SEVANODIFF: SEVERITY-AWARE DIFFUSION ADAPTATION FOR ZERO-SHOT INDUSTRIAL ANOMALY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial anomaly detection relies on diverse anomaly images with pixel-level annotations, yet real industrial anomalies are rare and costly to collect. Existing anomaly generation methods mainly synthesize isolated samples while not modeling how the anomaly changes across various severity stages. First, we construct SevADSet, an industrial anomaly dataset with four-stage ordinal trajectories from normal images to increasingly visible anomaly stages. Second, based on SevADSet, we propose SevAnoDiff, a diffusion-based method for zero-shot severity-aware anomaly generation. An Appearance Prior Adapter first learns transferable anomaly appearance from observed anomalies. A Severity Progression Module then models adjacent-stage changes using stage tokens, mask trajectories, and previous-stage visual references. Without using anomalous images or masks from the target domain, SevAnoDiff generates realistic anomaly image-mask pairs with clear severity progression on target domain normal images. Extensive experiments on MVTec-AD and VisA demonstrate the effectiveness of SevAnoDiff in anomaly generation and its utility for anomaly detection and localization.

1 INTRODUCTION

In industrial anomaly detection, real anomaly samples are scarce, diverse, and difficult to collect and annotate, particularly at the pixel level Bergmann et al. (2019); Zou et al. (2022); Zavrtanik et al. (2021). Such limited supervision can constrain the performance and generalization of deep learning models, especially for subtle or early-stage defects with only a few available examples. These challenges motivate few-shot and zero-shot settings, which aim to recognize anomalies from limited or no anomalous training samples Jeong et al. (2023); Zhou et al. (2024).

To mitigate the scarcity of real anomalies, synthetic anomaly generation has been widely studied for industrial anomaly detection. Early methods construct pseudo anomalies through cut-and-paste augmentation, texture corruption, or handcrafted perturbations Li et al. (2021); Zavrtanik et al. (2021); Schlüter et al. (2022). Recent diffusion-based methods further improve generation realism, diversity, and controllability, benefiting anomaly detection and segmentation Hu et al. (2024); Gui et al. (2024); Dai et al. (2025); Sun et al. (2025). However, most existing methods generate anomaly samples independently and do not explicitly model severity variations of the same anomaly pattern.

Industrial anomalies may exhibit varying degrees of severity. For example, a stain may gradually spread, and a dent may change from a slight deformation to an obvious flaw. In this work, we define visual severity as the relative strength of anomaly manifestations within the same anomaly type and image context, yielding an ordinal progression from subtle to pronounced defects. Modeling such severity progression provides controllable samples spanning subtle anomaly cues and pronounced anomaly patterns, helping detectors learn comprehensive anomaly representations. Moreover, generating all stages from a single normal image while preserving the background reduces variations in object appearance and context, enabling detectors to learn severity changes within the same anomaly category and potentially improving robustness and generalization.

In this work, we construct the SevADSet dataset and propose SevAnoDiff, a method for severity-aware anomaly generation. SevADSet treats each real anomaly image-mask pair as a reference stage and constructs a four-stage ordinal trajectory consisting of a normal stage and three increasingly se-

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.

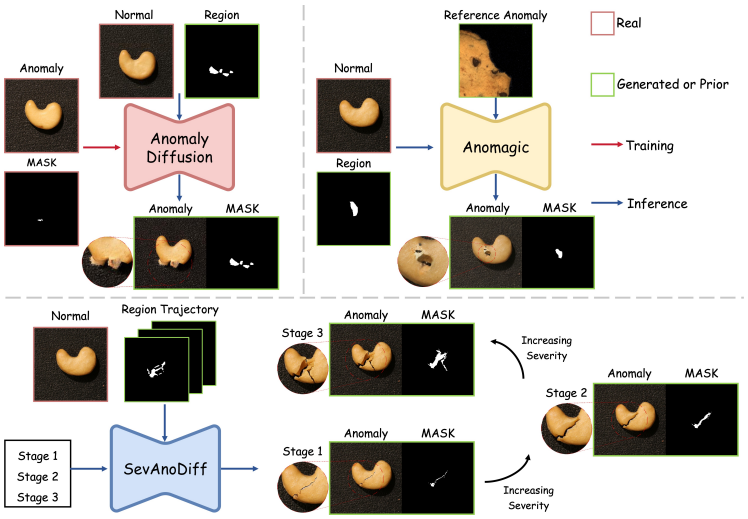

Figure 1: **Comparison of anomaly generation settings.** AnomalyDiffusion and Anomagic generate individual anomalies in few-shot and zero-shot settings, respectively. Our SevAnoDiff generates ordered anomaly stages from normal images with progressively increasing visual severity, without using target-domain anomaly samples.

vere anomaly stages. The real anomaly corresponds to the reference stage S^2, while S^1 and S^3 are constructed as visually attenuated and intensified variants, respectively. The stage labels denote relative positions within the same trajectory. Based on these trajectories, SevAnoDiff adapts a pretrained Stable Diffusion Inpainting model through two successive stages. The Appearance Prior Adapter first learns anomaly appearance from real anomalous regions. The Severity Progression Module then learns adjacent-stage transitions and leverages a Multi-Stage Mask Encoder to model dependencies across the severity trajectory. At inference, SevAnoDiff transfers the learned progression to normal images without target-domain anomaly samples. Experiments on MVTec-AD and VisA demonstrate competitive generation and downstream performance against existing anomaly generation methods, while ablations show that SevAnoDiff substantially improves ordinal severity consistency and downstream utility over the pretrained diffusion baseline. Our main contributions are as follows:

- We construct SevADSet, an industrial anomaly dataset containing 2,476 four-stage ordinal image-mask trajectories. Each trajectory comprises a normal image and three progressively more severe stages of the same anomaly category, allowing industrial anomaly generation to be studied as ordered severity progression rather than isolated sample synthesis.
- We propose SevAnoDiff, a two-stage diffusion adaptation framework that separates anomaly appearance learning from severity progression modeling, enabling ordered anomaly stages to be generated without target-domain anomaly samples.
- Extensive experiments on MVTec-AD and VisA demonstrate that SevAnoDiff achieves high-quality generation with ordered severity progression, and that its generated samples are effective for downstream anomaly detection and localization.

2 Related Work

2.1 Industrial Anomaly Detection

Industrial anomaly detection aims to recognize abnormal samples and localize defective regions. Since real anomaly samples are rare and expensive to annotate, whereas normal samples are much easier to collect, most methods were developed under unsupervised settings using only normal training images. Existing approaches mainly follow reconstruction, feature representation, and knowledge distillation paradigms. Reconstruction methods, such as DRAEM Zavrtanik et al. (2021) and

DeCo-Diff Beizaee et al. (2025), learned to restore normal visual patterns and detect anomalies from reconstruction discrepancies. Feature representation methods, such as PatchCore Roth et al. (2022), CFA Lee et al. (2022), and Dinomaly Guo et al. (2025), compared test features with normal feature distributions or memory banks. Knowledge distillation methods, such as RD4AD Deng & Li (2022) and EfficientAD Batzner et al. (2024), detected anomalies from teacher–student feature discrepancies. Despite their strong performance, these methods generally treat anomalies as static targets and rarely model severity explicitly. However, industrial anomalies are diverse in appearance, scale, and severity, making subtle early-stage defects and rare severe failures difficult to detect. Severity-aware modeling may provide richer supervision for learning fine-grained anomaly representations and improving detection robustness.

2.2 Anomaly Image Generation

Few-shot Anomaly Generation. Few-shot anomaly generation synthesizes realistic anomaly images from only a few defective examples, alleviating anomaly data scarcity in industrial inspection. DFMGAN Duan et al. (2023) introduced defect-aware feature manipulation into a StyleGAN2 backbone to generate realistic defects. More recent diffusion-based methods further improve generation fidelity and controllability. AnomalyDiffusion Hu et al. (2024) introduced spatial anomaly embedding and adaptive attention reweighting for mask-guided synthesis, while AnoGen Gui et al. (2024) learned anomaly distributions from a few anomaly samples to generate diverse defective images.

Zero-shot Anomaly Generation. Zero-shot anomaly generation aims to synthesize anomalies without using anomalous samples from that target domain. One line of work constructs pseudo anomalies from normal images through handcrafted perturbations or learned simulation, including CutPaste Li et al. (2021), DRAEM Zavrtanik et al. (2021), NSA Schlüter et al. (2022), and RealNet Zhang et al. (2024). More recent methods exploited diffusion models to improve the realism and diversity of generated anomalies. AnomalyAny Sun et al. (2025) leveraged Stable Diffusion priors for category-agnostic anomaly generation. Anomagic Jiang et al. (2026) introduced AnomVerse as an anomaly-mask-caption resource and combined visual and textual cues through cross-modal prompt encoding for anomaly inpainting. These methods have improved the realism, diversity, and image-mask consistency of generated anomalies. However, they do not explicitly model visual consistency and severity progression within the same anomaly trajectory. In contrast, our SevAnoDiff generates anomaly trajectories in a zero-shot setting through lightweight diffusion adaptation.

2.3 Image Editing and Conditional Adaptation

Diffusion models have been widely applied to image generation and editing. Latent Diffusion Models Rombach et al. (2022) perform generation in a compressed latent space and provide a practical foundation for text-guided inpainting. Blended Diffusion Avrahami et al. (2022) and InstructPix2Pix Brooks et al. (2023) showed that text conditions and editing regions can guide localized image modification. Recent models such as Qwen-Image-Edit Wu et al. (2025) and FLUX.1 Kontext Labs et al. (2025), further improved the generality and quality of instruction-based editing. ControlNet Zhang et al. (2023) and T2I-Adapter Mou et al. (2024) injected spatial conditions into frozen diffusion models, while Textual Inversion Gal et al. (2022), DreamBooth Ruiz et al. (2023), and LoRA Hu et al. (2021) enabled parameter-efficient customization. While these methods mainly target generic editing, spatial control, or subject personalization, our work focuses on localized industrial anomaly editing that preserves normal regions and follows an ordered severity progression.

3 Approach

Our approach consists of two parts. First, we construct SevADSet, a dataset of four-stage ordinal anomaly trajectories. Second, we propose SevAnoDiff, a two-stage diffusion adaptation framework for severity-aware anomaly generation.

3.1 SevADSet

We construct SevADSet by transforming real anomaly image-mask pairs into severity trajectories:

$$\mathcal{T}_{n,y} = \{(I_n^i, M_n^i, S^i)\}_{i=0}^{3}, \qquad S^0 < S^1 < S^2 < S^3,$$

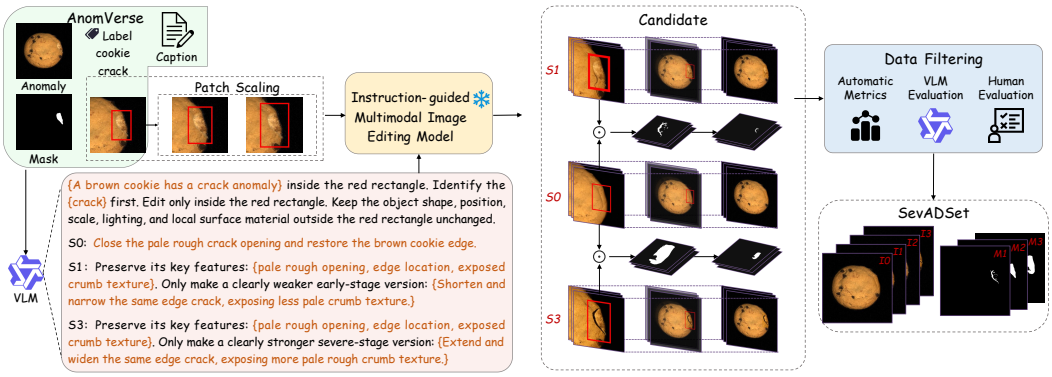

Figure 2: **Overview of SevADSet construction.** The pipeline combines visual and semantic cues with stage-specific editing and multiple filtering steps to construct ordinal image-mask trajectories for each anomaly category.

where n indexes the reference sample, y denotes its anomaly type, and I_n^i, M_n^i, and S^i denote the image, mask, and severity stage, respectively. Here, severity denotes the relative ordering of visual anomaly manifestations within the same anomaly type and trajectory. Each real anomaly serves as the reference stage S^2, while S^0, S^1, and S^3 denote its normal, weaker, and more severe variants.

