DeltaDeno: Zero-Shot Anomaly Generation via Delta-Denoising Attribution

Chaoran Xu[1,2], Chengkan Lv[1,2,3], Qiyu Chen[1,2], Yunkang Cao[4], Feng Zhang[1,2,3]⋆, and Zhengtao Zhang[1,2,3]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] CASI Vision Technology CO., LTD., Luoyang, China
[4] School of Artificial Intelligence and Robotics, Hunan University, Changsha, China
{xuchaoran2024,chengkan.lv,chenqiyu2021,feng.zhang,zhengtao.zhang}@ia.ac.cn
caoyunkang@ieee.org

Abstract. Anomaly generation is often framed as few-shot fine-tuning with anomalous samples, which contradicts the scarcity that motivates generation and tends to overfit category priors. We tackle the setting where no real anomaly samples or training are available. We propose Delta-Denoising (**DeltaDeno**), a training-free zero-shot anomaly generation method that localizes and edits defects by contrasting two diffusion branches driven by a minimal prompt pair under a shared schedule. By accumulating per-step denoising deltas into an image-specific localization map, we obtain a mask to guide the latent inpainting during later diffusion steps and preserve the surrounding context while generating realistic local defects. To improve stability and control, DeltaDeno performs token-level prompt refinement that aligns shared content and strengthens anomaly tokens, and applies a spatial attention bias restricted to anomaly tokens in the predicted region. Experiments on public datasets show that DeltaDeno achieves great generation, realism and consistent gains in downstream detection performance. Code will be made publicly available at `https://github.com/CROVO1026/DeltaDeno`.

Keywords: Anomaly generation · Anomaly detection · Zero-shot

1 Introduction

Automated visual inspection [2] is a cornerstone of modern smart manufacturing and safety-critical operations. In principle, high-performance anomaly detection (AD) systems can be trained with sufficient defective data under supervised learning [3, 8, 9, 15, 30, 38]. In practice, defects are rare, diverse, and costly to curate. Early-stage production lines and high-yield semiconductor processes exhibit exceptionally low defect rates. Many defects emerge only after process changes or tool aging and pixel-accurate masks also require expert annotation [14, 20, 37, 41].

⋆ Corresponding author.

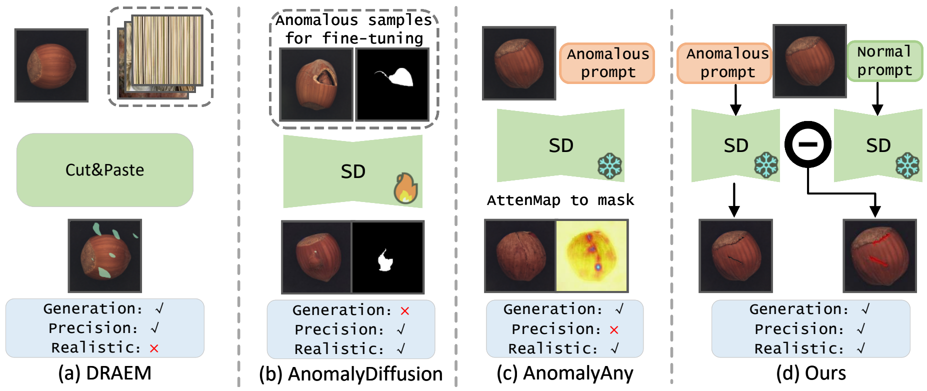


Fig. 1: Comparison between visual anomaly generation methods. Compared with prior approaches, **DeltaDeno** delivers cross-category generalization, precise masks, and high realism while requiring no fine-tuning.

Beyond scarcity, domains evolve over time through changes in materials, suppliers, and lighting, which creates persistent distribution shifts that degrade deployed AD [21]. These factors motivate *anomaly generation (AG)* as a complementary route to provide realistic and controllable abnormal samples and to stress-test detectors without halting production [20, 35, 39].

Early AG efforts centered on cut-paste composition [25, 26, 39] and GAN-based generators [13, 40]. With the rapid advances in diffusion models [12, 18, 19, 29, 31], diffusion-based AG methods have since emerged and become a major line of work that are usually guided by masks, bounding boxes, or short text prompts [11, 33, 37]. This line of research is primarily driven by two objectives. The first is *realism* [21, 35], meaning that the generated region should be photometrically and geometrically consistent with the background. The second is *controllability* [33], meaning that one can specify where the change occurs, how strong it is, and what type of defect is inserted.

Despite steady progress, many current pipelines adopt *few-shot anomaly generation (FSAG)* [13,20,21,33,35]. A pretrained generator is adapted with a small number of real anomalous samples and masks, and the adapted model is then used to expand the training set. This practice runs against the motivation for AG, which is to reduce dependence on rare and costly annotations. In many deployments, anomalous samples do not exist at scale, are expensive to collect, or are restricted by safety and privacy rules. FSAG also binds the generator to category-specific priors such as texture, lighting, and geometry, which encourages overfitting and weakens transfer across categories. As a result, gains on category A often fail to carry over to category B without anomalous samples. *Zero-shot anomaly generation (ZSAG)* aims to remove this dependence on anomalous pixels and additional training. AnomalyAny [37] is a representative ZSAG example that shows good cross-category generalization, yet it also exposes practical limits. As shown in Fig. 1, DRAEM follows the cut-paste paradigm and shows strong generalization ability but very limited realism [39]. The representative FSAG approach, AnomalyDiffusion, relies on supervised fine-tuning with real anomalous samples and masks, which makes it difficult to generalize when no downstream anomalies are available [20]. AnomalyAny builds masks from attention maps, but their low resolution and inaccurate localization remain problematic.

