaAD: Exploring state space models for multi-class unsupervised anomaly detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 71162–71187, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2022.

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*, 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*, pp. 19606–19616, 2023.

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

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

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.

Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A. Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10743–10752, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn 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.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Scholkopf, 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*, pp. 14318–14328, 2022.

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.

Hannah M. Schluter, 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, 2022.

Qishan Wang, Jia Guo, Shuyong Gao, Haofen Wang, Li Xiong, Junjie Hu, Hanqi Guo, and Wenqiang Zhang. Search is all you need for few-shot anomaly detection. *arXiv preprint arXiv:2504.11895*, 2025. URL `https://arxiv.org/abs/2504.11895`.

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*, pp. 2524–2534, 2021.

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

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, 2024.

APPENDIX

A IMPLEMENTATION DETAILS AND BASELINES

A.1 IMPLEMENTATION DETAILS

We implement ***DecoGen*** on the pretrained `text2img-large` latent diffusion backbone. The U-Net has 320 base channels and receives a 1280-dimensional context from the BERT text encoder. Each anomaly condition contains $K = 4$ learnable appearance vectors and $K = 4$ mask-derived spatial vectors from the shared pSp-style encoder. Unless stated otherwise, source training uses five real anomalies per fabric style, sampling uses DDIM, and all images are resized to 256×256. When attribute supervision is enabled, we set $(\lambda_{\text{sem}}, \beta) = (1.0, 0.5)$ in Eq. 13; the regression and classification terms therefore follow the decomposition in Sec. 3.

Table 5 summarizes the implementation used for the controlled cross-style experiments. ***DecoGen*** starts from the pretrained `text2img-large` latent diffusion model. The VAE, BERT conditioner, and U-Net remain frozen during source-domain personalization; optimization updates the four-vector appearance embeddings, the shared mask encoder, and, when enabled, the attribute aligner. Joint appearance–spatial training uses Adam. The configuration admits a larger maximum number of iterations, but the ablation driver explicitly stops each run after 10,001 optimization steps.

Table 5: Verified implementation settings for the controlled ***DecoGen*** runs.

Component	**Setting**
Diffusion backbone	pretrained `text2img-large` LDM
Image / latent resolution	256×256 / $4 \times 32 \times 32$
U-Net / text context	320 base channels / 32-layer BERT, $d = 1280$
Appearance / spatial vectors	$K = 4$ / $K = 4$
Spatial encoder	shared IR-50 pSp-style encoder
Attribute aligner	hidden size 256, 4 heads, 2 layers, dropout 0.1
Optimizer	Adam
Source training	batch 4, 2 workers, at most 5 anomalies per style
Ablation control	10,001 steps, seed 23
Generation / evaluation batch	8 / 32; evaluation workers 4

The tuned component study uses $(\text{lr}, \lambda_{\text{gen}}, s) = (0.01, 0.05, 7.5)$. The one-factor sensitivity study uses $(0.005, 0.10, 1)$ as the repeated reference point and changes one coordinate at a time. This distinction is important: the component table measures the complete model at its tuned setting, whereas the parameter sweeps measure local sensitivity around a common reference.

A.2 STRUCTURED PROMPT BANK

The prompt bank represents every anomaly with the same semantic schema. This shared structure makes prototypes comparable across product styles and prevents the transfer rule from relying on a coarse defect name alone. Table 6 groups the twelve slots used in the implementation. A prototype stores candidate phrases for each slot together with a short identity summary. During prompt rendering, four to six salient slots are sampled in a fixed semantic order and inserted into a context template describing the product, camera, and illumination.

For ZJU Leaper, masked CLIP-ViT features are clustered and summarized into the eight fabric prototypes in Table 7. For MVTec AD, the bank follows the native category–defect taxonomy and contains 71 prototypes. The same schema is retained, with *fiber state* interpreted more generally as the state of local material, edges, or thin parts. This keeps the alignment and transfer paths identical across the two datasets.

Table 6: Semantic slots in the structured anomaly prompt bank.

Factor	Slots and role
Physical cause	Substrate relation and mechanism describe how a defect interacts with the product and how it is formed.
Visual appearance	Chromaticity, morphology, boundary, intensity, scale, and contrast mode describe directly observable appearance.
Local material state	Surface texture and fiber state describe local roughness, yarn, edge, or thin-part changes.
Structural effect	Structure state and residue state distinguish physical damage from deposited or attached material.

Table 7: ZJU Leaper anomaly prototypes and their dominant semantic cues.

Prototype	Dominant cue
Pink-red smear	Reddish transferred residue with a soft boundary and preserved weave.
Pulled-fiber snag	Protruding strands and local weave distortion without strong color contamination.
Yellowish stain patch	Patch-shaped discoloration with a soft irregular contour.
Yellowish smear mark	Slender dragged morphology with light deposited residue.
Frayed tear or hole	Structural rupture with an opening, broken threads, and a ragged border.
Surface fuzz or pilling	Subtle local roughening and fine protruding fibers.
Dark ink smear	Elongated dark pigment with a stroke-like appearance.
Foreign-residue smear	Attached material that partially covers the fabric surface.

A.3 Prompt–image semantic consistency

The structured prompt bank represents each defect with a shared set of physical attribute slots. To inspect whether these semantics are reflected in the generated result, Fig. 4 shows four generated anomaly crops from the Blue Plaid category of ZJU Leaper together with their prompt-bank fields. The examples cover color-dominant contamination, structural rupture, and fiber-pulling defects. The semantic panel exposes the key–value fields (e.g., mechanism, chromaticity, morphology, structure state, residue state, and contrast mode) used by the prompt bank, making the correspondence auditable rather than relying on a free-form caption alone.