Data Construction Pipeline. To build SevADSet, we develop a pipeline illustrated in Fig. 2. We use real anomaly images, masks, labels, and captions from AnomVerse Jiang et al. (2026) as references, and construct new four-stage severity trajectories. We crop a local patch around the anomaly region according to the mask and expand it when necessary to provide sufficient surrounding context. A vision language model extracts the anomaly identity, key visual attributes, and the changes required to produce normal, weaker, and more severe variants relative to the reference anomaly from the anomaly label, caption, local patch, and mask. These outputs are combined with a shortened AnomVerse caption and a fixed editing instruction to form the prompt for each target stage. An instruction-guided multimodal image editing model then generates multiple candidates for S^0, S^1, and S^3. We set $M^0 = \mathbf{0}$ and retain the original annotation as M^2. The masks M^1 and M^3 are extracted from differences with the generated normal image, followed by thresholding and morphological refinement. Candidates are first evaluated using a background preservation score. A vision language model then evaluates anomaly identity, image quality, mask alignment, and the ordinal relation $I^1 < I^2 < I^3$. Human evaluation further verifies the overall quality and ordinal consistency of the trajectories. After filtering, one severity trajectory with paired images and masks is retained for each valid reference sample. Specifically, we use Qwen3-VL Bai et al. (2025) for prompt construction and VLM evaluation, and Qwen-Image-Edit Wu et al. (2025) for stage generation.

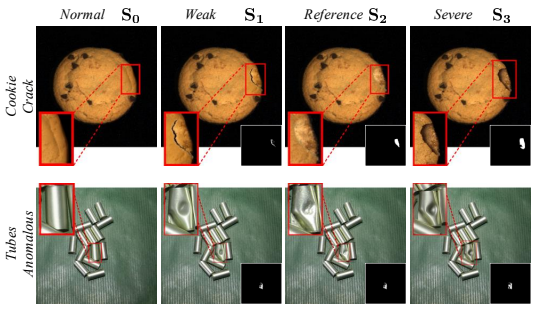

Figure 3: **Visualization of SevADSet with four stages.** The red boxes highlight the changing defective regions.

Dataset Statistics. We use real anomaly images, pixel-level masks, and captions from AnomVerse Jiang et al. (2026) to construct SevADSet. We focus on common anomaly types that show clear visual changes across severity levels, including scratches, cracks, cuts, holes, contamination, rust, color abnormalities, deformation, breakage, and other generic anomalies. For each selected reference sample, we generate five candidate four-stage trajectories. After automatic filtering and manual verification, SevADSet contains 2,476 four-stage trajectories and 9,904 image-mask pairs.

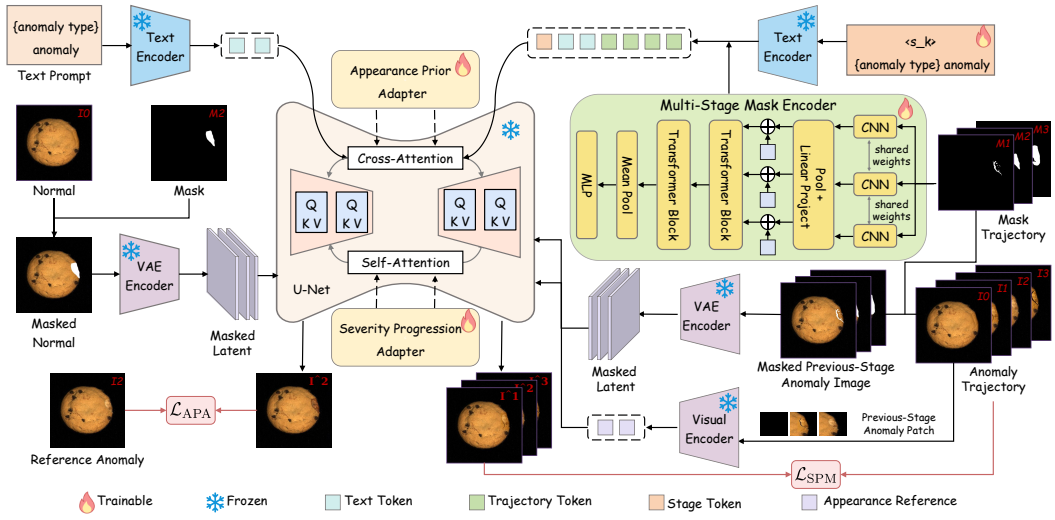

Figure 4: **Overview of SevAnoDiff.** In the first stage (left), the Appearance Prior Adapter (APA) is trained in the cross-attention layers to learn an anomaly appearance prior from a normal-to-reference transition. In the second stage (right), APA and the diffusion backbone are frozen, while the Severity Progression Adapter, stage tokens, and Multi-Stage Mask Encoder are trained to model adjacent-stage changes using the mask trajectory and the visual reference from the previous stage.

3.2 SevAnoDiff

We develop SevAnoDiff, a method for generating severity-controllable industrial anomalies from normal images. As shown in Fig. 4, it adapts a pretrained Stable Diffusion Inpainting model in two stages: The Appearance Prior Adapter (APA) first learns anomaly appearance from anomalous regions. The Severity Progression Module (SPM) then learns transitions between adjacent stages.

3.2.1 Appearance Prior Adapter.

Given a normal image I^0, the reference anomaly image I^2, its mask M^2, and an anomaly prompt p_a, APA learns anomaly appearance through the transition $I^0 \rightarrow I^2$. Implemented as LoRA in the U-Net cross-attention layers, APA is the only trainable module, while the VAE encoder $\mathcal{E}$, text encoder τ, and base denoising U-Net remain frozen. The source and target images are encoded as

$$z_0 = \mathcal{E}(I^0), \qquad z_2 = \mathcal{E}(I^2). \tag{1}$$

We add noise to z_2 at timestep t to obtain z_t and encode the masked source as $z_0^{\text{mask}} = \mathcal{E}\big(I^0 \odot (1 - M^2)\big)$. The U-Net input is $u_t = \big[z_t, M_\ell^2, z_0^{\text{mask}}\big]$, where M_ℓ^2 denotes M^2 resized to the latent resolution. The frozen text encoder maps the anomaly prompt to the conditioning feature $c_a = \tau(p_a)$. To avoid small anomalies being overwhelmed by background pixels, we design a *Region-Balanced Denoising Loss*:

$$\mathcal{L}_{\text{BD}} = \frac{\|M_\ell^2 \odot r_t\|_2^2}{2\sum M_\ell^2} + \frac{\|(1 - M_\ell^2) \odot r_t\|_2^2}{2\sum(1 - M_\ell^2)}, \tag{2}$$

where $r_t = \epsilon - \epsilon_{\text{APA}}(u_t, t, c_a)$. Let $\hat{z}_2$ denote the latent recovered from the predicted noise. The *Anomaly-Region Reconstruction Loss* reconstructs the target inside the editable region:

$$\mathcal{L}_{\text{AR}} = \frac{\big\|M_\ell^2 \odot (\hat{z}_2 - z_2)\big\|_1}{\sum M_\ell^2}. \tag{3}$$

The *Background Preservation Loss* preserves the normal-stage content outside the editable region:

$$\mathcal{L}_{\text{BP}} = \frac{\big\|(1 - M_\ell^2) \odot (\hat{z}_2 - z_0)\big\|_1}{\sum(1 - M_\ell^2)}. \tag{4}$$

The overall APA training objective is

$$\mathcal{L}_{\text{APA}} = \lambda_{\text{bd}}\mathcal{L}_{\text{BD}} + \lambda_{\text{ar}}\mathcal{L}_{\text{AR}} + \lambda_{\text{bp}}\mathcal{L}_{\text{BP}}, \tag{5}$$

where λ_{bd}, λ_{ar}, and λ_{bp} are parameters to balance the three loss terms.

3.2.2 SEVERITY PROGRESSION MODULE.

Given an ordered image trajectory $\{I^k\}_{k=0}^3$ with masks $\{M^k\}_{k=1}^3$, SPM learns the three transitions $I^{k-1} \rightarrow I^k$ for $k \in \{1,2,3\}$. A Severity Progression Adapter is inserted into the U-Net self-attention layers. It is trained together with the stage tokens and the Multi-Stage Mask Encoder. APA and the remaining modules stay frozen to preserve the appearance prior. For each transition, we define the editable region as

$$E_1 = M^1, \qquad E_k = M^{k-1} \cup M^k, \quad k \in \{2,3\}. \tag{6}$$

This region allows the model to modify both the existing anomaly and newly changed areas. Let $z_k = \mathcal{E}(I^k)$ for $k \in \{0,1,2,3\}$. After adding noise to z_k at timestep t, the masked source latent is $z_{k-1}^{\text{mask}} = \mathcal{E}\big(I^{k-1} \odot (1 - E_k)\big)$. The U-Net input is $u_t^k = \left[z_t^k, E_{k,\ell}, z_{k-1}^{\text{mask}}\right]$, where $E_{k,\ell}$ denotes E_k resized to the latent resolution. For each transition, a learnable stage token $\langle s_k \rangle$ specifies the relative target position. The resulting prompt is $p_k = \langle s_k \rangle$ y anomaly, and is encoded as $C_{\text{text}}^k = \tau(p_k)$. The text encoder remains frozen except for the three stage token embeddings.

Multi-Stage Mask Encoder. We develop a Multi-Stage Mask Encoder to map the ordered mask trajectory $\mathcal{M} = \{M^1, M^2, M^3\}$ to trajectory tokens $C_{\mathcal{M}}$. Each stage mask is encoded by a shared CNN and projected to the conditioning dimension. A learnable stage embedding is added to preserve the identity of each severity stage. Transformer blocks model relations among the three stages, followed by an MLP that projects the resulting features to the conditioning dimension. The transition mask E_k specifies the editable region, while $C_{\mathcal{M}}$ represents the spatial progression across all stages. The text and trajectory tokens form the cross-attention condition $C_k = [C_{\text{text}}^k; C_{\mathcal{M}}]$. To preserve anomaly identity across transitions, a frozen visual encoder extracts an appearance reference C_k^{ref} from the local anomaly region of the preceding stage for $k \in \{2,3\}$. This reference is not used for the normal to weak transition. The denoiser predicts $\hat{\epsilon}_k = \epsilon_{\text{SPM}}\left(u_t^k, t, C_k, C_k^{ref}\right)$, from which the clean target latent $\hat{z}_k$ is recovered.

Algorithm 1: Inference of SevAnoDiff

Input: Normal image I^0, anomaly type y, region bank $\mathcal{B}_y$
Output: $\{(\hat{I}^k, \hat{M}^k)\}_{k=1}^3$
$O \leftarrow \text{SAM}(I^0)$ *# valid object region*
$(R^1, R^2, R^3) \sim \mathcal{B}_y$ *# sample region trajectory*
$(M^1, M^2, M^3) \leftarrow \text{Place}(R^1, R^2, R^3, O)$ *# region placement*
$\mathcal{M} \leftarrow (M^1, M^2, M^3)$
$X^0 \leftarrow I^0$
for $k = 1$ to 3 **do**
  $E_k \leftarrow \begin{cases} M^1, & k = 1, \\ M^{k-1} \cup M^k, & k > 1 \end{cases}$ *# transition region*
  $p_k \leftarrow$ "$\langle s_k \rangle$ y anomaly" *# stage prompt*
  if $k = 1$ **then**
    $\hat{I}^k \leftarrow \mathcal{G}(X^{k-1}, E_k, p_k, \mathcal{M})$
  else
    $C_{\text{ref}}^k \leftarrow \text{Ref}(X^{k-1}, M^{k-1})$ *# visual reference*
    $\hat{I}^k \leftarrow \mathcal{G}(X^{k-1}, E_k, p_k, \mathcal{M}, C_{\text{ref}}^k)$
  end if
  $\hat{M}^k \leftarrow \text{Refine}(I^0, \hat{I}^k, M^{1:k}, O)$ *# anomaly mask*
  $X^k \leftarrow \hat{I}^k$
end for
return $\{(\hat{I}^k, \hat{M}^k)\}_{k=1}^3$

SPM applies the three APA objectives to each transition by replacing M_ℓ^2, z_2, and z_0 with $E_{k,\ell}$, z_k, and z_{k-1}, respectively. The overall SPM objective averages the transitions:

$$\mathcal{L}_{\text{SPM}} = \frac{1}{3}\sum_{k=1}^{3}\left(\lambda'_{\text{bd}}\mathcal{L}_{\text{BD}}^k + \lambda'_{\text{ar}}\mathcal{L}_{\text{AR}}^k + \lambda'_{\text{bp}}\mathcal{L}_{\text{BP}}^k\right). \tag{7}$$

3.2.3 INFERENCE.

Given a target-domain normal image I^0 and anomaly type y, we sample a three-stage region trajectory from the category-specific region bank. The region bank is constructed from SevADSet mask trajectories and contains only spatial shape trajectories. The sampled region trajectory provides the spatial progression prior, while SevAnoDiff models the stage specific appearance changes within these regions. SevAnoDiff then recursively generates $\{\hat{I}^k\}_{k=1}^3$, using each generated stage as the source for the next transition. The anomaly mask $\hat{M}^k$ is refined from the difference between $\hat{I}^k$ and I^0 within the dilated cumulative region restricted to the object area, followed by adaptive thresholding and morphological refinement.