To address these limitations, we present *DeltaDeno*, a training-free ZSAG framework that localizes and generates defects without any anomalous samples or model adaptation. We use a minimal pair of prompts that differ only by an anomaly descriptor to drive two synchronized diffusion branches under the same schedule. At each step, we compare the two branches' denoiser predictions and accumulate their differences to build an image-specific anomaly attribution map as mask. Later, we perform mask-guided latent inpainting, which refines the initially coarse anomaly regions during later time-steps while preventing the synthesized defects from spilling into normal areas. We also apply **normal-reference guided initialization**, which starts from a partially noised encoding of the normal image to keep synthesis aligned with the original content distribution and to shorten the reverse trajectory. To further sharpen localization, we introduce lightweight **token-level prompt refinement** and **spatial attention biasing**, which strengthen the anomaly-token discrepancy across the two denoising branches and suppress the influence of non-anomaly tokens, yielding a more precise mask with fewer spurious edits. This attribution-guided design ties localization to the model's own denoising dynamics and produces sharper, more realistic anomalies without fine-tuning. The main contribution of this paper can be summarized as follows:

- We propose a zero-shot, training-free anomaly generation framework DeltaDeno that requires no anomalous samples, no fine-tuning, and no task-specific retraining.
- We introduce *delta-denoising attribution* for ZSAG, using stepwise denoising discrepancies between normal and anomaly prompts to localize and generate defects, while lightweight prompt refinement and spatial attention biasing improve realism and mask precision.
- We demonstrate strong localization fidelity, photorealism, and cross-category generalization across industrial and other domains, yielding competitive improvements in downstream AD performance.

2 Related Work

2.1 Diffusion-based Image Editing

Text-guided diffusion enables high-quality and wide-coverage image editing, with Stable Diffusion [31] and its latent variants serving as scalable backbones for reconstruction and control. Some methods focus on preserving structural consistency while steering appearance. For example, SDEdit [27] adjusts denoising schedules to balance fidelity and edit strength, while Prompt-to-Prompt [17] constrains cross-attention so that edits follow token semantics without disturbing the spatial layout. Attend-and-Excite [7] further amplifies weak tokens to improve semantic faithfulness. Other approaches enhance localization by tightening the reconstruction path, as in Null-text inversion [28] and Imagic [22], or by confining edits with masks as demonstrated in blended latent diffusion [1]. DiffEdit [10] derives an edit mask by contrasting source and target prompts under a shared schedule, while subsequent methods such as MasaCtrl [6] and InstructPix2Pix [5] extend diffusion control to multi-prompt and instruction-driven editing. Inspired by these works, we edit normal images with anomaly semantics, enabling training-free, context-consistent anomaly generation.

2.2 Anomaly Generation

Anomaly generation provides synthetic abnormal samples for training, evaluation, and stress testing of anomaly detection systems. Early practical approaches rely on cut-paste composition, such as DRAEM, which transfers defect patches onto normal images but often suffers from lighting and material inconsistency [39]. GAN-based AG [13, 40] improves texture realism and diversity through adversarial learning, while diffusion-based backbones further enhance fidelity and controllability by introducing masks, boxes, or text prompts to guide synthesis. A common trend is FSAG [11, 20, 21], which fine-tunes a pretrained generator on a small set of real anomalies and pixel masks to expand training data. Although FSAG produces detailed and category-specific defects, it depends on scarce annotations and tends to overfit to class priors, limiting cross-domain generalization. To remove this dependence, ZSAG leverages the intrinsic guidance of pretrained diffusion models and text prompts. Recent ZSAG methods, such as AnomalyAny [37] and AnoStyler [34], broaden category coverage, yet still face limitations in generation realism and mask precision to varying degrees. In contrast, we aggregate step-wise denoising discrepancies between normal and anomaly prompts to directly attribute semantic divergence, enabling training-free, high-quality local anomaly generation with matched mask.

3 Method

We present *DeltaDeno*, a training-free zero-shot anomaly generation framework that localizes and generates defects through delta-denoising attribution, without any anomalous pixel labels or model adaptation, as shown in Fig. 2. We contrast the denoiser predictions of two synchronized diffusion branches, driven by normal and anomaly prompts, to capture semantic divergences as anomaly masks. We then refine these coarse regions via mask-guided latent inpainting, preserving contextual consistency and preventing spill-over into clean areas. To maintain alignment with the normal content distribution, we apply **normal reference guided initialization**, which starts from a partially noised normal encoding and shortens the reverse trajectory. Furthermore, lightweight **token-level prompt refinement** and **spatial attention biasing** sharpen localization by emphasizing anomaly-related tokens and suppressing irrelevant ones.

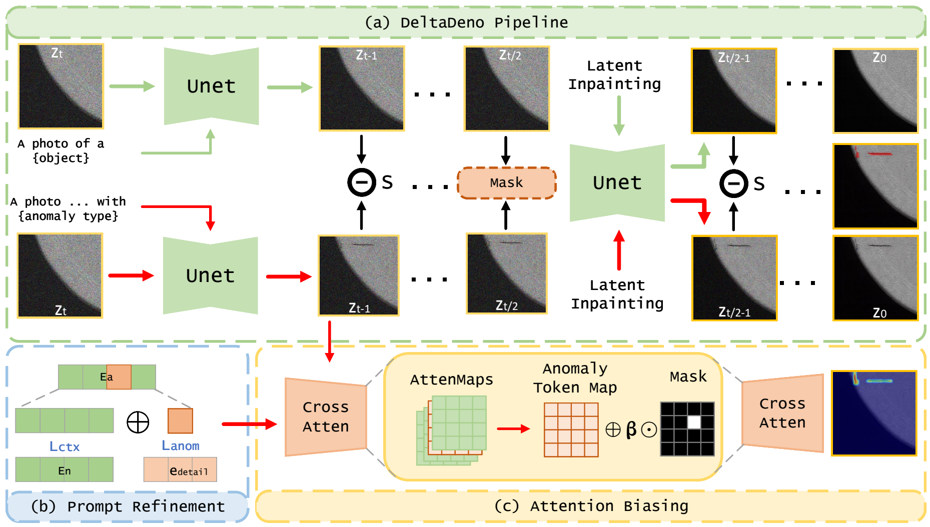


Fig. 2: Overview of the DeltaDeno framework integrating delta-denoising localization, prompt refinement, and attention biasing.

3.1 Preliminaries

Latent diffusion models. Denoising diffusion models learn the data distribution by predicting the noise added at each step. We follow latent diffusion, where an encoder $\mathcal{E}$ maps an image x to a latent $z_0 = \mathcal{E}(x)$ and a decoder $\mathcal{D}$ reconstructs images from latents [23]. Given a noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and a scheduler with (α_t, σ_t), the noisy latent is

$$z_t \;=\; \alpha_t\, z_0 \;+\; \sigma_t\, \epsilon. \tag{1}$$

A U-Net [32] ϵ_θ predicts the added noise at step t, and the training objective of LDM is

$$\mathcal{L}_{\mathrm{LDM}} \;=\; \mathbb{E}_{z_0=\mathcal{E}(x),\, \epsilon\sim\mathcal{N}(0,\mathbf{I}),\, t}\Big[\,\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\,\Big]. \tag{2}$$

At inference, the reverse process starts from $z_T \sim \mathcal{N}(0, \mathbf{I})$ and gradually denoises to z_0, followed by $x' = \mathcal{D}(z_0)$.