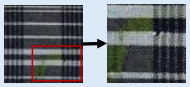
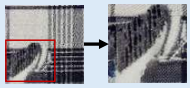
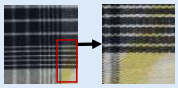
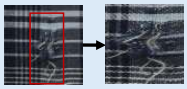
Prompt—Image Semantic Consistency				
Generated Figures				
Prototype key	greenish stain	frayed tear hole	yellowish stain	pulled fiber snag
Mechanism	surface contamination	structural rupture	stain transfer	fiber pulling
Chromaticity	greenish-yellow	low color change	pale yellow	low color change
Morphology	local patch	irregular opening	broad patch	loose curled strands
Boundary	soft irregular edge	ragged frayed edge	soft diffuse edge	irregular rough edge
Structure state	weave intact	weave borken	weave intact	local weave distortion
Residue state	color residue	no residue	stain residue	no residue
Contrast mode	color contrast	structure contrast	color contrast	fiber contrast

Figure 4: Qualitative prompt–image semantic consistency on the Blue Plaid category of ZJU Leaper. Each column shows a generated anomaly crop and the corresponding key–value fields from the structured prompt bank. The four examples cover greenish stain (surface contamination), frayed tear hole (structural rupture), yellowish stain (stain transfer), and pulled fiber snag (fiber pulling); their mechanism, chromaticity, and structure-state descriptions agree with the visible defect morphology and the surrounding plaid context.

The first and third columns show color or residue dominated defects: the greenish stain and yellowish stain patches are described by surface contamination or stain-transfer mechanisms with greenish-yellow and pale yellow chromaticity, while the weave structure remains largely intact (weave intact) and exhibits color contrast and color/stain residue. The second column shows a rupture-dominant defect (frayed tear hole) characterized by structural rupture, low color change, an irregular opening with a ragged frayed edge, broken weave, no residue, and structure contrast. The fourth column shows a fiber-pulling defect (pulled fiber snag) whose main evidence is local weave distortion and exposed loose curled strands rather than a strong color change, exhibiting irregular rough edge, no residue, and fiber contrast. These examples provide qualitative evidence of semantic faithfulness; they complement, but do not replace, the quantitative generation and downstream detection metrics.

A.4 Complete alignment objective

The Appendix uses exactly the loss decomposition defined in the main text. For a masked anomaly region, the attribute aligner predicts the multi-hot attribute target $\mathbf{a}$ and the prototype label P_i. The two semantic terms are

$$\begin{aligned} \mathcal{L}_{\text{reg}} &= \text{BCE}\big(\hat{\mathbf{a}}, \mathbf{a}\big), \\ \mathcal{L}_{\text{cls}} &= \text{CE}\big(\{s_n\}_{n=1}^{N}, P_i\big), \end{aligned} \tag{A.1}$$

where $\hat{\mathbf{a}}$ contains the slot-level attribute logits and s_n is the similarity between the region embedding and prototype n. They are combined into the semantic alignment loss

$$\mathcal{L}_{\text{sem}} = \mathcal{L}_{\text{reg}} + \beta \mathcal{L}_{\text{cls}}. \tag{A.2}$$

The complete training objective is therefore

$$\mathcal{L} = \mathcal{L}_{\text{D}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}}. \tag{A.3}$$

All Appendix experiments use $(\lambda_{\text{sem}}, \beta) = (1.0, 0.5)$ when attribute supervision is enabled. When the attribute-alignment component is removed, $\mathcal{L}_{\text{sem}}$ is disabled and the objective reduces to $\mathcal{L} = \mathcal{L}_{\text{D}}$. No additional ranking, generated-region, or feedback loss is used in the reported objective.

A.5 Mechanism Behind Zero-Shot Cross-Style Transfer

The transfer rule separates the physical defect mechanism from the target background. Appearance tokens are balanced across source styles so that a prototype is not dominated by one product texture.

At inference, the target provides only a normal reference and a mask; its anomaly token is retrieved or composed from the source prototype bank. The following equations give the exact operators used for observed prototypes and unseen attribute compositions. Let $\mathcal{I} = \{(p, n) : \exists i,\ s_i = p \text{ and } y_i = n\}$ and $\mathcal{S}_n = \{p : (p, n) \in \mathcal{I}\}$. For an observed prototype, the style-balanced appearance token is

$$\overline{\mathbf{E}}_a(n) = \frac{1}{|\mathcal{S}_n|} \sum_{p \in \mathcal{S}_n} \mathbf{E}_a^{p,n} \in \mathbb{R}^{K \times d}. \tag{A.4}$$

The target token for an observed request $n^\star$ is

$$\mathbf{E}_a^{\mathrm{tgt}}(n^\star) = \overline{\mathbf{E}}_a(n^\star). \tag{A.5}$$

For an unseen attribute composition, partition $\mathbf{a}_n = [\mathbf{a}_n^{(1)}; \ldots; \mathbf{a}_n^{(S)}]$ by slots. Its slot-balanced similarity to an observed prototype is

$$q_n = \frac{1}{S} \sum_{s=1}^{S} \frac{\langle \mathbf{a}^{\star(s)}, \mathbf{a}_n^{(s)} \rangle}{\|\mathbf{a}^{\star(s)}\|_2 \, \|\mathbf{a}_n^{(s)}\|_2 + \varepsilon}. \tag{A.6}$$

Let $\mathcal{N}_k$ contain the k highest-scoring observed prototypes and $\omega_n = \exp(q_n/\tau) / \sum_{j \in \mathcal{N}_k} \exp(q_j/\tau)$. The unseen composition token is

$$\mathbf{E}_a^{\mathrm{tgt}}(\mathbf{a}^\star) = \sum_{n \in \mathcal{N}_k} \omega_n \overline{\mathbf{E}}_a(n). \tag{A.7}$$

A.6 Prompt rendering and target-token registration

For each source sample, the renderer combines the assigned prototype with phrases for product appearance, camera, and illumination while preserving the learnable placeholder. Target registration then proceeds as follows:

1. retrieve all learned pattern–prototype tables $\{\mathbf{E}_a^{p,n} : (p, n) \in \mathcal{I}\}$ and group them by prototype;
2. for an observed target prototype $n^\star$, average only the tables in $\mathcal{S}_{n^\star}$ using Eq. A.4;
3. for an unseen attribute composition, compute the slot-balanced similarities and top-k weights of Eqs. A.6 and A.7;
4. register the resulting $K \times d$ embedding as a target token; and
5. invoke the same mask-guided DDIM sampling path used for source tokens.