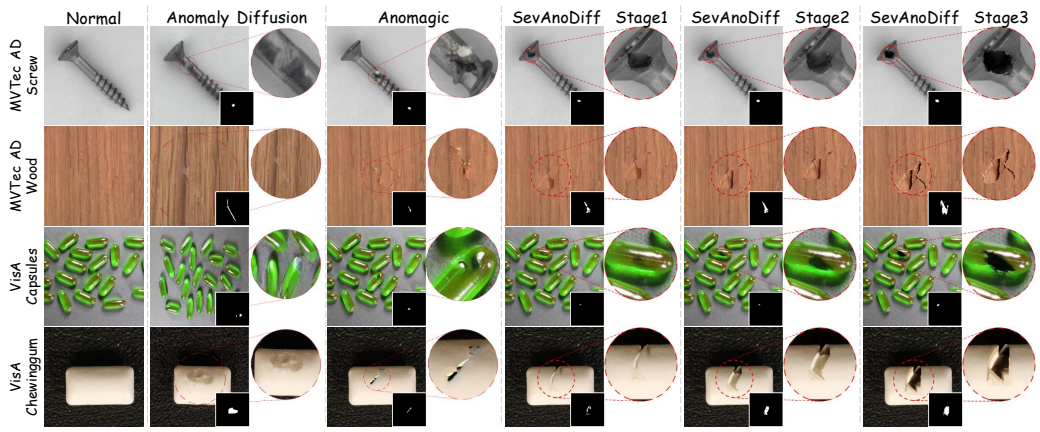

Figure 5: **Qualitative results of anomaly image generation.** Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages for hole, crack, color, and cut anomalies.

4 Experiments

4.1 Experiment Settings

Dataset and Evaluation Protocol. Following Anomagic, we use MVTec-AD Bergmann et al. (2019) and VisA Zou et al. (2022) only for evaluation and construct SevADSet from the other 11 datasets in AnomVerse. No anomalous images or masks from the target datasets are used for SevADSet construction or model training. At inference, SevAnoDiff generates anomalies on target domain normal images using five generic anomaly types, including hole, crack, color, scratch, and cut, together with source derived region trajectories.

Implementation Details. SevAnoDiff is built upon the pretrained Stable Diffusion v1.5 Rombach et al. (2022), with all images resized to 512×512. We use AdamW Loshchilov & Hutter (2019) for both training stages. In the first stage, APA is trained with a learning rate of 1×10^{-4}. In the second stage, APA is frozen, while the Severity Progression Adapter, stage tokens, and Multi-Stage Mask Encoder are optimized. The Severity Progression Adapter and Multi-Stage Mask Encoder use a learning rate of 4×10^{-5}, while the stage tokens use 1×10^{-4}. Transitions from the same trajectory share the diffusion timestep and noise realization. The loss weights are learned using uncertainty-based weighting Kendall et al. (2018). At inference, we generate 1000 three-stage trajectories per object category using DDIM Song et al. (2021) sampling. Details are provided in Appendix A.3.

Baselines. We compare with zero-shot anomaly synthesis baselines, including CutPaste Li et al. (2021), DRAEM Zavrtanik et al. (2021), NSA Schlüter et al. (2022), RealNet Zhang et al. (2024), AnomalyAny Sun et al. (2025), Anomagic Jiang et al. (2026) and AnoStyler So & Kang (2026). We also compare with few-shot anomaly synthesis baselines which use one-third of the target-domain anomalies for training, including DFMGAN Duan et al. (2023), AnomalyDiffusion Hu et al. (2024), and AnoGen Gui et al. (2024).

Metrics. Following Hu et al. (2024); Jiang et al. (2026), we report Inception Score (IS) Salimans et al. (2016) and intra-cluster pairwise LPIPS (IC-LPIPS) Ojha et al. (2021) for generation quality and diversity, respectively. For downstream anomaly detection and localization, we report AUROC, average precision (AP), and F1-score at the image and pixel levels. We further develop Severity Progression Accuracy (SPA), a VLM-based metric that measures whether the generated stages follow the intended ordinal severity order. For each trajectory, a frozen Gemini 2.5 Pro Comanici et al. (2025) ranks the three stages under three random permutations. A trajectory is counted as correct only when all three stage pairs follow the intended severity order after majority voting. Appendix A.4 provides the formal definitions of SPA and the complementary Pairwise Severity Accuracy.

Table 1: **Comparison of anomaly generation on MVTec-AD and VisA.** We report IS, IC-LPIPS and SPA. The best and second-best results are shown in **bold** and underlined. SPA is not reported for baselines because they generate individual anomalies rather than trajectories.

Dataset	Metric	Few-Shot			Zero-Shot							
		DFMGAN	AnomalyDiffusion	AnoGen	CutPaste	DRAEM	NSA	RealNet	AnomalyAny	AnoStyler	Anomagic	Ours
MVTec-AD	IS	1.72	1.80	1.77	1.76	1.76	1.44	1.64	2.02	2.04	1.84	**2.07**
	IC-LPIPS	0.20	0.32	0.27	0.22	0.25	0.26	0.22	**0.33**	0.32	0.28	0.27
	SPA	–	–	–	–	–	–	–	–	–	–	**89.11%**
VisA	IS	1.48	1.50	1.40	1.52	1.50	1.42	1.53	1.41	1.55	1.44	**1.58**
	IC-LPIPS	0.28	0.29	0.22	0.26	0.25	0.19	0.29	0.19	0.32	0.32	**0.33**
	SPA	–	–	–	–	–	–	–	–	–	–	**88.63%**

Table 2: **Comparison of anomaly detection on MVTec-AD and VisA.** We report AUROC, AP, and F_1 at the image level (I-AUC, I-AP, and I-F_1) and pixel level (P-AUC, P-AP, and P-F_1). Within each setting, the best and second-best results are shown in **bold** and underlined, respectively.

Method	MVTec-AD						VisA					
	I-AUC	I-AP	I-F_1	P-AUC	P-AP	P-F_1	I-AUC	I-AP	I-F_1	P-AUC	P-AP	P-F_1
Few-Shot Anomaly Generation												
DFMGAN	87.2	94.8	94.7	90.0	62.7	62.1	83.7	85.7	80.3	90.6	31.0	34.6
AnomalyDiffusion	**99.2**	**99.7**	**98.7**	**99.1**	**81.4**	**76.3**	86.9	89.1	85.4	**93.2**	**33.0**	**36.8**
AnoGen	98.7	99.6	97.7	96.9	73.6	66.7	**90.4**	**92.4**	**87.2**	89.1	31.6	33.4
Zero-Shot Anomaly Generation												
CutPaste	89.8	92.1	89.8	88.2	51.9	50.7	86.3	86.9	87.1	88.4	32.2	39.6
DRAEM	94.6	97.0	94.4	92.2	54.1	53.1	91.8	92.9	88.6	91.4	29.5	37.2
NSA	93.0	95.6	91.6	92.0	52.6	52.5	87.3	89.8	84.2	92.6	26.1	34.2
RealNet	95.2	97.0	95.3	94.0	57.7	56.6	92.6	93.8	89.2	92.2	33.3	**41.0**
AnomalyAny	95.2	96.9	96.3	89.0	62.7	59.9	88.9	86.2	85.9	90.4	31.2	33.0
Anomagic	96.9	98.5	95.1	92.8	54.9	54.9	90.3	90.4	83.5	92.5	32.5	37.6
AnoStyler	98.0	**99.0**	**97.0**	94.4	62.9	60.7	93.9	**95.3**	**90.1**	**93.8**	31.4	36.4
Ours	**98.3**	98.9	96.9	**95.1**	**63.7**	**64.0**	**94.3**	93.6	87.8	93.7	**33.9**	40.1

4.2 Results on Anomaly Generation

As shown in Tab. 1, SevAnoDiff achieves the highest IS among the compared methods on both MVTec-AD and VisA. The moderate IC-LPIPS is consistent with our constrained generation design: ordered mask trajectories restrict anomaly changes to plausible object regions, while visual information inherited from the previous stage maintains anomaly continuity, which together may reduce the full-image perceptual variation measured by IC-LPIPS. Moreover, the high SPA demonstrates that SevAnoDiff produces clear and consistent severity progression across the generated stages. As shown in Fig. 5, SevAnoDiff generates realistic anomaly images with aligned masks and ordered progression across three stages. The screw defect develops into a darker cavity with more irregular boundaries; the wood crack extends and widens with increased branching; the capsule color anomaly shows stronger chromatic deviation within a largely aligned region; and the cut develops into a wider opening with more evident local deformation. These examples show that severity progression involves changes in spatial extent, appearance contrast, structural complexity, and deformation rather than mask area expansion alone, while preserving the object structure and background.

4.3 Results on Anomaly Detection

Similar to Hu et al. (2024); So & Kang (2026), we use 500 generated anomaly image-mask pairs for each target category and train a U-Net Ronneberger et al. (2015) for anomaly detection and localization. As shown in Tab. 2, SevAnoDiff achieves the best I-AUC and all three pixel-level metrics on MVTec-AD in the zero-shot setting. On VisA, it achieves the best I-AUC and P-AP. It also outperforms all few-shot baselines across all metrics on VisA without using target-domain anomaly samples. These results show that the generated anomaly image-mask pairs provide effective training data for downstream anomaly detection and localization.

Table 3: **Ablation study of appearance and severity progression modeling.**

APA	SPM	MVTec-AD					VisA				
		SPA	I-AUC	I-AP	P-AUC	P-AP	SPA	I-AUC	I-AP	P-AUC	P-AP
		6.0%	88.7	93.1	87.0	37.6	11.67%	82.8	80.6	83.3	18.4
✓		38.67%	94.9	97.3	91.8	52.5	36.67%	89.9	88.4	89.5	21.9
✓	✓	**89.11%**	**98.3**	**98.9**	**95.1**	**63.7**	**88.63%**	**94.3**	**93.6**	**93.7**	**33.9**

Table 4: **Ablation study of the severity progression conditions.**

Method	MVTec-AD					VisA				
	SPA	I-AUC	I-AP	P-AUC	P-AP	SPA	I-AUC	I-AP	P-AUC	P-AP
w/o Stage Token	76.00%	94.1	97.0	88.8	46.1	70.00%	86.8	85.1	85.7	15.2
w/o Mask Encoder	66.67%	92.8	96.0	89.4	44.8	51.11%	87.8	86.0	86.7	18.6
w/o Visual Reference	70.00%	92.8	95.9	89.1	46.0	64.29%	89.0	87.9	87.3	17.5
Full	**89.11%**	**98.3**	**98.9**	**95.1**	**63.7**	**88.63%**	**94.3**	**93.6**	**93.7**	**33.9**

4.4 Ablation Studies

Appearance and Severity Progression Modeling. We progressively add APA and SPM to the frozen Stable Diffusion Inpainting backbone. APA introduces the Appearance Prior Adapter, while the full model further adds SPM with stage tokens, mask-trajectory conditioning, and visual references. The two baselines receive explicit textual severity prompts and use the same target domain normal images, three stage masks, recursive inference strategy, and parameters as the full model. Tab. 3 shows the effectiveness of APA and SPM. Additionally, both baselines achieve substantially lower SPA than the full model, despite using the same prescribed spatial trajectories. This shows that the spatial trajectories alone are insufficient to produce consistent severity ordering, while explicit progression modeling substantially improves trajectory consistency and downstream utility.

