GroundingAnomaly: Spatially-Grounded Diffusion for Few-Shot Anomaly Synthesis

Yishen Liu[1], Hongcang Chen[1,2], Pengcheng Zhao[3], Yunfan Bao[1], Yuxi Tian[1], Jieming Zhang[4], Hao Chen[4], Zhi Zheng[4], Yongchun Liu[4], Ying Li[1⋆], and Dongpu Cao[5⋆]

[1] Beijing Institute of Technology, Beijing, China
ying.li@bit.edu.cn
[2] Hong Kong Polytechnic University, Hong Kong, China
[3] Beijing Jiaotong University, Beijing, China
[4] Li Auto, China
[5] Tsinghua University, Beijing, China
dongpu.cao@tsinghua.edu.cn

Abstract. The performance of visual anomaly inspection in industrial quality control is often constrained by the scarcity of real anomalous samples. Consequently, anomaly synthesis techniques have been developed to enlarge training sets and enhance downstream inspection. However, existing methods either suffer from poor integration caused by inpainting or fail to provide accurate masks. To address these limitations, we propose **GroundingAnomaly**, a novel few-shot anomaly image generation framework. Our framework introduces a **Spatial Conditioning Module** that leverages per-pixel semantic maps to enable precise spatial control over the synthesized anomalies. Furthermore, a **Gated Self-Attention Module** is designed to inject conditioning tokens into a frozen U-Net via gated attention layers. This carefully preserves pretrained priors while ensuring stable few-shot adaptation. Extensive evaluations on the MVTec AD and VisA datasets demonstrate that GroundingAnomaly generates high-quality anomalies and achieves state-of-the-art performance across multiple downstream tasks, including anomaly detection, segmentation, and instance-level detection. Code will be made publicly available from this URL.

Keywords: Image Generation · Anomaly Detection · Anomaly Generation

1 Introduction

Recently, visual anomaly inspection has demonstrated significant potential in industrial quality control in manufacturing [2]. Nevertheless, anomalous samples are scarce in real-world industrial production, which constrains the performance of anomaly inspection. To mitigate this, many methods adopt unsupervised

⋆ Corresponding author.

learning on abundant normal samples, detecting anomalies as deviations from the learned distribution [5, 10, 11, 16, 22, 27, 31, 38]. However, such approaches exhibit limited localization accuracy and do not provide class-aware anomaly information.

Such limitation motivates the development of anomaly generation techniques to augment existing datasets. Early model-free methods [21,37] synthesize pseudo-anomalies through data augmentation but suffer from low fidelity and defect appearance remains limited without specific classification. More recently, generative models [8, 13, 29, 32] have been adopted for anomaly synthesis. By leveraging large-scale pretrained priors, these models significantly enhance both visual realism and anomaly diversity. Existing model-based approaches fall into two categories. **Anomaly Generation (AG)** methods that synthesize isolated defect patches and edit them onto real backgrounds [9, 15], while they preserve background coherence, inpainting often yields poor integration and misaligned masks. **Anomaly Image Generation (AIG)** methods jointly synthesize objects and defects to improve global realism [6, 17, 40], but they still struggle to produce precise masks.

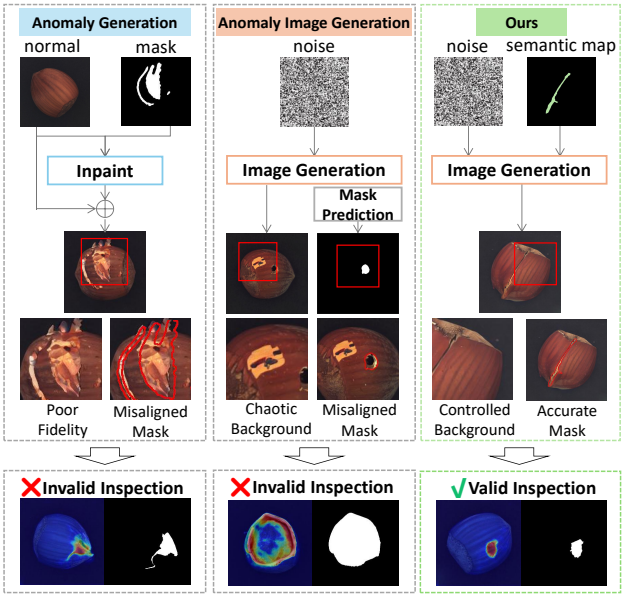

Fig. 1: Anomaly Generation methods inpaint anomalies onto normal images; **Anomaly Image Generation** methods jointly generate anomalies with products and predict masks after generation; our **GroundingAnomaly** grounds anomalies with semantic maps and generates the whole anomalous images.

These limitations motivate us to develop a diffusion framework that provides precise spatial **grounding** for **anomaly** image generation while preserving high visual fidelity. In this work, we propose **GroundingAnomaly**, an AIG framework that learns anomaly appearance and position representation in a few samples and generates realistic anomalous images with precise spatial control. GroundingAnomaly first learns a set of product and anomaly tokens and encodes a pixel-wise semantic map; these are merged in the Spatial Conditioning Module (SCM) to produce conditioning tokens. The conditioning tokens are injected into the diffusion U-Net and fused with the visual tokens in the transformer blocks via a Gated Self-Attention Module (GSM). The modules are trained on mixed batches of normal and anomalous images to leverage cross-domain appearance priors, enabling the model to generate diverse, high-fidelity defects with precise spatial grounding while maintaining global product consistency, which effectively enhances downstream anomaly inspection tasks.

Extensive experiments demonstrate that the proposed GroundingAnomaly generates high-quality anomalies with precise masks and improves downstream anomaly inspection performance reaching a pixel-level 99.3% AUROC and 85.9% AP score in anomaly segmentation on MVTec AD [1] dataset and 98.2% AUROC and 67.2% AP score on VisA [42] dataset.