This mechanism transfers a defect prior without optimizing on a real target anomaly. It also provides an explicit audit trail: the generator records the selected target prototype, contributing (p, n) tokens, aggregation coefficients, and any unseen-composition weights for every output.

A.7 End-to-end training and synthesis procedure

Source-domain training learns tokens jointly from pattern1–pattern7. Each style contributes at most five anomaly image–mask pairs. The mask is resized to the 32×32 latent grid for the denoising loss and is also encoded into four spatial vectors. Appearance embeddings, the spatial encoder, and the optional attribute aligner are updated, while the diffusion backbone remains fixed. Checkpoints store the appearance embeddings, spatial encoder, and attribute aligner separately.

At synthesis time, pattern8–pattern15 are never used with real anomalies. For each target style, the generator cycles through five fixed normal references, samples a target anomaly prototype, registers its composed token, and performs 200-step DDIM sampling with $\eta = 1$. At every denoising step, Eq. 14 replaces the region outside the mask with the correspondingly noised normal latent. The reported cross-style runs use generation batch size eight and disable random-mask fallback. Each target style contains 200 generated image–mask–normal triplets, giving 1,600 triplets per experimental variant.

Table 8: Reproducibility artifacts and their roles.

Artifact	Role
`configs/latent-diffusion/....yaml`	Model, data, and alignment configuration.
`prompt_bank_zju_leaper_fabric_star.json`	Structured anomaly prototypes and slot values.
`token_attr_map_template.json`	Source-token to prototype mapping.
`train_source_domain.py`	Joint source-token and spatial-encoder training.
`generate_crossdomain_zjuleaper2.py`	Target registration, sampling, metadata, and resumable export.
`ablation_experiments.py`	Isolated module/parameter runs and manifests.
`evaluate_ablation_v2.py`	Name-agnostic integrity checks and metrics.
`metrics_v2.json` / summary CSV	Per-style metrics and macro averages.

Table 9: Dataset roles and controlled sample construction. A dash indicates that the quantity is not part of the cross-style protocol.

Dataset / split	Styles or categories	Anomaly shots	Normal shots	Generated per style	Role
ZJU source (pattern1–7)	7	≤ 5 each	–	–	Source-prior learning
ZJU target (pattern8–15)	8	0	5 each	200	Cross-style synthesis
MVTec AD	15	few-shot protocol	category-specific	–	Industrial benchmark

A.8 Baselines and downstream protocol

The generation comparison preserves each baseline's method-specific training procedure while standardizing the 256×256 output resolution and metric implementation. All generated sets are evaluated through the same metric code; experiment names are not exposed to the controlled ablation evaluator. This separates output-level fairness from method-specific optimization choices.

For downstream detection, synthetic image–mask pairs are added to the training or adaptation data, while the detector architecture and real evaluation split remain unchanged. The two-dataset comparison uses AnomalyCLIP, MambaAD, and VisionAD; the cross-style experiment uses AnomalyCLIP, PatchCore, and WinCLIP. The consistency scores used in the controlled generation ablation are never reused as detector predictions. Thus, Tables 4 and 3 answer different questions: whether a generated pair is internally consistent, and whether generated supervision improves a separately trained detector, respectively.

A.9 Artifacts, resumption, and output integrity

Table 8 maps the principal reproducibility artifacts to their roles. All paths are relative to the project root. The experiment driver writes every run into an isolated directory and stores the resolved commands and status in a JSON manifest.

Resumption is based on complete triplets, not on file counts alone. For each numeric index, the generator intersects the saved anomaly-image, valid-mask, and normal-reference indices, retains metadata only for the complete prefix, and resumes from the first missing index. Empty, full, or unreadable masks are not counted as valid. This rule prevents a power interruption or a partially written batch from being labeled complete.

B More Details About Datasets and Evaluation

B.1 Dataset construction and split

Table 9 records the data roles relevant to reproduction. ZJU Leaper provides the primary cross-style setting. Patterns 1–7 supply source anomaly image–mask pairs, whereas patterns 8–15 supply only normal references during synthesis. The eight target styles are Twill Plaid, Blue Plaid, Brown Plaid, Gray Plaid, Red Plaid, and Floral Print 1–3. No target anomaly participates in target-token construction. MVTec AD (Bergmann et al., 2019) provides 15 object and texture categories and is used as the complementary industrial benchmark.

The five normal target references are fixed within a run and cycled by the data loader, so the 200 outputs do not imply 200 distinct normal backgrounds. The generation process instead varies diffusion noise, target anomaly prototype, and its source-token combination. In all quantitative cross-style ablations, an experiment is eligible only after all eight target directories contain 200 valid triplets.

B.2 Generation and consistency metrics

In this study, IC-L is the mean pairwise distance between normalized Inception features. Image-level anomaly scores are cosine distances to the paired-normal feature centroid, while the pixel score combines RGB, local-SSIM, and coarse residuals with fixed weights $(0.45, 0.35, 0.20)$. Metrics are computed per style and macro-averaged. The AUROC/AP/F_1-max/IoU values therefore measure generated image–mask consistency and are distinct from the downstream detector results in Table 10.

