IN-PLACE DEFECT EDITING VIA TIMESTEP-CONSISTENT MASK DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot defect generation mitigates the scarcity of labeled anomalies in industrial visual inspection by synthesizing realistic and diverse defects from a few real examples. Two paradigms emerge: good-to-defect diffusion renders a defect onto a clean template, carrying correspondence risk and domain gap; defect-to-defect editing preserves context but rearranges observed pixels. We find that neither fully preserves defect realism, and we recast augmentation as *in-place defect-to-defect editing*, re-synthesizing the defect at its original location under diffusion-governed fidelity rather than template correspondence. Mask-guided diffusion, however, trains a soft region weighting but samples a hard binary cut, so the imposed boundary is one the model never learned to respect. We couple the mask to the diffusion timestep: a Timestep-adaptive Dynamic Mask Region (TDMR) loss softens the constraint at high noise and sharpens it at low noise, while stepwise blended sampling re-anchors the background at every step with the same mask. Treating the mask as an evolving schedule shared between training and sampling, we identify a critical yet under-explored principle: training–sampling mask consistency, whose violation yields boundary artifacts even when the mask is exact, and whose hard boundary is the limit of the softened training schedule. On MVTec AD, a detector trained solely on our synthetic data attains the best localization accuracy among protocol-matched augmentation baselines, consistent with this principle; mixing with real data yields further gains.

1 INTRODUCTION

Defect detection and pixel-level anomaly segmentation play a central role in industrial visual inspection, yet defective samples are rare, expensive to label, and strongly imbalanced against normal ones (Cheng et al., 2026; Zhu et al., 2026). Generative data augmentation offers a principled remedy: from a few real defective examples, a generator synthesizes realistic, diverse, and spatially controllable defects that enlarge the training pool of a downstream detector.

Two diffusion-based paradigms exist. The dominant one, *distribution-level good-to-defect generation*, renders a defect onto an otherwise clean image (Hu et al., 2024; Dai et al., 2025; Jin et al., 2025; Song et al., 2025); it presupposes a defect-free template and carries a mask–image correspondence risk and domain gap. We study the complementary *instance-level defect-to-defect* setting, where a defective image and its ground-truth mask are given and the defect is re-synthesized in place, inheriting the true location, background, and mask fidelity. The existing defect-to-defect literature, however, relies mostly on non-learned transformations (CutPaste (Li et al., 2021), perturbations), which re-arrange observed pixels and stay within the few-shot appearance manifold; learned resynthesis has received far less attention, with the closest diffusion works operating in pixel space via erasing-and-inpainting (Tao, 2024) or timestep-invariant masks (Pan et al., 2024; Tai et al., 2025). Our editing instead operates in latent space at the defect's original location, with no erasing step to propagate artifacts.

Mask-guided local editing is known to suffer from weak mask–content alignment and boundary artifacts (Lugmayr et al., 2022; Avrahami et al., 2022; 2023; Manukyan et al., 2025). We find these persist even when the mask is exact, because training and sampling treat the mask differently: training uses a soft region weighting the denoiser learns to obey, while inference applies a hard binary cut to the latent trajectory. A model trained under smooth weighting has never learned to respect the sharp boundary inference forces, yielding a measurable gap (Section 4.4) that we name **training–sampling**

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.

mask inconsistency. We propose a timestep-coupled fix: the Timestep-adaptive Dynamic Mask Region (TDMR) loss makes effective mask softness a scheduled function of the timestep—relaxed at high noise, hardened into the binary mask at low noise—and stepwise blended latent sampling re-anchors the background at *every* reverse step with the same latent-resolution mask. Training and sampling thus instantiate one notion of mask fidelity: the hard sampling mask is the limit of the soft training schedule as $t \to 0$ (Figure 1, Section 3).

Our contributions are:

1. **In-place latent editing as defect augmentation.** We reformulate defect augmentation as an editing problem rather than a synthesis problem, re-generating the defect at its original location in latent space while reusing the observed background through the same latent-resolution mask that supervises training; a protocol-matched downstream comparison confirms the benefit (Section 4.2).
2. **The Timestep-adaptive Dynamic Mask Region (TDMR) loss.** A self-normalized, timestep-scheduled loss makes mask softness a sigmoid function of the diffusion timestep, implementing a coarse-to-fine curriculum that remains stable across defects of very different sizes without per-category tuning (Ablation A).
3. **A train–inference consistency principle for mask-guided diffusion.** We identify the objective–operator mismatch underlying mask-guided generation—a soft, timestep-varying training weighting vs. a hard, timestep-invariant inference mask—and instantiate the principle that faithful sampling must mirror the mask semantics of the training objective; violating it inherits boundary artifacts even with an exact mask, while restoring it yields complementary evidence profiles (Section 4.4).

2 Related Work

Few-shot defect synthesis: distribution-level vs. instance-level. Defect generators fall into two paradigms. *Distribution-level good-to-defect synthesis* hallucinates a defect onto a defect-free base image: AnomalyDiffusion learns separate spatial and defect-type embeddings (Hu et al., 2024), the Dual-Interrelated model jointly generates image–mask pairs (Jin et al., 2025), SeaS separates defect-specific and background-preserving branches (Dai et al., 2025), and DefectFill fine-tunes an inpainting model with masked auxiliary losses (Song et al., 2025); GAN methods manipulate defect-aware features (Duan et al., 2023; Chen et al., 2026) and recent variants operate under frozen backbones (Tai et al., 2025; Zhang et al., 2025), including controllable mask-guided frameworks built on inpainting diffusion (Fan et al., 2026; Cho et al., 2026). *Instance-level defect-to-defect editing*, which re-synthesizes an observed defect in place, has received far less attention: non-learned copy-paste and perturbation transforms remain the default baseline (Li et al., 2021), and the closest diffusion attempts are discussed in Section 1 (Tao, 2024; Pan et al., 2024; Tai et al., 2025).

Mask-guided local editing across training and sampling. Mask-guided local editing originated at the sampling side: RePaint re-injects a noised copy of the unmasked region (Lugmayr et al., 2022); the blended-(latent-)diffusion family composes a static mask at every reverse step (Avrahami et al., 2022; 2023); object-aware masking retains static composition (Manukyan et al., 2025); and recent mask-conditioned synthesis keeps the mask static in training (Song et al., 2025; Tai et al., 2025; Zhang et al., 2023; Ali et al., 2024). On the training side, objectives are reweighted only by global, timestep-dependent scalars such as Min-SNR (Ho et al., 2020; Hang et al., 2023), which cannot express spatial distinctions. Prior work thus moves only one side—static sampling-side composition, or spatially blind training-side reweighting—whereas our stepwise blended sampling couples to a timestep-adaptive training-side mask schedule (Section 1) through a shared latent-resolution mask.

Parameter-efficient adaptation and protocols. Low-rank adaptation keeps the backbone frozen and injects trainable low-rank updates into attention projections (Hu et al., 2022), enabling controllable defect augmentation under limited data (Zhang et al., 2025). We adopt category-specific adapters on a frozen Stable Diffusion U-Net (Rombach et al., 2022), validating downstream with a reconstruction-based detector (Zavrtanik et al., 2021) on the full MVTec AD benchmark (Bergmann et al., 2019) under matched generation budgets.

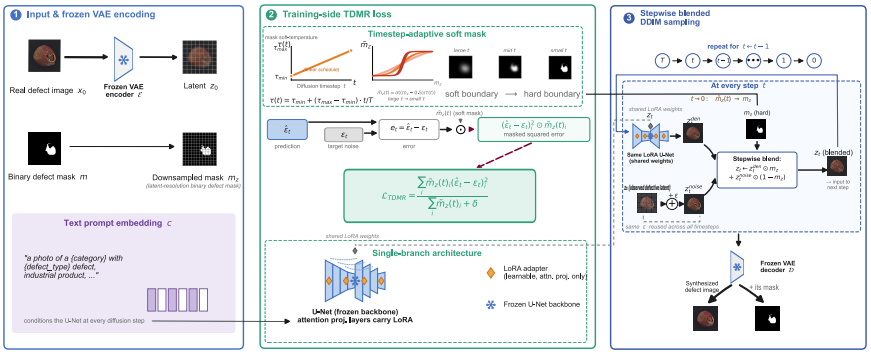


Figure 1: Overview of TDMR. A real defective image and its binary mask are encoded by a frozen VAE to a shared latent resolution. The timestep-adaptive soft-mask loss (TDMR) concentrates training on the defect region with softness scheduled by $\tau(t)$; stepwise blended DDIM re-anchors the background at every reverse step with the same latent-resolution mask. The hard sampling mask is the limit of the softened training schedule as $t \to 0$.

3 METHOD

We build on a pre-trained text-to-image diffusion model and couple two design choices. On the training side, a timestep-adaptive mask loss (Section 3.1) concentrates the gradient on the defect region, soft at high noise and sharp at low noise, with per-category lightweight LoRA adapters (Section 3.2); on the sampling side, a stepwise blended latent procedure (Section 3.3) re-anchors the background at every reverse step with the same latent-resolution mask, so the mask is treated consistently across the pipeline. Given a defective image and its mask, the pipeline regenerates the masked region in place while the outside region remains consistent with the input (Figure 1).

We adopt the DDPM formulation (Ho et al., 2020). Let $x_0 \in \mathbb{R}^{H\times W\times 3}$ denote a real product image containing a defect and $m \in \{0,1\}^{H\times W}$ its binary ground-truth mask. A frozen VAE encoder maps x_0 to a latent $z_0 = \mathcal{E}(x_0)$, and the binary mask is downsampled by nearest-neighbor interpolation to the latent resolution, $m_z \in \{0,1\}^{h\times w}$. The forward diffusion process (Ho et al., 2020) defines noisy latents $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$ for $t \in \{1,\dots,T\}$, and a U-Net (Ronneberger et al., 2015) $\epsilon_\theta(z_t, t, c)$ conditioned on a text embedding c describing the product category and defect type is trained to predict ϵ.