Cross-Attention in U-Net. Let $E \in \mathbb{R}^{Z\times d}$ be the text token embeddings (with Z tokens) and $F^{\,l} \in \mathbb{R}^{N\times d}$ be the flattened image features at layer l (with $N = r^2$ image tokens). For a cross-attention in U-Net, the projections are

$$Q^l = F^{\,l} W_q^{(l)}, \qquad K^l = E\, W_k^{(l)}, \tag{3}$$

and the cross-attention map is

$$A^l = \mathrm{softmax}\!\left(\frac{Q^l (K^l)^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{N\times Z}, \tag{4}$$

where $A^l \in \mathbb{R}^{N\times Z}$ represents the attention weights from all N image tokens to the Z text tokens.

Latent inpainting. For localized edits and background preservation, inference often blends a proposed edited latent $\tilde{z}_{t-1}$ with a source latent z^{src}_{t-1} using a spatial mask $M \in [0,1]^{H_z\times W_z}$ at the latent resolution:

$$z_{t-1} \;=\; M \odot \tilde{z}_{t-1} \;+\; (1-M) \odot z^{\mathrm{src}}_{t-1}. \tag{5}$$

This confines changes to the masked region while keeping the outside region consistent with the source.

3.2 Delta-Denoising Localization

Normal reference guided initialization. To ensure that the synthesized defective images do not deviate completely from the distribution of the original normal samples, we do not start the generation from pure Gaussian noise. We encode the guidance image x^{normal} to z_0^{normal} and partially noise it to an intermediate level t_{start}:

$$z^{\mathrm{normal}}_{t_{\mathrm{start}}} = \sqrt{\bar{\alpha}_{t_{\mathrm{start}}}}\, z_0^{\mathrm{normal}} + \sqrt{1-\bar{\alpha}_{t_{\mathrm{start}}}}\, \epsilon, \tag{6}$$

where $t_{\mathrm{start}}{=}\gamma T$ and $\epsilon \sim \mathcal{N}(0, I)$. Sampling then starts from t_{start} and performs the remaining reverse steps ($t_{\mathrm{start}} \to 0$) for both branches under the same schedule. This warm start preserves geometry and illumination, reduces computational cost, and prevents the class-typical drift observed when starting from pure noise.

Prompt design. We use a minimal pair of prompts that differ only by an anomalous attribute. For example, the *normal prompt* is "`a photo of a {object}`", and the *anomaly prompt* is "`a photo of a {object} with {anomaly type}`". Additionally, a *description prompt* is introduced for the anomaly token, which provides a refined semantic embedding (e.g., adding shape or intensity cues) during the subsequent prompt refinement stage. This descriptor is not inserted into the sentence but distilled into the anomaly token to enhance its representation to minimize off-target semantic drift.

Synchronized denoising and delta-attribution. Let T be the total number of inference steps. We define $t \in [T, T/2]$ as the early denoising stage and $t \in [T/2, 0]$ as the late denoising stage. We run two synchronized trajectories under the same schedule and guidance: a normal branch with prompt p_{n} and an anomaly branch with p_{a}. Denote their latents at step t by $z_t^{(\mathrm{n})}$ and $z_t^{(\mathrm{a})} \in \mathbb{R}^{H_z\times W_z\times C}$. A single reverse step of the U-Net is

$$z^{(\cdot)}_{t-1} \;=\; \Phi\big(z^{(\cdot)}_t,\, \hat{\epsilon}_{\mathrm{cfg}}(z^{(\cdot)}_t, t, e_{(\cdot)}),\, t\big), \tag{7}$$

where Φ is the reverse-step operator, $\hat{\epsilon}_{\mathrm{cfg}}$ is the classifier-free guided prediction, and $e_{(\cdot)}$ is the text embedding of p_{n} or p_{a}. We measure the per-step discrepancy between the two branches after each reverse update:

$$d_{t-1}(u) \;=\; \big\|\, z^{(\mathrm{n})}_{t-1}(u) - z^{(\mathrm{a})}_{t-1}(u)\,\big\|_2, \tag{8}$$

where u indexes a spatial position in the latent feature map, and the ℓ_2 norm is applied over channels. The accumulator integrates the discrepancies along the reverse trajectory as

$$S_{t-1}(u) \;=\; S_t(u) + d_{t-1}(u), \qquad S_{t_{\text{start}}}(u) = 0, \tag{9}$$

where t_{start} denotes the starting timestep. At the t_{mid}=$\lfloor T/2 \rfloor$, we lightly smooth and normalize $S_{t_{\text{mid}}}$ to obtain a continuous delta map $\hat{S}$, and threshold it to generate a binary mask:

$$M_{\text{mid}}(u) \;=\; \mathbf{1}\Big\{\, \hat{S}(u) > \tau_{\text{mid}} \,\Big\}. \tag{10}$$

After extracting M_{mid}, the accumulator S is reset to zero, and a new aggregation begins for the late denoising stage, focusing on boundary refinement and local details.

During the late denoising stage, the image-specific mask M_{mid} serves as a latent inpainting mask that confines subsequent edits to the coarse anomaly region discovered earlier. The update is

$$z_{t-1} \leftarrow M_{\text{mid}} \odot \tilde{z}_{t-1} + (1 - M_{\text{mid}}) \odot z^{\text{src}}_{t-1}, \tag{11}$$

where the z^{src}_{t-1} is obtained by forward noising the normal reference latent z_0^{normal} to the same timestep via DDIM [36]. This operation refines defect synthesis inside the mask while preserving the normal-image distribution outside, ensuring local detail enhancement and global consistency.

3.3 Prompt Refinement and Attention Biasing

Diffusion-based generation in Stable Diffusion is tightly coupled with the text condition, where each token dynamically interacts with visual features throughout the denoising process. Given our delta-denoising formulation, the semantic discrepancy between the two diffusion branches should be localized exclusively to the anomaly token, rather than spreading across the entire prompt. Therefore, we introduce two complementary mechanisms, *Token-level prompt refinement* and *Spatially-biased cross-attention*.