IS is computed from Inception-v3 class probabilities with ten splits. For generated Inception features $\{\mathbf{f}_i\}_{i=1}^n$, IC-L is the mean pairwise distance between normalized features:

$$\text{IC-L} = \frac{1}{n(n-1)} \sum_{i \neq j} \sqrt{\max\Big(2 - 2\hat{\mathbf{f}}_i^\top \hat{\mathbf{f}}_j, 0\Big)}. \tag{B.1}$$

Higher IC-L indicates greater feature-space diversity. The main generation comparison additionally uses KID, global IC-LPIPS, and anomaly-region IC-LPIPS(a), as described in Sec. 4.

The controlled ablation evaluator receives only a generated dataset path and never its experiment label. At image level, generated and paired-normal Inception features are scored by cosine distance to the normalized normal feature centroid $\hat{\boldsymbol{\mu}}_N$:

$$s_{\text{img}}(\mathbf{x}) = 1 - \hat{\mathbf{f}}(\mathbf{x})^\top \hat{\boldsymbol{\mu}}_N. \tag{B.2}$$

At pixel level, the evaluator compares each generated image $\mathbf{x}$ with its paired normal background $\mathbf{x}^N$:

$$\mathbf{S}_{\text{pix}} = 0.45\,\mathbf{R}_{\text{RGB}} + 0.35\,\mathbf{R}_{\text{SSIM}} + 0.20\,\mathbf{R}_{\text{coarse}}. \tag{B.3}$$

Here $\mathbf{R}_{\text{RGB}}$ is the channel-averaged absolute residual, $\mathbf{R}_{\text{SSIM}}$ uses an 11×11 local window, and $\mathbf{R}_{\text{coarse}}$ compares 9×9 average-pooled images. A 3×3 average filter smooths the final score map.

Image and pixel AUROC and AP are threshold-independent. F_1-max is the maximum over thresholds from the precision–recall curve, and IoU is evaluated at that same threshold. Metrics are computed independently for every target style and macro-averaged over the eight styles, preventing a style with more pixels from dominating the result.

B.3 Evaluation eligibility and consistency checks

For every target style, the evaluator requires exactly 200 shared numeric stems across `test/bad`, `ground_truth/bad`, and `train/good`. Masks are resized by nearest-neighbor interpolation and thresholded at 127. Empty or full masks cause the run to fail instead of contributing trivial labels. Image-level labels must include paired normal and generated examples, and pixel labels must contain both background and foreground. Metrics are saved per pattern before the unweighted macro average is computed. A completion marker or 1,600 image files alone is therefore insufficient evidence of a valid run; the metrics JSON and successful integrity checks are both required.

C Additional Experimental Results

C.1 Complete downstream anomaly detection results

Table 2 summarizes the gains from DecoGen augmentation across the three detector families. Averaged over AnomalyCLIP, MambaAD, and VisionAD, it improves every reported metric relative to training without synthetic anomalies. On ZJU Leaper, image-level AUROC/AP/F_1-max increase by 4.88/4.56/6.12 points, while pixel AP, F_1-max, and IoU improve by 5.62, 4.62, and 4.22 points. On MVTec AD, the corresponding image-level gains are 6.42/4.95/5.22 points, and pixel AP/F_1-max/IoU improve by 8.48/8.69/8.73 points. DecoGen gives the strongest augmentation average

Table 10: Downstream anomaly detection results on ZJU Leaper and MVTec AD. Synthetic image–mask pairs from each generator are used with the same detector training protocol. Bold indicates the best result within each detector block and among the averaged results. All metrics are reported in percent.