3.1 THE TIMESTEP-ADAPTIVE DYNAMIC MASK REGION (TDMR) LOSS

A hard split between a freely generated defect region and an untouched background leaves a discontinuity at the mask boundary. Rather than adding a boundary-specific loss term, which would reintroduce the multi-term balancing that this design avoids, we make the softness of the mask a scheduled function of the diffusion timestep.

Given the latent-resolution binary mask m_z, we soften it via a sigmoid transform whose temperature is scheduled as a function of the sampled timestep t:

$$\tau(t) = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \frac{t}{T}, \qquad \tilde{m}_z(t) = \sigma\left(\frac{m_z - 0.5}{\tau(t)}\right), \tag{1}$$

where σ is the logistic sigmoid, $T = 1000$ is the total number of diffusion timesteps, and $[\tau_{\min}, \tau_{\max}] = [0.05, 0.5]$ is set by the sensitivity study in Appendix A.1. As in standard DDPM training, the timestep t is sampled uniformly at random for each training example, so the objective below is an expectation over the full noise-level range:

$$\mathcal{L}_{\text{TDMR}} = \mathbb{E}_{t\sim\mathcal{U}\{1,\dots,T\},\,\epsilon\sim\mathcal{N}(0,I)}\left[\frac{\sum_i \tilde{m}_z(t)_i \cdot \big(\epsilon_\theta(z_t,t,c)_i - \epsilon_i\big)^2}{\sum_i \tilde{m}_z(t)_i + \delta}\right]. \tag{2}$$

At large t, where latents are heavily noised and denoising governs coarse structure, the effective mask has a wider, softer transition band; at small t, where fine detail is committed, the mask approaches

the hard binary boundary. Formally, since σ is the logistic sigmoid, $\tilde{m}_z(t)$ converges pointwise to the hard binary mask as the temperature vanishes: $\lim_{\tau \to 0^+} \sigma\big((m_z - 0.5)/\tau\big) = \mathbf{1}[m_z > 0.5] = m_z$, with a residual transition band of width $O(\tau(t))$; the training schedule thus asymptotically recovers the hard sampling mask, and the boundary the sampler imposes at inference is exactly the limit of the softened objective used in training (Section 3.3). This progression instantiates a coarse-to-fine curriculum along the timestep axis, mirroring the inductive bias of denoising networks. The soft mask is broadcast across channels and weights the noise-prediction loss, where $\delta = 10^{-4}$ is a small constant for numerical stability, several orders of magnitude below the smallest mask mass $\sum_i \tilde{m}_z(t)_i$ across all categories. The self-normalizing denominator makes the effective per-pixel weighting invariant to the mask area, in contrast to unnormalized masked losses whose magnitude scales with mask size (Song et al., 2025).

Relationship to DefectFill. DefectFill (Song et al., 2025) combines timestep-invariant masked auxiliary losses with fixed scalar weights. In contrast, TDMR subsumes the masked terms in a single self-normalized loss whose inside/outside weighting varies along the denoising trajectory through $\tilde{m}_z(t)$; the softening is itself a continuous function of t, rather than a constant fixed a priori. This timestep-coupling of the mask schedule is what distinguishes the training-side objective from fixed-weight losses.

3.2 Category-specific LoRA fine-tuning

The training-side objective is deployed per product category: defect morphologies are category-specific, yet each category contributes only a handful of real examples (as few as 9 images in the training pool), so adapting the shared backbone directly would be prone to overfitting and would discard the general image-generation prior it provides. We therefore keep the pre-trained Stable Diffusion U-Net and text encoder (Rombach et al., 2022) frozen and fine-tune lightweight per-category adapters: LoRA modules (Hu et al., 2022) are injected into the query, key, value, and output projections of every cross- and self-attention block, with rank $r = 8$ and scaling factor $\alpha = 8$. Each adapter is trained on only that category's defect images and masks, with a text prompt describing the product category and defect subtype. Trainable parameters amount to 0.185% of the frozen backbone. This is an architectural constant, hence identical across all categories, while the frozen backbone retains its general image prior throughout.

Optimization follows standard practice with a cosine learning-rate schedule (Loshchilov & Hutter, 2017) and EMA-smoothed early stopping; checkpoint selection relies only on held-out training-set diagnostics, never on test-set metrics (full hyperparameters in Appendix A.1).

3.3 Stepwise blended latent sampling

At inference, a real base image x with its defect mask is encoded to a latent $z_0 = \mathcal{E}(x)$, and image-to-image DDIM sampling (Song et al., 2021) in the spirit of SDEdit (Meng et al., 2022) is performed starting from an intermediate noise level determined by strength $s = 0.75$ with 50 DDIM steps. At every reverse step, the one-step denoised latent z_t^{den} is computed under classifier-free guidance with scale $w = 6.0$ (Ho & Salimans, 2022). The original latent is independently re-noised to the same timestep, $z_t^{\text{noise}} = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$, using the same noise sample ϵ drawn once at the start of sampling. The two latents are then blended with the (hard) latent-resolution mask:

$$z_t \leftarrow z_t^{\text{den}} \odot m_z + z_t^{\text{noise}} \odot (1 - m_z). \tag{3}$$

This per-step re-anchoring prevents background drift from accumulating, in contrast to a single post-hoc replacement after the full reverse pass, and concentrates the gradient on the defect region alone. The blend belongs to the blended (latent) diffusion family (Avrahami et al., 2023), with a distinctive coupling: the sampling-time blend uses the hard binary mask at every step, while the network was trained under the softened schedule of Eq. (1), so the boundary the sampler imposes is exactly the boundary the model learned to respect. Each generation call additionally applies spatial and stochastic shape augmentation of the mask (Appendix A.1), whose contribution we isolate in Section 4.4 (Ablation C). Table 1 makes the update-equation differences from RePaint, Blended Diffusion, and Blended Latent Diffusion precise: all prior methods compose a static mask on the sampling side with no counterpart on the training side, whereas TDMR couples the stepwise re-anchoring to the timestep-adaptive training schedule through the same latent-resolution mask.

Table 1: Update-equation comparison of prior blending methods and TDMR.

Method	Sampling-side update	Training-side mask schedule	Timestep-adaptive softness
RePaint	hard-mask inpainting + resampling	none (pre-trained)	×
Blended Diffusion	static soft mask m fixed for all t, pixel space	none (pre-trained)	×
Blended Latent	static soft mask m fixed for all t, latent space	none (pre-trained)	×
DefectFill	static hard-mask replacement	fixed timestep-invariant masked auxiliary losses ($\mathcal{L}_{def}$, $\mathcal{L}_{obj}$)	×
TDMR (Ours)	stepwise blend with the *same* mask, same latent resolution, re-anchored each step	TDMR loss with $\tau(t)$ softness schedule	✓ (wide transition at large $t \rightarrow$ hard at small t)

The stepwise blend itself adds no extra reverse steps: it is an element-wise mask composition of two latents already in memory, whose cost is negligible relative to a single denoising forward pass. Algorithm 1 (Appendix A.2) summarizes the full generation procedure.

4 Experiments

4.1 Setup

We evaluate on MVTec AD (Bergmann et al., 2019), which contains 15 categories (10 object and 5 texture types) spanning 73 anomaly types, and follow the few-shot protocol of prior work (Hu et al., 2024; Song et al., 2025): the first 33% of defective images of each subtype form the training pool, the remaining 67% the evaluation set, and normal images are retained in full on both sides. From each subtype we synthesize $N_{\text{gen}} = \text{clamp}(8 \times N_{\text{train}}, 30, 100)$ image–mask pairs (about 3,120 in total) and train a DRAEM-style segmentation network (Zavrtanik et al., 2021) on the synthetic defects paired 1:1 with real normal images, evaluating on the real held-out split. For detection we report image- and pixel-level AUC (area under the ROC curve), image AP, pixel AP, pixel F1-max, IoU@BestF1, and PRO (Bergmann et al., 2019) (Table 2); table headers abbreviate these as AUC-I/AP-I (image-level) and AUC-P/AP-P/F1-P/IoU-P (pixel-level). Per-category values are given in Appendix A.3.

4.2 Main results: downstream detection

We include only baselines fairly reproducible from public datasets under the identical protocol, anchoring the two paradigm directions in defect augmentation—CutPaste for the classical augmentation direction and AnomalyDiffusion for diffusion-based generation. All methods share an identical detector architecture, training protocol, and test split, and are evaluated from the same number of generated images per category (for AnomalyDiffusion we randomly sample from its released repository the same count as TDMR produces), so the comparison isolates data quality from data quantity.

TDMR is the only method that improves image-level detection and localization together, attaining the best accuracy on all localization-oriented metrics while remaining competitive on the image-level ones (Table 2); the per-category comparisons are in Appendix A.3. The single column in which the strongest baseline leads, pixel AUROC, is also the one most sensitive to the hardest low-frequency categories (e.g., *tile*), where the mask schedule is least effective; the gap of 1.3 points is a signature predicted by our mechanism analysis rather than a general weakness (Appendix A.6). Mixing the synthesized data with the real few-shot pool improves detection further (Section 4.3).

Table 2: Protocol-matched downstream detection performance, mean over the complete MVTec AD benchmark. AUC-I/AP-I: image-level; AUC-P/AP-P/F1-P/IoU-P: pixel-level; PRO: per-region overlap. Real-only denotes the lower bound trained without synthetic data; bold marks the best of the three per column.

Method	AUC-I	AP-I	AUC-P	AP-P	F1-P	IoU-P	PRO
Real-only	0.8662	0.8854	0.8489	0.5385	0.5889	0.4572	0.7118
CutPaste	0.9397	0.9657	0.8516	0.5879	0.6099	0.4692	0.7310
AnomalyDiffusion	0.9029	0.9491	**0.9221**	0.6200	0.6098	0.4782	0.7736
TDMR (Ours)	**0.9488**	**0.9708**	0.9093	**0.6459**	**0.6477**	**0.5035**	**0.8143**

Table 3: Effect of mixing TDMR synthetic defects with the real few-shot pool: absolute pixel-level improvements (percentage points) over the real-only floor; TDMR denotes synthetic-only training and TDMR + real denotes 1:1 mixing with real normal images; absolute values in Table 2. AUC-P/AP-P/F1-P/IoU-P: pixel-level; PRO: per-region overlap. Bold marks the best configuration per column.