Token-level prompt refinement. To convey fine-grained defect semantics while minimizing semantic differences outside the anomaly token, we perform a lightweight token-level refinement on the anomaly prompt embedding before denoising. Let the normal and anomaly prompts be tokenized as $E_{\text{n}} = \{e_{\text{n}}^1, \dots, e_{\text{n}}^{Z_{\text{n}}}\}$ and $E_{\text{a}} = \{e_{\text{a}}^1, \dots, e_{\text{a}}^{Z_{\text{a}}}\}$, respectively. We selectively update E_{a} with two complementary objectives.

(i) Anomaly Semantic Distillation. Embeddings of anomaly-specific tokens (e.g., *crack*, *scratch*, *damage*) are steered toward a detailed semantic anchor e_{detail} distilled from a more descriptive phrase:

$$\mathcal{L}_{\text{anom}} = 1 - \cos(e_{\text{a}}^j, e_{\text{detail}}) + \lambda \| e_{\text{a}}^j - e_{\text{detail}} \|_2^2.$$

(ii) Context Alignment for non-anomaly tokens. For the remaining Z' non-anomaly tokens in the anomaly prompt, we compute their mean embedding as a context centroid and encourage each token to stay close to it:

$$\bar{e}_{\text{ctx}} = \frac{1}{Z'} \sum_{j=1}^{Z'} e_{\text{a}}^{j}, \qquad \mathcal{L}_{\text{ctx}} = \frac{1}{Z'} \sum_{j=1}^{Z'} \| e_{\text{a}}^{j} - \bar{e}_{\text{ctx}} \|_2^2.$$

This simple MSE regularization stabilizes the shared context and prevents semantic drift among non-anomaly tokens. We jointly optimize the two objectives as

$$\mathcal{L}_{\text{prompt}} = \mathcal{L}_{\text{anom}} + \mathcal{L}_{\text{ctx}},$$

and update the anomaly prompt embedding as

$$E_{\text{a}} \leftarrow E_{\text{a}} - \eta \nabla_{E_{\text{a}}} \mathcal{L}_{\text{prompt}}, \tag{12}$$

where η is the step size. This training-free refinement runs for a few iterations with a fixed step size, ensuring that semantic changes remain localized to the anomaly token while other tokens retain the original context.

Spatially-biased cross-attention. We use a two-stage spatial prior for attention. During the early denoising stage, a coarse foreground mask M_{fg} (from SAM [24]) activates plausible surface regions. During the late denoising stage, we switch to the M_{mid} from Sec. 3.2 to focus the defect area. At each step, we inject the mask M into the cross-attention logits for the anomaly token only:

$$A^l = \text{softmax}\left(\frac{Q^l (K^l)^\top + \beta\, M\, o_a^\top}{\sqrt{d_h}} \right), \tag{13}$$

Here M represents the current spatial prior mask, either M_{fg} in the early denoising stage or M_{mid} in the late denoising stage, and o_a denotes the one-hot column selector that specifies the anomaly token $j{=}a$. The bias term $\beta\, M\, o_a^\top$ injects spatial awareness into the anomaly token's attention map, so that the model focuses on masked regions when attending to the anomaly description, while leaving other tokens unaffected to maintain overall consistency.

3.4 Inference

At inference, ZSAG generates a realistic defect image and its mask from a single normal input. Given a normal image I, a normal prompt p_{n}, and an anomaly prompt p_{a}, we: (i) run Token-level prompt refinement for p_{a} (ii) encode I into latent z_0 and get normal reference guided initialization to two synchronized diffusion branches conditioned on p_{n} and p_{a}; (iii) accumulate latent discrepancies S during the early denoising stage to form M_{mid} and reset S at t_{mid}; (iv) continue the late denoising stage with attention biasing and inpainting guided by M_{mid}; (v) at the final step, re-threshold S to get the final mask M, and decode the anomaly branch as the defect image I'.

Algorithm 1 DeltaDeno Inference

$M_{\text{fg}} \leftarrow \text{SAM}(I)$, $z_0 \leftarrow \text{VAE-enc}(I)$, $z_t \leftarrow$ normal reference guided(z_0)
Token-level prompt refinement for p_{a}
Run two diffusion branches with same schedule and seed: $(p_{\text{n}}, p_{\text{a}})$
for each step t **do**
  if early denoising stage **then**
    attention biasing with M_{fg}; update S
  else if $t = t_{\text{mid}}$ **then**
    $M_{\text{mid}} \leftarrow$ threshold(clean(normalize(S))); reset S
  else
    attention biasing and latent inpainting with M_{mid}; update S
  end if
end for
$M \leftarrow$ threshold(clean(normalize(S))), $I' \leftarrow$ VAE-dec(anomaly branch)
return (I', M)

4 Experiments

4.1 Experiment Setup

Datasets and Baselines. Following the latest representative ZSAG and FSAG methods [11, 21, 35], we mainly conduct experiments on the widely used MVTec AD [2] and VisA [42], which provide a comprehensive benchmark for both generation and detection. We compare with some existing few-shot and zero-shot methods including DRAEM [39], Crop-Paste [26], DFMGAN [13], AnomalyDiffusion [20], AnoGen [16], AnomalyAny [37] and AnoStyler [34].

Evaluation Metrics. To quantitatively evaluate the anomaly generation quality, we employ the Inception Score (IS) to assess image realism and the intra-cluster pairwise LPIPS distance (IC-LPIPS) to measure generation diversity. Following previous FSAG and ZSAG works, we also validate the usefulness of our generated anomalies by testing their effectiveness in anomaly detection frameworks. For anomaly localization and detection performance, we report both image-level and pixel-level metrics, including the Area Under the Receiver Operating Characteristic (AUROC) and the Average Precision (AP). These metrics jointly reflect the localization precision and global detection quality of generated anomalies under a consistent evaluation protocol.