Severity Progression Conditions. We further ablate three progression conditions in SPM: the stage token, Multi-Stage Mask Encoder, and visual reference. As shown in Tab. 4 and Fig. 6, removing any condition degrades both SPA and downstream performance. Since the editable regions and previous-stage image inputs provide basic progression cues, removing any condition causes degradation but does not eliminate severity progression. Without the stage token, adjacent-stage differences become less distinct; removing the Mask Encoder causes the largest SPA drop, with less coherent spatial progression; without the visual reference, anomaly appearance becomes less consistent across stages. These results support their complementary roles in severity progression modeling.

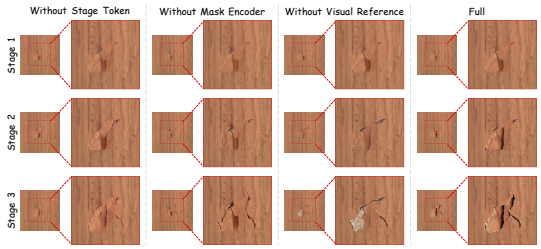

Figure 6: **Qualitative ablation of the severity progression conditions.** The full model shows clearer stage differences, spatial progression, and cross-stage appearance consistency.

5 Conclusion

We have developed a novel approach to industrial anomaly detection by studying ordered anomaly progression across visual severity levels. For this new study, we have constructed SevADSet, an industrial anomaly dataset with ordered severity trajectories. Then we have developed SevAnoDiff, a zero-shot severity-aware anomaly generation method. Our SevAnoDiff learns transferable anomaly appearance priors and models adjacent-stage changes using stage tokens, mask trajectories, and previous-stage visual references. Without using target-domain anomalous images or masks, SevAnoDiff generates realistic anomaly image–mask pairs with clear severity progression. Experiments on MVTec-AD and VisA have demonstrated the effectiveness of our approach in anomaly generation and downstream detection and localization. These results also indicate the potential of severity-aware generation for fine-grained industrial inspection.

AI USE STATEMENT

In this work, we used generative AI tools for synthetic dataset construction, dataset filtering, and AI-assisted evaluation of generated samples. Specifically, Qwen3-VL was used to extract anomaly attributes, construct stage-specific editing instructions, and evaluate candidate trajectories during the construction of SevADSet. Qwen-Image-Edit was used to generate candidate normal, weaker, and more severe anomaly stages. Gemini 2.5 Pro was used as the VLM-based evaluator for the Severity Progression Accuracy (SPA) metric. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results. Tasks involving formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to edit the research paper to improve readability. We have reviewed all AI-assisted work. Specifically, the AI-generated samples used in the construction of SevADSet were subject to the automatic filtering and human verification procedures described in the paper, and all AI-assisted edits to the manuscript were reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18208–18218, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 128–138, 2024.

Farzad Beizaee, Gregory A Lodygensky, Christian Desrosiers, and Jose Dolz. Correcting deviations from normality: A reformulated diffusion model for multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19088–19097, 2025.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021.

Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Proceedings of the Asian Conference on Computer Vision*, pp. 3586–3602, 2022.

Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9737–9746, 2022.

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

Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagnucco, and Yang Song. Manta: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25518–25527, 2025.

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

Guan Gui, Bin-Bin Gao, Jun Liu, Chengjie Wang, and Yunsheng Wu. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *European Conference on Computer Vision*, pp. 210–226. Springer, 2024.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20405–20415, 2025.

Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. *International Journal of Computer Vision*, 134(4):175, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36(1):85–96, 2020.

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.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pp. 66–71. IEEE, 2021.

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

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *arXiv preprint arXiv:2206.04325*, 2022.

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.

Wenqiao Li, Bozhong Zheng, Xiaohao Xu, Jinye Gan, Fading Lu, Xiang Li, Na Ni, Zheng Tian, Xiaonan Huang, Shenghua Gao, et al. Multi-sensor object anomaly detection: Unifying appearance, geometry, and internal properties. In *Proceedings of the computer vision and pattern recognition conference*, pp. 9984–9993, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. *arXiv preprint arXiv:2104.10036*, 2021.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

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 Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 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.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

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

Javier Lidiano Silvestre Blanes, Teresa Albero Albero, Ignacio Miralles, Rubén Pérez Llorens, and Jorge Moreno. A public fabric database for defect detection methods and results. 2019.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

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

Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, volume 6, pp. 11, 2007.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

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

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

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, volume 2024, pp. 49705–49737, 2024.

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

A APPENDIX

A.1 OVERVIEW

This supplementary material consists of:

- Details of SevADSet construction and dataset statistics (Sec. A.2).
- More implementation details of SevAnoDiff (Sec. A.3).
- More details of Severity Progression Accuracy (SPA) (Sec. A.4).
- More results on anomaly generation (Sec. A.5).
- More results on downstream anomaly detection (Sec. A.6).
- More ablation studies (Sec. A.7).
- Further discussion and future directions (Sec. A.8).

A.2 DETAILS OF SEVADSET CONSTRUCTION

A.2.1 FOUR-STAGE TRAJECTORY CONSTRUCTION

Given a real anomaly image I^2 with mask M^2, category label, and anomaly type, we use the observed anomaly as the reference stage of a four-stage trajectory:

$$I^0 \rightarrow I^1 \rightarrow I^2 \rightarrow I^3.$$

The local editing region is obtained from the anomaly mask or its bounding box and is used to constrain both image editing and VLM evaluation. Only this region is modified, while the object structure and background outside it are preserved.

The normal stage I^0 is obtained by removing the anomaly, I^1 retains the same anomaly type with weaker visual manifestation, and I^3 retains the same anomaly type with stronger visual manifestation. The reference stage I^2 is the original real anomaly and is not generated.

A.2.2 VLM-GUIDED CONSTRUCTION

The anomaly type is determined by the source label and a canonical defect mapping rather than inferred freely by the VLM. Given the source image, the annotated anomaly region, category and defect labels, and the corresponding caption, the VLM identifies the anchor appearance and its key visual features. It then produces stage specific edit instructions for I^0, I^1, and I^3. Each editing prompt combines a fixed template with the VLM generated instruction. The fixed template restricts editing to the marked region, preserves the content outside it, and requires the anomaly identity to remain unchanged. The VLM instruction specifies how the observed anomaly should change at each stage. For example, for a crack, it may request closing the crack for I^0, shortening and narrowing it for I^1, and extending and widening it for I^3. The original anomaly image is retained as the reference stage I^2.

A.2.3 QUALITY CONTROL

Prompt Constraints. Before generation, a fixed prompt template restricts editing to the marked region and preserves the object and background outside it. The generated stages must retain the same anomaly type, local position, and key visual features, with only the anomaly severity changed. Candidates are ranked according to a background preservation score. This step filters out candidates with poor background preservation and retains the top two candidates for each generated stage. The VLM jointly evaluates the trajectory and selects one candidate for each generated stage. It checks whether the normal stage removes the anomaly, whether the weak and severe stages preserve the anomaly identity and remain weaker and stronger than the real reference, respectively, and whether all stages retain the same object and local anomaly region. Candidates with box remnants, editing artifacts, category changes, or unrelated anomalies are rejected. The selected stages may come from different candidate indices. After VLM selection, two reviewers with experience in industrial anomaly analysis independently examine the retained trajectories. They check anomaly identity, severity ordering, object and background preservation, editing artifacts, and image mask consistency.

Postprocessing and Mask Validation. After selection, the annotation box is removed and the edited patch is mapped back to the full image. The normal stage uses an empty mask, and the real reference stage keeps its original source mask. The masks for the weak and severe stages are refined from their differences relative to the restored normal image, together with the source mask prior, editing region, and morphological processing. The marked region is used only to control editing and is not treated as the final anomaly mask.

A.2.4 Dataset Statistics

We use anomaly images, pixel level masks, labels, and captions from AnomVerse Jiang et al. (2026) as references for constructing SevADSet. For each selected reference sample, we construct a new four stage trajectory through our editing, filtering, and verification pipeline. AnomVerse contains 12,987 anomaly images collected from 13 public datasets. MVTec AD Bergmann et al. (2019) and VisA Zou et al. (2022) are included among these source datasets. We exclude all samples originating from MVTec AD and VisA during SevADSet construction, region bank preparation, and model training. SevADSet is constructed from selected samples in the remaining 11 source datasets. No target-domain anomalous images and masks are used during construction, region-bank preparation, training, or inference. Table 5 summarizes the source datasets in AnomVerse and the samples retained for constructing SevADSet. The original AnomVerse captions are used together with the anomaly labels, local anomaly patches, and masks. Qwen3-VL processes this information to construct stage-specific editing instructions for the normal, weaker, and more severe variants.

Table 5: **Source data and generated trajectories in SevADSet.** The numbers of source images are taken from AnomVerse Jiang et al. (2026). The last column reports the four-stage trajectories retained after construction and filtering.

Dataset	Source Images	Trajectories
AITEX Silvestre Blanes et al. (2019)	103	26
BTAD Mishra et al. (2021)	290	41
Eyecandies Bonfiglioli et al. (2022)	240	175
MPDD Jezek et al. (2021)	282	233
MTD Huang et al. (2020)	2,676	231
MVTec 3D-AD Bergmann et al. (2021)	948	615
KolektorSDD2 Božič et al. (2021)	894	142
MulSen AD Li et al. (2025)	360	127
MVTec AD-2 Heckler-Kram et al. (2026)	705	376
DAGM Wieler & Hahn (2007)	1,054	183
MANTA Fan et al. (2025)	2,977	327
Total	10,529	2,476

A.3 Implementation Details

A.3.1 Training Details

SevAnoDiff is built upon Stable Diffusion v1.5 Inpainting. The main experiments are conducted on an NVIDIA 5090 32GB GPU. All images and masks are resized to 512×512. Both training stages use AdamW with a constant learning rate, a weight decay of 0.01, and gradient clipping at 1.0. The training step requires only 40 epochs.

In the first stage, APA is optimized with a learning rate of 1×10^{-4}, a batch size of 4, and two gradient accumulation steps. In the second stage, the diffusion backbone and APA are frozen. The Severity Progression Adapter and Multi-Stage Mask Encoder are optimized with a learning rate of 4×10^{-5}, while the stage token embeddings use 1×10^{-4}. The batch size is 3 with two gradient accumulation steps. Both adapters use a LoRA rank and scaling parameter of 64. APA is applied to the cross-attention layers, while the Severity Progression Adapter is applied to the self-attention layers. The LoRA modules adapt the query, key, value, and output projections of the corresponding attention layers. Both stages are optimized using the region-balanced denoising, anomaly-region reconstruction, and background preservation losses.

Both training stages use the Region-Balanced Denoising Loss, Anomaly-Region Reconstruction Loss, and Background Preservation Loss defined in the main paper. The three terms are combined using uncertainty-based weighting:

$$\mathcal{L} = \sum_i \exp(-s_i)\mathcal{L}_i + s_i,$$

where s_i is a learnable log-variance parameter initialized to zero.

APA and SPM maintain separate uncertainty parameters. For SPM, the three adjacent-stage transitions share the same set of loss-weighting parameters. The Multi-Stage Mask Encoder receives masks resized to 64×64, uses a hidden dimension of 256, and contains two Transformer layers with four attention heads. It produces four trajectory tokens. The previous-stage visual reference is implemented using a frozen IP-Adapter.

