Quality-Aware Language-Conditioned Local Auto-Regressive Anomaly Synthesis and Detection

Long Qian[1,2], **Bingke Zhu**[1,2*], **Yingying Chen**[1,2], **Ming Tang**[1,2], **Jinqiao Wang**[1,2,3]

[1]Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Future Technology, University of Chinese Academy of Sciences, Beijing, China
[3]Objecteye Inc., Beijing, China
qianlong2024@ia.ac.cn, {bingke.zhu,yingying.chen,tangm,jqwang}@nlpr.ia.ac.cn

Abstract

Despite substantial progress in anomaly synthesis, existing diffusion-based and coarse inpainting pipelines commonly suffer from structural deficiencies such as micro-structural discontinuities, limited semantic controllability, and inefficient generation. To overcome these limitations, we introduce *ARAS*, a language-conditioned, auto-regressive anomaly synthesis approach that precisely injects local, text-specified defects into normal images via token-anchored latent editing. Leveraging a hard-gated auto-regressive operator and a training-free, context-preserving masked sampling kernel, *ARAS* significantly enhances defect realism, preserves fine-grained material textures, and provides continuous semantic control over synthesized anomalies. Integrated within our Quality-Aware Re-weighted Anomaly Detection (*QARAD*) framework, we propose a dynamic weighting strategy that emphasizes high-quality synthetic samples by computing an image-text similarity score with a dual-encoder model. Extensive experiments across three datasets, *MVTec AD*, *VisA*, and *BTAD*, demonstrate that our *QARAD* outperforms **SOTA** methods in both image- and pixel-level anomaly detection tasks, achieving improved accuracy, robustness, and a 5× synthesis speedup compared to diffusion-based alternatives.

Code — https://github.com/neymarql/QARAD
Extended version — https://arxiv.org/abs/2508.03539

Introduction

Anomaly detection remains critically challenged by the persistent and fundamental issue of *data imbalance*: while massive amounts of normal data are collected, anomalous samples are rare, and difficult to gather comprehensively in real-world environments. To bridge this gap, recent efforts have increasingly turned toward anomaly synthesis, utilizing powerful generative frameworks such as *Augmentation-based* (Li et al. 2021; Liu et al. 2023; Schlüter et al. 2022), *GANs* (Perera et al. 2019; Akcay, Atapour-Abarghouei, and Breckon 2018), and diffusion-based models (Wyatt et al. 2022; He et al. 2023). Despite significant progress, recent synthesis methods suffer from several pervasive limitations.

Recent diffusion-based anomaly synthesis methods (Zhang, Xu, and Zhou 2024; Wyatt et al. 2022) typically operate *globally* or via *coarse-grained inpainting*. A representative pipeline encodes the entire image (or masked crop) into a noisy latent, progressively denoises at a reduced working resolution, and finally upsamples and composites the region back into the original image. Traditional methods (Zavrtanik, Kristan, and Skočaj 2022, 2021; Tien et al. 2023) similarly rely on low-frequency residual modulation or random blob. While effective for gross structural corruption, these schemes struggle to *precisely respect the micro-structure of the underlying material*. See Fig. 1(a) which illustrates several recurring *symptoms* (*e.g.*, texture breaks and seam artifact), but the deeper *causal mechanisms* matter for designing better generators.

Recent methods suffer from high-noise overwriting and a resolution bottleneck. Coarse inpainting overwrites masked pixels with high-variance noise latents, denoises at a spatially compressed resolution, then upsamples. Each stage discards sub-pixel structure (*e.g.*, metal grain, fiber weave). When the patch is fused back, phase mis-match emerges at the boundary. This is a *structural* failure: fine-scale material statistics never propagate through the noisy, low-res bottleneck. As seen in Fig. 1(a), diffusion-based and traditional methods exhibit such artifacts. *Template discretization restricts the semantic granularity of synthesized defects.* Many pipelines provide only categorical control (*e.g.*, "scratch", "stain", or random blob masks (Li et al. 2024)). This discretizes a *continuous* defect manifold (*e.g.*, color shift, alignment to texture flow) into a handful of coarse prototypes *uninformed by the actual local micro-structure*. As a result, synthesized anomalies rarely conform to material-dependent cues that real detectors must learn. *Uniform training weights amplify the influence of poor-quality synthetic data.* Current detectors typically treat all synthetic samples equally during optimization. However, generations that deviate from their intended semantics (ambiguous prompt, collapsed texture) are often easier to fit and can produce disproportionately large gradients, steering the model toward artifactual cues. This *"bad drives out good"* effect degrades detection reliability. Fig. 1(b) visualizes our motivation of *QAW*.

Motivated by these critical gaps, we introduce a language-conditioned auto-regressive patch editor that synthesize anomaly region: ***A**uto-**R**egressive **A**nomaly **S**ynthesis* (***ARAS***), seamlessly integrated into our broader ***Q**uality-**A**ware **R**e-weighted **A**nomaly **D**etection* (***QARAD***)

*Corresponding author.