Training data	AUC-P	AP-P	F1-P	IoU-P	PRO
Real-only	0.8489	0.5385	0.5889	0.4572	0.7118
TDMR	+6.04	+10.74	+5.88	+4.63	+10.25
TDMR + real	**+7.56**	**+14.85**	**+8.91**	**+7.83**	**+13.76**

4.3 Synergy with the real few-shot pool

The synthesized data is complementary to, not a substitute for, real samples. Training the same DRAEM detector on the real few-shot pool together with TDMR's output (1:1 with real normal images) improves *all five* pixel-level localization metrics simultaneously (Table 3): the mixed configuration is best in every column, and lifts pixel AUROC past the strongest single-method baseline of Table 2. Mixing beats the real-only floor in every category and both ingredients in most categories, indicating that the synthetic defects carry information orthogonal to the real pool; the per-category values (Appendix A.4) show the largest gains precisely where the real pool alone is weakest (e.g., *transistor*, *toothbrush*).

4.4 Mechanism analysis

We isolate the two mechanisms that constitute our training–sampling consistency claim, independent of any downstream detector, under identical hyperparameters on the complete benchmark. If the principle is correct, the timestep-adaptive schedule should matter most where the training–sampling mismatch is largest, while stepwise blending should help uniformly.

Ablation A: static vs. timestep-adaptive mask softness. Freezing the schedule at the mean temperature $\bar{\tau}$ degrades the held-out high-noise diagnostic in all but three categories (Table 4); the largest losses fall on the categories where the training–sampling mismatch is largest (e.g., *toothbrush*, *tile*), exactly as the consistency principle predicts.

Ablation B: single-step vs. stepwise blending. Replacing the per-timestep re-blending with a single post-hoc blending (trained weights and guidance fixed) raises the boundary-region pixel L2 error in every category, by more than a tenth on average (Table 4); a control region well outside the mask is nearly unchanged, confirming that the gain reflects the blending strategy rather than a global shift.

Ablation C: isolating the source of within-defect diversity. Our recipe couples the timestep-adaptive supervision and stepwise blending (contributions 2–3) with the stochastic shape augmentation of the mask (erosion, blur-based re-thresholding, internal holes). Table 5 ablates the shape augmentation on three categories (capsule, tile, leather) with identical adapters, base images, and initialization across the two branches. Removing it reduces within-defect diversity substantially (across all categories, Table 9 in Appendix A.5), showing that within-defect diversity is *primarily* driven by the mask perturbation. Augmentation also incurs a disclosed fidelity cost at the boundary (Table 5), which the downstream gains of Table 2 justify. The components play complementary, non-overlapping roles:

Table 4: Mechanism ablations across all categories, shown as two independent side-by-side tables: (left) Ablation A, held-out high-noise diagnostic reduction (%) with dynamic $\tau(t)$ vs. fixed $\bar{\tau} = 0.275$; (right) Ablation B, boundary-region pixel L2 with stepwise vs. single-step blending.

Ablation A: diagnostic (%)

Category	Fixed $\bar{\tau}$	Dynamic $\tau(t)$	Δ
bottle	12.07	14.64	+2.57
cable	12.72	15.06	+2.35
capsule	22.32	20.42	−1.90
carpet	4.85	3.90	−0.96
grid	7.71	11.49	+3.78
hazelnut	14.66	15.08	+0.42
leather	3.30	3.65	+0.35
metal_nut	11.39	14.92	+3.53
pill	15.25	17.53	+2.28
screw	9.57	11.27	+1.70
tile	−6.57	−2.79	+3.79
toothbrush	2.07	9.45	+7.38
transistor	13.44	16.41	+2.98
wood	4.28	3.00	−1.28
zipper	7.74	7.80	+0.07
Mean	**8.99**	**10.79**	**+1.80**

Ablation B: boundary L2

Category	Single-step	Stepwise	Reduction
bottle	16.650	15.073	9.5%
cable	29.829	25.496	14.5%
capsule	14.887	13.759	7.6%
carpet	42.139	39.092	7.2%
grid	25.542	20.160	21.1%
hazelnut	25.841	21.665	16.2%
leather	24.855	21.115	15.1%
metal_nut	20.132	16.831	16.4%
pill	11.055	10.318	6.7%
screw	20.326	17.910	11.9%
tile	31.901	26.728	16.2%
toothbrush	24.332	20.950	13.9%
transistor	16.186	14.506	10.4%
wood	19.520	17.320	11.3%
zipper	20.138	19.372	3.8%
Mean	**22.889**	**20.020**	**12.5%**

Table 5: Diversity-source ablation and background preservation on three representative categories (capsule, leather, tile), shown as two independent side-by-side tables. Left: shape-augmentation on/off (identical adapters, base images, and initialization across the two branches); IC-LPIPS measures within-defect diversity with the clustering protocol of Appendix A.5; Boundary-L2 uses the width-5 ring of the original mask, RGB L2, matching Ablation B. Right: background-preservation control (pixel L2, mean over 20 samples per defect type); FREE = same pipeline without the mask-blending constraint; VAE = encoder–decoder reconstruction of the source image (the reconstruction floor). Per-defect-type values in Appendix A.5 (Tables 9, 10).

Ablation C: shape augmentation

Category	IC-LPIPS+	IC-LPIPS−	Bnd+	Bnd−
capsule	0.119	0.045	15.7	12.2
leather	0.229	0.040	27.0	22.0
tile	0.345	0.104	46.6	27.0
Mean	**0.231**	**0.063**	**29.8**	**20.4**

Background preservation

Category	TDMR	FREE	VAE	Fg-L2
capsule	8.7	31.0	6.3	43.4
leather	19.8	71.2	13.7	55.5
tile	42.9	76.6	11.1	73.3
Mean	**20.4**	**63.3**	**9.8**	**55.5**

mask perturbation drives morphological variation, whereas the proposed mechanisms contribute boundary fidelity and train–inference consistency (Ablations A and B).

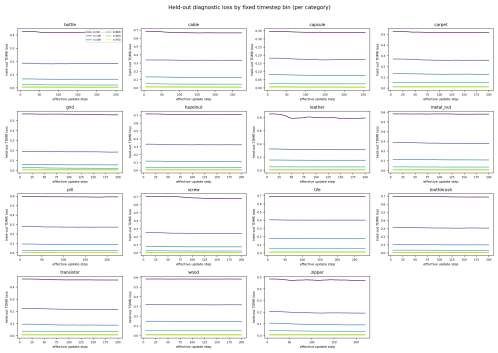
The two mechanisms have complementary evidence profiles (uniform across categories for stepwise blending, concentrated in the hardest categories for the adaptive schedule), exactly as the consistency principle predicts. Figure 2 visualizes the training-side effect: the held-out denoising diagnostic improves most at the high-noise bins where the softened mask is active, and per category the high- and low-noise regimes improve together with *tile* the sole outlier (Appendix A.6). Across the non-*tile* categories the high-noise bins, where the TDMR schedule is most relaxed, improve substantially; the low-noise bins improve in every category but by a smaller margin. Training thus concentrates capacity exactly where the softened mask is active, without sacrificing the fine-detail regime the hard mask governs at sampling. The schedule itself is robust to its operating point: doubling the learning rate alone and widening $\tau_{\max}$ alone each improve the held-out high-noise loss, while combining both yields more than the sum of the two individual changes, indicating that the two adjustments act complementarily rather than independently (Table 6; full configuration grid in Appendix A.1).

The boundary-fidelity gain is visible qualitatively. Figure 3 shows representative categories (*grid*, *wood*, *bottle*, *carpet*), each panel paired with a magnified view of the region boxed in red: CutPaste leaves a visible pasted-patch seam, AnomalyDiffusion drifts most on texture categories, and TDMR's

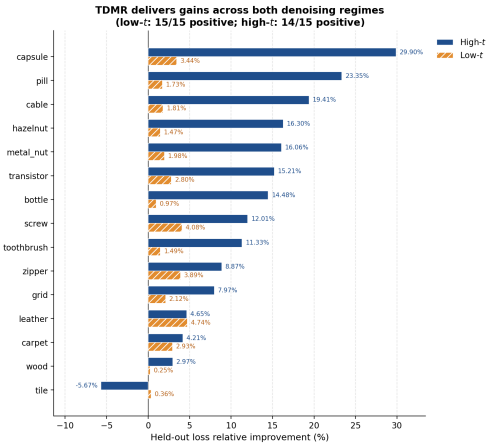
Relative to the baseline schedule ($\tau_{\max} = 0.2$, lr 1×10^{-4}), widening the temperature range alone, doubling the learning rate alone, and combining both each improve the held-out high-noise diagnostic. The combined configuration improves by more than the sum of the two individual changes, indicating that the two adjustments act complementarily rather than independently: a wider high-noise band exposes more of the softened objective to optimization, while a higher learning rate accelerates its convergence, so the two knobs interact instead of acting as independent contributors. Bold marks the best configuration.

Table 6: Hyperparameter sensitivity of the adaptive schedule: held-out high-noise improvement relative to the baseline schedule.

Configuration	$\tau_{\max}$	lr	Held-out
Baseline	0.2	1×10^{-4}	1.42%
Widen range	0.5	1×10^{-4}	2.96%
Double lr	0.2	2×10^{-4}	6.70%
Both (final)	**0.5**	2×10^{-4}	**12.92%**



(a) High-noise bins improve most.



(b) High- vs. low-noise gains per category.

Figure 2: Training-side evidence for the adaptive schedule, configuration C. (a) Held-out denoising diagnostic at six fixed timestep bins ($t \in \{50, \ldots, 950\}$); high-noise bins improve most. (b) Per-category improvement at high- ($t \in \{800, 950\}$) vs. low-noise ($t \in \{50, 200\}$) timesteps; they improve together, with *tile* the sole outlier (its checkpoint diagnostic is deferred to Appendix A.6).

blended output best matches the real background at the mask boundary. An additional qualitative comparison covering all but one category is provided in Appendix A.7, and the failure regime of the hard-mask assumption is discussed in Appendix A.6.