A.3.2 Inference Details

We use DDIM sampling with 20 denoising steps, a guidance scale of 4.0 and a denoising strength of 0.85. For each transition k, the target prompt is "$\langle s_k \rangle$ {anomaly type} anomaly", while no source-side text prompt is used. The editable region is dilated by 2 pixels and smoothed with a Gaussian filter of $\sigma = 1.5$. For $k \in \{2, 3\}$, the anomaly region from the previous stage is used as the visual reference with a conditioning scale of 0.5. Given a normal image, SevAnoDiff recursively generates a three-stage anomaly trajectory $\{\hat{I}^1, \hat{I}^2, \hat{I}^3\}$. The anomaly type remains fixed across stages, while the sampled region trajectory controls the spatial progression. For target domain generation, we use five generic anomaly types: hole, crack, color, scratch, and cut. These anomaly types are fixed before evaluation and do not rely on target anomaly annotations.

A.3.3 Mask Trajectory Generation and Refinement

Mask Trajectory Generation. For each anomaly type y, we construct an anomaly-type-specific region bank $\mathcal{B}_y$ from the three-stage mask trajectories in SevADSet. The region bank stores only spatial shape trajectories and does not contain source-image appearance or target-domain anomaly information. At inference, a trajectory $\mathcal{R} = \{R^1, R^2, R^3\}$ is sampled from $\mathcal{B}_y$. The same random geometric transformation T, including scaling, rotation, flipping, and translation, is applied to all three stages: $M^k = T(R^k), \quad k \in \{1, 2, 3\}$. For transition k, the editable region follows the definition:

$$E_1 = M^1, \qquad E_k = M^{k-1} \cup M^k, \quad k \in \{2, 3\}.$$

Before inpainting, E_k is dilated and Gaussian-smoothed to obtain the soft editable mask $\widetilde{E}_k$ used by the inpainting model.

Mask Refinement. The editable region determines where the diffusion model can modify the input image, but it does not necessarily match the visible anomaly. We therefore refine a pixel-level mask after generating each stage. Given the normal image I^0, the generated image $\hat{I}^k$, and the editable region $\widetilde{E}_k$, we compute the pixel-wise color difference between I^0 and $\hat{I}^k$ in the Lab color space. The refinement is restricted to the local support of $\widetilde{E}_k$ to exclude unrelated image changes. We then apply adaptive dual-threshold filtering. Pixels above the high threshold form seed regions, while pixels above the low threshold are retained only when connected to a seed region. The resulting mask $\hat{M}^k$ is paired with $\hat{I}^k$ as its pixel-level annotation. Figure 7 shows the difference between the editable region and the refined visible mask for a crack trajectory from S1 to S3.

A.3.4 Details of Multi-Stage Mask Encoder

The Multi-Stage Mask Encoder takes the complete mask trajectory $\mathcal{M} = \{M^1, M^2, M^3\}$ as input. Each mask is resized to 64×64 and processed by a shared CNN encoder consisting of four convolutional layers with Group Normalization and SiLU activation. The resulting mask features are augmented with learnable stage-index embeddings and processed jointly by a two-layer Transformer encoder. The Transformer outputs are mean-pooled into a global trajectory representation and projected into a set of mask trajectory condition tokens:

$$C_{\mathcal{M}} \in \mathbb{R}^{N_{\mathcal{M}} \times d_c},$$

where $N_{\mathcal{M}} = 4$ is the number of condition tokens and $d_c = 768$ is the cross-attention dimension. The tokens are appended to the text tokens and supplied to the cross-attention layers. The condition tokens are parallel global representations projected from the complete mask trajectory and do not correspond individually to the three severity stages. Transition-specific information is provided separately by the learned stage token and the editable region E_k, while $C_{\mathcal{M}}$ encodes the global spatial progression of the complete mask trajectory.

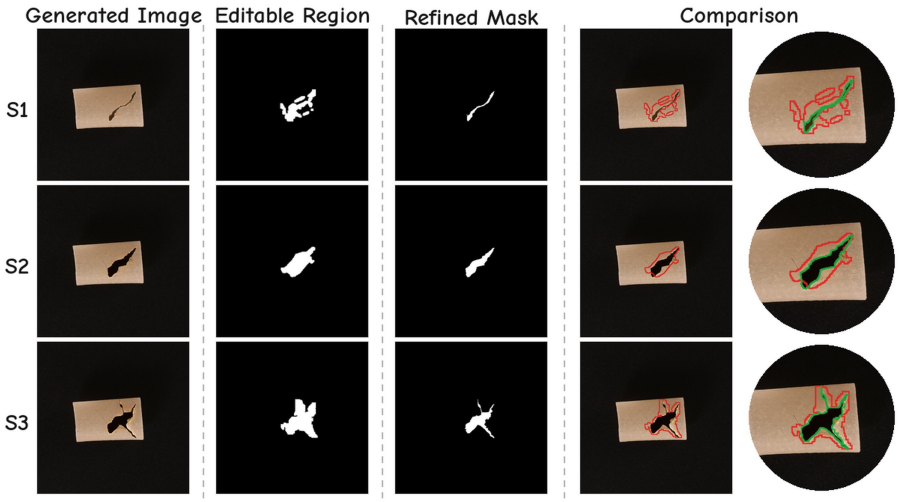

Figure 7: **Qualitative comparison between editable regions and refined anomaly masks across severity stages.** Red and green contours denote the editable region and refined mask, respectively. The editable region provides a coarse spatial constraint for generation, while the refined mask aligns the anomaly that appears in the generated image.

A.4 Details of Severity Progression Accuracy

Severity Progression Accuracy (SPA) measures whether the generated anomaly trajectory follows the intended severity order. For each trajectory, the three images are independently permuted and presented to the VLM in three evaluation rounds. The images are labeled only as A, B, and C, without revealing their stage identities. In each presentation, the VLM ranks the images from the least to the most severe anomaly.

Each ranking provides one vote for the three stage pairs (S_1, S_2), (S_1, S_3), and (S_2, S_3). A pair is judged correct when at least two of the three rankings agree with its intended order. For trajectory n, let $s_{n,ij} = 1$ when the pair (S_i, S_j) passes majority voting, and $s_{n,ij} = 0$ otherwise. For N generated trajectories, SPA is computed as

$$\mathrm{SPA} = \frac{100}{N} \sum_{n=1}^{N} \mathbb{I}\left[s_{n,12} s_{n,13} s_{n,23} = 1\right].$$

A trajectory contributes to SPA only when all three stage pairs pass majority voting.

Pairwise Severity Accuracy. We also report Pairwise Severity Accuracy (PSA), which measures the proportion of correctly ordered stage pairs:

$$\mathrm{PSA} = \frac{100}{3N} \sum_{n=1}^{N} \left(s_{n,12} + s_{n,13} + s_{n,23}\right).$$

Unlike SPA, PSA gives partial credit when only some stage pairs follow the intended severity order.

A.5 Additional Results on Anomaly Generation

A.5.1 Quantitative Results of Per-Category Anomaly Generation

Tables 6 and Tables 7 report the per-category anomaly generation results on MVTec-AD and VisA, respectively. IS and IC-LPIPS evaluate the quality and diversity of the generated anomaly images, while PSA and SPA evaluate the ordinal consistency of the complete three-stage trajectories.

Table 6: **Per-category anomaly generation performance of SevAnoDiff on MVTec AD.**

Category	IS↑	IC-L↑	PSA↑	SPA↑
bottle	1.99	0.11	98.61%	95.83%
cable	2.00	0.38	90.48%	77.14%
capsule	1.56	0.16	93.21%	85.19%
carpet	1.52	0.23	96.21%	88.64%
grid	2.40	0.37	90.74%	86.11%
hazelnut	2.12	0.30	98.96%	96.88%
leather	2.94	0.31	94.44%	86.11%
metal nut	1.48	0.23	95.50%	86.49%
pill	2.23	0.24	95.83%	87.50%
screw	1.16	0.31	95.35%	88.37%
tile	2.91	0.43	97.98%	93.94%
toothbrush	1.60	0.15	97.73%	93.18%
transistor	1.51	0.29	96.75%	90.24%
wood	3.60	0.34	96.19%	88.57%
zipper	1.99	0.22	97.50%	92.50%
Average	2.07	0.27	95.70%	89.11%

Table 7: **Per-category anomaly generation performance of SevAnoDiff on VisA.**

Category	IS↑	IC-L↑	PSA↑	SPA↑
candle	1.40	0.18	97.56%	92.68%
capsules	1.57	0.59	91.87%	80.49%
cashew	1.78	0.38	95.70%	90.32%
chewinggum	1.98	0.36	91.92%	87.88%
fryum	1.32	0.28	96.21%	90.91%
macaroni1	1.26	0.28	95.35%	86.05%
macaroni2	2.10	0.45	95.50%	86.49%
pcb1	1.26	0.34	94.29%	85.71%
pcb2	1.22	0.32	95.24%	91.43%
pcb3	1.45	0.26	96.67%	93.33%
pcb4	1.29	0.30	94.12%	88.24%
pipe fryum	2.29	0.25	95.56%	90.00%
Average	1.58	0.33	95.00%	88.63%

A.5.2 Qualitative Results of Per-Category Anomaly Generation

We provide per-category qualitative results on MVTec AD (Fig. 8–Fig. 12) and VisA (Fig. 13–Fig. 16). For each category, we visualize the generated severity trajectory from the normal stage to progressively stronger anomaly manifestations. The results show that SevAnoDiff preserves the object appearance and local anomaly identity while producing visual severity changes.

A.6 Additional Results on Downstream Anomaly Detection

Following the downstream evaluation protocol of AnomalyDiffusion Hu et al. (2024) and anostyler So & Kang (2026), we train an object-specific discriminative U-Net for anomaly detection. The detector takes a 256×256 image as input. For each target object category, real normal images from the target training set are used as normal samples, while 500 generated image-mask pairs are used as anomalous samples and are only used for detector training. The generated samples are selected approximately uniformly from the three severity stages, with 167 samples from Stage 1, 166 samples from Stage 2, and 167 samples from Stage 3. The detector is optimized using pixel-wise focal loss. We use Adam with a learning rate of 1×10^{-4} and a batch size of 8 for 200 epochs. All reported downstream results are computed on real test images from MVTec AD and VisA. We report AUROC, average precision, and F_1 at both image and pixel levels.

A.6.1 Per-Category Downstream Results

Table 8 and Table 9 report the per-category downstream performance on MVTec AD and VisA. SevAnoDiff achieves consistently strong image-level detection across most categories.

A.7 Additional Ablation Studies

A.7.1 Appearance and Severity Progression Modeling

In addition to SPA and anomaly inspection performance described in Sec. 4.4, we report Pairwise Severity Accuracy (PSA) to measure the correctness of individual stage pairs. As shown in Table 10, the Base model achieves 21.78% PSA and 6.00% SPA on MVTec-AD, and 32.50% PSA and 11.67% SPA on VisA, indicating limited ordinal consistency from severity prompts and progressively changing masks alone. Adding APA increases PSA to 64.00% and 64.44%, and SPA to 38.67% and 36.67% on MVTec-AD and VisA, respectively. It also provides I-AUC and P-AP gains of 6.2% and 14.9% on MVTec-AD, and 7.1% and 3.5% on VisA. Adding SPM further increases PSA to 95.70% on MVTec-AD and 95.00% on VisA, while SPA reaches 89.11% and 88.63%. The

Table 8: **Per-category anomaly detection performance of SevAnoDiff on MVTec AD.**

Category	I-AUROC	I-AP	I-F_1	P-AUROC	P-AP	P-F_1
bottle	100.0	100.0	100.0	96.0	70.7	70.2
cable	94.2	95.9	91.5	85.9	62.1	64.8
capsule	96.8	99.1	96.1	94.2	30.1	38.5
carpet	96.5	98.7	95.9	97.7	79.3	74.4
grid	99.4	99.7	98.8	98.4	54.6	54.6
hazelnut	99.9	99.9	99.0	99.4	80.4	78.2
leather	100.0	100.0	100.0	99.7	71.8	70.7
metal nut	98.2	99.5	97.6	96.4	74.4	73.2
pill	97.8	99.4	96.8	98.9	81.3	76.0
screw	94.2	96.4	92.6	95.0	22.0	32.0
tile	100.0	100.0	100.0	98.1	77.9	79.7
toothbrush	100.0	100.0	100.0	99.3	71.5	70.9
transistor	97.5	95.2	87.7	75.0	35.0	42.1
wood	100.0	100.0	100.0	95.7	72.3	66.4
zipper	99.6	99.8	97.6	97.1	71.9	68.3
Average	98.3	98.9	96.9	95.1	63.7	64.0