Segmentation Models	ZJU Leaper							MVTec AD						
	Image-level			Pixel-level				Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	IoU	AUROC	AP	F_1-max	AUROC	AP	F_1-max	IoU
AnomalyCLIP	92.57	93.71	87.72	84.01	37.84	42.89	28.02	86.31	93.94	90.31	91.35	36.92	39.54	26.57
+ AnoDiff	93.59	94.50	89.28	**84.32**	39.56	44.60	29.62	88.34	94.02	91.24	91.46	39.66	39.27	27.58
+ SeaS	91.72	92.78	87.43	82.20	35.53	40.96	27.23	87.16	94.15	90.89	92.24	38.56	40.28	28.03
+ MAGIC	90.83	92.36	86.20	83.04	35.97	41.42	26.71	89.48	92.78	89.76	91.67	39.12	39.18	26.69
+ DualAnoDiff	89.16	90.44	85.47	78.98	30.15	36.12	22.54	86.90	93.75	90.58	90.88	**43.21**	37.88	27.64
+ Ours	**94.71**	**95.68**	**90.92**	83.49	**39.82**	**46.14**	**30.59**	**89.98**	**94.87**	**92.30**	**92.27**	39.81	**41.72**	**28.43**
MambaAD	82.12	84.74	76.46	87.96	44.55	46.46	31.01	78.66	67.56	70.33	91.81	25.50	29.71	19.24
+ AnoDiff	83.72	86.13	78.45	90.32	49.98	50.99	35.00	79.50	69.68	72.96	92.41	28.00	31.83	20.90
+ SeaS	83.72	86.13	78.45	90.32	49.98	50.99	35.00	89.20	79.96	81.34	96.14	39.06	42.76	29.10
+ MAGIC	86.69	89.03	80.85	90.89	51.67	52.11	35.89	88.17	78.47	79.71	95.92	38.17	42.66	28.82
+ DualAnoDiff	86.51	88.69	80.83	90.66	52.05	52.31	36.26	88.86	79.10	80.47	96.06	38.26	41.57	28.64
+ Ours	**88.48**	**90.56**	**82.86**	**91.91**	**54.74**	**54.67**	**38.41**	**90.17**	**80.79**	**81.77**	**96.55**	**39.88**	**43.00**	**29.96**
VisionAD	90.76	93.15	86.46	89.98	56.74	56.62	39.92	95.10	99.18	97.03	94.16	61.62	58.95	42.58
+ AnoDiff	93.48	96.59	89.20	91.44	57.02	56.46	40.52	97.68	99.53	98.48	96.21	62.40	62.02	45.53
+ SeaS	96.22	98.63	94.90	90.07	57.42	56.88	41.16	**99.36**	**99.94**	99.22	95.78	62.81	64.54	44.67
+ MAGIC	95.63	97.56	91.76	90.62	59.32	57.71	40.25	98.29	99.81	98.47	97.16	**71.77**	68.08	52.17
+ DualAnoDiff	96.18	98.71	94.59	90.73	**62.07**	58.64	42.23	98.10	99.83	98.31	96.54	69.79	**71.46**	55.93
+ Ours	**96.88**	**99.02**	**95.24**	**91.58**	61.44	**59.04**	**42.59**	99.17	99.86	**99.27**	**97.53**	69.80	69.54	**56.18**
Average	88.48	90.53	83.55	87.32	46.38	48.66	32.98	86.69	86.89	85.89	92.44	41.35	42.73	29.46
+ AnoDiff	90.26	92.41	85.64	88.69	48.85	50.68	35.05	88.51	87.74	87.56	93.36	43.35	44.37	31.34
+ SeaS	90.55	92.51	86.93	87.53	47.64	49.61	34.46	91.91	91.35	90.48	94.72	46.81	49.19	33.93
+ MAGIC	91.05	92.98	86.27	88.18	48.99	50.41	34.28	91.98	90.35	89.31	94.92	49.69	49.97	35.89
+ DualAnoDiff	90.62	92.61	86.96	86.79	48.09	49.02	33.68	91.29	90.89	89.79	94.49	**50.42**	50.30	37.40
+ Ours	**93.36**	**95.09**	**89.67**	**88.99**	**52.00**	**53.28**	**37.20**	**93.11**	**91.84**	**91.11**	**95.45**	49.83	**51.42**	**38.19**

Table 11: Detector-wise gains from ***DecoGen*** augmentation over the corresponding unaugmented detector. All entries are percentage-point changes; positive values indicate improvement. The comparison uses the same training protocol within each detector and dataset.

Dataset	Detector	Image-level gain			Pixel-level gain			
		ΔAUROC	ΔAP	ΔF_1-max	ΔAUROC	ΔAP	ΔF_1-max	ΔIoU
ZJU Leaper	AnomalyCLIP	+2.14	+1.97	+3.20	-0.52	+1.98	+3.25	+2.57
	MambaAD	+6.36	+5.82	+6.40	+3.95	+10.19	+8.21	+7.40
	VisionAD	+6.12	+5.87	+8.78	+1.60	+4.70	+2.42	+2.67
MVTec AD	AnomalyCLIP	+3.67	+0.93	+1.99	+0.92	+2.89	+2.18	+1.86
	MambaAD	+11.51	+13.23	+11.44	+4.74	+14.38	+13.29	+10.72
	VisionAD	+4.07	+0.68	+2.24	+3.37	+8.18	+10.59	+13.60

on 13 of the 14 metrics; only MVTec AD pixel AP is higher for DualAnoDiff. The consistent gains across CLIP-, state-space-, and searched architectures indicate that the synthetic pairs provide detector-agnostic supervision.

Table 10 reports the complete detector-wise results underlying the averages in Table 2. We retain all image- and pixel-level metrics for AnomalyCLIP, MambaAD, and VisionAD, together with the averaged results, to enable comparison across both generators and detector architectures.

C.2 Cross-detector consistency of synthetic augmentation

Table 10 reports absolute detector performance. To make the effect of synthetic supervision easier to compare across architectures, we additionally compute the paired change from each unaugmented detector to the same detector augmented with ***DecoGen***. No model is selected or tuned for this analysis; every entry in Table 11 is a direct subtraction of the corresponding entries in Tables 10 and 2.

Table 12: Per-style cross-domain gains from ***DecoGen*** augmentation. ΔI and ΔP denote percentage-point changes in I-AUROC and P-AUROC, respectively. Mean columns average the paired gains over AnomalyCLIP, PatchCore, and WinCLIP.

Unseen target style	**AnomalyCLIP**		**PatchCore**		**WinCLIP**		**Mean gain**	
	ΔI	ΔP	ΔI	ΔP	ΔI	ΔP	ΔI	ΔP
Twill Plaid	+2.04	+10.39	+5.84	+11.90	+2.22	+4.59	+3.37	+8.96
Blue Plaid	+1.82	+3.98	+1.89	+2.54	+1.49	+4.26	+1.73	+3.59
Brown Plaid	+3.16	+2.88	+1.62	+10.48	+2.12	+4.50	+2.30	+5.95
Gray Plaid	+4.57	+9.99	+6.06	+2.82	+1.73	+0.63	+4.12	+4.48
Red Plaid	+5.37	+13.22	+11.56	+0.29	+1.99	+3.01	**+6.31**	+5.51
Floral Print 1	+0.39	+3.16	+6.92	+5.08	+4.88	+6.42	+4.06	+4.89
Floral Print 2	+9.14	+6.79	+4.40	+15.63	+0.79	+4.90	+4.78	**+9.11**
Floral Print 3	+1.85	+7.50	+15.10	+11.94	−1.28	+4.66	+5.22	+8.03

DecoGen improves 41 of the 42 dataset–detector–metric combinations. The only negative entry is ZJU Leaper pixel AUROC for AnomalyCLIP (-0.52 point), while its pixel AP, F_1-max, and IoU still increase by 1.98, 3.25, and 2.57 points. MambaAD obtains the largest overall gains on both datasets, including 13.23 image AP and 14.38 pixel AP on MVTec AD. VisionAD starts from a stronger baseline but still benefits substantially, especially in MVTec AD localization. This consistency across CLIP-, state-space-, and searched architectures indicates that the generated pairs are not specialized to a single detector family.