Implementation Details. Unless otherwise specified, we adopt the pretrained Stable Diffusion v1.5 model as our default backbone, with $T = 100$ denoising steps and a partial-noise ratio of $\gamma = 0.3$ for a fair comparison with AnomalyAny. During generation, we extract the mask at t_{mid} using a threshold of 0.6, and the final mask at the last step with a threshold of 0.35 by default. Unless otherwise stated, all other parameters and implementation settings follow the standard Stable Diffusion configuration. As for the prompt design, we use simple, dataset-agnostic prompt templates during downstream generation. The normal prompt is "`a photo of a {object}`", while the anomaly prompt extends this to "`a photo of a {object} with {anomaly type} on it`". For

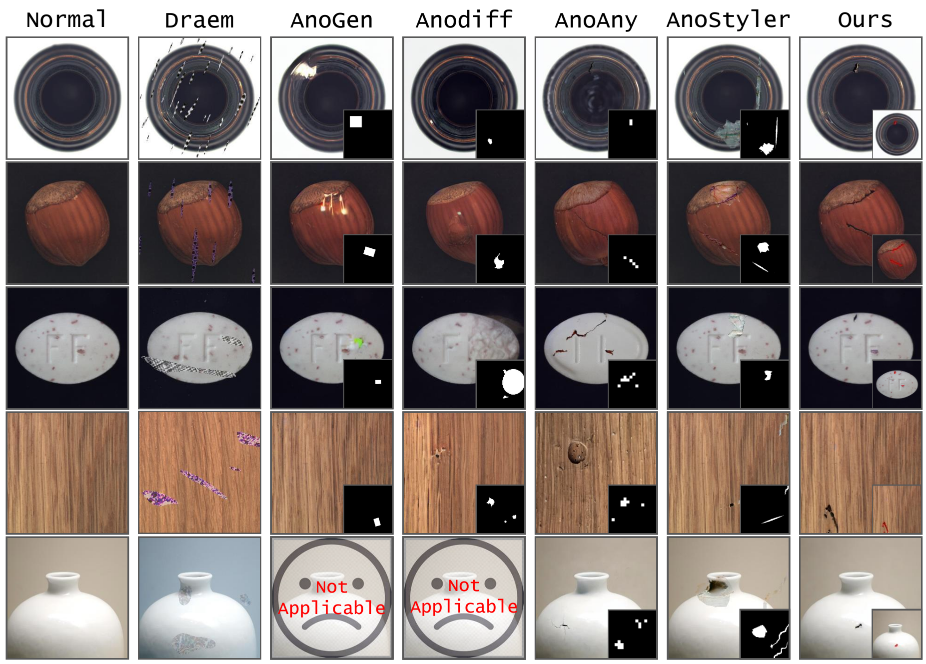
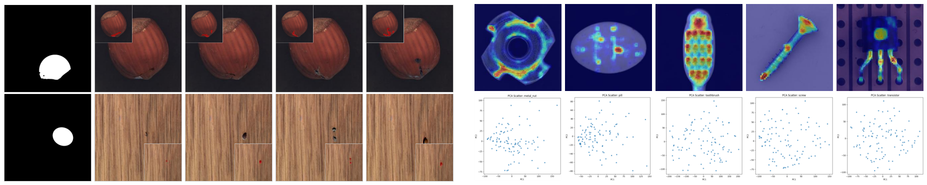


Fig. 3: Qualitative comparison with existing anomaly generation methods. Rows 1–4 correspond to categories from MVTec AD, while the last row shows ceramic bottle images collected from the Internet.

token-level refinement, we additionally use the description prompt in the simple form of ‘‘`some large and deep {anomaly type}`’’ or similar variants.

4.2 Anomaly Generation Results

Qualitative visualization. We conduct a qualitative comparison on major MVTec AD categories and unseen real industrial samples (Fig. 3). DRAEM often exhibits illumination mismatch and visible boundary seams due to cut-paste composition. FSAG methods (AnomalyDiffusion, AnoGen) rely on few-shot fine-tuning with real defects and masks, performing well on seen categories but failing to generalize to unseen ones. AnomalyAny produces coarse masks derived from low-resolution cross-attention, leading to blurry boundaries and inaccurate localization. AnoStyler sometimes generates defects that are semantically inconsistent with the target category, reducing realism. In contrast, our method produces visually sharp boundaries and context-consistent defects across both seen and unseen categories. To further illustrate the generation quality, Fig. 4(a) demonstrates basic spatial location control, where the generated anomalies follow the specified masks. Fig. 4(b) presents the spatial frequency maps and PCA projections of generated defect masks across multiple categories, revealing diverse spatial distributions and structural variations.



(a) (b)

Fig. 4: (a) Illustration of basic spatial location control. The leftmost column shows the manually specified control masks. **(b)** Spatial frequency maps and PCA visualizations of defect masks generated across multiple categories.

Table 1: Comparison with existing anomaly generation methods on generation quality and diversity.

Category	Crop-Paste [26]		DFMGAN [13]		AnomalyDiff [20]		AnomalyAny [37]		AnoStyler [34]		Ours	
	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑
bottle	1.43	0.04	1.62	0.12	1.58	0.19	1.73	0.17	1.55	0.21	2.36	0.27
cable	1.74	0.25	1.96	0.25	2.13	0.41	2.06	0.41	1.53	0.39	2.49	0.41
capsule	1.23	0.05	1.59	0.11	1.59	0.21	2.16	0.23	1.96	0.20	1.82	0.23
carpet	1.17	0.11	1.23	0.13	1.16	0.24	1.10	0.34	1.28	0.26	1.27	0.45
grid	2.00	0.12	1.97	0.13	2.04	0.44	2.31	0.38	2.52	0.39	2.26	0.39
hazelnut	1.74	0.21	1.93	0.24	2.13	0.31	2.55	0.32	1.83	0.32	2.62	0.38
leather	1.47	0.14	2.06	0.17	1.94	0.41	2.26	0.41	2.94	0.45	1.77	0.44
metal_nut	1.56	0.15	1.49	0.32	1.96	0.30	1.82	0.27	2.33	0.30	2.00	0.31
pill	1.49	0.11	1.63	0.16	1.61	0.26	2.91	0.30	1.65	0.24	2.26	0.33
screw	1.12	0.16	1.12	0.14	1.28	0.30	1.33	0.32	1.39	0.32	1.29	0.37
tile	1.83	0.20	2.39	0.22	2.54	0.55	2.66	0.53	2.92	0.52	2.30	0.50
toothbrush	1.30	0.08	1.82	0.18	1.68	0.21	1.64	0.22	1.46	0.17	1.99	0.24
transistor	1.39	0.15	1.64	0.25	1.57	0.34	1.66	0.28	1.57	0.29	1.97	0.32
wood	1.95	0.23	2.12	0.35	2.33	0.37	1.93	0.41	2.91	0.39	2.19	0.44
zipper	1.23	0.11	1.29	0.27	1.39	0.25	2.14	0.33	2.77	0.29	2.09	0.37
Avg.	1.51	0.14	1.72	0.20	1.80	0.32	2.02	0.33	2.04	0.32	**2.05**	**0.36**