Our main contributions are summarized as follows:

- We propose **GroundingAnomaly**, a few-shot AIG framework that achieves precise spatial and semantic control over defect synthesis while preserving coherent backgrounds.
- We introduce the **Spatial Conditioning Module**, which fuses disentangled product and anomaly tokens with a pixel-wise semantic map, and the **Gated Self-Attention Module**, which injects the spatial conditioning into a frozen U-Net.
- GroundingAnomaly is evaluated on MVTec AD [1] and VisA [42] using multiple downstream models for anomaly detection, segmentation and instance-level anomaly detection. Experimental results demonstrate state-of-the-art generation quality and substantial improvements in downstream anomaly inspection.

2 Related Works

2.1 Anomaly Inspection

Anomaly Inspection is critical for maintaining product quality in modern manufacturing [12]. A common approach is to train **supervised** object-detection or segmentation models on annotated anomalous images to perform instance-level detection or anomaly segmentation [3,18,28,33,34,36]. However, real anomalous samples are rare and highly scarce in practice, so supervised approaches often suffer from poor generalization and are constrained by the substantial cost of collecting pixel-accurate annotations.

Unsupervised methods have been proposed to alleviate this issue. These methods are trained solely on normal examples and detect anomalies by measuring deviations from the learned data distribution [5, 10, 11, 31, 38]. Recent few-shot variants further exploit large-scale pretrained models to perform inspection with only a few exemplars [16, 16, 22, 25, 27, 35, 41]. Nevertheless, these approaches only segment anomalies from images without class information, and their performances are still constrained by the representation learned from normal images.

2.2 Anomaly Synthesis

Many anomaly synthesis methods have been developed to generate realistic anomalous data for training inspection models. Early model-free methods [21,37] synthesize pseudo-anomalies through data augmentation but suffer from low fidelity and inconsistent defect patterns. More recently, generative models [8, 13, 29, 32] have been adopted for anomaly synthesis. By leveraging large-scale pretrained priors, they substantially improve visual realism and diversity.

Existing model-based approaches can be broadly grouped into two categories: **Anomaly Generation** methods synthesize isolated defect patches that are composited onto real normal backgrounds. For example, AnomalyDiffusion [15] uses Textual Inversion [7] to learn anomaly appearance and spatial priors and then synthesizes defects within masked regions of normal images. Although they preserve background coherence, such inpainting-based approaches often fail to integrate anomalies seamlessly with the surrounding context and produce misaligned anomaly masks.

Anomaly Image Generation methods synthesize anomalies together with their host products to ensure coherence and realism. DFMGAN [6] pretrains a StyleGAN-based generator [19] on normal data and fine-tunes it on a few anomalous exemplars. DualAnoDiff [17] uses a dual-branch U-Net to separately model normal and anomalous content, significantly improving the fidelity. SeaS [40] finetunes a shared U-Net with unbalanced abnormal embeddings to preserve global consistency while maintaining anomaly diversity. Despite improving image coherence, existing AIG methods struggle to generate precise masks because they derive masks from upsampled low-resolution attention maps or post-hoc segmentation, which lack pixel-level precision. We propose an AIG framework that generates coherent anomalous images with precise, pixel-aligned spatial grounding.

2.3 Diffusion Models

Diffusion models are probabilistic generative models that learn a data distribution $p_{data}(x)$ by reversing a fixed noising Markov chain of length T. For image synthesis [13, 32], U-Net backbones are implemented to predict the noise injected at each timestep, enabling reconstruction of the denoised image. More recently, Stable Diffusion [29] conditions generation via cross-attention to text or other modalities, and ControlNet [39] augments pretrained U-Net with a

lightweight, zero-initialized copy to enable precise spatial conditioning. Moreover, GLIGEN [23] introduces gated attention layers that inject layout tokens into a pretrained diffusion backbone, enabling controllable layout-to-image generation. Our GroundingAnomaly adopts gated self-attention layers and strengthens spatial grounding and few-shot adaptation via the following designs.

3 Method

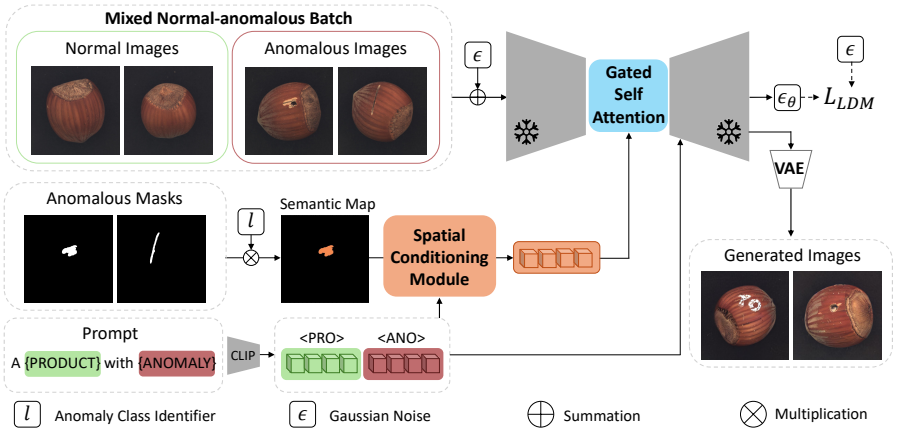

Fig. 2: Proposed framework of **GroundingAnomaly**: (i) **Spatial Conditioning Module** that encodes a pixel-wise semantic map fuses with disentangled product and anomaly tokens; (ii) **Gated Self-Attention Module**, which injects the spatial conditioning into a frozen U-Net. (iii) The framework is trained on mixed batches of normal and anomalous images to leverage cross-domain appearance priors, and it generates diverse, high-fidelity anomaly images.

Synthesizing diverse, high-fidelity anomalies with spatially accurate control from limited examples poses two key challenges: (i) learning disentangled appearance and position representations for each anomaly type, and (ii) enabling stable few-shot adaptation without degrading pretrained priors. Our method is based on a pretrained Stable Diffusion v1.4 backbone [29] and addresses both challenges by introducing two core modules: (i) **Spatial Conditioning Module** (Sec. 3.2) which encodes a pixel-wise semantic map and fuses it with learned product and anomaly tokens to provide explicit, per-pixel conditioning for appearance and localization control and (ii) **Gated Self-Attention Module** (Sec. 3.3) which injects the fused conditioning tokens into the U-Net transformer blocks via gated self-attention, enabling expressive few-shot adaptation while preserving pretrained knowledge. Modules are trained on both normal and anomalous images to exploit cross-domain appearance representations between products

and defects, thereby improving generation fidelity and diversity; training is performed under the conditional diffusion objective (Eq. (3)).