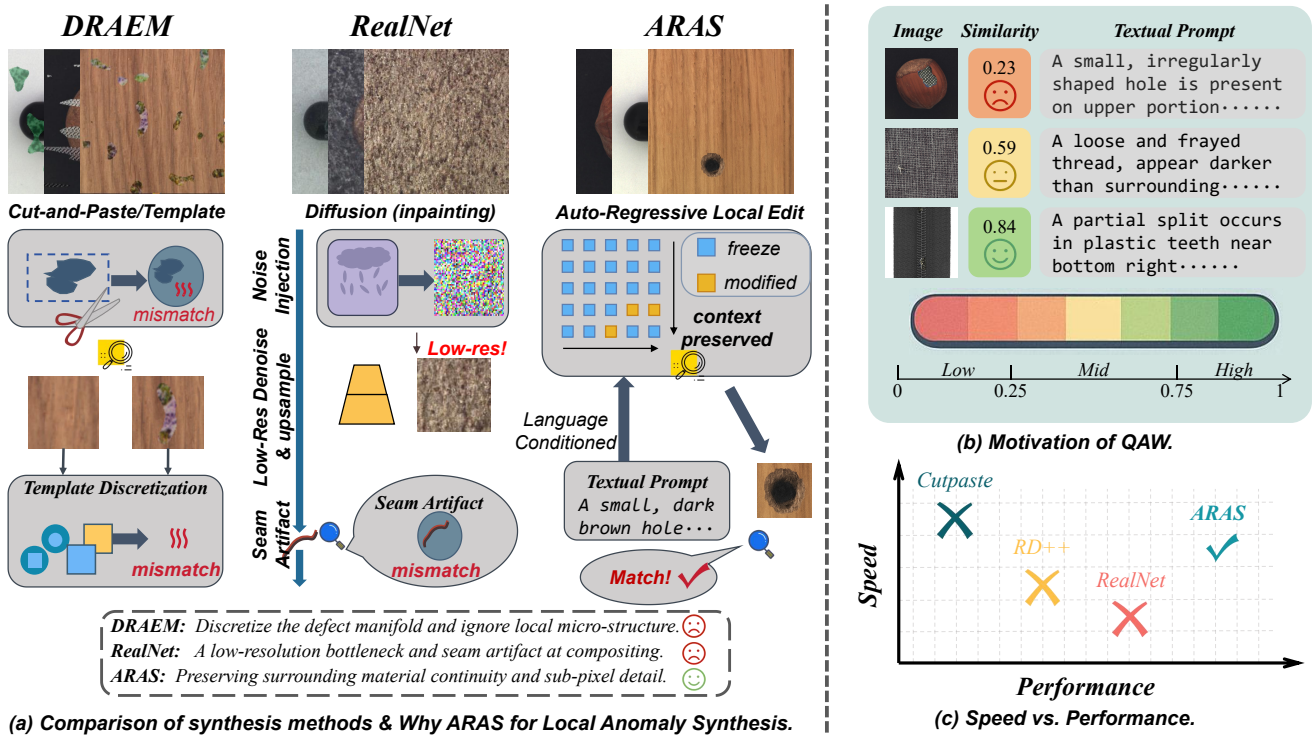


Figure 1: Motivation of *ARAS* and Quality-Aware Weighting (*QAW*) (a) Coarse anomaly synthesis vs. *ARAS*. *Cut-Paste/Template* and diffusion inpainting pipelines either discretize the defect manifold and ignore local micro-structure or pass through a noisy, low-resolution denoising bottleneck that yields seam artifacts when composited. Our *ARAS* freezes surrounding context tokens and inserts fine-grained defects aligned with material texture. (b) Sample quality varies. Text-image alignment reveals that many synthetic samples fail to express their prompts. (c) Speed vs. performance. *ARAS* achieves substantially faster synthesis than diffusion-based methods while improving downstream anomaly detection accuracy.

framework. At its core, *ARAS* leverages the *Infinity* architecture (Han et al. 2024), a powerful *8B AR* model operating on vector-quantized variational autoencoder (*VQ-VAE*) latent tokens. In contrast to prior coarse methods, *ARAS* precisely locks all tokens outside user-specified anomaly masks, ensuring only desired local pixels are synthesized. Coupled with detailed linguistic prompts, *ARAS* enables semantic and spatial control, synthesizing defects described by textual descriptions, while faithfully preserving the intricate textures and micro-structural fidelity of original images. Unlike previous approaches that implicitly learn textures through iterative refinement, *ARAS* explicitly leverages context-aware latent anchoring and linguistic prompts, effectively encoding material-specific priors into the anomaly synthesis process.

Embedded within our *QARAD* framework, we propose a novel *Quality-Aware Weighting* (*QAW*) strategy. Rather than uniformly treating all synthetic anomalies, *QAW* dynamically adjust the training weight of each synthetic sample based on the semantic consistency between its linguistic description and generated visual representation. Specifically, leveraging *CLIP*-based similarity scores, *QAW* systematically down-weight low-quality synthetic samples, mitigating their negative impact on the anomaly detector's training which substantially enhances detection stability and accuracy. This continuous re-weighting couples synthesis and detection in a single loop, improving robustness without discarding data diversity. Moreover, by adaptively calibrating the gradient landscape during training, *QAW* encourages the detector to prioritize subtle, high-fidelity cues over misleading coarse artifacts, yielding detectors that generalize significantly better to real-world anomalies.

Across *MVTec AD*, *VisA*, and *BTAD* benchmarks, *QARAD* consistently improves image- and pixel-level AUROC over diffusion-style synthesis baselines and traditional methods. Because we avoid iterative denoising, *ARAS* synthesizes a 1024^2-resolution defect image in a single forward pass, reducing per-sample generation latency by at least 5 times relative to diffusion inpainting (see Fig. 1(c)).

In summary, our contributions are three-fold:

- We propose *ARAS*, the first auto-regressive, language-driven anomaly synthesis method capable of precise, pixel-level local editing of defects in high-resolution industrial images, significantly surpassing traditional approaches in realism, controllability, and efficiency.
- We introduce the comprehensive anomaly detection framework *QARAD*, which integrates *ARAS* with a novel Quality-Aware Weighting scheme to selectively emphasize high-quality synthetic samples, thus stabilizing training and improving detection performance.

- Evaluations across datasets, *MVTec AD*, *VisA* and *BTAD*, demonstrate that *QARAD* achieves **SOTA** performances.

Related Work

Synthesis-Based Industrial Anomaly Detection. Because defective samples are scarce in manufacturing, many *AD* systems learn from synthetic anomalies composited onto normal imagery. Early but still influential industrial baselines such as *CutPaste* (Li et al. 2021) paste randomly cropped texture patches to simulate defects, improving data diversity for one-class detectors. *DRAEM* (Zavrtanik, Kristan, and Skočaj 2021) extends this idea with dual reconstruction and segmentation branches trained on synthetic corruptions blended into normal images, establishing a strong pixel-level *AD* benchmark. Recent work has pushed toward more realistic or targeted corruptions: *RD++* (Tien et al. 2023) augments residual defects at multiple scales to sharpen boundary cues and improve localisation. Diffusion models have recently been adapted for industrial anomaly synthesis; *RealNet* (Zhang, Xu, and Zhou 2024) leverages a diffusion-driven synthetic defect generator paired with a feature selection network to better match real failure texture statistics. *AnomalyDiffusion* (Hu et al. 2024) further conditions a denoising diffusion process on structural priors to inject class-aware yet diverse anomalous patterns for industrial inspection. Other approaches like *SPADE* (Yoon et al. 2022), *DiffusionAD* (Zhang et al. 2023), *TransFusion* (Fučka, Zavrtanik, and Skočaj 2024), *GLAD* (Yao et al. 2024) also demonstrate the utility of synthetic data, yet most operate through stochastic global corruption or coarse inpainting of large regions, which can blur fine material microstructure that is critical for anomalies, motivating our focus on mask-local, high-fidelity synthesis.