Table 9: **Per-category anomaly detection performance of SevAnoDiff on VisA.**

Category	I-AUROC	I-AP	I-F_1	P-AUROC	P-AP	P-F_1
candle	94.4	94.0	87.7	90.0	14.0	26.5
capsules	93.4	94.7	87.5	94.4	33.2	41.5
cashew	94.5	96.7	90.2	96.0	62.0	60.9
chewinggum	98.4	99.1	96.3	98.8	79.2	74.2
fryum	97.0	97.9	93.9	89.2	28.5	34.4
macaroni1	95.5	92.6	87.6	99.2	24.7	33.3
macaroni2	88.9	88.2	80.0	98.3	16.6	22.9
pcb1	89.6	85.0	80.5	94.5	56.6	58.3
pcb2	95.2	93.5	85.3	87.9	12.8	24.3
pcb3	90.4	89.6	80.0	82.5	8.2	18.5
pcb4	97.5	94.6	92.4	96.4	26.1	35.3
pipe fryum	96.7	97.5	92.4	96.8	44.3	50.7
Average	94.3	93.6	87.8	93.7	33.9	40.1

Table 10: **Extended ablation of appearance adaptation and severity progression modeling. In addition to the metrics reported in the main paper, we report PSA to evaluate pairwise severity.**

Method	MVTec-AD						VisA					
	PSA	SPA	I-AUC	I-AP	P-AUC	P-AP	PSA	SPA	I-AUC	I-AP	P-AUC	P-AP
Base	21.78%	6.00%	88.7	93.1	87.0	37.6	32.50%	11.67%	82.8	80.6	83.3	18.4
APA	64.00%	38.67%	94.9	97.3	91.8	52.5	64.44%	36.67%	89.9	88.4	89.5	21.9
APA+SPM (Full)	**95.70%**	**89.11%**	**98.3**	**98.9**	**95.1**	**63.7**	**95.00%**	**88.63%**	**94.3**	**93.6**	**93.7**	**33.9**

substantially larger gains in SPA than PSA indicate that SPM particularly improves the consistency of the complete three-stage ordering, beyond correcting individual pairwise relations. SPM also improves I-AUC and P-AP by 3.4% and 11.2% on MVTec-AD, and by 4.4% and 12.0% on VisA.

Table 11: **Extended ablation of severity progression conditions. In addition to the metrics reported in the main paper, we report PSA to evaluate pairwise severity.**

Method	MVTec-AD						VisA					
	PSA	SPA	I-AUC	I-AP	P-AUC	P-AP	PSA	SPA	I-AUC	I-AP	P-AUC	P-AP
w/o Stage Token	85.11%	76.00%	94.1	97.0	88.8	46.1	82.73%	70.00%	86.8	85.1	85.7	15.2
w/o Multi-Stage Mask Encoder	81.33%	66.67%	92.8	96.0	89.4	44.8	75.56%	51.11%	87.8	86.0	86.7	18.6
w/o Visual Reference	80.00%	70.00%	92.8	95.9	89.1	46.0	79.52%	64.29%	89.0	87.9	87.3	17.5
Full model	**95.70%**	**89.11%**	**98.3**	**98.9**	**95.1**	**63.7**	**95.00%**	**88.63%**	**94.3**	**93.6**	**93.7**	**33.9**

A.7.2 Severity Progression Conditions

For the condition ablations, all remaining components and experimental settings are kept unchanged. When removing the Multi-Stage Mask Encoder, the current-stage editing region is still provided, while removing the visual reference retains the same recursive generation strategy. This isolates the contribution of each progression condition while preserving the remaining stage-wise generation process. In addition to the SPA and downstream metrics reported in Sec. 4.4, Table 11 reports PSA for pairwise severity ordering. Removing the stage token reduces PSA from 95.70% to 85.11% on MVTec-AD and from 95.00% to 82.73% on VisA. Without the Multi-Stage Mask Encoder, PSA decreases to 81.33% and 75.56%, while removing the visual reference yields 80.00% and 79.52%, respectively. Thus, all three conditions contribute to pairwise severity ordering, although their relative effects vary across datasets. Compared with PSA, SPA exhibits larger degradation for all three ablations. For example, removing the Multi-Stage Mask Encoder yields 81.33% PSA but 66.67% SPA on MVTec-AD, and 75.56% PSA but 51.11% SPA on VisA. This gap indicates that maintaining a consistently ordered trajectory is more challenging than preserving individual pairwise relations.

A.8 Further Discussion and Future Directions

In this work, we study a new problem of visual severity, defined as the relative strength of anomaly manifestations within the same anomaly type and image context. We first construct SevADSet by using real anomalies as reference states and creating visually weaker and stronger variants around them, providing controlled ordinal supervision for severity modeling. Based on SevADSet, we further develop SevAnoDiff, a severity-aware anomaly generation framework that produces diverse synthetic anomalies across different severity levels and provides richer training samples for downstream anomaly detection and localization. Visual severity captures observable changes in anomaly appearance, but does not necessarily reflect the complete physical severity of an industrial defect. Physical severity may further depend on material properties, structural damage, operating conditions, and temporal evolution. Future work could incorporate multimodal measurements, longitudinal inspection data, and expert defined severity criteria to connect visual manifestations with physical defect severity and extend severity modeling to more complex industrial scenarios.

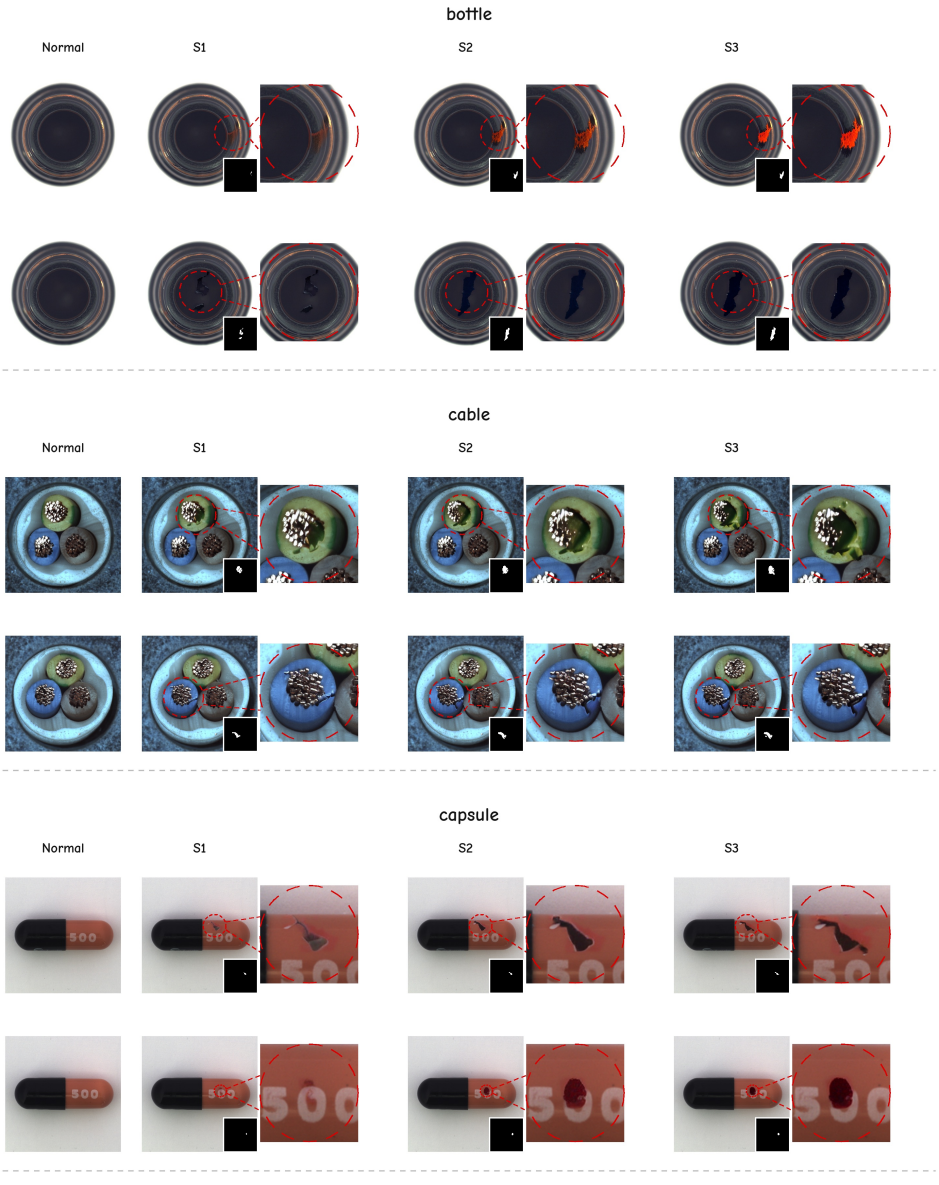

Figure 8: Qualitative results on MVTec AD bottle, cable, and capsule. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.

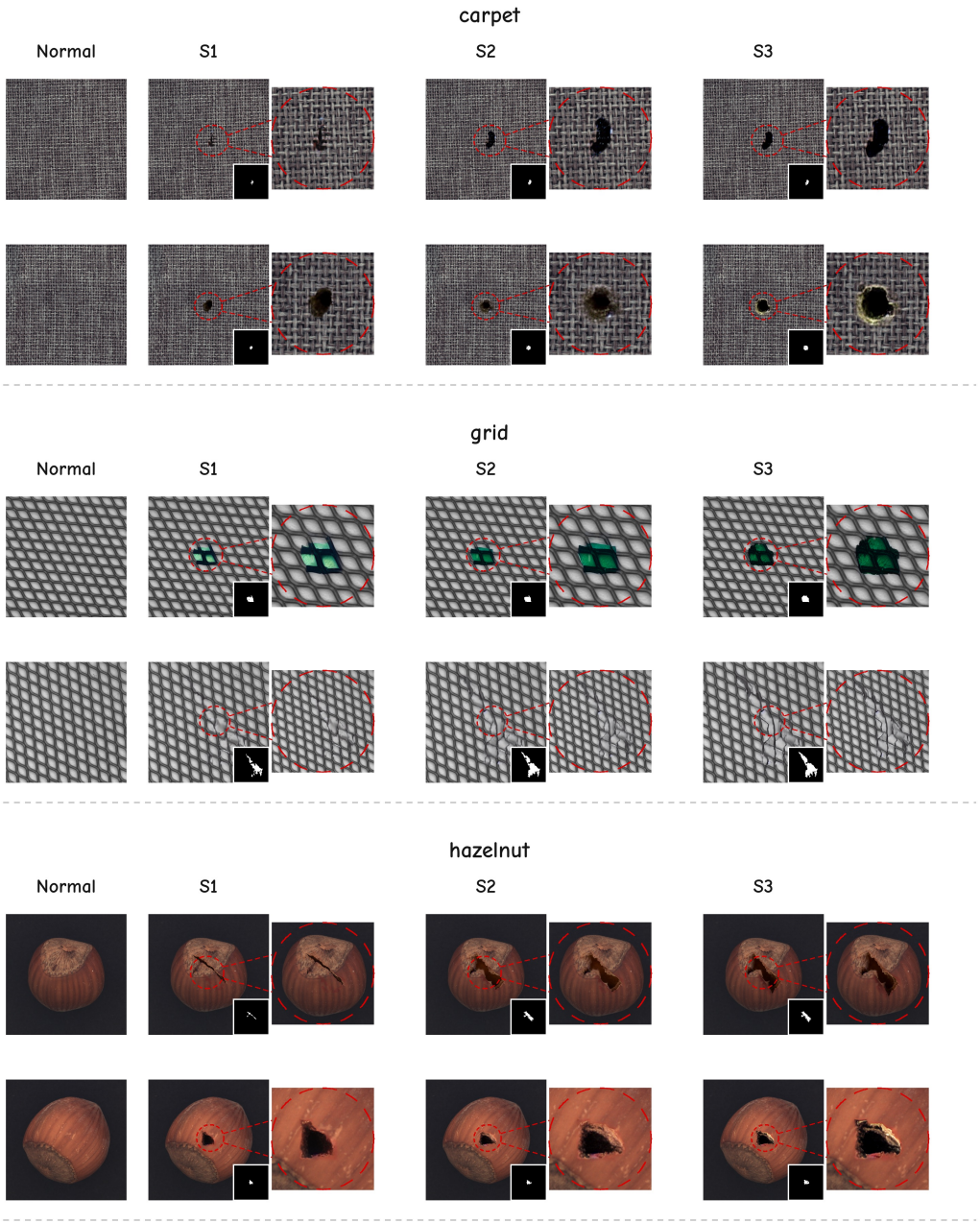

Figure 9: Qualitative results on MVTec AD carpet, grid, and hazelnut. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.