3.1 Preliminaries

Conditioning Diffusion Models. Latent Diffusion Model [29] (LDM) learns data representations in the latent space of a variational auto-encoder (VAE) [20]. Given the latent representation $z_0 = \mathcal{E}(x_0)$, the forward process injects Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ over T timesteps to produce the sequence $\boldsymbol{z} = \{z_1, \ldots, z_T\}$. The reverse denoising process is parameterized by a neural network $\epsilon_\theta(z_t, t)$ trained to predict the injected noise at each timestep, where z_t denotes the noised latent at timestep t, thereby enabling recovery of z_0 from z_T.

Various conditioning strategies such as Stable Diffusion [29] have been proposed to guide the denoising process. Let $\mathbf{v} = \{v_1, \ldots, v_M\}$ denote the visual feature tokens of an image in the transformer blocks, the LDM applies residual self-attention on the visual tokens followed by cross-attention with conditioning tokens $\mathbf{h}$, the two residual updates are:

$$\mathbf{v} \leftarrow \mathbf{v} + \mathrm{SelfAttn}(\mathbf{v}), \tag{1}$$

$$\mathbf{v} \leftarrow \mathbf{v} + \mathrm{CrossAttn}(\mathbf{v}, \mathbf{h}). \tag{2}$$

Accordingly, the learning objective of a conditional LDM can be written as

$$\mathcal{L}_{\mathrm{LDM}} \;=\; \mathbb{E}_{x \sim p_{\mathrm{data}},\, \epsilon \sim \mathcal{N}(0,I),\, t} \left\| \epsilon - \epsilon_\theta(z_t, t, \mathbf{h}) \right\|_2^2, \tag{3}$$

where p_{data} denotes the data distribution of x. At inference, sampling begins from $z_T \sim \mathcal{N}(0, I)$ and the reverse denoising chain produces z_0, which is decoded by the VAE to produce the final image.

3.2 Spatial Conditioning Module

Disentangled Token Learning. To enable few-shot adaptation, we learn compact token embeddings that disentangle anomaly appearance from product identity. Motivated by the observation that anomalies manifest on products rather than as independent entities, we design the prompt template:

`"A photo of a {PRODUCT} with {ANOMALY}"`,

where `{PRODUCT}` and `{ANOMALY}` are placeholders for learnable token sets $\{\langle \mathrm{pro}_n \rangle\}_{n=1}^{N}$ and $\{\langle \mathrm{ano}_k \rangle\}_{k=1}^{K}$, representing the host product and anomaly class, respectively. Concretely, we maintain a compact set of N product tokens $\{\langle \mathrm{pro}_n \rangle\}_{n=1}^{N}$ for each product and K anomaly tokens $\{\langle \mathrm{ano}_k \rangle\}_{k=1}^{K}$ for each anomaly class.

Unlike Textual Inversion [7], which optimizes a single global token per concept from holistic images, our method learns spatially grounded tokens: each embedding is explicitly aligned with semantic maps, enabling precise control over both anomaly location and local appearance.

Semantic Map Representation. Binary masks, while indicating anomaly presence, fail to encode anomaly types. To achieve precise spatial and appearance grounding, we adopt a dense semantic map that jointly encodes anomaly location and class identity. Let C denote the number of anomaly classes for a given product. We assign each anomaly type a unique identifier $l \in \{1, \ldots, C\}$ and represent per-pixel semantics by an integer-valued map:

$$S(x,y) = \begin{cases} 0, & \text{if pixel } (x,y) \text{ is background;} \\ l, & \text{if pixel } (x,y) \text{ belongs to anomaly class } l. \end{cases} \tag{4}$$

Where $(x,y) \in \{1,\ldots,H_S\} \times \{1,\ldots,W_S\}$, H_S and W_S represent the height and width of the semantic map, respectively.

Spatial-Textual Feature Fusion. The semantic map S is encoded by a pretrained ConvNeXt-Tiny encoder [24] into a spatial feature map $F_{\mathrm{S}} \in \mathbb{R}^{H_f \times W_f \times D_f}$, where $H_f = W_f = 8$ and $D_f = 768$ aligns with the CLIP text embedding space. To align textual semantics with spatial layout, we construct a textual feature map F_{T} that mirrors the spatial grid of F_{S}. Specifically, the learned tokens $\{\langle \mathrm{pro}_n \rangle\}_{n=1}^{N}$ and $\{\langle \mathrm{ano}_k \rangle\}_{k=1}^{K}$ are concatenated into a vector of dimension $N \times D_{\mathrm{tok}}$ (or $K \times D_{\mathrm{tok}}$) and projected to dimension D_f via a multi-layer perceptron (MLP), then spatially broadcasted according to the textual semantic map S_f. The textual feature map $F_{\mathrm{T}} \in \mathbb{R}^{H_f \times W_f \times D_f}$ is constructed by assigning each spatial location (i,j) the embedding corresponding to its semantic label in the textual semantic map S_f: Let $\tilde{e}_{\mathrm{pro}}$ and $\{\tilde{e}_{\mathrm{ano},l}\}_{l=1}^{C}$ denote the product and anomaly embeddings after MLP projection to dimension D_f. The construction is as follows:

$$F_{\mathrm{T}}(i,j) = \begin{cases} \tilde{e}_{\mathrm{pro}}, & \text{if } S_f(i,j) = 0; \\ \tilde{e}_{\mathrm{ano},l}, & \text{if } S_f(i,j) = l, \text{ for } l \in \{1,\ldots,C\}. \end{cases} \tag{5}$$

The textual map F_T is concatenated with the spatial map F_S along the channel dimension (D_f), yielding

$$F_{\mathrm{cat}} = [F_S; F_T] \in \mathbb{R}^{H_f \times W_f \times 2D_f}. \tag{6}$$

The concatenated feature map F_{cat} is projected to the U-Net's transformer token dimension D_v via an MLP, then flattened into a sequence of N_c conditioning tokens $\mathbf{c} = \{c_1, \ldots, c_{N_c}\}$ where $N_c = H_f \times W_f$.