4.5 Background Preservation

Contribution 1 asserts that editing preserves the observed background by construction rather than re-synthesizing it. Ablation B already rules out a global shift, but does not quantify how much the background actually moves. We measure this directly against two controls: **FREE**, the same pipeline with the mask-blending constraint removed (background re-synthesized freely), and **VAE**, the encoder–decoder reconstruction of the source image, which bounds the background change achievable by any latent-diffusion method. Identical base images, prompts, LoRA adapters, and initialization are used across the three variants; background is defined as pixels outside the original mask.

The measurement confirms the claim directly: TDMR changes the background markedly less than the unmasked pipeline (pixel L2 on three representative categories; per-defect-type values in Appendix A.5), approaching the VAE reconstruction floor on capsule and leather—beyond the unavoidable autoencoder reconstruction error, the background is reused, not re-synthesized. The single deviation, *tile*, is fully explained by the mask-displacement artifact measured in Ablation C; even there, TDMR halves the background change relative to FREE. Together with the foreground change (Table 5), the three metrics show one coherent picture: TDMR concentrates its edit on the defect region and preserves the surrounding background without a visible seam.

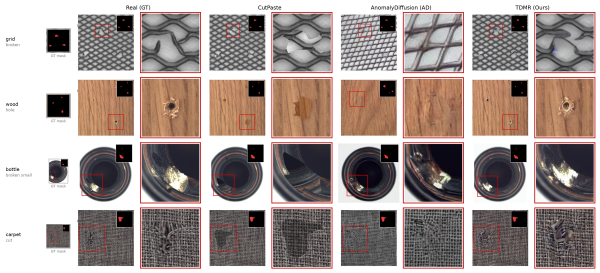


Figure 3: Qualitative comparison with local magnification on four representative categories (grid, wood, bottle, carpet). Columns: real defective image with its ground-truth mask, CutPaste, AnomalyDiffusion, TDMR (ours). CutPaste and TDMR are mask-conditioned; AnomalyDiffusion does not consume the reference mask, serving as a realism reference rather than a boundary-fidelity one.

5 Conclusion

We reformulate few-shot defect augmentation as in-place latent editing of the observed defective image and instantiate it through a single train–inference consistency principle: the TDMR loss couples mask softness to the diffusion timestep via a scheduled sigmoid temperature, and stepwise blended latent sampling re-anchors the background at every reverse step with the same mask used in training. The coupling is not incidental: each mechanism is effective exactly where the other is active. A causal ablation across all categories substantiates the principle from both sides (uniform boundary-fidelity gains from stepwise blending; held-out high-noise diagnostic gains from the adaptive schedule), a downstream detector trained on TDMR's output attains the best localization accuracy among protocol-matched methods, and mixing with the real few-shot pool improves this further. A direct control shows the observed background changes far less than an unmasked pipeline, approaching the VAE reconstruction floor on capsule and leather: the background is reused, not re-synthesized. Because the principle is not defect-specific, we expect the timestep-adaptive softening schedule and the per-step re-anchoring to transfer to other mask-guided diffusion settings that pair a soft training-time region loss with a hard inference-time mask.

Two directions frame our future work. We evaluate on a single benchmark, MVTec AD, following the convention of prior few-shot defect synthesis (Hu et al., 2024; Song et al., 2025); extending the per-mechanism attribution to structural (VisA (Zou et al., 2022)) and logically constrained (MVTec LOCO AD (Bergmann et al., 2022)) benchmarks is the immediate next step. We assume a defect mask as input, as is standard for mask-guided editing; relaxing this assumption is a natural direction. Among the categories, *tile* is the single one whose held-out high-noise diagnostic is non-monotonic, so it alone deviates from the uniform final-step checkpoint rule (Appendix A.6).

Ethics Statement

This work addresses synthetic defect generation for industrial anomaly detection. The synthesized defects are generated from public MVTec AD data and used exclusively to benchmark detection and segmentation algorithms; the method involves no human subjects, personal data, or content that could cause social harm. Synthetic defect images could in principle be misused to fabricate inspection evidence, so we restrict the intended use to research and quality-control validation.

References

Musawar Ali, Nicola Fioraio, Samuele Salti, and Luigi Di Stefano. Anomalycontrol: Few-shot anomaly generation by controlnet inpainting. *IEEE Access*, 12:192903–192914, 2024. doi: 10.1109/ACCESS.2024.3520002. URL https://ieeexplore.ieee.org/document/10806704.

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. doi: 10.1109/CVPR52688.2022.01767.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):149:1–149:11, 2023. doi: 10.1145/3592450. URL https://doi.org/10.1145/3592450.

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. doi: 10.1109/CVPR.2019.00982.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022. doi: 10.1007/s11263-022-01578-9.

Junchao Chen, Yahui Cheng, Guojun Wen, and Xingyue Liu. Bmmc-gan: A bi-directional mapping and multi-category controlled generative adversarial network for diverse industrial defect images generation. *Neurocomputing*, 672:132859, 2026. doi: 10.1016/j.neucom.2026.132859.

Yuqi Cheng, Yunkang Cao, Haiming Yao, Wei Luo, Cheng Jiang, Hui Zhang, and Weiming Shen. A comprehensive survey for real-world industrial surface defect detection: Challenges, approaches, and prospects. *Journal of Manufacturing Systems*, 84:152–172, 2026. doi: 10.1016/j.jmsy.2025.11.022.

Jaebong Cho, Dohyeon Kong, Jihoon Nam, and Hyunbo Cho. A synthetic defect image generation method for enhanced industrial defect detection based on inpainting diffusion and context-aware mask generation. *Journal of Intelligent Manufacturing*, 2026. doi: 10.1007/s10845-026-02915-2.

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.

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. doi: 10.1609/aaai.v37i1.25132.

Kang Fan, Kun Qian, Jingxuan Xu, Bowei Liu, Yuao Gao, Nianyin Zeng, and Zheng Fang. Few-shot mask-guided controllable anomaly image generation via inpainting diffusion for surface defect detection. *Journal of Intelligent Manufacturing*, 2026. doi: 10.1007/s10845-026-02893-5.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7441–7451, 2023. URL https://arxiv.org/pdf/2303.09556.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

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

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

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeefYf9.

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. doi: 10.1609/aaai.v38i8.28696.

Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Jiafu Wu, Hao Chen, Haoxuan Wang, Wenbing Zhu, Mingmin Chi, Jun Liu, and Yabiao Wang. Dual-interrelated diffusion model for few-shot anomaly image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30420–30429, 2025. doi: 10.1109/CVPR52734.2025.02832.

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. doi: 10.1109/CVPR46437.2021.00954.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1608.03983.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Hrant Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=qioDi3afqm.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Zhihong Pan, Rahul Shenoy, Kaushik Balakrishnan, Qisen Cheng, Janghwan Lee, Yongmoon Jeon, Deokyeong Jeong, and Jaewon Kim. Improving qd backplane defect image generation using automatic masking in diffusion models. In *SID Symposium Digest of Technical Papers*, volume 55, pp. 409–412, 2024.

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)*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Jaewoo Song, Daemin Park, Kanghyun Baek, Sangyub Lee, Jooyoung Choi, Eunji Kim, and Sungroh Yoon. Defectfill: Realistic defect generation with inpainting diffusion model for visual inspection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18718–18727, 2025. doi: 10.1109/CVPR52734.2025.01744.

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.

Yichun Tai, Zhenzhen Huang, Tao Peng, and Zhijiang Zhang. Deffiller: Mask-conditioned generation with diffusion prior for saliency-based defect detection. *The Visual Computer*, 41(13):10947–10962, 2025.

Huanjie Tao. Erasing-inpainting-based data augmentation using denoising diffusion probabilistic models with limited samples for generalized surface defect inspection. *Mechanical Systems and Signal Processing*, 208:111082, 2024.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dræm—a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021. doi: 10.1109/ICCV48922.2021.00822.

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. 3813–3824, 2023. doi: 10.1109/ICCV51070.2023.00355.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Zhe Zhang, Yuhang Zhou, and Jie Ma. Defect-lora: Controllable defect data augmentation based on low-rank adaptation for surface defect recognition under limited data. *Expert Systems with Applications*, 268:126251, 2025.

Jiaqi Zhu, Yunfeng Fan, Geng Han, Xiang Shi, Fang Deng, and Jie Chen. Generative anomaly detection: A comprehensive review of modeling principles, advances, and future opportunities. *Artificial Intelligence Review*, 59(9):193, 2026. doi: 10.1007/s10462-026-11591-w.

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 (ECCV)*, pp. 392–408, 2022. doi: 10.1007/978-3-031-20056-4_23.

REQUIRED: AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of this paper, to assist with retrieval and discovery of related work, and to draft sections of the manuscript. The initial draft was generated with the aid of AI tools and subsequently substantially revised by the authors. We have not used generative AI tools for research ideation or execution, for generating synthetic datasets, or for proving mathematical claims, and these uses are not applicable to this work. We have reviewed all AI-assisted work. Specifically, the authors reviewed and verified the AI-assisted content, and all references were checked against authoritative sources before submission. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A APPENDIX

We organize the appendix around the claims made in the main paper and the questions an ICLR reviewer is likely to raise, rather than as a raw dump of tables: **Appendix A** makes every training and inference number reproducible; **Appendix B** positions stepwise blending against the prior blending lineage at the update-equation level; **Appendix C–E** provide the complete per-category evidence behind the headline numbers of the main tables; **Appendix F** reports the failure regime of the hard-mask assumption (the categories in which TDMR gains least); **Appendix G** provides an additional qualitative comparison over all but one category so that the qualitative evidence is auditable beyond the four categories of Figure 3.