Language-Conditioned Anomaly Synthesis and Detection. Natural-language supervision offers a scalable channel for specifying rare or long-tail defect semantics beyond fixed class taxonomies. *PromptAD* (Li et al. 2024) explores using textual prompts to guide anomaly generation and improve category transferability in industrial *AD* settings. More broadly, open-vocabulary and *multimodal AD* systems like *AnomalyGPT* (Gu et al. 2023) are emerging: *AnomalyAny* (Sun et al. 2025) leverages large text–image models to support free-form prompt queries across multiple industrial datasets and shows that textual conditioning can retrieve or localise defect evidence without per-class retraining. Despite this progress, current language-aware pipelines (*e.g.*, *AdaCLIP* (Cao et al. 2024), *AnomalyCLIP* (Zhou et al. 2025), *AACLIP* (Ma et al. 2025) and *VCPCLIP* (Qu et al. 2024)) either map prompts to coarse class tokens or apply text conditioning at image-global scale; none tightly couple fine-grained spatial masks with rich textual attributes (size, color shift), which our *ARAS* framework addresses.

Auto-regressive Visual Generative Models. Large-scale auto-regressive (*AR*) decoders have recently re-emerged as competitive, scalable image generators. *VAR* (Tian et al. 2024) shows that token-factorized visual *AR* models achieve strong fidelity and flexible conditioning, and further explores visual auto-regressive modelling for detection tasks. *Infinity* (Han et al. 2024) scales this paradigm to multi-billion-parameter transformers operating directly over vector-quantized visual tokens, supports dynamic resolution schedules, and exposes interfaces for selective token resampling properties that make it naturally suited to local masked editing at test time. Our *ARAS* synthesis engine builds on these *AR* advances: by freezing all non-masked VQ tokens and sampling only the masked subset under language guidance, we produce high-detail, spatially targeted industrial defects substantially faster than iterative diffusion pipelines while preserving surrounding context.

Method

Overview

Fig. 2 depicts our *two-stage yet fully decoupled* pipeline. Let

$$x \in \mathbb{R}^{H\times W\times 3}, \quad m \in \{0,1\}^{H\times W}, \quad \tau \in \mathcal{T},$$

denote an original image, a binary anomaly mask, and a fine-grained text description, respectively. We introduce two deterministic operators:

$$\underbrace{\mathcal{G}_{\theta^\star}:\ (x,m,\tau)\ \longmapsto\ \widehat{x}}_{ARAS} \quad \text{and} \quad \underbrace{\mathcal{W}:\ (\widehat{x},\tau)\ \longmapsto\ w\in\mathbb{R}_+}_{QARAD},$$

where $\mathcal{G}_{\theta^\star}$ is a frozen large-scale auto-regressive editor that *replaces* the token subset indexed by m while *identity-mapping* the complement; $\mathcal{W}$ assigns a reliability weight to each synthesis via an image–text alignment score.

Denoting by $\mathcal{D}_\mathrm{r}$ the distribution of real samples and by $\mathcal{D}_\mathrm{s}$ the sampling measure over (x,m,τ) triplets, called $\mathcal{A}$, the detector f_ϕ is obtained by minimising the *weighted risk*

$$\min_\phi \mathbb{E}_{\mathcal{A}\sim\mathcal{D}_\mathrm{s}}\Big[\underbrace{\mathcal{W}\big(\mathcal{G}_{\theta^\star}\mathcal{A},\ \tau\big)}_{w}\cdot \ell\big(f_\phi(\mathcal{G}_{\theta^\star}\mathcal{A}),\ 1\big)\Big] + \mathbb{E}_{(x,y)\sim\mathcal{D}_\mathrm{r}}\big[\ell(f_\phi(x),\,y)\big] \tag{1}$$

with loss ℓ. Eq. (1) formalises the **decoupled design Philosophy**: *Generator modularity*, $\theta^\star$ is frozen; any stronger *AR* editor may replace $\mathcal{G}$ without touching detector training. *Quality adaptivity*, $\mathcal{W}$ down-weights prompt–inconsistent syntheses while preserving their diversity.

ARAS: Token-Anchored, Training-Free Local Edit

Token lattice. A normal image x is discretised by a fixed vector-quantiser $E:\mathbb{R}^{H\times W\times 3}\to\mathbb{Z}^{h\times w}$, yielding a latent lattice $z=E(x)=\{z_{ij}\}_{i,j}$. Let the binary anomaly mask be $m\in\{0,1\}^{H\times W}$ and define the token index sets

$$\mathcal{C}\ =\ \big\{(i,j)\mid m_{ij}=0\big\},\quad \mathcal{M}\ =\ \big\{(i,j)\mid m_{ij}=1\big\}.$$

Hard-gated auto-regressive operator. Denote by p_ϑ a *frozen* auto-regressive decoder over the token space. We introduce a *hard-gating* functional

$$\mathsf{Gate}_{\mathcal{C}}(q)\ :=\ \prod_{(i,j)\in\mathcal{C}}\delta\big[q(z_{ij})-z_{ij}\big], \tag{2}$$

where δ is the *Dirac mass* and q an arbitrary joint distribution on z. Eq. (2) forces unit probability on context tokens.