3.3 Gated Self-Attention Module

Few-shot anomaly synthesis poses a trade-off: prior methods either freeze the U-Net and train only textual embeddings [15], which limits adaptation capacity, or fine-tune the U-Net [17,40], which risks overwriting pretrained knowledge. To address this, we keep the U-Net frozen and introduce a **Gated Self-Attention Module** that injects spatial–textual conditioning into transformer blocks in the U-Net via gated self-attention layers, enabling expressive adaptation without

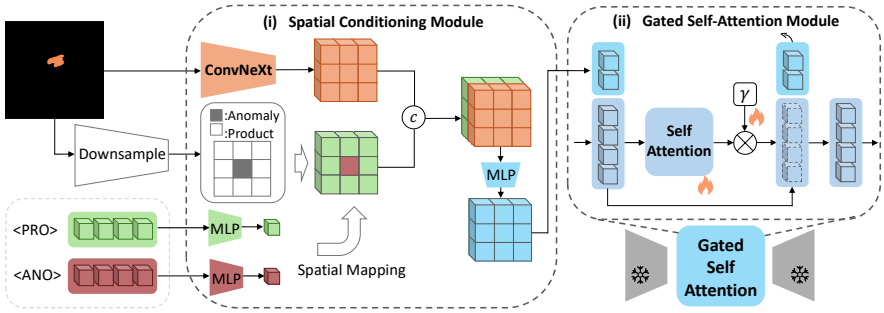

Fig. 3: (i) Spatial Conditioning Module; (ii) Gated Self-Attention Module.

modifying original U-Net weights. The conditioning tokens from the SCM are concatenated with the visual tokens to form $\mathbf{h_c} = [\mathbf{v}; \mathbf{c}]$, and self-attention is applied to the resulting sequence. The gated fusion updates the visual tokens as:

$$\mathbf{v} \leftarrow \mathbf{v} + \tanh(\gamma) \cdot \text{Select}_v\big(\text{SelfAttn}(\mathbf{h_c})\big), \tag{7}$$

where γ is a scalar parameter initialized to 0, and $\text{Select}_v(\cdot)$ extracts the first M tokens (corresponding to visual features) from the attention output while discarding the conditioning tokens. The gated attention is performed between the standard self-attention (Eq. (1)) and cross-attention (Eq. (2)) in each transformer block. Initialize $\gamma = 0$ to preserve the pretrained U-Net at the beginning of the training; the self-attention layers are augmented with LoRA [14] for low-rank updates, enabling stable few-shot adaptation.

3.4 Normal-Guided Training and Synthesis

Mixed Normal-anomalous Training. While GSM enables effective few-shot adaptation, fine-tuning solely on limited anomalous exemplars risks severe overfitting, as their normal regions lack visual diversity. To mitigate this domain collapse, we propose a **mixed normal-anomalous training** (MNT) strategy. By incorporating abundant normal images, we regularize the model to preserve high-fidelity product priors while learning anomaly-specific patterns.

Concretely, a unified model is trained per product across all anomaly types to enable feature sharing and reduce per-class data requirements. Normal images are conditioned with all-zero semantic maps $S = 0$ and the prompt `"A photo of {PRODUCT}"`, teaching the model to reconstruct defect-free appearances. At each iteration, mini-batches are sampled from both normal and anomalous sets. This mixed sampling significantly improves anomaly fidelity and preserves global background consistency.

Normal-prior Denoising Initialization. To further bridge the domain gap between synthesized defects and real backgrounds, we introduce **normal-prior denoising initialization** (NDI) to leverage the visual priors of available normal

images during the generation phase. Instead of initializing the reverse diffusion process from pure Gaussian noise $z_T \sim \mathcal{N}(0, I)$, we utilize the latent representation z_0 of a real normal image.

Specifically, we inject noise into z_0 via the forward process to an intermediate timestep $t' < T$, yielding a partially noised latent $z_{t'}$. The reverse process then starts from $z_{t'}$, performing t' denoising steps conditioned on the target semantic map and prompt to yield the final anomalous image. This initialization strategy offers two key advantages: (i) it yields higher generation quality by seamlessly integrating anomalies into realistic, highly-detailed backgrounds, and (ii) it significantly accelerates inference by reducing the required denoising steps from T to t'.

3.5 Mask Generation

Spatial control requires diverse anomaly masks as input, yet real datasets typically provide very few masks per anomaly type. This scarcity limits the spatial variability of synthesized defects, reducing the effectiveness of downstream anomaly detectors trained on generated data. To ensure a sufficient and varied supply of masks for our pipeline, we adopt a standard approach by learning the underlying mask distribution via Textual Inversion [7, 15].

Specifically, we learn a dedicated token embedding, e_m, that encapsulates the concept of an "anomaly mask." For each anomaly type, this embedding is optimized by minimizing the standard diffusion loss with respect to the available real masks:

$$e_m^* = \arg\min_{e_m} \mathbb{E}_{m \sim p_{\text{masks}},\, \epsilon \sim \mathcal{N}(0,I),\, t} \left\| \epsilon - \epsilon_\theta(z_t, t, e_m) \right\|_2^2, \tag{8}$$

where z_t is the noised latent of a real mask m. Once optimized, this embedding e_m^* can be used as a conditional text prompt to generate novel masks that preserve the shape and size characteristics of the original distribution while introducing spatial variations.

4 Experiments

4.1 Experiment Settings

Datasets. Experiments are conducted on MVTec AD [1] and VisA [42] datasets. In Secs. 4.2 and 4.3, following the protocol of SeaS [40], 60 normal images per product and one-third of anomalous images per anomaly type are used for training, with the remaining two-thirds reserved for testing. In Sec. 4.4, following the setup of SeaS [40], which shows that using more than 2 images per anomaly type can induce train–test overlap (the minimum number of abnormal training images is 2 in previous setting), we conduct experiments under the 1-shot and 2-shot settings.