A.1 IMPLEMENTATION AND REPRODUCIBILITY DETAILS

All generation uses a frozen Stable Diffusion 1.4 VAE and U-Net (Rombach et al., 2022); category-specific LoRA modules (rank 8, scaling 8) are trained on the attention projections of every cross- and self-attention block. Training uses AdamW (Loshchilov & Hutter, 2019) (peak learning rate 2×10^{-4}, cosine decay, 50-step warmup), bf16, gradient clipping at 0.5, effective batch size 16, and EMA-smoothed early stopping (decay 0.9, 200-step minimum, 60-step patience). Training-time augmentation applies horizontal flip (50%), affine perturbation ($\pm 8^{\circ}$ rotation, scale 0.92–1.08, translation $\pm 3\%$), and color jitter (brightness 0.2, contrast 0.2, saturation 0.1, hue 0.02; applied with 70% probability), jointly applied to the image and its mask. A pool-size-adaptive gradient-accumulation rule (4/3/2 steps for pools of < 15 / 15–24 / ≥ 25 images) normalizes the per-task effective batch size without per-category tuning. Each category trains for 200–280 optimizer steps. **Hyperparameter sensitivity.** Relative to the baseline schedule ($\tau_{\max} = 0.2$, learning rate 1×10^{-4}, $\sim$60–66 steps; held-out high-noise improvement 1.42%), widening the temperature range alone ($\tau_{\max} = 0.5$, $\sim$60–100 steps) improves by 2.96%, doubling the learning rate alone (2×10^{-4},

200–280 steps) improves by 6.70%, and the combined configuration improves by 12.92%, exceeding the sum of the individual effects.

Checkpoint selection rule. Selection relies only on held-out training-set diagnostics, never on test-set metrics. For all but one category the standard post-early-stopping checkpoint is used; *tile* is the single exception—its held-out high-noise diagnostic is non-monotonic (improves over the first $\sim$40 steps, then degrades), so for this one category we select the step-40 checkpoint that minimizes the held-out high-noise ($t \in \{800, 950\}$) diagnostic loss (Appendix A.6), which is 5.49% better than the final-step checkpoint. We highlight this rule explicitly because it is deterministic, diagnostic-driven, and reproducible; no test-set metric is ever consulted during selection.

Inference. DDIM sampling (Song et al., 2021) with 50 steps, strength $s = 0.75$, classifier-free guidance $w = 6.0$ (Ho & Salimans, 2022). Beyond resampling the diffusion noise, each generation call applies spatial augmentation (horizontal flip and $\pm 5^\circ$ rotation, jointly applied to image and mask; horizontal flip is disabled for the orientation-sensitive categories *capsule* and *pill*, whose printed text would otherwise appear mirrored) and stochastic shape augmentation of the mask itself (erosion, Gaussian-blur re-thresholding, or small internal holes, selected by relative defect area). For *capsule* and *pill*, whose masks frequently overlap printed text, an overlap-strength parameter (0.30) down-weights the free-generation region where it intersects a text-detection heuristic, preserving original characters instead of hallucinating glyphs. Per-category fine-tuning completes in 14–35 minutes on a single A100. We place strong emphasis on reproducibility: upon acceptance we will publicly release the full source code of our method, the per-category generation scripts, and implementations of all baseline methods used in the paper, together with scripts to reproduce every experiment and evaluation protocol reported here.

A.2 Relationship to prior blending methods

The central claim of the paper is that the *training side* and the *sampling side* of mask-guided diffusion should speak the same language. A reviewer may ask whether our stepwise blended latent sampling is simply RePaint (Lugmayr et al., 2022), Blended Diffusion (Avrahami et al., 2022), or Blended Latent Diffusion (Avrahami et al., 2023) with new names. The update-equation view makes the difference precise.

The two prior blending methods fix a single mask softness m across the entire reverse trajectory and have no counterpart on the training side: the model never sees time-dependent mask softness during training, so at sampling time the hard mask is applied post-hoc to a model that was never trained under it. TDMR removes precisely this inconsistency: the training loss schedules mask softness through $\tau(t)$, and the sampling update re-applies the same mask at the same resolution throughout. The stepwise blend is therefore not an additional heuristic stacked on an existing recipe; it is the inference-side instantiation of the same principle, and it is effective only in the regime where the training-side schedule is active (Ablation B in Appendix A.5).

Formal reconciliation: how the softened training schedule converges to the hard sampling cut. The training-side schedule and the sampling-side operator are two ends of one curve rather than two independent objects.

Proposition 1 (Training–sampling mask consistency). *Let $\tilde{m}_z(t) = \sigma\big((m_z - 0.5)/\tau(t)\big)$ be the softened mask of Eq. (1) with $\tau(t) = \tau_{\min} + (\tau_{\max} - \tau_{\min})\, t/T$ and $[\tau_{\min}, \tau_{\max}] = [0.05, 0.5]$. For every latent location i, the softened mask converges pointwise to the hard binary mask as $t \to 0$:*

$$\lim_{t \to 0} \tilde{m}_z(t)_i = \lim_{t \to 0} \sigma\left(\frac{m_{z,i} - 0.5}{\tau_{\min} + (\tau_{\max} - \tau_{\min})\, t/T} \right) = \begin{cases} 1, & m_{z,i} = 1 \\ 0, & m_{z,i} = 0 \end{cases}, \tag{4}$$

hence the hard mask composed at every sampling step (Eq. (3)) equals $\lim_{t \to 0} \tilde{m}_z(t)$: the boundary the sampler imposes is the small-t limit of the boundary the model was trained to respect.

Proof. Since $\tau(t)$ is continuous and strictly increasing in t with $\tau(t) \to \tau_{\min}$ as $t \to 0$, the argument $(m_{z,i} - 0.5)/\tau(t)$ diverges to $+\infty$ when $m_{z,i} = 1$ and to $-\infty$ when $m_{z,i} = 0$ as $t \to 0$. Because the logistic sigmoid satisfies $\sigma(x) \to 1$ as $x \to +\infty$ and $\sigma(x) \to 0$ as $x \to -\infty$, the limit follows. The residual transition band has width $O(\tau(t))$, vanishing linearly with $\tau(t)$. □

Input: source image x, binary mask m, strength s, guidance scale w
Output: generated latent $\hat{z}_0$

$z_0 \leftarrow \mathcal{E}(x)$; $m_z \leftarrow \text{downsample}(m)$
$t_{\text{start}} \leftarrow \lceil sT \rceil$; draw $\epsilon \sim \mathcal{N}(0, I)$
for $t = t_{\text{start}}$ down to 1 **do**
  $z_t^{\text{den}} \leftarrow \text{denoise}(z_t, t, c;\, w)$ (one DDIM step, classifier-free guidance)
  $z_t^{\text{noise}} \leftarrow \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$ (re-noise the source latent)
  $z_t \leftarrow z_t^{\text{den}} \odot m_z + z_t^{\text{noise}} \odot (1 - m_z)$ (mask re-anchoring, Eq. 3)
end for
return $\hat{z}_0$

Algorithm 1: Stepwise blended latent sampling.

Table 7: Per-category pixel-level AP (AP-P, %), CutPaste vs. AnomalyDiffusion vs. TDMR.

Category	CutPaste	AnomalyDiffusion	TDMR
bottle	85.52	78.02	**86.10**
cable	26.47	**73.03**	64.62
capsule	10.37	32.02	**32.60**
carpet	60.37	23.50	**72.54**
grid	24.18	6.97	**30.18**
hazelnut	62.68	**88.95**	76.69
leather	72.24	63.40	**74.27**
metal_nut	97.17	**97.38**	96.78
pill	85.11	**93.40**	89.79
screw	**27.38**	10.95	21.28
tile	92.52	91.17	**94.43**
toothbrush	42.66	**55.54**	26.34
transistor	**76.58**	64.07	57.91
wood	65.53	**76.47**	67.23
zipper	53.14	75.11	**78.09**
Mean	58.79	62.00	**64.59**

Thus the hard binary mask m_z composed at every sampling step is exactly $\lim_{t\to 0} \tilde{m}_z(t)$: training and sampling instantiate a single notion of mask fidelity rather than two unrelated operators. In the opposite regime (large t), the model is trained under a deliberately relaxed mask, which is why it can freely synthesize coherent defect structure before the hard boundary is imposed at low noise. This pointwise limit is the formal content of the training–sampling consistency principle; no additional assumption is introduced.

A.3 Complete downstream comparison

Claim. TDMR is the best among protocol-matched few-shot generation methods on the detection-oriented metrics that reward localization (Pixel AP, F1-max, IoU, PRO), while being competitive on image-level metrics; the single metric on which AnomalyDiffusion leads (Pixel AUROC) is a rank-based threshold-sensitive quantity, and the gap is 1.3 points (0.9093 vs. 0.9221).

Per-category AP-P (%). Bold marks the best of the three per row.

Reading of the per-category table. The gains are not concentrated in one class: TDMR wins 9 of the 15 categories outright, and loses by more than 5 points to the best method on only *cable* (-8.4 vs. AD), *toothbrush* (-29.2 vs. AD), *transistor* (-18.7 vs. CutPaste), and *wood* (-9.2 vs. AD). These four are precisely the failure regime discussed in Appendix A.6: cable/transistor/wood are high-frequency-texture or specular categories, and toothbrush is the category with the smallest visual defect area in the dataset.

Table 8: Per-category pixel-level AP (AP-P, %), real-only vs. TDMR (synthetic-only) vs. TDMR + real (mixed training).

Category	Real-only	TDMR	TDMR + real
bottle	84.15	86.10	**91.81**
cable	52.79	**64.62**	63.56
capsule	8.17	32.60	**36.08**
carpet	38.33	**72.54**	69.70
grid	7.34	**30.18**	27.14
hazelnut	74.14	**76.69**	74.31
leather	72.63	74.27	**79.62**
metal_nut	97.93	96.78	**97.94**
pill	74.34	**89.79**	82.54
screw	29.83	21.28	**31.35**
tile	93.24	**94.43**	93.28
toothbrush	11.35	26.34	**41.20**
transistor	17.44	57.91	**86.83**
wood	70.59	67.23	**71.60**
zipper	75.54	78.09	**83.85**
Mean	53.85	64.59	**68.70**

A.4 Mixing with the real few-shot pool

Claim. The synthesized data is not merely a substitute for real samples; it is complementary. Training on the real pool together with TDMR's output (1:1 with real normal images) yields 68.70% mean pixel AP, above synthetic-only (64.59%) and well above the real-only floor (53.85%). Mixing improves over the real-only floor in every category and over both ingredients simultaneously in most categories, indicating that TDMR's synthesized defects carry information orthogonal to the real pool rather than redundant to it.