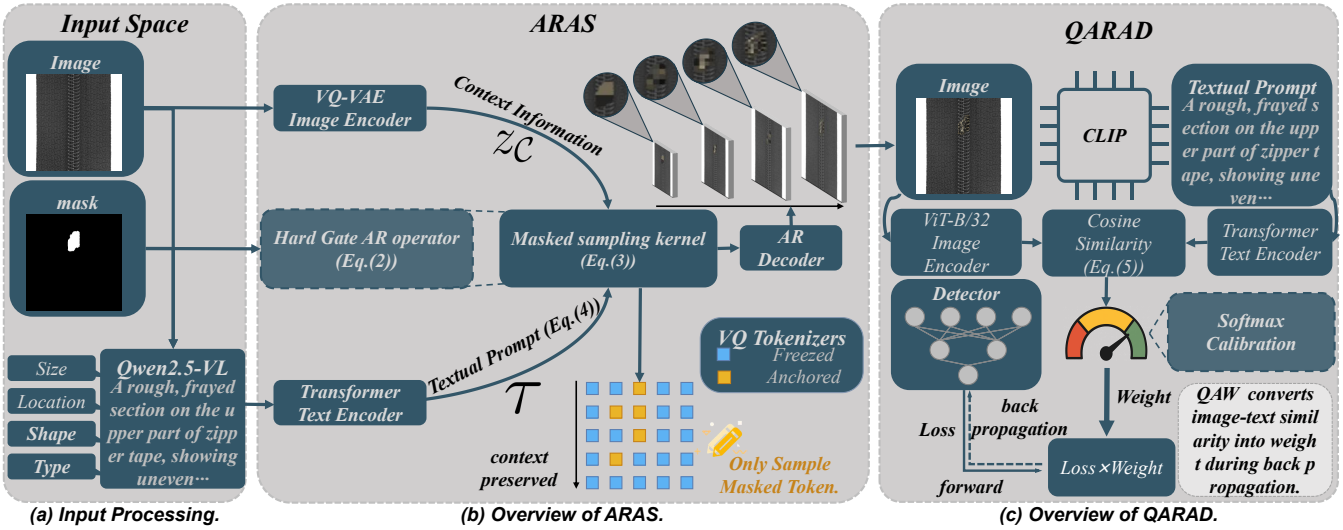


Figure 2: End-to-End Pipeline of *ARAS* and *QARAD* Framework. (a) *Input Processing*: raw image–mask-text triplets are generated and act as the subsequent input of our *ARAS* and *QARAD*; (b) *ARAS* inserts linguistically-specified local defects by freezing all context *VQ-tokens* and sampling only those indexed by the anomaly mask. (c) *QARAD* down-weights prompt-inconsistent synthesis images by scaling each sample's loss with a *CLIP*-based image–text similarity score.

Masked-AR kernel. Let π be a bijection $\{1,\dots,|\mathcal{M}|\}\to\mathcal{M}$. We define the *token-anchored sampling kernel*

$$\mathcal{K}_{\tau,\mathcal{C}} \;=\; \mathsf{Gate}_{\mathcal{C}}\Big(\prod_{t=1}^{|\mathcal{M}|} p_\vartheta\big(z_{\pi(t)} \mid z_{\mathcal{C}}, z_{\pi(<t)}, \tau\big)\Big) \qquad (3)$$

which leaves $z_{\mathcal{C}}$ untouched and successively samples masked positions under prompt τ, context $z_{\mathcal{C}}$ and tokens $z_{\pi(<t)}$. Sampling produces $\widehat{z} = \mathcal{K}_{\tau,\mathcal{C}}(z)$, subsequently decoded by D to obtain the locally edited image $\widehat{x} = D(\widehat{z})$.

Context-invariance guarantee. Immediate from the definition of $\mathsf{Gate}_{\mathcal{C}}(\cdot)$ in Eq. (2), for any (x, m, τ) and for any stochastic realisation of $\widehat{z}\sim\mathcal{K}_{\tau,\mathcal{C}}(z)$, $\widehat{z}_{ij} = z_{ij}\ \ \forall\,(i,j)\in\mathcal{C}$. Hence $\widehat{x}_{\mathcal{C}} = x_{\mathcal{C}}$, we guarantee context-invariance.

Prompt-conditioned micro-structural field. Let the text prompt be encoded as a point $\boldsymbol{\tau}\in\mathcal{T}\subset\mathbb{R}^d$, a smooth *semantic manifold*. Eq. (3) induces a stochastic operator

$$\Phi:\ \mathcal{T}\times\mathbb{Z}^{h\times w} \;\longrightarrow\; \mathcal{P}\big(\mathbb{Z}^{|\mathcal{M}|}\big), \qquad (\boldsymbol{\tau},\, z_{\mathcal{C}}) \;\mapsto\; p_{\boldsymbol{\tau},z_{\mathcal{C}}},$$

where $\mathcal{P}(\cdot)$ denotes the probability simplex. Under the cross-modal key–value projection of the *AR* decoder, Φ is *Lipschitz-continuous* in $\boldsymbol{\tau}$; an infinitesimal prompt edit (*e.g.*, defect length $+\varepsilon$ mm) perturbs the token logits by at most $L\varepsilon$ for some $L<\infty$. Sampling

$$\widehat{z}_{\mathcal{M}} \;\sim\; p_{\boldsymbol{\tau},z_{\mathcal{C}}} \qquad\Longrightarrow\qquad \widehat{z}_{\mathcal{M}} = \mathscr{F}_\tau(z_{\mathcal{C}}), \qquad (4)$$

therefore delivers a **semantically differentiable** field that *inherits* the stationary high-frequency statistics (grain, weave, gloss) of the frozen context $z_{\mathcal{C}}$.

In contrast, any method that re-encodes the entire patch (downsample $\to$ denoise $\to$ upsample) breaks this coupling and forfeits micro-structural coherence.

Our *ARAS* converts a generic auto-regressive decoder into a *local, prompt-selective anomaly synthesis editor* via the hard-gating mechanism in Eq. (2) and the masked-kernel operator of Eq. (3). The result is a ***training-free***, ***micro-structure preserving*** anomaly injector that enjoys exact locality and continuous attribute control through τ. These properties jointly address the structural deficiencies of augmentation-based methods (semantic-granularity loss) and diffusion-based methods (phase discontinuity).

QARAD: Quality-Aware Re-Weighting

Synthetic anomalies, while diverse, vary in their fidelity to the user-specified semantics. To prevent low-consistency outliers from dominating training, *QARAD* endows each synthetic sample with a continuous *reliability weight* w_i based on its image–text alignment, and integrates these weights into an importance-weighted risk estimator.