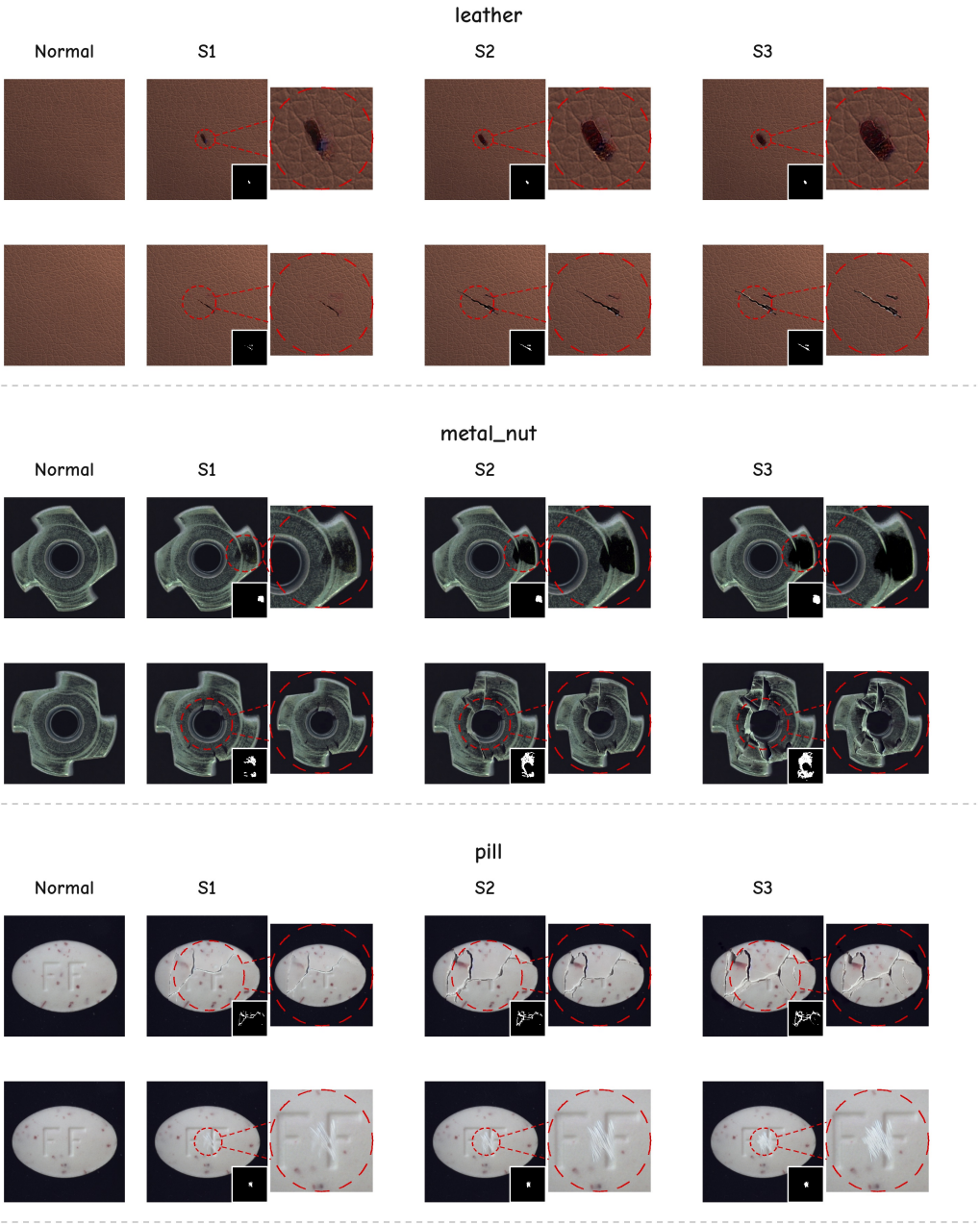

Figure 10: Qualitative results on MVTec AD leather, metal nut, and pill. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.

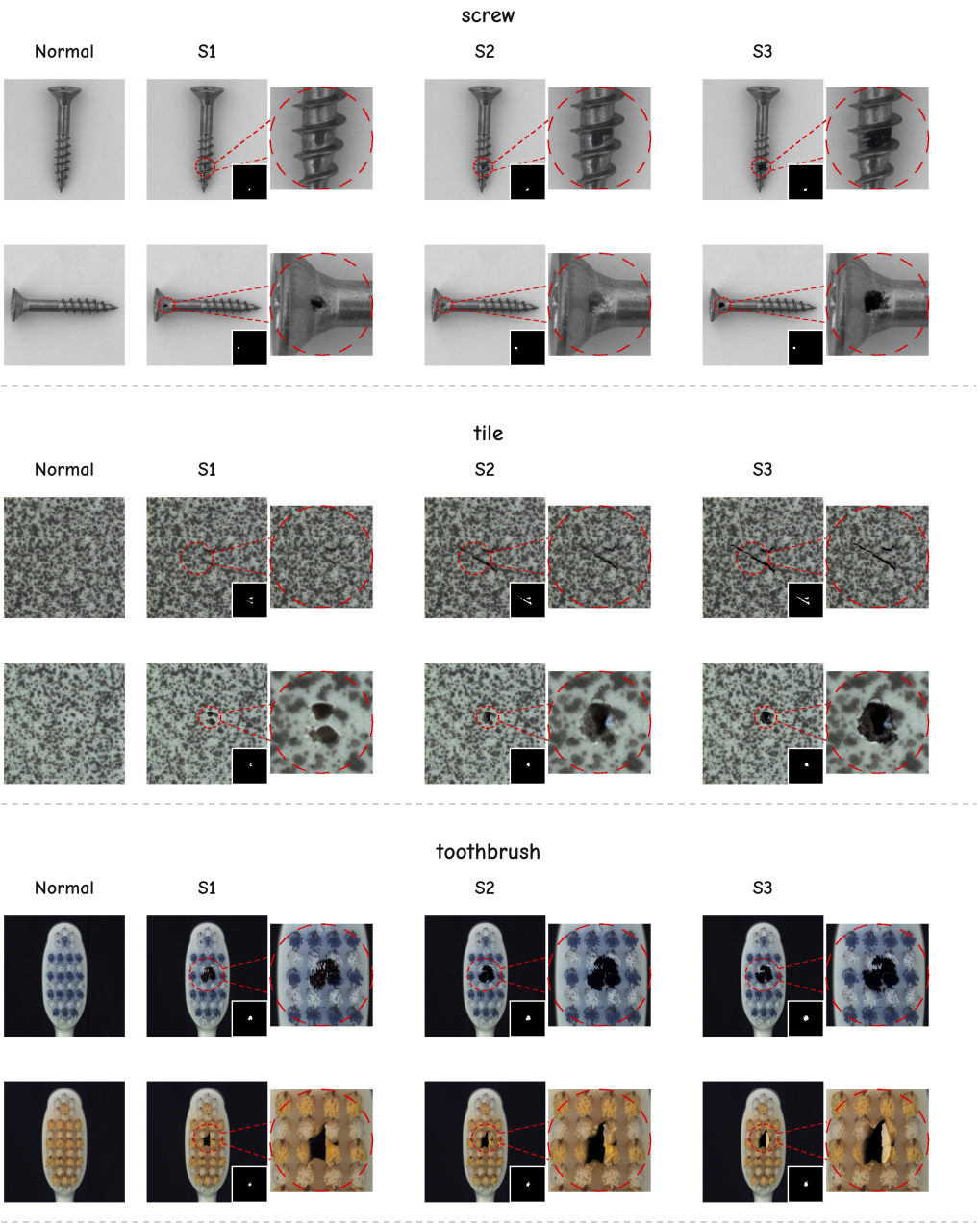

Figure 11: Qualitative results on MVTec AD screw, tile, and toothbrush. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.

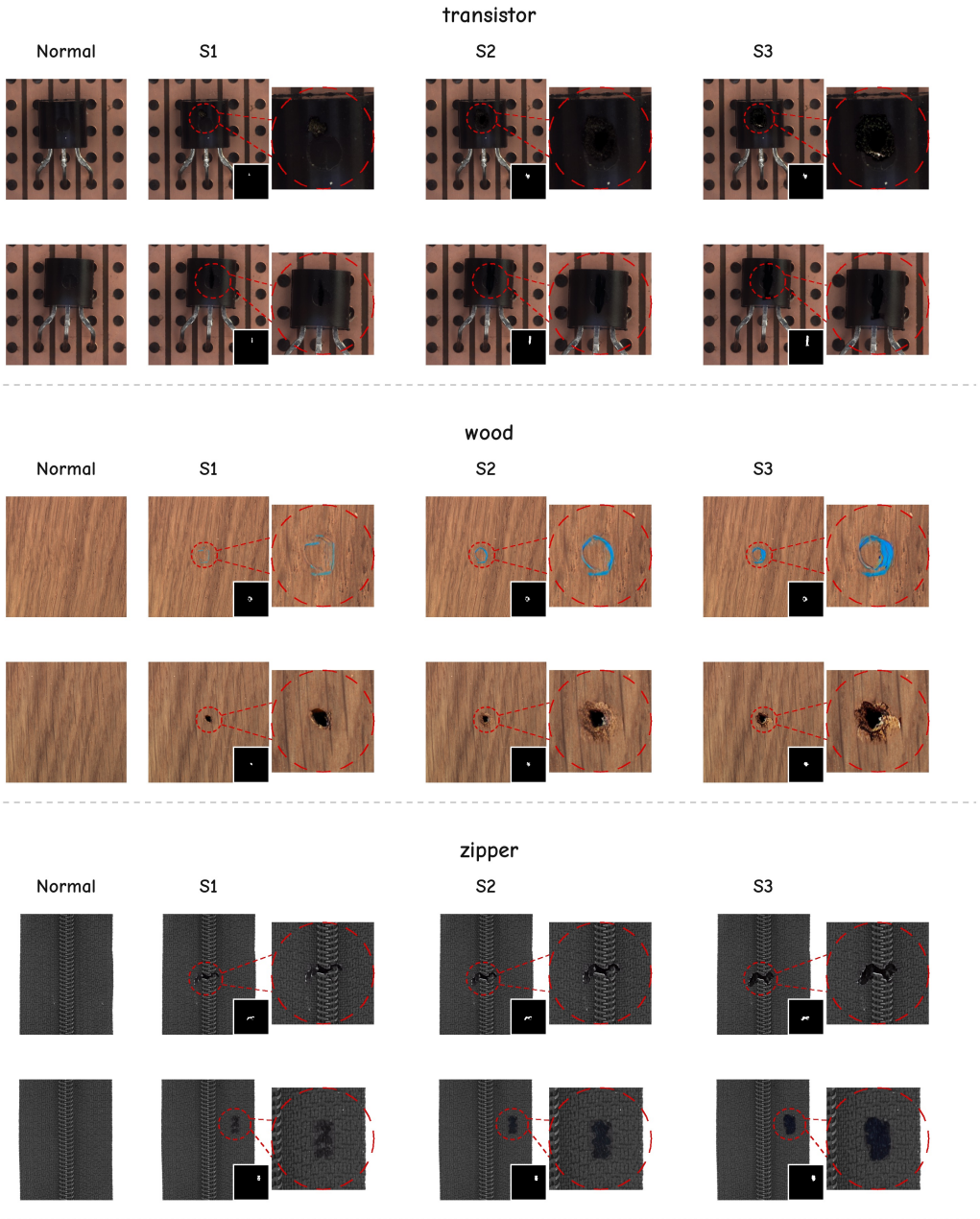

Figure 12: Qualitative results on MVTec AD transistor, wood, and zipper. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.