Per-category AP-P (%): real vs. TDMR vs. mixed training. Table 8 reports the per-category values underlying the 68.70% mean of Table 3 in the main paper, together with the two ingredient configurations; the mean row therefore reproduces the headline mixing number exactly. Bold marks the best of the three per row; the final row is the 15-category mean.

Reading of the per-category table. Mixing improves over the real-only floor in every one of the 15 categories (15/15), confirming the complementary claim in the strictest form. Against synthetic-only training, mixing improves 9 of 15 categories; the 6 categories in which mixing is slightly below TDMR synthetic-only are *cable* (-1.06), *carpet* (-2.84), *grid* (-3.04), *hazelnut* (-2.38), *pill* (-7.25), and *tile* (-1.15), where the real pool contributes limited additional information over the synthetic defects. The most striking gains are on *transistor* ($+28.92$ over synthetic-only, $+69.39$ over real-only), *toothbrush* ($+14.86$), and *capsule* ($+3.48$), the categories where the few-shot real pool alone is weakest.

A.5 Full core-mechanism ablation

Claim. Both mechanisms are individually necessary and jointly sufficient: the dynamic schedule beats the fixed-softness baseline on the held-out high-noise diagnostic (Ablation A), and stepwise blending reduces boundary-region pixel L2 across all 15 categories, by 12.5% on average (Ablation B). Where the dynamic schedule does not help (capsule, leather, wood), the boundary reduction is still present, showing the two mechanisms address different failure modes.

Ablation C: shape-augmentation on/off (per-defect-type values underlying Table 5 in the main paper). IC-LPIPS measures within-defect diversity (cluster centers = first $\lfloor |\text{real}|/3 \rfloor$ real images, pairwise LPIPS within clusters, VGG backbone); Boundary-L2 uses the width-5 ring of the original mask, RGB L2, matching Ablation B. Identical adapters, base images, and initialization are used across the two branches.

Background-preservation control (per-defect-type values underlying the background-preservation discussion in Section 4.5 of the main paper). Background is defined as pixels

Table 9: Ablation C: shape-augmentation on/off, per-defect-type values.

Defect type	IC-LPIPS w/ aug	IC-LPIPS w/o aug	Boundary-L2 w/ aug	Boundary-L2 w/o aug
capsule/crack	0.1155	0.0431	17.54	10.41
capsule/faulty_imprint	0.1283	0.0461	18.84	15.99
capsule/poke	0.1175	0.0433	13.15	12.29
capsule/scratch	0.1250	0.0424	13.22	11.52
capsule/squeeze	0.1095	0.0513	15.67	10.93
tile/crack	0.2959	0.0555	54.34	30.31
tile/glue_strip	0.4195	0.1435	49.53	26.66
tile/gray_stroke	0.3274	0.0530	40.31	26.36
tile/oil	0.3577	0.1561	43.93	25.57
tile/rough	0.3222	0.1109	44.86	26.10
leather/color	0.2292	0.0353	23.96	22.02
leather/cut	0.2201	0.0397	25.95	21.21
leather/fold	0.2384	0.0456	28.38	20.28
leather/glue	0.2018	0.0403	22.20	19.68
leather/poke	0.2548	0.0391	34.60	26.64
Mean	**0.231**	**0.063**	**29.8**	**20.4**

Table 10: Background-preservation control, per-defect-type values.

Defect type	TDMR bg-L2	FREE bg-L2	VAE bg-L2	TDMR fg-L2
capsule/crack	8.69	32.55	6.34	43.40
capsule/faulty_imprint	8.85	31.89	6.21	39.44
capsule/poke	8.55	31.68	6.28	63.84
capsule/scratch	8.60	31.77	6.29	17.05
capsule/squeeze	7.82	27.24	6.21	37.43
tile/crack	43.46	78.82	11.23	78.96
tile/glue_strip	37.75	76.33	11.04	76.36
tile/gray_stroke	39.62	69.22	11.22	66.53
tile/oil	46.28	78.22	11.10	73.27
tile/rough	47.14	80.48	11.14	57.67
leather/color	19.73	72.17	15.17	45.95
leather/cut	20.38	68.13	15.02	66.62
leather/fold	19.98	72.46	13.56	48.17
leather/glue	16.82	65.72	11.33	59.66
leather/poke	21.83	77.42	13.61	40.92
Mean	**20.4**	**63.3**	**9.8**	**55.5**

outside the original mask; pixel L2 is computed between the source image and the generated image, mean over 20 samples per defect type. FREE = same pipeline without the mask-blending constraint; VAE = encoder–decoder reconstruction of the source image (the reconstruction floor). Identical base images, prompts, LoRA adapters, and initialization across the three variants.

Generation-quality metrics. Following the few-shot protocol we use IC-LPIPS (Zhang et al., 2018) and FID (Heusel et al., 2017) to assess generative fidelity independently of the downstream detector. TDMR attains substantially higher within-defect diversity than CutPaste (IC-LPIPS 0.2068 vs. 0.0298) while matching its distributional fidelity (FID 52.89 vs. 52.01), and far surpasses AnomalyDiffusion on both axes (IC-LPIPS 0.3004, FID 111.36). Per-category FID values below show that TDMR attains the best per-category FID in 11 of the 15 categories; CutPaste wins only on four texture-dominated categories, where its lower FID reflects the collapsed distribution of copy-pasted patches (its IC-LPIPS is an order of magnitude below TDMR's) rather than superior fidelity.

Table 11: Per-category FID (lower is better).

Category	CutPaste FID	AnomalyDiffusion FID	TDMR FID
bottle	55.96	109.78	**41.05**
cable	61.60	124.50	**58.10**
capsule	29.24	44.81	**19.34**
carpet	**34.79**	118.12	52.88
grid	46.39	114.33	**42.72**
hazelnut	44.75	78.44	**39.82**
leather	**44.33**	172.68	68.97
metal_nut	**31.56**	127.20	48.54
pill	43.01	63.57	**26.31**
screw	13.99	37.09	**13.25**
tile	**84.46**	224.36	122.27
toothbrush	79.28	89.55	**75.79**
transistor	53.25	127.40	**52.79**
wood	115.70	149.98	**102.22**
zipper	41.78	88.64	**29.28**
Mean	52.01	111.36	52.89

A.6 Failure regime of the hard-mask assumption

We do not cherry-pick only favorable outputs. The categories on which TDMR leaves the most visible residual artifacts are exactly those with the smallest per-category gains in Appendix A.5: synthesized text-like defects (e.g., *capsule*, *pill*) can exhibit slight glyph distortion, and strong specular reflections (e.g., *metal_nut*) can bleed across the mask boundary. These cases bound the regime in which the learned hard-mask guidance remains approximate (Section 5 of the main paper): when the defect region contains highly structured content that the frozen prior cannot easily re-synthesize faithfully (text glyphs, specular highlights), the hard-mask composition at small t can leak.

Training curves. The per-step training loss is highly variable for every category (isolated spikes 2–4$\times$ the local trend for *capsule*, *pill*, *carpet*, *wood*), which motivates the EMA-smoothed early-stopping criterion (Section 4.4). Figure 4 plots the raw and EMA-smoothed ($\alpha = 0.9$) per-step TDMR training loss for all 15 categories under the final configuration.

The tile checkpoint diagnostic. *tile* is the single category whose held-out high-noise diagnostic is non-monotonic (improves over the first $\sim$40 steps, then degrades); we select the step-40 checkpoint for this one category (a 5.49% improvement over the final-step checkpoint), while all other 14 categories use the standard post-early-stopping checkpoint. Figure 5 shows the held-out high-noise loss trajectory with the selected checkpoint marked, decomposed across all six timestep bins.

A.7 Additional qualitative results and diversity

Additional qualitative results. The main-paper comparison (Figure 3) shows one defect type for each of four categories; Figures 6–13 extend it to one held-out sample per defect type, covering 14 of the 15 MVTec AD categories (all except *toothbrush*, covered quantitatively in Appendix A.3). Each row is labeled with its category and defect subtype directly on the figure. Within each row, the mask-conditioned methods (CutPaste, TDMR) synthesize the defect on the same normal image under the reference mask; TDMR additionally applies rotation/erosion-style mask augmentation, so its defect location may deviate slightly from the reference mask, whereas the AnomalyDiffusion panel is an officially released sample produced by its own good$\rightarrow$defect procedure, so it serves as a defect-realism reference, not a boundary-fidelity comparison. Each panel insets the reference binary mask; AnomalyDiffusion does not consume it. Column order matches Figure 3 (real, CutPaste, AnomalyDiffusion, TDMR).

Diversity. The per-seed samples complement the within-defect diversity evidence of Ablation C: across different sampling seeds TDMR varies the defect texture and shape while preserving the same semantic defect type and the same clean background, whereas the higher IC-LPIPS of AnomalyDiffusion corresponds to content drift rather than within-defect diversity.