Implementation details. Following AnomalyDiffusion [15], 1,000 anomaly image–mask pairs per anomaly type are generated to train inspection models. A

single generative model is trained per product to cover all anomaly types. Additional experimental details are provided in the supplementary material.

Metrics. For generation quality, Inception Score (IS) and intra-cluster pairwise LPIPS (IC-LPIPS) are reported. For anomaly segmentation, Area Under the ROC curve (AUROC), Average Precision (AP), and F_1-max are used to evaluate image-level and pixel-level performance on MVTec and VisA; for instance-level anomaly detection, mean Average Precision (mAP) is reported.

Baselines. Our method is compared with state-of-the-art anomaly-synthesis approaches: DFMGAN [6], AnomalyDiffusion [15], DualAnoDiff [17], and SeaS [40]. GroundingAnomaly is also compared against recent anomaly detection methods: DRÆM [37], GLASS [4], WinCLIP+ [16], PromptAD [22], ReMP-AD [27], PatchCore [31], ViTAD [38], MambaAD [11], INP-Former [26] and Dinomaly2 [10].

4.2 Anomaly Generation Quality

GroundingAnomaly is compared with baselines on anomalous image generation quality (IS) and diversity (IC-LPIPS) on MVTec AD and VisA (see Tab. 1). The results demonstrate that our method achieves the highest quality and diversity among the evaluated approaches. Generated examples for MVTec AD and VisA are shown in Fig. 4. Compared to prior methods, images produced by GroundingAnomaly exhibit significantly improved visual fidelity with more controlled backgrounds and tighter alignment between anomaly masks and defect regions.

4.3 Anomaly Generation for Anomaly Inspection

Anomaly image generation for anomaly detection and segmentation. To evaluate downstream anomaly detection and segmentation performance, a simple U-Net [30] is trained per product using the image–mask pairs synthesized by each method. Pixel-level segmentation outputs are converted to image-level anomaly confidence scores via global average pooling. MVTec AD results are reported in Tab. 2. GroundingAnomaly substantially improves segmentation performance, notably on the *grid* and *screw* categories with AP gains of 5.9% and 2.7%, respectively, and achieves the highest AP (85.9%) and F_1-max (81.8%) on MVTec AD. VisA results are presented in Tab. 3. Our method yields AP improvements of 16.6% and 15.2% on the *capsules* and *macaroni2* categories,

Table 1: Comparison on IS and IC-LPIPS on MVTec AD and VisA. Bold indicates the best performance.

Dataset	DFMGAN [6]		AnoDiff [15]		DualAnoDiff [17]		SeaS [40]		**Ours**	
	IS	IC-L	IS	IC-L	IS	IC-L	IS	IC-L	IS	IC-L
MVTec AD	1.72	0.20	1.80	0.32	1.91	0.37	1.95	0.34	**1.99**	**0.41**
VisA	1.25	0.26	1.26	0.25	1.25	0.25	1.27	0.26	**1.29**	**0.31**

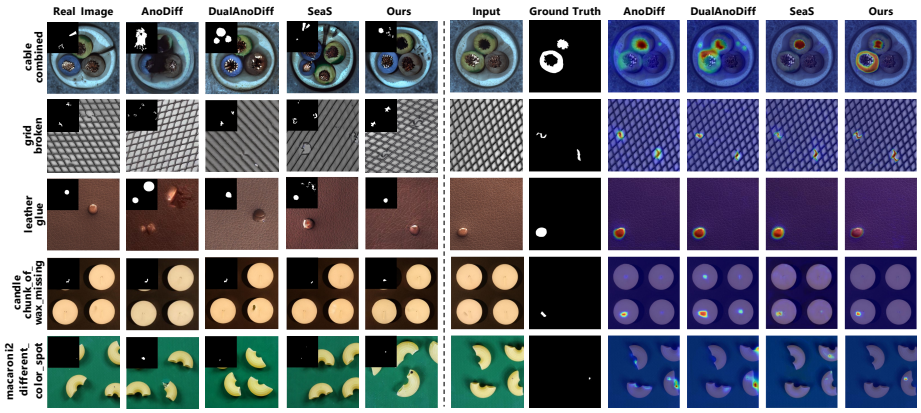

Fig. 4: Left: Comparison on the generation results on MVTec AD and VisA. Cable, grid and leather belong to MVTec AD and candle and macaroni2 belong to VisA. Our method generates high-quality anomaly images that are accurately grounded with masks. **Right:** Visualization of U-Net segmentation results on MVTec AD and VisA. The detection model trained on GroundingAnomaly-generated data exhibits more robust performance.

Table 2: Comparison on pixel-level segmentation and image-level detection on MVTec AD dataset by training U-Nets on the generated data from DRÆM, DFMGAN, AnomalyDiffusion, DualAnoDiff, SeaS and our proposed GroundingAnomaly.