C.3 Per-style cross-domain gain breakdown

Table 12 resolves the cross-domain averages in Table 3 into paired gains for each unseen target style. For every style and detector, we subtract the base I-AUROC and P-AUROC from the corresponding result with ***DecoGen*** augmentation. The final two columns average these paired gains over the three detectors.

The mean image- and pixel-level gains are positive for all eight unseen styles. The largest mean image-level improvement occurs on Red Plaid ($+6.31$ points), whereas Floral Print 2 obtains the largest mean pixel-level gain ($+9.11$ points). Twill Plaid and Floral Print 3 also show large mean localization gains of 8.96 and 8.03 points. WinCLIP loses 1.28 I-AUROC points on Floral Print 3, but its P-AUROC improves by 4.66 points and the cross-detector mean remains strongly positive. Thus, the aggregate cross-domain result is supported by broad per-style improvements rather than by one easy target style.

C.4 Component variants and interpretation

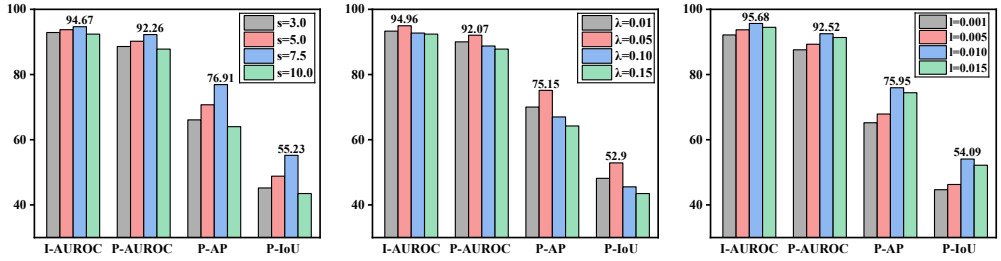
The three variants correspond to distinct functional removals. The Structured Prompt Bank decomposes each defect into a prototype identity and shared physical attribute slots, providing comparable semantic targets and routing source appearance tokens. Attribute Alignment maps the masked anomaly ROI to these attributes and prototype space, constraining the learned appearance to remain semantically faithful and reducing entanglement with background texture and mask-derived spatial support. Prototype-Aware Adaptation uses the target prototype to assign similarity-based weights to source embeddings, so that transfer emphasizes compatible defect mechanisms instead of averaging unrelated appearances. In the strict prompt-bank ablation, a single learnable anomaly vector is used in place of the entire bank; consequently, no prototype identity, attribute vector, or prototype-conditioned routing is available. This setting isolates the benefit of structured, compositional semantics rather than merely reducing the number of parameters.

The component study changes one semantic mechanism at a time while preserving the data, seed, training schedule, and sampling configuration:

- **w/o Attribute Alignment** removes the attribute head and all semantic losses while retaining the prompt bank and weighted transfer;
- **w/o Prototype Adaptation** removes prototype-aware target initialization and replaces it with the uniform mean of source embeddings;
- **w/o Structured Prompt Bank** replaces the entire bank with one shared learnable anomaly vector; prototype identities, attribute targets, and prototype-conditioned routing are therefore unavailable; and
- **Full *DecoGen*** enables the prompt bank, alignment branch, and prototype-aware target initialization.

The full model is best on eight of nine metrics in Table 4. The only exception is pixel AP, for which the prototype-ablation value exceeds the full model by approximately 0.01 percentage point before rounding. This difference is much smaller than the drops in IS, IC-L, all image-level metrics, pixel F_1-max, and IoU. We therefore interpret prototype adaptation as improving the balance across generation, separability, and localization rather than maximizing one metric.

C.5 Parameter sensitivity



(a) Guidance scale (b) Generated-feedback weight (c) Learning rate

Figure 5: Parameter sensitivity on eight unseen ZJU Leaper styles. Panels (a)–(c) compare guidance scale s, generated-feedback weight λ_{gen}, and learning rate, respectively. Each panel reports I-AUROC, P-AUROC, P-AP, and P-IoU in percent; bar colors distinguish the four settings in its legend. Numerical annotations show the scores for the best setting within each panel.

Analysis. Figure 5 compares four settings for each of three parameters. Increasing guidance from 3.0 to 7.5 raises pixel AP from 66.10 to 76.91 and pixel IoU from 45.21 to 55.23, gains of 10.81 and 10.02 percentage points. Both metrics decrease at $s = 10.0$. The generated-feedback weight $\lambda_{\text{gen}} = 0.05$ and learning rate 0.010 achieve the highest values for all four plotted metrics within their respective groups. These single-seed comparisons describe sensitivity within each sweep; the individual best settings do not establish a globally optimal joint configuration.

The parameter study varies classifier-free guidance $s \in \{3.0, 5.0, 7.5, 10.0\}$, generated-feedback weight $\lambda_{\text{gen}} \in \{0.01, 0.05, 0.10, 0.15\}$, and learning rate $\text{lr} \in \{0.001, 0.005, 0.010, 0.015\}$. Every group uses seed 23, at most five source anomalies per source style, 200 valid triplets per target style, and the same evaluator.

Figure 5 visualizes four representative consistency metrics. Table 13 reports all nine metrics, with the image- and pixel-level metrics in percent; bold marks the best setting in each group, and † marks the selected generated-feedback weight.