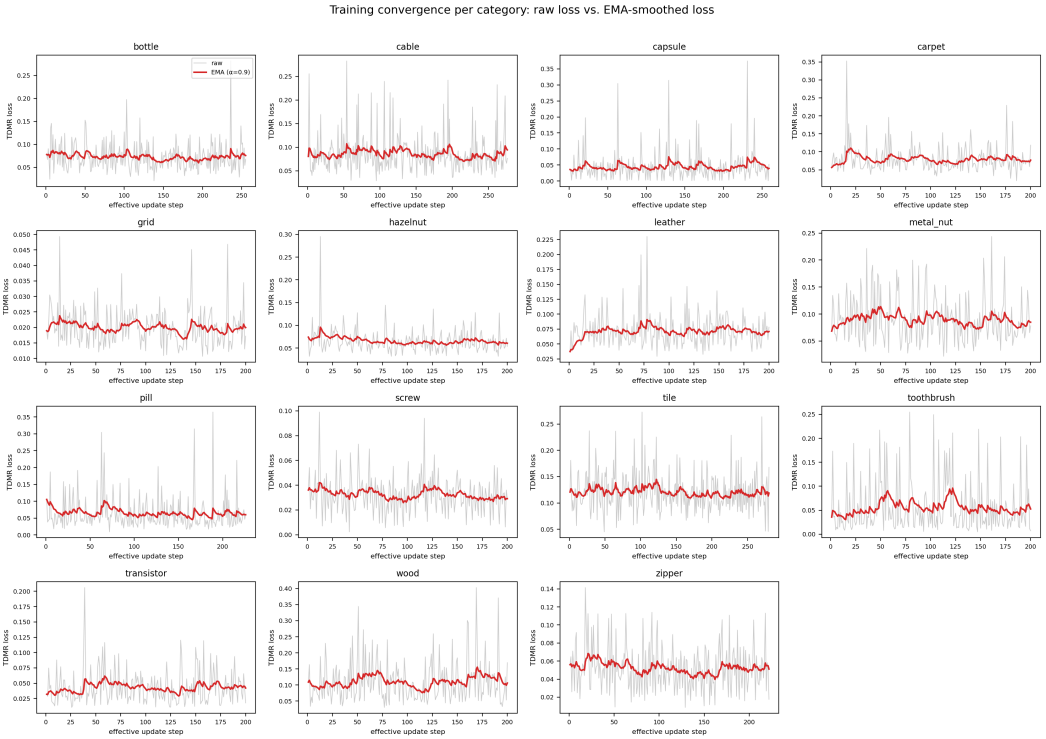


Figure 4: Per-step TDMR training loss (raw and EMA-smoothed, $\alpha = 0.9$) for all 15 categories, configuration C.

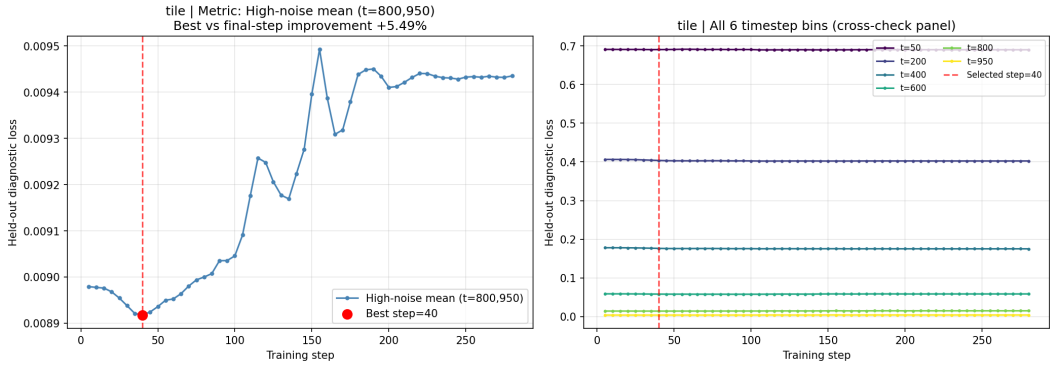


Figure 5: *tile* checkpoint diagnostic: the held-out high-noise loss improves over the first ∼40 steps then degrades, so this single category uses the step-40 checkpoint (marked), 5.49% better than the final-step checkpoint.

A.8 Reproducibility Statement

All implementation details required to replicate our results are provided in the main paper and the appendix. Generation runs at 512×512 resolution and the downstream detector at 256×256; raw generation metrics use the standard 299×299 Inception-V3 input size. All experiments were run on a single NVIDIA A100 GPU, taking 14–35 minutes per category. The experiments rely on publicly available pre-trained models, specifically the Stable Diffusion latent diffusion U-Net, identified in Section 3.2; full hyperparameters and training details are given in Appendix A.1, and the data protocol is specified in Section 4.1.

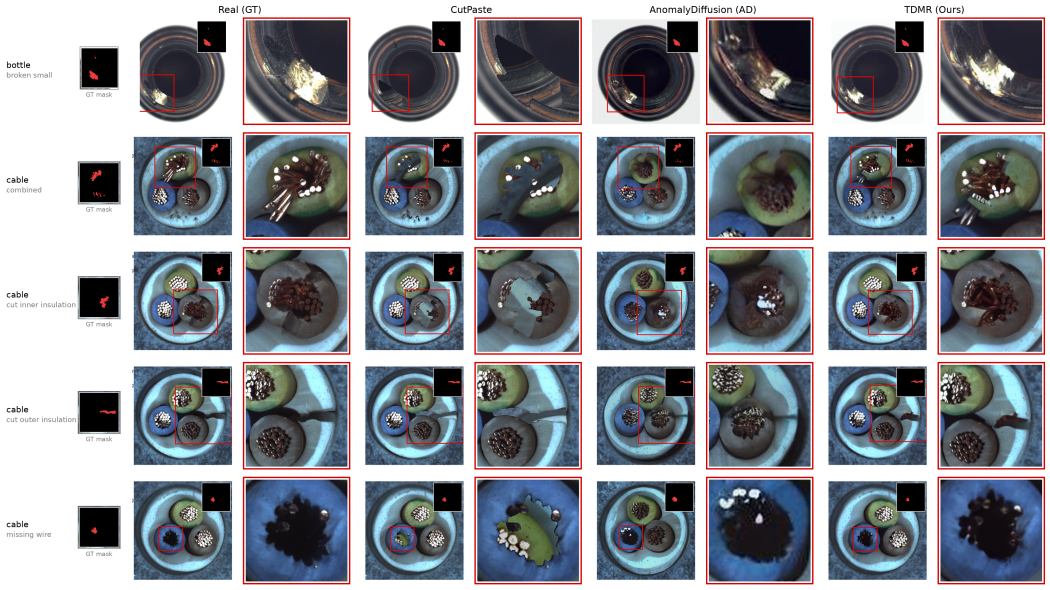


Figure 6: Additional qualitative comparison (1/8).

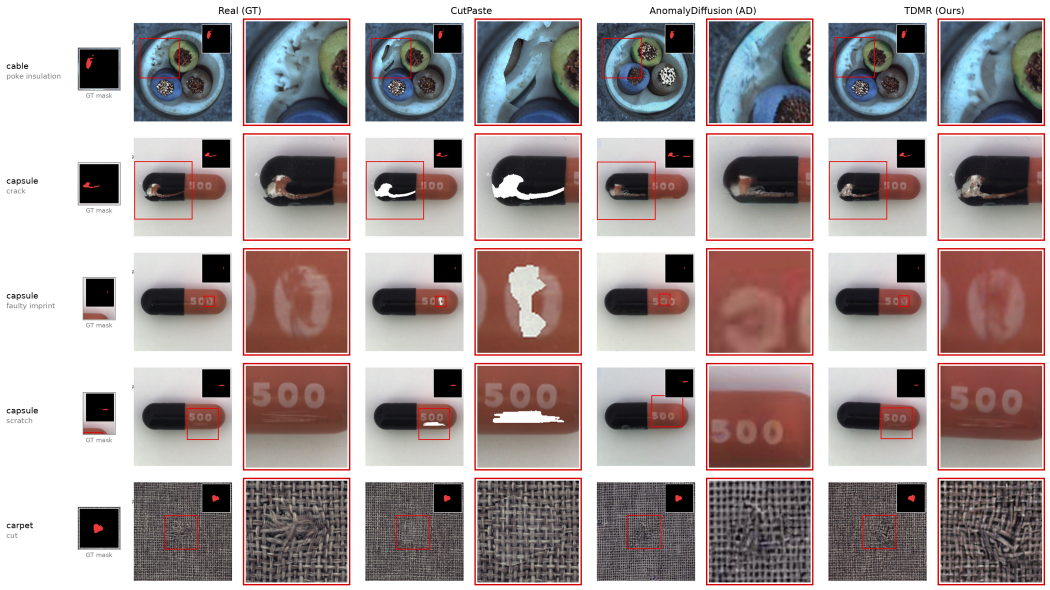


Figure 7: Additional qualitative comparison (2/8).

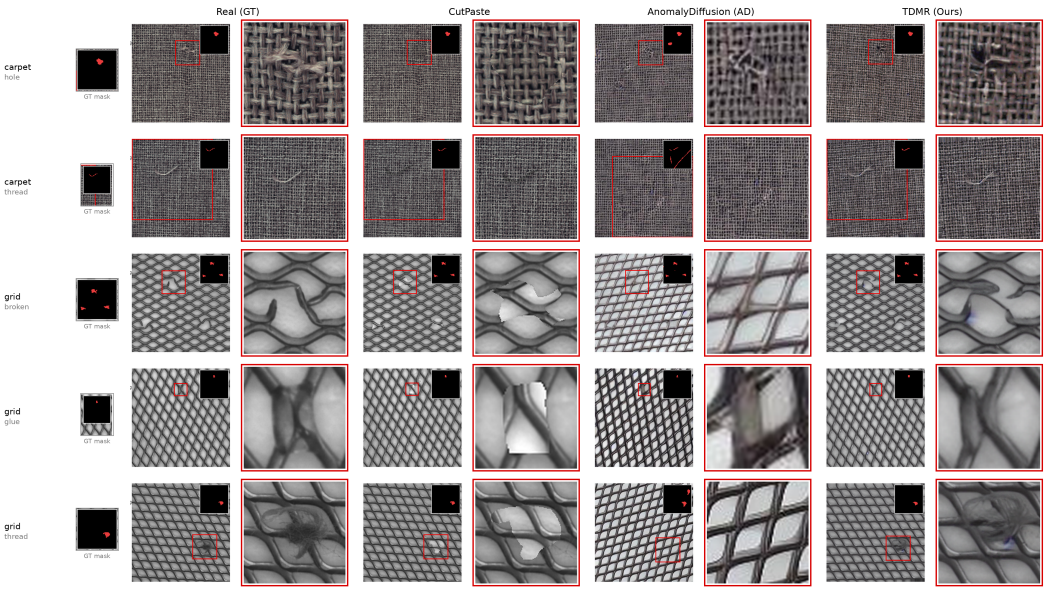


Figure 8: Additional qualitative comparison (3/8).