Category	DRÆM [37]				DFMGAN [6]				AnomalyDiffusion [15]				DualAnoDiff [17]				SeaS [40]				Ours			
	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I
bottle	96.7	80.2	74.0	99.8	98.9	90.2	83.9	99.8	99.4	94.1	87.3	99.9	99.5	93.4	85.7	**100.0**	**99.7**	**95.5**	87.9	**100.0**	99.6	94.1	**88.1**	**100.0**
cable	80.3	21.8	28.3	83.2	97.2	81.0	75.4	97.8	**99.2**	**90.8**	**83.5**	**100.0**	97.5	82.6	76.9	98.3	94.5	77.8	75.0	98.5	97.9	81.0	78.2	99.5
capsule	76.2	25.5	32.1	98.7	79.2	26.0	35.0	98.5	98.8	57.2	59.8	**99.9**	**99.5**	**73.2**	**67.0**	99.2	95.7	51.7	52.4	99.6	98.3	65.5	59.4	98.8
carpet	92.6	43.0	41.9	98.7	90.6	33.4	38.1	98.5	98.6	81.2	74.6	98.8	99.4	89.1	80.2	**99.9**	**99.5**	84.5	76.5	99.1	**99.5**	**89.3**	**86.3**	99.6
grid	99.1	59.3	58.7	99.9	75.2	14.3	20.5	90.4	98.3	52.9	54.6	99.5	98.5	57.2	54.9	99.7	99.6	67.7	62.1	99.9	**99.7**	**73.6**	**69.9**	**100.0**
hazelnut	98.8	73.6	68.5	**100.0**	99.7	95.2	89.5	**100.0**	**99.8**	96.5	90.6	99.9	**99.8**	**97.7**	**92.8**	**100.0**	99.7	91.4	84.3	**100.0**	99.6	94.4	89.4	**100.0**
leather	98.5	67.6	65.0	**100.0**	98.5	68.7	66.7	**100.0**	99.8	79.6	71.0	**100.0**	**99.9**	88.8	78.8	**100.0**	99.2	69.1	64.9	99.9	**99.9**	**89.2**	**88.1**	**100.0**
metal nut	96.9	84.2	74.5	99.6	99.3	98.1	94.5	99.8	99.8	98.7	94.0	**100.0**	99.6	98.0	93.0	99.9	99.8	98.8	94.3	**100.0**	**99.9**	**99.3**	**95.7**	**100.0**
pill	95.8	45.3	53.0	98.9	81.2	67.8	72.6	91.7	**99.8**	**97.0**	**90.8**	99.6	99.6	95.8	89.2	99.0	99.6	90.2	82.7	**99.8**	99.2	95.0	89.7	98.8
screw	91.0	30.1	35.7	96.3	58.8	2.2	5.3	64.7	97.0	51.8	50.9	**97.9**	98.1	57.1	**56.1**	95.0	**98.3**	55.2	54.7	97.6	97.8	**59.8**	55.3	91.9
tile	98.5	93.2	87.8	**100.0**	99.5	97.1	91.6	**100.0**	99.2	93.9	86.2	**100.0**	99.7	97.1	91.0	**100.0**	**99.8**	**97.2**	91.7	**100.0**	99.5	96.2	**93.1**	**100.0**
toothbrush	93.8	29.5	28.4	99.8	96.4	75.9	72.6	**100.0**	99.2	76.5	73.4	**100.0**	98.2	68.3	68.6	99.7	96.1	57.5	58.8	95.9	**99.3**	**79.2**	**75.7**	**100.0**
transistor	76.5	31.7	24.2	80.5	96.2	81.2	77.0	92.5	99.3	**92.6**	85.7	**100.0**	98.0	86.7	79.6	93.7	96.9	80.5	77.6	**100.0**	**99.6**	92.3	**88.7**	**100.0**
wood	98.8	87.8	80.9	**100.0**	95.3	70.7	65.8	99.4	98.9	84.6	74.5	99.4	99.4	**91.6**	83.8	99.9	99.4	87.2	79.2	**100.0**	**99.5**	90.7	**86.7**	**100.0**
zipper	93.4	65.4	64.7	**100.0**	92.9	65.6	64.9	99.9	99.4	86.0	79.2	**100.0**	**99.6**	**90.7**	**82.7**	**100.0**	99.3	85.8	79.0	**100.0**	**99.6**	88.7	82.2	**100.0**
Average	92.5	55.9	54.5	97.0	90.6	64.5	63.6	95.5	99.1	82.2	77.1	**99.7**	99.1	84.5	78.7	99.0	98.5	79.3	74.7	99.4	**99.3**	**85.9**	**81.8**	99.2

respectively, attaining the highest overall AP (67.2%) on VisA and outperforming the second-ranked DualAnoDiff by 6.3% AP.

Visualized results are presented in Fig. 4, which shows that segmentation U-Nets trained on image–mask pairs generated by GroundingAnomaly exhibit better alignment with the ground truth with fewer false positives and higher prediction confidence. GroundingAnomaly is also compared with state-of-the-art anomaly inspection methods in Tab. 4, which shows that simple U-Nets trained on GroundingAnomaly-generated data outperform these baselines by 11.8% AP on MVTec AD and 12.5% AP on VisA, while achieving competitive AUROC.

Anomaly image generation for instance-level anomaly detection. To assess instance-level detection utility, object detectors (Faster R-CNN [28], DETR [3] and YOLOv5 [18]) are trained separately for each product on the synthesized image–mask pairs; instance bounding boxes are derived from tight bound-

Table 3: Comparison on pixel-level segmentation and image-level detection on VisA dataset by training U-Nets on the generated data from DRÆM, DFMGAN, AnomalyDiffusion, DualAnoDiff, SeaS and our proposed GroundingAnomaly.