Guidance $s = 7.5$, $\lambda_{\text{gen}} = 0.05$, and learning rate 0.01 are the group-wise winners in the measured sweeps. The generated-feedback sweep is reported as a separate implementation sensitivity study; it is not an additional term in the objective defined above. Because each factor is varied separately, these results do not prove that the three individual winners are a globally optimal joint configuration.

C.6 Randomness and uncertainty

All controlled ablations use one fixed seed. Consequently, the results support directional conclusions when several metrics move consistently, but they do not provide confidence intervals. The 0.01-point

Table 13: One-factor parameter sensitivity on the reference model. Detection-style metrics are reported in percent. Bold denotes the best setting within each parameter group; † marks the selected nonzero generated-feedback weight.

Parameter	Value	Generation		Image-level			Pixel-level			
		IS	IC-L	AUROC	AP	F_1-max	AUROC	AP	F_1-max	IoU
Guidance s	3.0	1.849	0.605	92.89	93.40	91.32	88.61	66.10	60.96	45.21
	5.0	1.909	0.621	93.76	94.33	92.02	90.21	70.72	64.53	48.83
	7.5	**1.976**	**0.640**	**94.67**	**95.23**	**92.73**	**92.26**	**76.91**	**70.37**	**55.23**
	10.0	1.834	0.602	92.40	92.98	91.51	87.81	64.22	59.27	43.50
λ_{gen}	0.01	1.879	0.613	93.32	94.04	91.77	90.02	70.03	63.86	48.17
	$0.05^{\dagger}$	**2.010**	**0.636**	**94.96**	**95.38**	**93.10**	**92.07**	**75.15**	**68.30**	**52.90**
	0.10	1.887	0.609	92.72	93.40	91.53	88.76	66.99	61.24	45.56
	0.15	1.834	0.602	92.40	92.98	91.51	87.81	64.22	59.27	43.50
Learning rate	0.001	1.827	0.598	92.15	92.47	91.06	87.58	65.22	60.15	44.68
	0.005	1.876	0.614	93.72	94.33	92.04	89.29	67.89	61.98	46.29
	0.010	**1.987**	**0.635**	**95.68**	**96.22**	**93.43**	**92.52**	**75.95**	**69.24**	**54.09**
	0.015	1.982	0.633	94.48	95.19	92.20	91.37	74.43	67.63	52.18

pixel-AP gap between Full ***DecoGen*** and w/o Prototype Adaptation should not be interpreted as statistically significant. Multi-seed means and standard deviations remain necessary before making claims about differences of this scale. We report this limitation explicitly rather than extrapolating variance from unrelated experiments.

C.7 Extended comparison on ZJU Leaper

Figure 6 provides the expanded comparison corresponding to Figure 3 in the main text. It shows six generated outputs for each of the five methods on the Blue Plaid category of ZJU Leaper, complementing the selected outputs and enlarged defect regions in the main figure. This larger set allows inspection of background preservation and variation in defect appearance across multiple examples from the same style.

Background preservation and defect appearance. The expanded examples illustrate the behavior discussed with Figure 3. Several AnomalyDiffusion outputs contain broad patches with abrupt transitions to the plaid texture, while some SeaS outputs disrupt the local weave over a large area. MAGIC often produces small, similar-looking marks, and several DualAnoDiff examples contain repeated streak-like patterns. The ***DecoGen*** row includes changes in color, local texture, and thread structure, with the surrounding plaid remaining recognizable. These examples support the qualitative comparison of local integration and within-style variation in the main text.

Relation to the quantitative results. The IC-LPIPS column provides the dataset-level diversity results alongside the Blue Plaid examples. IC-LPIPS measures perceptual variation and does not establish realism or background preservation by itself. Those properties are assessed visually here; the complete numerical generation comparison, including IS, KID, and IC-LPIPS(a), is reported in Table 1.

C.7.1 Examples across unseen target styles

Figure 7 shows one manually selected representative generated anomaly for each unseen textile style. These tiles are presentation examples, not substitutes for the evaluator's triplet and mask checks. They illustrate the central transfer requirement: anomaly appearance changes locally while plaid, weave, or floral structure remains recognizable outside the mask.