Consistency scoring as pseudo-density. Define a dual-encoder mapping into the unit hypersphere,

$$g_{\text{img}} : X \to \mathbb{S}^{d-1}, \quad g_{\text{txt}} : \mathcal{T} \to \mathbb{S}^{d-1}.$$

For each pair (x_i, τ_i), compute the cosine similarity

$$s_i \;=\; \big\langle g_{\text{img}}(x_i),\, g_{\text{txt}}(\tau_i)\big\rangle \;\in [-1,1]. \qquad (5)$$

We interpret s_i as an unnormalised proxy for the *likelihood ratio* between the true conditional defect distribution and the sampler's output, thereby guiding the weighting mechanism to prioritise semantically faithful samples.

Monotonic calibration φ. We then pass $\{s_i\}$ through a smooth, strictly increasing calibration function

$$\varphi:\ [-1,1] \;\longrightarrow\; \mathbb{R}_+,$$

to obtain weights

$$w_i \;=\; \frac{\varphi(s_i)}{\frac{1}{N}\sum_{j=1}^{N}\varphi(s_j)}, \qquad \mathbb{E}[w_i] = 1. \qquad (6)$$

Category	*DSR*	*SimpleNet*	*DRAEM*	*RD++*	*RealNet*	*QARAD (Ours)*
Candle	86.4/79.7	92.3/97.7	91.8/96.6	96.4/98.6	96.1/99.1	**97.8/99.9**
Capsules	93.4/74.5	76.2/94.6	74.7/98.5	92.1/99.4	93.2/98.7	**97.4/99.8**
Cashew	85.2/61.5	94.1/99.4	95.1/83.5	97.8/95.8	97.8/98.3	**98.6/99.9**
Chewinggum	97.2/58.2	97.1/97.0	94.8/96.8	96.4/99.0	**99.9**/99.8	**99.9/100**
Fryum	93.0/65.5	88.0/93.5	97.4/87.2	95.8/94.3	97.1/96.2	**99.4/99.9**
Macaroni1	91.7/57.7	84.7/95.4	97.2/**99.9**	94.0/99.7	**99.8/99.9**	99.4/**99.9**
Macaroni2	79.0/52.2	75.0/83.8	85.0/99.2	88.0/87.7	95.2/99.6	**97.4/99.8**
PCB1	89.1/61.3	93.4/99.1	47.6/88.7	97.0/75.0	**98.5/99.7**	**98.5**/99.4
PCB2	96.4/84.9	90.0/94.8	89.8/91.3	97.2/64.8	97.6/98.0	**99.4/99.9**
PCB3	97.0/79.5	91.3/98.2	92.0/98.0	96.8/95.5	**99.1**/98.8	**99.1/99.9**
PCB4	98.5/62.1	99.1/94.5	98.6/96.8	**99.8**/92 .8	99.7/98.6	99.7/**99.7**
Pipe fryum	94.3/80.5	89.0/95.3	**100**/85.8	99.6/92.0	99.9/99.2	99.9/**99.9**
Avg.	91.8/68.1	89.2/95.3	88.7/93.5	95.9/90.1	97.8/98.8	**98.9/99.8**

Table 1: Performance comparison across different SOTA methods on *VisA* dataset. Bold text indicates the best performance among all method. The values in the form of *xx/xx* represent *image-level AUROC / pixel-level AUROC*.

Two effective choices are:

$$\varphi_{\text{soft}}(s) = \exp(\gamma s), \quad \varphi_{\text{hinge}}(s) = \max\{0,\, s - \beta\},$$

each offering a tunable parameter (γ or β) that sharpens the weight distribution around high-quality samples. We empirically adopt the *softmax* form because it down-weights low-consistency samples *smoothly* while granting them non-zero influence, thus retaining sample diversity and stabilising early optimization, whereas *hinge* can discard too much gradient signal when the quality score distribution is narrow.

Variance reduction guarantee. We will prove it as follows. Let $\ell_i = \ell\big(f_\phi(\mathcal{G}(x_i, m_i, \tau_i)), 1\big)$ and w_i be defined as in Eq. (6). If the loss ℓ_i is *positively correlated* with any monotone function of the similarity score s_i (*i.e.*, $\mathrm{Cov}\big(\ell_i, f(s_i)\big) > 0$), then classical importance-sampling theory (Owen 2013) immediately yields

$$\mathrm{Var}\big[w_i\, \ell_i\big] \;\le\; \mathrm{Var}[\ell_i],$$

with strict inequality whenever ℓ_i and s_i are not independent. Hence, weighting by $\varphi(s)$ *always reduces, or in the worst case, leaves unchanged, the variance of the synthetic risk term*, providing a formal justification for the empirical stability we observe, and guarantee variance reduction.

The proposed weighting scheme is *unbiased*, because the normalization $\mathbb{E}[w_i] = 1$ preserves the expected risk; *adaptive*, since high-consistency samples are automatically assigned larger gradient amplitudes and low-consistency ones are softly down-scaled; and *decoder-agnostic*, as any image–text dual encoder can replace $(g_{\text{img}}, g_{\text{txt}})$ without altering the formulation, *thereby providing a lightweight drop-in upgrade with virtually few computational overhead*.

Together, *ARAS* and *QARAD* form a fully *training-free generator + quality-adaptive learner* pipeline: *ARAS* supplies high-fidelity, micro-structure-aware anomalies, while *QARAD* ensures that only those samples which faithfully realize their textual intent dominate detector optimization.

Experiments

Experimental Setup

Datasets. We conduct all experiments on three public industrial anomaly detection benchmarks: *MVTec AD* (15 categories, $\sim$5.3 k images) (Bergmann et al. 2019), *VisA* (12 categories, $\sim$10.8 k images) (Zou et al. 2022), and *BTAD* (3 categories, $\sim$2.5 k images) (Mishra et al. 2021).

Evaluation metrics. Performance is reported with the two community standards: *Image-level AUROC* and *Pixel-level AUROC*. Higher is better for both evaluation metrics.