Category	DRÆM [37]				DFMGAN [6]				AnomalyDiffusion [15]				DualAnoDiff [17]				SeaS [40]				Ours			
	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I
candle	86.2	18.7	23.7	77.5	87.1	24.7	31.4	85.9	89.1	28.7	33.9	87.2	93.7	48.1	46.5	89.0	89.4	45.1	44.1	87.8	**98.9**	**61.2**	**58.3**	**94.4**
capsules	90.7	50.4	53.8	85.1	91.3	55.0	58.6	86.5	94.1	60.5	62.3	90.7	95.8	61.8	64.5	92.0	97.2	61.2	64.2	88.7	**99.3**	**78.4**	**72.1**	**92.3**
cashew	89.0	45.2	47.1	91.1	92.2	65.7	65.1	92.3	95.1	82.1	80.9	95.0	96.5	83.1	**87.1**	**96.1**	97.0	80.9	81.3	93.3	**99.3**	**87.1**	84.2	95.3
chewinggum	94.3	59.7	65.0	97.0	96.0	67.4	63.0	96.7	97.3	80.5	75.3	95.5	98.8	80.3	77.1	98.8	98.5	81.2	74.8	99.3	**99.5**	**84.6**	**78.2**	**99.4**
fryum	88.2	41.3	34.1	81.9	90.4	45.7	43.2	89.3	93.4	57.1	52.9	92.1	94.8	62.9	56.8	94.4	**95.8**	**63.8**	**59.0**	93.6	91.5	55.1	51.3	**96.4**
macaroni1	94.3	38.9	34.3	88.0	96.2	42.1	46.1	93.3	97.2	48.7	51.3	94.8	98.7	53.9	53.7	98.0	98.1	47.7	50.8	93.5	**99.8**	**65.4**	**59.4**	**99.6**
macaroni2	85.7	30.1	32.9	80.4	89.3	32.0	35.7	81.4	94.1	35.2	41.1	85.3	97.1	39.3	45.7	88.0	96.7	39.1	44.4	87.7	**99.3**	**54.5**	**50.9**	**91.9**
pcb1	91.1	72.2	71.4	95.4	93.0	79.4	74.3	97.1	96.3	80.7	77.1	97.9	97.1	81.4	**78.0**	**98.9**	**97.5**	**81.9**	77.9	98.8	96.7	81.8	75.7	97.8
pcb2	88.2	27.4	21.7	94.9	90.4	32.7	34.3	95.1	92.9	42.5	43.7	96.9	95.0	48.0	47.2	**97.3**	94.7	47.7	46.5	97.2	**97.5**	**54.0**	**55.4**	96.1
pcb3	93.3	38.1	39.5	92.1	93.8	50.1	36.7	94.3	94.2	51.9	47.3	94.9	96.7	53.2	55.9	**95.8**	95.9	48.3	50.2	95.0	**97.7**	**61.6**	**61.0**	**95.8**
pcb4	90.1	48.3	43.1	92.1	93.3	50.4	44.9	93.7	95.3	51.7	53.2	94.9	98.7	**58.3**	**58.1**	98.0	97.8	57.2	52.8	**98.7**	**99.4**	45.1	45.9	97.8
pipe fryum	84.4	41.7	47.1	82.1	88.3	44.3	48.3	85.7	92.2	55.4	62.4	86.5	96.4	60.3	65.7	90.1	98.2	64.0	66.5	88.3	**99.5**	**77.4**	**68.7**	**95.7**
Average	89.6	42.7	42.8	88.1	91.8	49.1	48.5	90.9	94.3	56.3	56.8	92.6	96.6	60.9	61.4	94.7	96.4	59.8	59.3	93.5	**98.2**	**67.2**	**63.5**	**96.0**

ing boxes of masks. Mixed-type (category *combined*) anomalies are excluded to avoid label ambiguity. The averaged mAP are reported in Tab. 5, showing average mAP improvements of 1.52% (MVTec AD) and 2.34% (VisA). Our method achieves superior performance by generating anomalies with more class-aligned appearances.

Table 4: Comparison on pixel-level segmentation and image-level detection with anomaly inspection methods.

Model	MVTec AD				VisA			
	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I
Unsupervised								
PatchCore [31]	98.4	56.1	58.9	98.6	98.4	48.6	49.7	94.8
ViTAD [38]	97.7	55.3	58.7	98.3	98.2	36.6	41.1	90.5
MambaAD [11]	97.7	56.3	59.2	93.1	98.5	39.4	44.0	94.3
WinCLIP+ [16]	96.2	62.7	60.1	98.8	97.4	50.2	47.9	89.8
PromptAD [22]	96.3	63.4	60.3	98.9	97.5	50.3	47.6	91.4
ReMP-AD [27]	96.7	64.8	61.2	99.2	97.9	51.2	49.1	94.8
INP-Former [26]	98.5	71.0	69.7	99.9	98.9	51.2	54.7	99.0
Dinomaly2 [10]	98.6	70.3	69.9	**100.0**	**99.2**	54.7	57.1	**99.3**
Synthesize-based								
DRÆM [37]	97.9	67.9	66.1	98.0	92.9	17.2	23.0	86.3
GLASS [4]	**99.3**	74.1	70.4	99.9	98.5	45.6	48.4	97.7
Ours+U-Net	**99.3**	**85.9**	**81.8**	99.2	98.2	**67.2**	**63.5**	96.0

Table 5: Comparison on trained supervised instance-level anomaly detection models on MVTec AD and VisA.

Model	DFMGAN avg mAP	AnoDiff avg mAP	DualAnoDiff avg mAP	SeaS avg mAP	**Ours** avg mAP
MVTec					
Faster R-CNN [28]	35.44	37.13	40.17	40.44	**41.96**
DETR [3]	37.88	38.91	40.79	42.81	**43.17**
YOLOv5 [18]	41.51	44.35	47.73	46.71	**49.40**
Average	38.28	40.13	42.90	43.32	**44.84**
VisA					
Faster R-CNN [28]	28.76	34.19	**37.73**	36.77	37.41
DETR [3]	30.17	32.07	35.16	33.95	**38.77**
YOLOv5 [18]	29.88	33.05	36.83	37.11	**40.54**
Average	29.60	33.10	36.57	35.94	**38.91**

4.4 Ablation Study

Ablation on model components. We evaluate each component via ablations: (i) without Disentangled Token Learning (fixed text tokens, w/o DTL); (ii) without Spatial–Textual Feature Fusion (w/o SFF); (iii) without GSM, injecting conditioning into U-Net cross-attention (w/o GSM); (iv) trained only on anomalous images (w/o MNT); (v) initialize generation from random noise (w/o NDI); and (vi) the full GroundingAnomaly. We use these models to generate 1,000 anomaly image-mask pairs per anomaly type and train a U-Net per product. The results are provided in Tab. 6, which shows that removing any proposed module degrades generation quality and anomaly inspection performance. Note that a high IC-LPIPS combined with a low IS (w/o GSM on MVTec) indicates poor fidelity and chaotic generation rather than diversity.

Ablation on few-shot generation. In previous experiments, we follow the setting in AnomalyDiffusion [15] and employed 1/3 of anomaly data for comparison. However, most industrial applications require synthesizing anomalies

from only a few exemplars. Here we analyze GroundingAnomaly's extreme few-shot capability. Following the setup of SeaS [40], which shows that using more than 2 images per anomaly type can introduce train–test overlap (the minimum number of abnormal training images is 2 in previous setting), we conduct experiments under the 1-shot and 2-shot settings, results are provided in Tab. 7, more qualitative results can be found in the supplementary material. In the 1-shot setting, GroundingAnomaly can generate images of satisfactory quality, but downstream detection performance is severely degraded by limited mask diversity, performance improves modestly in the 2-shot setting.