Generation quality. We quantitatively evaluate anomaly image quality on MVTec AD using Inception Score (IS) for fidelity and IC-LPIPS for diversity (Tab. 1). Our method consistently achieves higher IS and IC-LPIPS than prior works, indicating improved photorealistic fidelity while maintaining diverse defect patterns. These results confirm that our approach balances high-quality synthesis with cross-category generalization in a fully training-free setting.

4.3 Anomaly Detection Results

We evaluate the effectiveness of the generated anomalous data via a downstream detection task on MVTec AD and VisA. Following AnomalyAny, we generate 100 anomalous images per category from the single normal image to ensure a fair comparison and train a simple U-Net on the synthesized data. We report performance on the real MVTec AD and VisA test split (Tab. 2, Tab. 3). For ZSAG methods, we further control the source by generating from the same single normal image per category to isolate the generation effect. No anomalous samples from the real datasets are used during our synthesis or U-Net training. All results are obtained with the same U-Net architecture and training schedule without test-time adaptation.

Table 2: Comparison of downstream detection performance on MVTec AD from ZSAG methods. Bold numbers indicate the best results.

	AnomalyAny (CVPR'25)				AnoStyler (AAAI'26)				DeltaDeno (SD1.5)				DeltaDeno (FLUX)			
Category	AUC_i	AP_i	AUC_p	AP_p	AUC_i	AP_i	AUC_p	AP_p	AUC_i	AP_i	AUC_p	AP_p	AUC_i	AP_i	AUC_p	AP_p
bottle	94.9	98.0	86.9	32.2	92.3	97.3	67.6	17.7	**97.2**	**98.8**	86.0	28.4	92.7	97.1	**91.1**	**45.1**
cable	44.1	49.0	82.2	11.1	62.0	63.3	65.3	4.4	70.9	74.9	76.2	8.0	**89.3**	**91.7**	**86.6**	**24.5**
capsule	59.4	84.6	78.2	7.4	80.9	**93.6**	84.8	13.0	54.9	79.0	**87.9**	5.8	**82.9**	93.1	78.4	**18.4**
carpet	79.4	90.2	80.9	30.9	81.7	92.7	90.1	36.2	85.9	94.0	90.1	41.7	**89.1**	**95.3**	**95.8**	**61.8**
grid	85.7	92.7	73.9	5.3	71.0	83.0	55.2	0.9	81.3	89.9	77.3	8.7	**86.5**	**93.5**	**87.4**	**24.3**
hazelnut	98.8	99.0	95.3	43.8	97.6	98.8	97.5	71.2	99.1	99.2	97.5	69.6	**99.2**	**99.4**	**99.1**	**77.5**
leather	92.6	97.0	94.1	45.0	**98.6**	**99.8**	94.2	36.4	96.5	98.3	94.6	**47.1**	86.2	94.1	**97.2**	45.1
metal_nut	54.3	80.9	77.1	41.0	65.8	87.6	56.7	23.7	73.8	90.0	83.2	38.5	**85.6**	**95.0**	**92.2**	**69.1**
pill	65.9	88.9	88.9	36.7	82.3	95.1	84.2	59.4	**89.7**	**97.3**	93.3	**65.1**	87.7	96.4	**95.4**	**65.1**
screw	60.9	75.7	92.1	2.3	84.9	93.2	93.2	**15.9**	73.3	85.8	**95.6**	4.7	**100.0**	**100.0**	46.6	0.2
tile	97.0	98.5	89.2	53.9	99.1	99.5	94.2	75.8	99.9	**100.0**	87.0	51.9	**100.0**	**100.0**	**98.7**	**88.3**
toothbrush	89.6	95.0	75.2	12.2	87.8	92.3	77.8	10.8	89.2	94.1	78.9	12.9	**90.0**	**95.3**	**86.6**	**19.9**
transistor	78.9	72.4	**65.2**	12.0	**84.6**	**84.4**	58.4	**17.0**	78.8	76.8	58.9	14.9	**84.6**	81.3	62.8	**17.0**
wood	**99.4**	**99.7**	84.4	42.1	94.4	98.0	**84.9**	**56.4**	99.0	99.6	80.7	42.2	96.9	98.6	81.7	44.0
zipper	63.6	84.8	85.7	21.8	**83.5**	**94.7**	68.4	14.8	80.8	91.7	**86.8**	18.5	74.9	90.7	81.8	**22.5**
Mean	77.6	87.1	83.3	26.5	84.4	91.5	78.2	30.2	84.7	91.3	84.9	30.5	**89.7**	**94.8**	**85.4**	**41.5**

Table 3: Comparison of downstream detection performance on VisA from ZSAG methods. Bold numbers indicate the best results.