Implementation details. For each normal training image we sample *6* anomaly masks from the *MaPhC2F* dataset (Qian et al. 2025) and generate a fine-grained prompt using QWEN2.5-VL-32B-INSTRUCT (Bai et al. 2025), which is then passed to our *training-free ARAS* editor, developed by the public INFINITY-8B auto-regressive architecture (Han et al. 2024); all context tokens are hard-gated and only the masked *tokens* are resampled by our *masked-AR Kernel*, yielding a locally edited anomaly image.

For anomaly detection we adopt *RealNet* as the backbone and inject our *QAW* module: *CLIP* encodes the anomaly image and textual prompts to compute the similarity score; weights are obtained via the *softmax* calibration.

Comparison with SOTAs

Tab. 1, 2, 3 benchmark our *QARAD* method against the most competitive methods on *VisA*, *MVTec AD* and *BTAD* datasets. Our *QARAD* achieves new state-of-the-art performance in the most of all individual categories and improves the *dataset-level mean* to **98.9/99.8** (*VisA*), **99.7/99.8** (*MVTec AD*), and **96.7/98.0** (*BTAD*) in *image- / pixel-level AUROC* respectively surpassing the previous SOTA methods by *+1.1/+1.0 points on VisA dataset*, *+0.1/+0.8 points on MVTec AD dataset* and *+0.6/+0.1 points on BTAD dataset*.

The largest gaps appear on highly textured or fine-grained classes such as *Fryum* (+2.3/+3.7) and *Wood*(+0.7/+1.6), corroborating that *ARAS*'s token anchoring preserves

Category	*DRAEM*	*DSR*	*CutPaste*	*AnomalyDiffusion*	*RD++*	*RealNet*	*QARAD (Ours)*
Bottle	99.2/97.8	99.6/98.8	**100**/99.1	99.8/99.4	**100**/98.8	**100**/99.3	**100/99.9**
Cable	91.8/94.7	95.3/97.7	96.4/96.2	**100**/99.2	99.3/98.4	99.2/98.1	99.8/**99.8**
Capsule	98.5/94.3	98.3/91.0	98.5/99.1	99.7/98.8	99.0/98.8	99.6/99.3	99.9/**99.9**
Hazelnut	**100**/99.7	97.7/99.1	**100**/99.0	99.8/99.8	**100**/99.2	**100**/99.8	99.8/**99.9**
Metal Nut	98.7/99.5	99.1/94.1	99.9/98.0	**100/99.8**	**100**/98.1	99.7/98.6	99.8/**99.8**
Pill	98.9/97.6	98.9/94.2	97.2/99.0	98.0/99.8	98.4/98.3	99.1/99.0	**99.8/99.9**
Screw	93.9/97.6	95.9/98.1	92.7/98.5	96.8/97.0	**98.9/99.7**	98.8/99.5	98.5/99.6
Toothbrush	**100**/98.1	**100/99.5**	99.2/98.9	**100**/99.2	**100**/99.1	99.4/98.7	99.5/**99.9**
Transistor	93.1/90.9	96.3/80.3	99.4/96.3	**100**/99.3	98.5/94.3	**100**/98.0	99.7/**99.8**
Zipper	**100**/98.8	98.5/98.4	99.6/98.0	99.9/99.4	98.6/98.8	99.8/99.2	99.7/**99.9**
Carpet	97.0/95.5	99.6/96.0	99.2/98.4	96.7/98.6	**100**/99.2	99.8/99.2	99.4/**99.8**
Grid	99.9/**99.7**	**100**/99.6	**100**/99.2	98.4/98.3	**100**/99.3	**100**/99.5	99.6/99.6
Leather	**100**/98.6	99.3/99.5	**100**/99.4	**100**/99.8	**100**/99.5	**100**/99.8	99.7/**99.9**
Tile	99.6/99.2	**100**/98.6	99.9/97.6	**100**/99.2	99.7/96.6	99.9/99.4	99.7/**99.8**
Wood	99.1/96.4	94.7/91.5	99.0/95.0	98.4/98.9	99.3/95.8	99.2/98.2	**99.9/99.8**
Avg.	98.0/97.2	98.2/95.8	98.7/98.1	98.5/97.7	99.4/98.3	99.6/99.0	**99.7/99.8**

Table 2: Performance comparison across different SOTA methods on *MVTec AD* dataset.

Category	*SimpleNet*	*SPADE*	*RD*	*RD++*	*RealNet*	*QARAD (Ours)*
01	96.4/90.3	91.4/97.3	97.9/98.3	96.8/96.2	**100**/98.2	**100/98.5**
02	75.2/48.9	71.4/94.4	86.0/96.2	90.1/96.4	88.6/96.3	**90.5/96.7**
03	99.3/97.2	99.9/99.1	99.7/94.2	**100/99.7**	99.6/99.4	99.5/98.9
Avg.	90.3/78.8	87.6/96.9	94.5/96.2	95.6/97.4	96.1/97.9	**96.7/98.0**

Table 3: Performance comparison across different SOTA methods on *BTAD* dataset.

micro-structure that traditional pipelines cannot reconstruct. These categories contain filament-scale fibres or glossy coating patterns whose local phase is easily destroyed by the noise-re-encode–upsample cycle of diffusion and the rigid paste geometry of template methods. Because *ARAS* freezes all *out-of-mask VQ-tokens* by our *Masked AR Kernel*, the spectral statistics of the surrounding material flow seamlessly into the resampled region, preventing the artifacts that confuse our *QARAD* detectors. On the *BTAD* dataset, our *QARAD* also outperform the previous SOTA methods.

The aggregate evidence demonstrates that our pipeline delivers a *holistic advance*: the training-free *ARAS* editor provides structurally consistent anomalies with linguistically controllable properties, while the quality-aware reweighting training scheme *QAW* in *QARAD* converts that realism into consistent performance gains over a spectrum of object geometries, texture regimes, and defect scales.

Representative qualitative results of *ARAS*-generated anomalies and *QARAD* results, are provided in Fig. 3.

Efficiency & Computational Complexity

Tab. 4 reveals that our *training-free ARAS* editor synthesises a masked anomaly $5.0\times$ faster than the diffusion-based *SDAS* pipeline used in *RealNet* (1.49s *vs.* 7.51s per image at 1024^2 resolution). This speed-up stems directly from the linear–token sampling rule in Eq. (3): runtime scales with the mask size only, whereas diffusion must iterate through at least 50–100 denoising steps irrespective of the edited area.