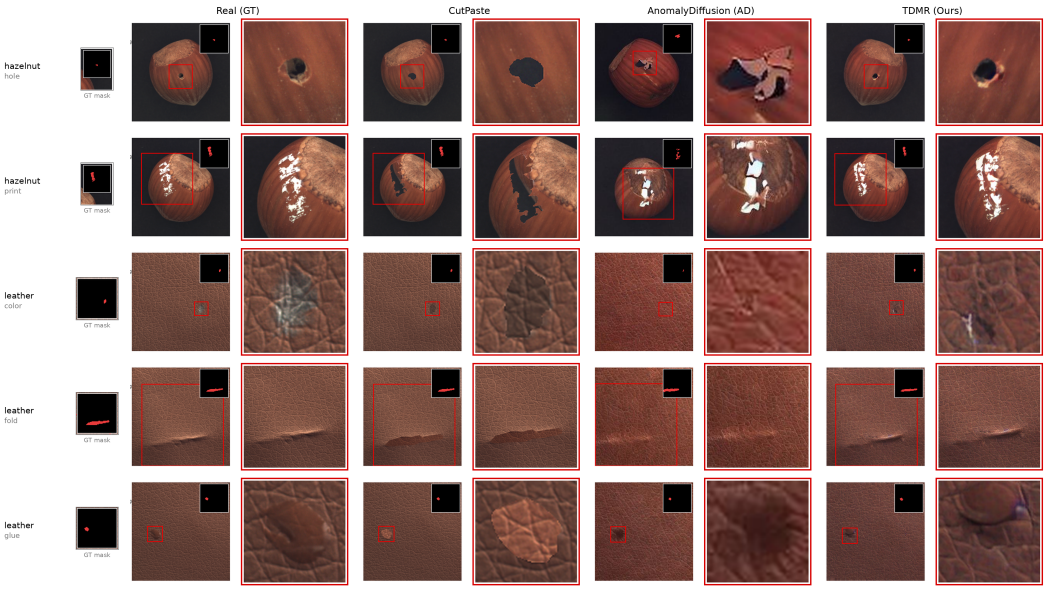


Figure 9: Additional qualitative comparison (4/8).

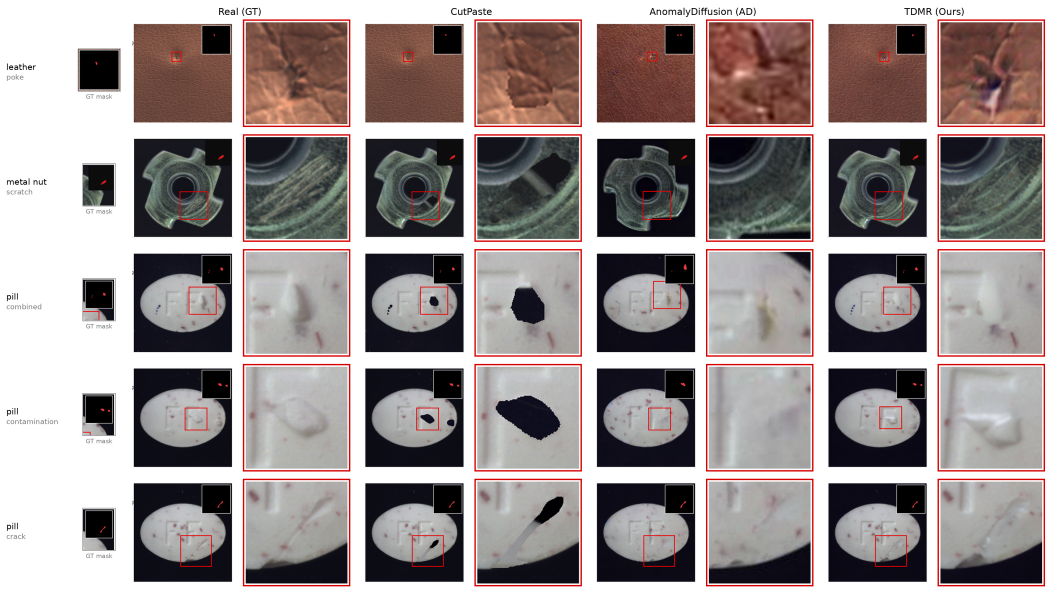


Figure 10: Additional qualitative comparison (5/8).

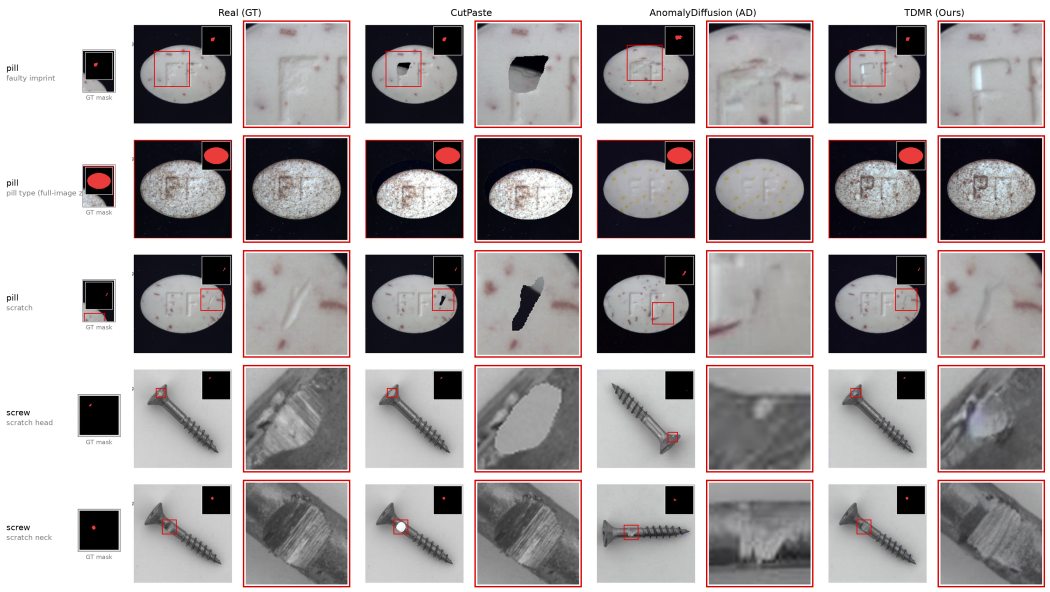


Figure 11: Additional qualitative comparison (6/8).

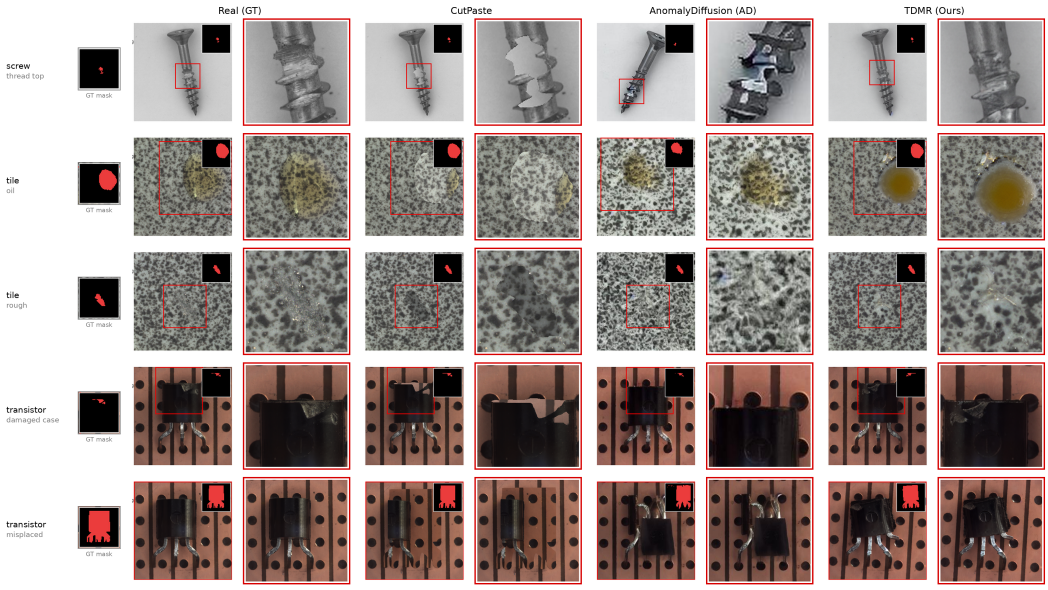


Figure 12: Additional qualitative comparison (7/8).

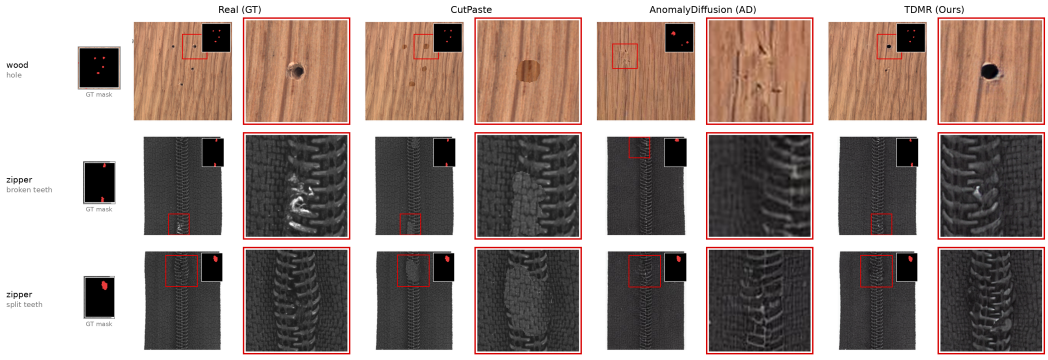


Figure 13: Additional qualitative comparison (8/8).