	AnomalyAny (CVPR'25)				AnoStyler (AAAI'26)				DeltaDeno (SD1.5)				DeltaDeno (FLUX)			
Category	AUC_i	AP_i	AUC_p	AP_p	AUC_i	AP_i	AUC_p	AP_p	AUC_i	AP_i	AUC_p	AP_p	AUC_i	AP_i	AUC_p	AP_p
candle	50.9	56.7	**87.8**	0.8	71.2	66.2	80.8	**16.5**	60.0	67.4	85.9	6.9	**81.9**	**82.7**	78.9	1.0
capsules	81.3	86.6	76.8	2.7	62.5	75.8	64.9	0.7	**91.6**	**94.1**	92.0	13.3	79.5	87.4	**92.2**	**21.5**
cashew	82.6	85.6	**90.1**	6.2	85.3	87.4	**90.1**	**7.8**	88.8	92.6	78.3	4.4	**95.8**	**97.7**	66.7	2.0
chewinggum	89.3	95.4	94.9	44.9	94.6	97.5	97.6	**71.1**	**95.4**	**97.9**	97.3	37.9	91.5	96.3	**98.0**	49.4
fryum	86.6	93.7	86.1	11.0	**90.2**	**94.0**	72.2	10.9	78.2	88.7	87.1	9.4	77.0	88.6	**92.2**	**20.6**
macaroni1	92.1	92.0	92.9	0.5	91.4	91.5	97.1	**25.0**	**94.8**	**94.9**	**98.3**	16.6	87.8	85.4	93.5	0.4
macaroni2	46.2	46.5	94.4	0.4	67.3	**69.7**	**97.4**	**20.6**	**70.5**	68.9	96.9	3.4	65.9	67.3	94.2	2.8
pcb1	69.4	69.6	90.9	3.1	49.2	48.3	73.2	0.8	72.3	77.4	86.5	**7.8**	**89.0**	**85.3**	**95.3**	6.2
pcb2	76.5	83.9	82.9	0.9	67.6	78.2	88.5	1.8	**85.6**	**86.1**	**90.7**	**9.1**	92.6	93.7	89.1	5.5
pcb3	49.9	49.9	84.7	1.1	81.5	84.4	84.0	**10.7**	72.1	76.5	**88.5**	2.2	**83.2**	**85.6**	87.3	7.9
pcb4	87.1	82.7	86.0	2.5	78.9	86.0	89.8	11.6	**96.6**	**95.4**	85.2	7.7	93.9	90.6	**95.6**	**13.3**
pipe_fryum	**84.2**	**92.7**	83.5	16.2	79.0	89.4	96.0	22.2	79.5	88.9	94.7	20.1	82.8	90.7	**98.9**	**49.6**
Mean	74.7	77.9	87.6	7.5	76.5	80.7	86.0	**16.6**	82.1	85.7	90.1	11.6	**85.1**	**87.6**	**90.2**	15.0

Results show that our synthesized data yield competitive or superior downstream detection scores compared to recent ZSAG methods across the majority of tested categories. Overall, these gains stem from jointly balancing localization precision, contextual fidelity, and morphological diversity during synthesis. Crucially, our delta-denoising framework is intrinsically model-agnostic, exhibiting strong transferability and generality. As demonstrated by the seamless transition from the default Stable Diffusion v1.5 to the more advanced FLUX.1-dev [4] backbone, our method scales effectively and yields further performance boosts without requiring architecture-specific tuning. Nevertheless, under the strict ZSAG setting, the synthesized defect-type distribution inevitably deviates from real-world data. Thus, for categories with highly complex or fine-grained normal patterns, detection gains may remain limited.

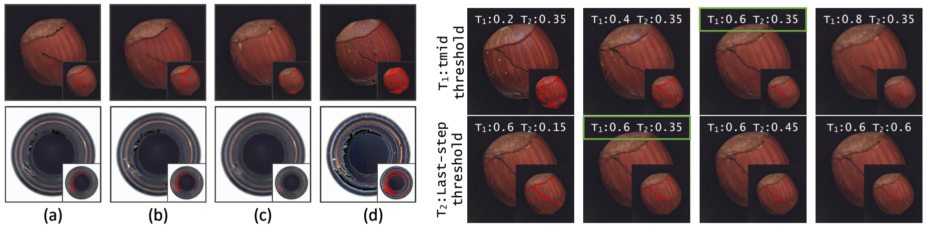


Fig. 5: Ablation study. Left: Module ablation of *DeltaDeno*. (a) Full model. (b) w/o L_{ctx}. (c) w/o spatial attention biasing. (d) w/o latent mask inpainting in the late denoising stage. **Right:** Threshold ablation showing the effect of different threshold values on generation results.

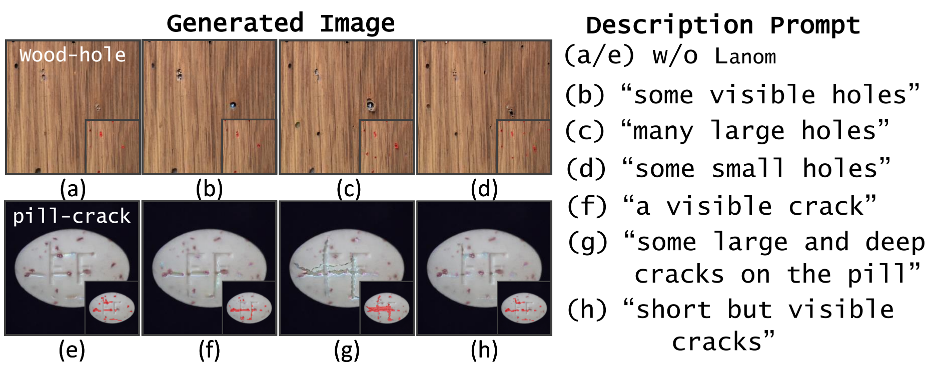


Fig. 6: Ablation visualization of the Anomaly Semantic Distillation module. Varying the description prompts steers the generated defect type and severity, yielding more controllable and diverse anomalies.

4.4 Ablation Studies

We conduct both qualitative and quantitative ablations on MVTec AD to analyze the contribution of each component and the sensitivity to key hyperparameters.

Module ablation. We first perform a qualitative module ablation as shown in Fig. 5 (left). Without the context alignment term L_{ctx}, non-anomaly tokens are more likely to drift, causing off-target edits that reduce mask localization accuracy. Removing latent-mask inpainting in the late denoising stage leads to noticeable spill-over edits and degraded background consistency, since the generation is no longer confined to the estimated defect region. Without spatial attention biasing, the anomaly-token semantics are not effectively activated in the target region, often leading to weak or even missing defect synthesis. Overall, the full model produces the sharpest boundaries and the most coherent defect synthesis.

Threshold tuning. We further study the sensitivity to the two thresholds used in our DeltaDeno (Fig. 5, right). The t_{mid} threshold mainly controls the size of the edited anomaly region during the late denoising stage, and thus has a larger

Table 4: Ablation on downstream detection (mean %). Higher is better. A: latent-mask inpainting, B: attention biasing, C: prompt refinement.