Model	**Inference Time (s/it)**
Diffusion-based (*e.g.*, *RealNet's SDAS*)	7.51
***ARAS* (Ours)**	1.49

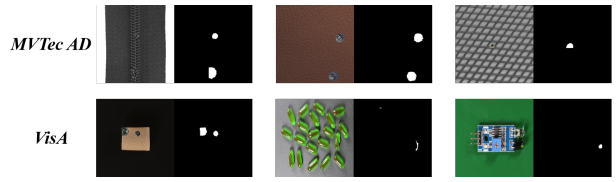
Table 4: Anomaly synthesis inference speed

Metric	*QARAD (w/o. QAW)*	***QARAD* (ours)**
Train Time (s/ep)	100.89	110.36 (+9.47)
Inference Time (s/it)	0.21	0.21
#Learnable Params (M)	524.11	524.11
#Total Params (M)	590.94	742.22 (+151.28)

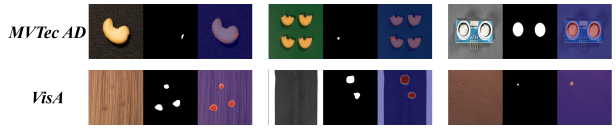
Table 5: Anomaly detection training & inference complexity

Turning to anomaly detection, Tab. 5 compares plain *RealNet* fine-tuning with our full *QARAD*. Integrating the *CLIP* backbone for image–text similarity scoring enlarges the total parameter budget by *151M*; however, these parameters are *frozen*, so the number of learnable weights remains unchanged *(524M)*. Consequently, the training time per epoch increases only slightly by *9.4s* (approximately *9%*), while the inference latency remains unchanged at *0.21s* per image, as the *QAW* module is exclusively utilized during training.

In aggregate, the proposed pipeline maintains *RealNet*'s



(a) *ARAS*-generated anomalies. Each shows: the locally edited image generated by *ARAS* and binary anomaly mask. Rows correspond to *MVTec AD* (top) and *VisA* (bottom) categories.



(b) *QARAD* detection results. From left to right: defective image, ground-truth mask, and *QARAD* anomaly heatmap.

Figure 3: Qualitative examples of (a) synthetic anomalies generated by *ARAS* and (b) anomaly detection results produced by *QARAD* on *MVTec AD* and *VisA* datasets.

real-time detection speed while accelerating anomaly synthesis by a factor of five, which is an attractive trade-off for industrial scenarios requiring both rapid data augmentation and swift deployment, without compromising performance.

Ablation Study

Method Variant	Image-AUROC	Pixel-AUROC	Comments
DREAM (DTD)	98.0	97.2	Original Synthesis Method
DREAM (ARSR)	98.6 ***(+0.6)***	97.9 ***(+0.7)***	Replaced with our *ARAS*
RealNet (SDAS)	99.6	99.0	Original Synthesis Method
RealNet (ARSR)	99.6 ***(+0.0)***	99.3 ***(+0.3)***	Replaced with our *ARAS*

Table 6: Ablation of *ARAS*.

QAW Variant	Image-AUROC	Pixel-AUROC	Comments
w/o. QAW	99.6	99.3	Uniform Weight
Hinge QAW	99.6***(+0.0)***	99.8***(+0.5)***	Clipped Linear Scaling
Softmax QAW	99.7***(+0.1)***	***99.9(+0.6)***	Softmax similarity weight

Table 7: Ablation of *QAW*.

Tab. 6, 7 disentangles the respective contributions of our anomaly synthesis method *ARAS* and our anomaly detection model *QARAD*'s training scheme *QAW*. All the ablation studies are experimented on the *MVTec AD* dataset.

Impact of *ARAS*. Replacing the augmentation-based and diffusion-inpainting pipelines in two representative baselines, *DRAEM* (originally *Perlin* masks with *DTD* texture blending) and *RealNet* (originally the *SDAS* synthesis method), with our training-free, token-anchored and language-conditioned editor *ARAS* produces consistent gains: *+0.6 image-AUROC* / *+0.7 pixel-AUROC* for *DRAEM*, and *+0.3 pixel-AUROC* for *RealNet* (Tab. 6). The larger margin on *DRAEM* is expected because its autoencoder backbone is highly susceptible to high-frequency seam artefacts that *ARAS* removes; by contrast, *RealNet* already achieves *image-level AUROC* very close to 100, so a modest additional improvement is both reasonable and informative. Qualitatively, *ARAS* yields synthetic defects whose spectral statistics align with those of the surrounding material, thereby suppressing spurious responses along texture boundaries and producing cleaner heat-maps.

Impact of *QAW*. Starting from *RealNet + ARAS*, we compare three weighting schemes (Tab. 7). Uniform weighting (*w/o. QAW*) serves as baseline. A *hinge* mapping that truncates low-similarity samples already lifts *pixel-AUROC* by *0.5*, indicating that text-image inconsistent anomalies indeed harm optimization. *Softmax* calibration delivers the best performance (*99.7 image- / 99.9 pixel-level AUROC*), demonstrating that *continuous attenuation* of moderate-quality samples is superior to hard rejection, preserving diversity while still focusing gradients on high-fidelity data.

Orthogonality of the two components. Because *ARAS* and *QAW* operate at distinct stages (data generation vs. loss re-weighting), their gains accumulate almost additively: the full *QARAD* configuration outperforms the *diffusion*-based by *+0.1 image-level AUROC / +0.6 pixel-level AUROC* on the *MVTec AD* dataset without increasing inference latency. This validates our design hypothesis that high-quality local edits and quality-aware training are mutually reinforcing.