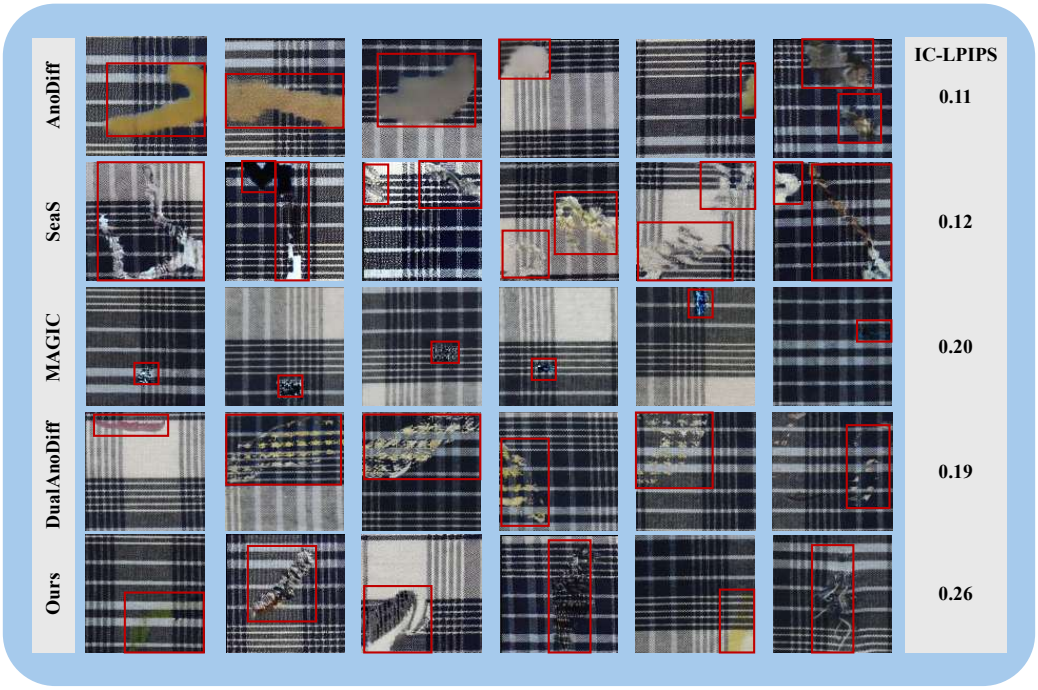


Figure 6: **Expanded comparison for Figure 3 on Blue Plaid (ZJU Leaper).** Each row contains six generated outputs. From top to bottom: AnomalyDiffusion (AnoDiff), SeaS, MAGIC, DualAnoDiff, and ***DecoGen*** (Ours). Red boxes highlight defect regions for visual inspection. The right column reproduces the ZJU Leaper IC-LPIPS scores from Table 1: 0.11, 0.12, 0.20, 0.19, and 0.26, respectively. These scores summarize the quantitative evaluation and are not computed only from the six displayed outputs.

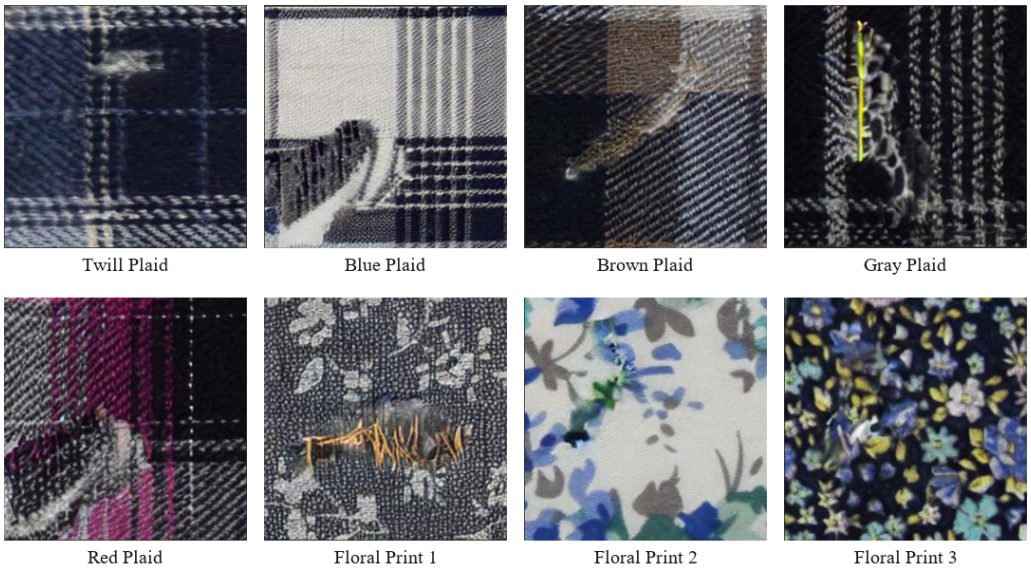


Figure 7: Manually selected ***DecoGen*** outputs for the eight unseen ZJU Leaper target styles. No real target anomaly is used to construct the displayed tokens. The figure is qualitative; numerical results are drawn only from runs that pass the complete-triplet and valid-mask checks.

C.7.2 Failure modes and interpretation

Four cases deserve particular attention. First, an incomplete source prototype bank cannot represent genuinely unseen defect physics. Second, slot-balanced similarity can still underweight synonyms or visually similar defects described with different words. Third, weak source masks can misalign the spatial token, while overly large masks permit background texture to change. Finally, strong classifier-free guidance can amplify contrast or damage periodic plaid structure near a mask boundary. Metadata and paired normal references make these cases auditable and allow low-quality pairs to be removed before detector training.

Archived qualitative export. An older local qualitative export for pattern9 (Blue Plaid) contains 200 empty masks. It is retained only as an example of why file counts are not sufficient for validation and must not be used for quantitative evaluation. The manually selected tile in Fig. 7 is shown only to visualize the appearance transfer. Quantitative tables are collected from experiment directories that pass the independent nonempty-mask checks described above; the invalid archived export is excluded.

D Limitations, Responsible Use, and LLM Assistance

Limitations. ***DecoGen*** assumes that source defects contain semantic primitives relevant to the target domain. The convex combination in Eq. A.7 cannot create a new physical mechanism absent from every source prototype. The transparent slot-balanced similarity also handles synonyms and higher-order semantic relations less effectively than a learned retriever. Generation depends on the quality of the source masks and normal target references, and the current controlled ablations provide no multi-seed uncertainty estimate. These limits motivate broader prototype coverage, semantic retrieval, and repeated trials in future work.

Responsible use. Synthetic anomalies can shorten the cold-start period for a new product style, but they do not replace real failure collection or production validation. A detector trained with ***DecoGen*** should be re-evaluated as genuine target defects become available. High-consequence

inspection decisions should retain human oversight, and metadata linking each generated image to its normal reference, mask, target prototype, and source weights should be preserved for auditing.

Use of large language models. Large-language-model tools were used to assist translation, grammar correction, document organization, and language polishing of the manuscript and appendix. The authors remain responsible for the scientific content and verified the method descriptions, equations, code-derived settings, and reported numerical results. LLMs were not used to fabricate measurements or replace execution of the experimental and evaluation pipelines.