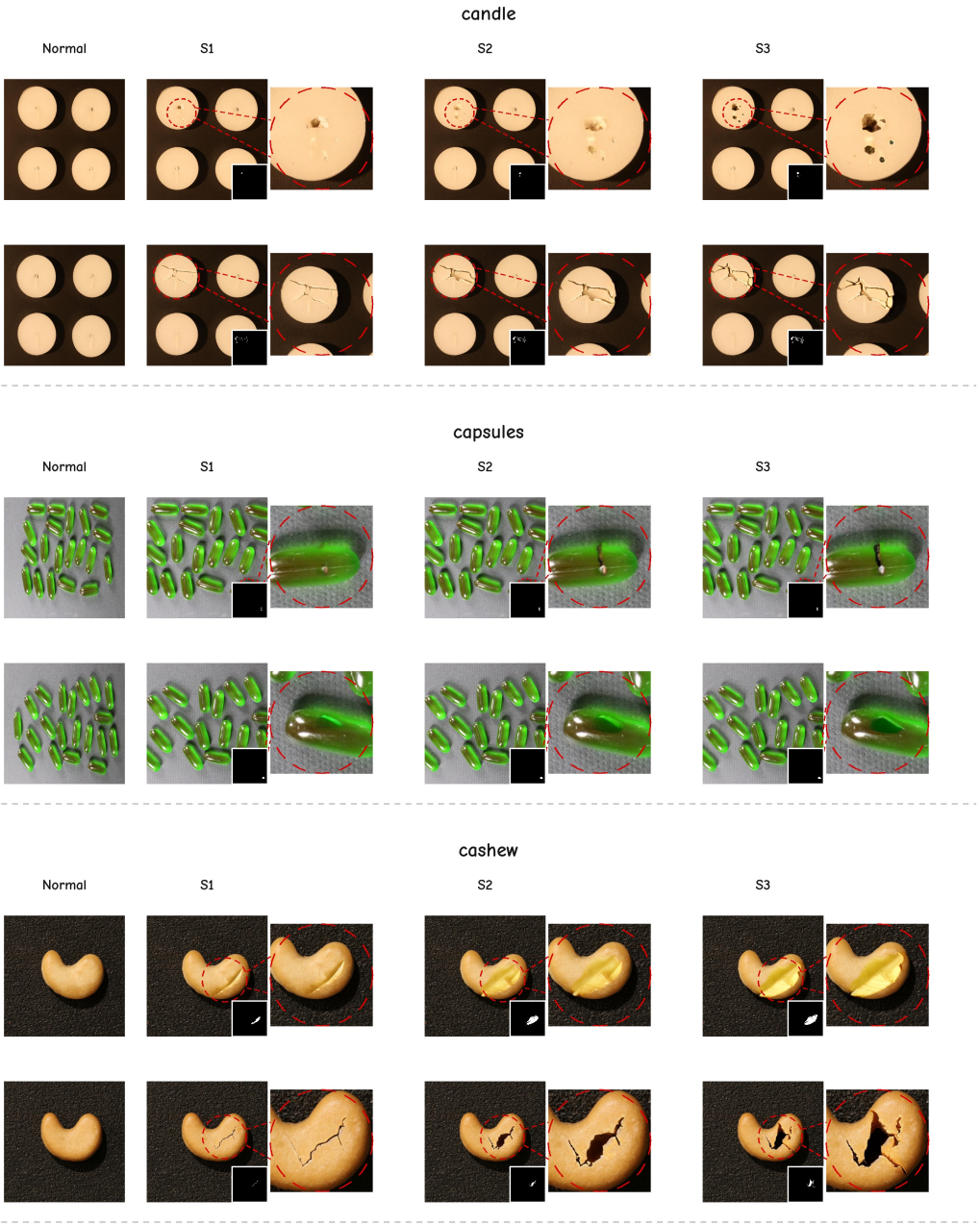

Figure 13: Qualitative results on VisA candle, capsules, and cashew. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.

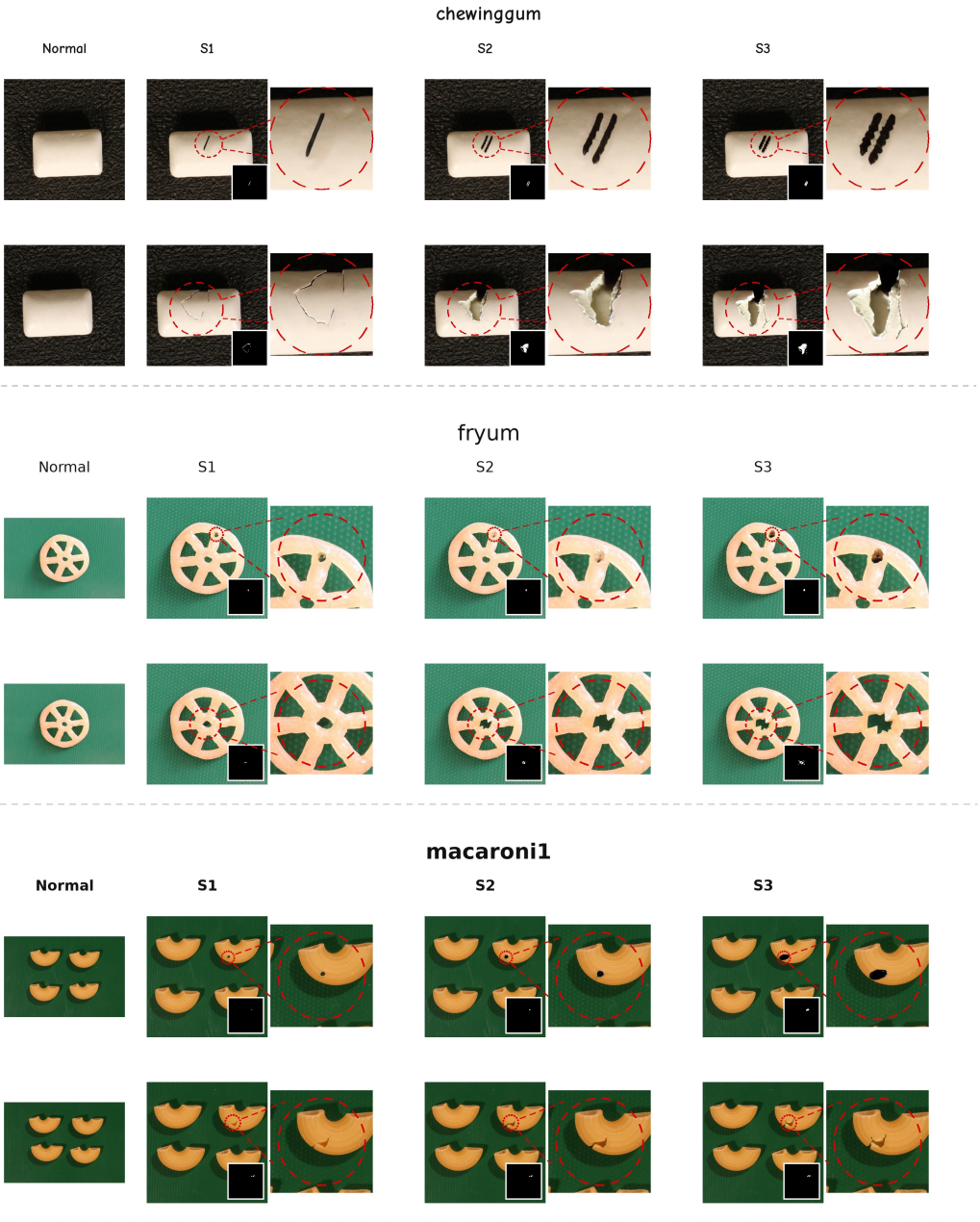

Figure 14: Qualitative results on VisA chewinggum, fryum, and macaroni1. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.

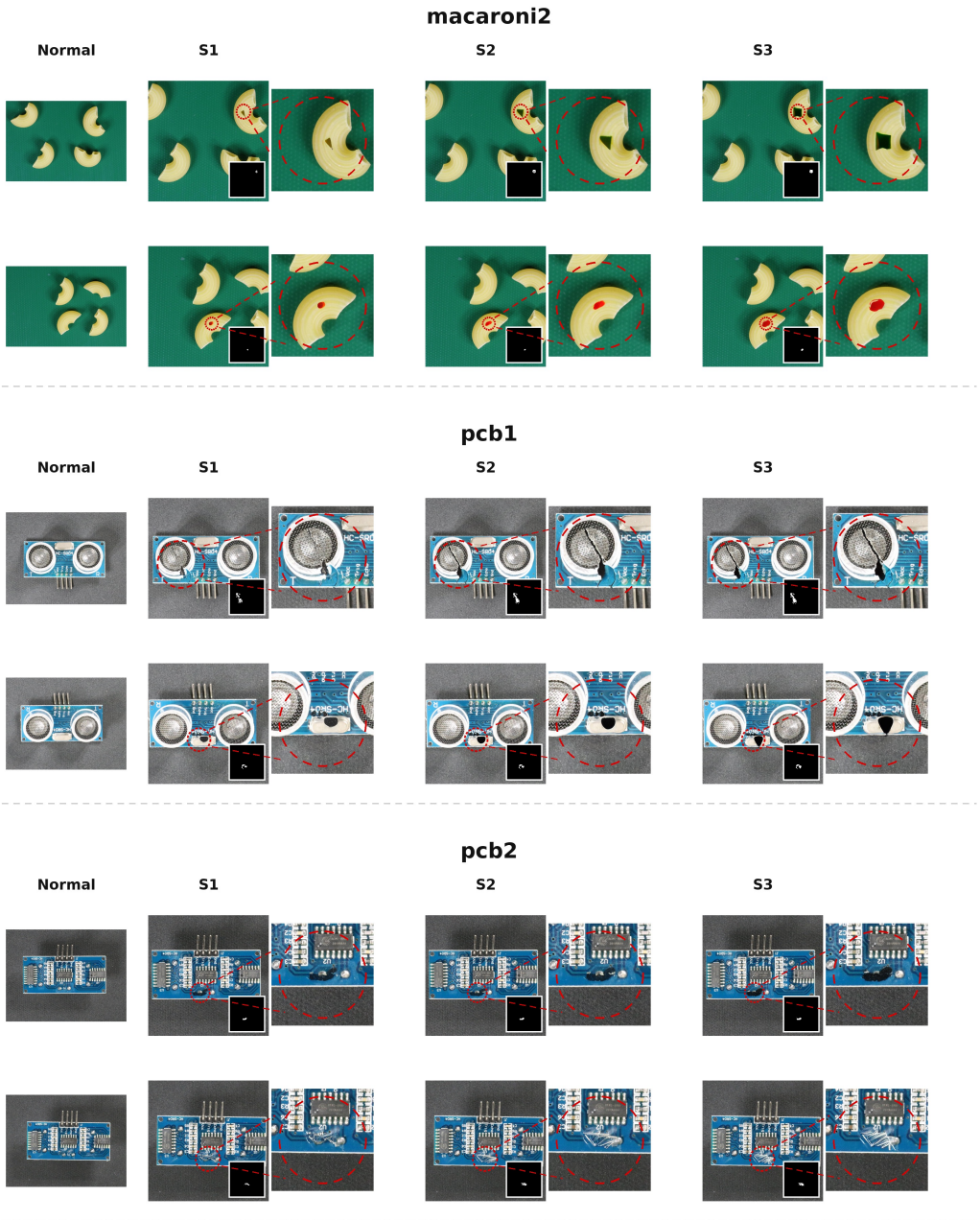

Figure 15: Qualitative results on VisA macaroni2, pcb1, and pcb2. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.



Figure 16: Qualitative results on VisA pcb3, pcb4, and pipe fryum. Our SevAnoDiff generates realistic anomaly images and masks with clear, ordered changes across three stages.