Method	AUC$_i$ ↑	AP$_i$ ↑	AUC$_p$ ↑	AP$_p$ ↑
w/o A	79.7	87.6	81.2	25.8
w/o B	79.5	88.2	83.4	28.3
w/o C	78.8	86.3	83.1	28.1
DeltaDeno (full)	**84.7**	**91.3**	**84.9**	**30.5**

influence on the final synthesized defect appearance. In contrast, the last-step threshold primarily affects the precision of the extracted mask, while having negligible impact on the visual quality of the generated image. Since the optimal thresholds vary across categories, we adopt a unified default setting of 0.6–0.35 for all categories in our main experiments based on this observation.

Anomaly semantic distillation. Fig. 6 visualizes the effect of the Anomaly Semantic Distillation module. By varying the description prompts, the anomaly token is anchored to different fine-grained attributes (e.g., length, width, texture, or severity), which steers the resulting defect morphology accordingly. This provides controllable generation while maintaining contextual consistency around the defect region.

Quantitative ablation. Finally, we evaluate the practical utility of each module via downstream detection (Tab. 4). All variants follow the same protocol: for each category we synthesize the same number of anomalies from the same normal guidance image, train a U-Net on the generated data, and evaluate on the MVTec AD test split. The full configuration achieves the best image-level and pixel-level metrics. Removing any component consistently degrades performance, confirming that latent-mask inpainting, attention biasing, and prompt refinement are all beneficial and complementary.

5 Conclusion

We introduced *DeltaDeno*, a training-free zero-shot anomaly generation framework that leverages delta-denoising to localize and edit defects in a normal sample. By contrasting two synchronized diffusion branches conditioned on normal and anomaly prompts, the method derives reliable masks and performs mask-guided latent inpainting to synthesize photorealistic, context-consistent anomalies. A normal-reference initialization keeps synthesis aligned with the distribution of normal content, while lightweight prompt refinement with spatially biased attention stabilizes localization without retraining. On public benchmarks, *DeltaDeno* attains higher generation quality and stronger downstream detection performance. The framework requires no anomalous labels or model adaptation and is practical for rapid data bootstrapping and stress testing when real defects are scarce.

Limitations and Future Work. Despite its effectiveness, *DeltaDeno* still faces several limitations. First, existing evaluation protocols for ZSAG are limited and may require MLLM-based comprehensive assessment in the future. Second, the controllability offered by semantic distillation remains limited, constraining fine-grained manipulation of defect attributes. Third, most generated anomalies exhibit structural or surface-level characteristics, while more abstract or logical defects remain challenging to produce. Future work can investigate these unresolved aspects to further improve the capability of zero-shot anomaly generation.

Acknowledgements

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0119400, the National Natural Science Foundation of China under Grant No. 62303458 and 62303461, and the Beijing Municipal Natural Science Foundation (China) under Grant 4252053.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM transactions on graphics (TOG) **42**(4), 1–11 (2023)
2. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9592–9600 (2019)
3. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: 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)
4. Black Forest Labs: FLUX. `https://github.com/black-forest-labs/flux` (2024), accessed: 25 June 2026
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
6. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
7. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
8. Chen, Q., Luo, H., Lv, C., Zhang, Z.: A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In: European Conference on Computer Vision. pp. 37–54. Springer (2024)
9. Chen, Q., Luo, H., Yao, H., Luo, W., Qu, Z., Lv, C., Zhang, Z.: Center-aware residual anomaly synthesis for multiclass industrial anomaly detection. IEEE Transactions on Industrial Informatics **21**(9), 7276–7286 (2025)

10. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
11. Dai, Z., Zeng, S., Liu, H., Li, X., Xue, F., Zhou, Y.: 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)
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
13. Duan, Y., Hong, Y., Niu, L., Zhang, L.: Few-shot defect image generation via defect-aware feature manipulation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 571–578 (2023)
14. Fang, Q., Su, Q., Lv, W., Xu, W., Yu, J.: Boosting fine-grained visual anomaly detection with coarse-knowledge-aware adversarial learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 16532–16540 (2025)
15. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Anomalygpt: Detecting industrial anomalies using large vision-language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 1932–1940 (2024)
16. Gui, G., Gao, B.B., Liu, J., Wang, C., Wu, Y.: Few-shot anomaly-driven generation for anomaly classification and segmentation. In: European Conference on Computer Vision. pp. 210–226. Springer (2024)
17. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Hu, T., Zhang, J., Yi, R., Du, Y., Chen, X., Liu, L., Wang, Y., Wang, C.: Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 8526–8534 (2024)
21. Jin, Y., Peng, J., He, Q., Hu, T., Wu, J., Chen, H., Wang, H., Zhu, W., Chi, M., Liu, J., et al.: Dual-interrelated diffusion model for few-shot anomaly image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30420–30429 (2025)
22. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6007–6017 (2023)
23. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
25. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: 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)
26. Lin, D., Cao, Y., Zhu, W., Li, Y.: Few-shot defect segmentation leveraging abundant defect-free training samples through normal background regularization and crop-and-paste operation. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE Computer Society (2021)

27. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
28. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6038–6047 (2023)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021)
30. Qu, Z., Tao, X., Shen, F., Zhang, Z., Li, T.: Investigating shift equivalence of convolutional neural networks in industrial defect segmentation. IEEE Transactions on Instrumentation and Measurement **72**, 1–17 (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
33. Shi, Q., Wei, J., Shen, F., Zhang, Z.: Few-shot defect image generation based on consistency modeling. In: European Conference on Computer Vision. pp. 360–376. Springer (2024)
34. So, Y., Kang, S.: Anostyler: Text-driven localized anomaly generation via lightweight style transfer. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
35. Song, J., Park, D., Baek, K., Lee, S., Choi, J., Kim, E., Yoon, S.: Defectfill: Realistic defect generation with inpainting diffusion model for visual inspection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18718–18727 (2025)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
37. Sun, H., Cao, Y., Dong, H., Fink, O.: Unseen visual anomaly generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25508–25517 (2025)
38. Xu, C., Lv, C., Chen, Q., Zhang, F., Zhang, Z.: MRAD: Zero-shot anomaly detection with memory-driven retrieval. In: The Fourteenth International Conference on Learning Representations (2026)
39. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8330–8339 (2021)
40. Zhang, G., Cui, K., Hung, T.Y., Lu, S.: 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)
41. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In: International Conference on Learning Representations (2024)
42. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European conference on computer vision. pp. 392–408. Springer (2022)