Conclusions

Our work presents a fully anomaly synthesis and detection pipeline for industrial anomaly detection that integrates two novel components: (i) ***ARAS***, a token-anchored auto-regressive editor that inserts linguistically controlled, micro-structure-preserving defects without re-encoding the surrounding image context, and (ii) ***QARAD***, a quality-aware re-weighting training scheme that leverages image–text similarity to modulate the influence of each synthetic sample during detector optimization. Extensive experiments on three public benchmarks, *MVTec AD*, *VisA*, and *BTAD*, demonstrate that our method establishes new SOTA performance at both *image-* and *pixel-level AUROC*, while reducing anomaly synthesis latency by a factor of five relative to diffusion-based approaches. Ablation studies confirm that *ARAS* and *QARAD* contribute complementary gains, with *softmax* calibration yielding the most stable and accurate results. Beyond accuracy and efficiency, *ARAS* delivers fine-grained semantic controllability, enabling users to steer defect attributes (size, orientation, material phase) via natural language. Representative qualitative results highlight sharper, seam-free edits and cleaner detection heatmaps, reinforcing the quantitative findings. In addition, we will publicly release the large-scale image–mask–text dataset generated by our *ARAS*. Overall, our study shows that high-fidelity local synthesis *plus* principled sample re-weighting can significantly narrow the realism gap in synthetic anomaly data, paving the way for faster, more reliable industrial anomaly detection systems.

Acknowledgments

This work was supported by National Key R&D Program of China under Grant No.2023ZD0120400, in part by National Natural Science Foundation of China (No. 62276260, 62472423), and Beijing Natural Science Foundation (L252036).

References

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 2018. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. In *Proc. of ACCV Workshops*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

Cao, Y.; Zhang, J.; Frittoli, L.; Cheng, Y.; Shen, W.; and Boracchi, G. 2024. *AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection*, 55–72. Springer Nature Switzerland. ISBN 9783031727610.

Fučka, M.; Zavrtanik, V.; and Skočaj, D. 2024. TransFusion – A Transparency-Based Diffusion Model for Anomaly Detection. In *Proc. of ECCV (35)*, 91–108.

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2023. AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models. arXiv:2308.15366.

Han, J.; Liu, J.; Jiang, Y.; Yan, B.; Zhang, Y.; Yuan, Z.; Peng, B.; and Liu, X. 2024. Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. arXiv:2412.04431.

He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; Wang, Y.; Wang, C.; and Xie, L. 2023. DiAD: A Diffusion-based Framework for Multi-class Anomaly Detection. arXiv:2312.06607.

Hu, T.; Zhang, J.; Yi, R.; Du, Y.; Chen, X.; Liu, L.; Wang, Y.; and Wang, C. 2024. AnomalyDiffusion: Few-Shot Anomaly Image Generation with Diffusion Model. In *Proc. of AAAI*, volume 38, 8526–8534.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9664–9674.

Li, X.; Zhang, Z.; Tan, X.; Chen, C.; Qu, Y.; Xie, Y.; and Ma, L. 2024. PromptAD: Learning Prompts with only Normal Samples for Few-Shot Anomaly Detection. arXiv:2404.05231.

Liu, Z.; Zhou, Y.; Xu, Y.; and Wang, Z. 2023. SimpleNet: A Simple Network for Image Anomaly Detection and Localization. arXiv:2303.15140.

Ma, W.; Zhang, X.; Yao, Q.; Tang, F.; Wu, C.; Li, Y.; Yan, R.; Jiang, Z.; and Zhou, S. K. 2025. AA-CLIP: Enhancing Zero-shot Anomaly Detection via Anomaly-Aware CLIP. arXiv:2503.06661.

Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; and Foresti, G. L. 2021. VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 01–06.

Owen, A. B. 2013. *Monte Carlo theory, methods and examples*.

Perera, P.; Nath, R.; Venkateswara, H.; Panchanathan, S.; and Patel, V. M. 2019. OCGAN: One-Class Novelty Detection Using GANs with Constrained Latent Representations. In *Proc. of CVPR*, 2898–2906.

Qian, L.; Zhu, B.; Chen, Y.; Tang, M.; and Wang, J. 2025. MathPhys-Guided Coarse-to-Fine Anomaly Synthesis with SQE-Driven Bi-Level Optimization for Anomaly Detection. arXiv:2504.12970.

Qu, Z.; Tao, X.; Prasad, M.; Shen, F.; Zhang, Z.; Gong, X.; and Ding, G. 2024. VCP-CLIP: A visual context prompting model for zero-shot anomaly segmentation. arXiv:2407.12276.

Schlüter, H. M.; Tan, J.; Hou, B.; and Kainz, B. 2022. Natural Synthetic Anomalies for Self-Supervised Anomaly Detection and Localization. In *Proc. of ECCV*, volume 13691 of *LNCS*, 474–489.

Sun, H.; Cao, Y.; Dong, H.; and Fink, O. 2025. Unseen Visual Anomaly Generation. arXiv:2406.01078.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. arXiv:2404.02905.

Tien, T. D.; Nguyen, A. T.; Tran, N. H.; Huy, T. D.; Duong, S. T.; Nguyen, C. D. T.; and Truong, S. Q. H. 2023. Revisiting Reverse Distillation for Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24511–24520.

Wyatt, J.; Leach, A. D.; Schmon, S. M.; and Willcocks, C. G. 2022. AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models Using Simplex Noise. In *Proc. of CVPR Workshops*, 650–656.

Yao, H.; Liu, M.; Wang, H.; Yin, Z.; Yan, Z.; Hong, X.; and Zuo, W. 2024. GLAD: Towards Better Reconstruction with Global and Local Adaptive Diffusion Models for Unsupervised Anomaly Detection. arXiv:2406.07487.

Yoon, J.; Sohn, K.; Li, C.-L.; Arik, S. O.; and Pfister, T. 2022. SPADE: Semi-supervised Anomaly Detection under Distribution Mismatch. arXiv:2212.00173.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021. DRÆM: A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection. In *Proc. of ICCV*, 8330–8339.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2022. DSR – A Dual Subspace Re-Projection Network for Surface Anomaly Detection. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 539–554. Cham: Springer Nature Switzerland. ISBN 978-3-031-19821-2.

Zhang, H.; Wang, Z.; Wu, Z.; and Jiang, Y.-G. 2023. DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection. *arXiv preprint arXiv:2303.08730*.

Zhang, X.; Xu, M.; and Zhou, X. 2024. 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 (CVPR)*.

Zhou, Q.; Pang, G.; Tian, Y.; He, S.; and Chen, J. 2025. AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. arXiv:2310.18961.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, 392–408. Springer.