4.5 Analysis

Analysis of Spatial-Textual Feature Fusion. While the quantitative ablation in Tab. 6 confirms the overall effectiveness of SFF, it does not explicitly reveal the source of this improvement. Qualitative analysis (Fig. 5) provides the answer: SFF primarily enhances the visual realism of the synthesized anomalies rather than their spatial grounding. As illustrated, SeaS [40] struggles to align masks with the generated defects. Conversely, our model without SFF (w/o SFF) achieves accurate spatial bounding but often yields artificial textures and blunt transitions. By utilizing spatially aligned tokens to learn both anomaly and product features, SFF significantly enhances the overall generation quality and the realism of the synthesized defects.

Unseen anomaly generation. Beyond reconstructing known defects, the proposed GroundingAnomaly can synthesize novel anomalies on unseen products through brief fine-tuning on normal samples. Fig. 5 illustrates this cross-domain generalization: a model trained on *wood* successfully projects learned anomaly characteristics onto *leather*, generating realistic defects that are entirely absent from the target dataset's real distribution. The generated results confirm that our approach deeply comprehends anomaly structures and can synthesize diverse, out-of-distribution defects across various datasets, rather than merely overfitting to training exemplars.

Table 6: Ablation study results.

Model	IS	IC-L	AUC-P	AP-P	AP-I
MVTec AD					
w/o DTL	1.88	0.39	98.7	81.7	98.6
w/o SFF	1.72	0.37	97.5	77.3	97.7
w/o GSM	1.63	**0.42**	95.7	75.7	96.1
w/o MNT	1.74	0.37	99.1	83.1	99.1
w/o NDI	1.94	0.40	99.2	84.7	99.1
Ours	**1.99**	0.41	**99.3**	**85.9**	**99.2**
VisA					
w/o DTL	1.21	0.27	95.2	62.3	94.8
w/o SFF	1.13	0.28	93.9	59.1	93.7
w/o GSM	1.09	0.30	93.7	58.1	93.4
w/o MNT	1.17	0.27	97.3	62.9	95.3
w/o NDI	1.25	0.29	97.6	64.9	95.9
Ours	**1.29**	**0.31**	**97.7**	**67.2**	**96.0**

Table 7: Ablation on few-shot generation.

Model	IS	IC-L	AUC-P	AP-P	AP-I
MVTec AD					
1-shot	1.77	0.39	96.1	72.3	92.7
2-shot	1.83	0.38	97.2	76.4	97.9
Ours	**1.99**	**0.41**	**99.3**	**85.9**	**99.2**
VisA					
1-shot	1.22	0.30	81.7	44.3	84.9
2-shot	1.19	0.30	83.1	49.3	85.7
Ours	**1.29**	**0.31**	**97.7**	**67.2**	**96.0**

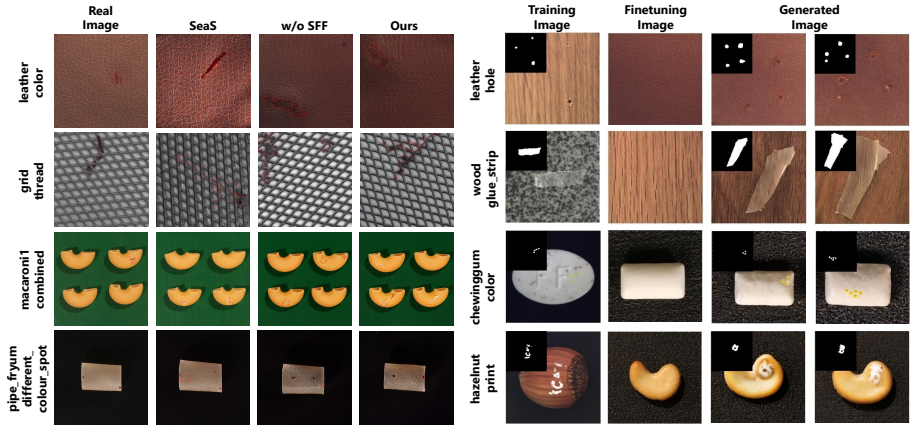

Fig. 5: Left: Qualitative analysis of SFF. The red outlines denote the ground-truth mask. While SeaS [40] suffers from spatial misalignment, our model without SFF (w/o SFF) achieves accurate bounding. Our full model (Ours) leverages SFF to achieve both precise spatial grounding and high-fidelity anomaly synthesis. **Right:** Unseen anomaly generation results on MVTec AD and VisA. GroundingAnomaly is trained on anomalous images and finetuned on normal images from another product.

Multi-Class anomaly generation. A key advantage of GroundingAnomaly is its ability to synthesize multiple, diverse defects of different classes within a single image via grounding anomalies with a semantic map, rather than treating defect combinations as a single rigid *combined* category. The semantic map is constructed by $S = S_1 + S_2$, where S_1 and S_2 represent semantic maps of different anomalies, respectively. And a prompt template

"A photo of a {PRODUCT} with {ANOMALY$_1$} and {ANOMALY$_2$}"

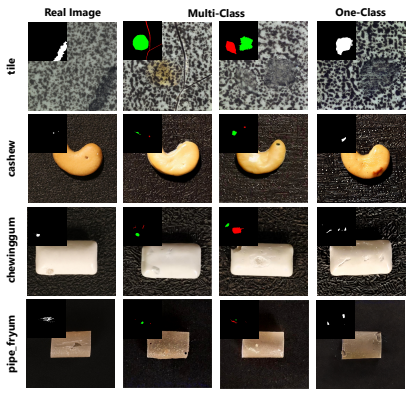

Fig. 6: Multi-anomaly generation result of GroundingAnomaly.

is employed. As illustrated in Fig. 6, our model effortlessly generalizes to these complex, multi-class scenarios on MVTec AD and VisA. Additional analyses are available in the supplementary material.

5 Conclusion

In this paper, we propose GroundingAnomaly, a novel framework for few-shot anomaly synthesis. Unlike prior approaches, our method achieves spatially grounded

anomaly generation through a spatial conditioning module, jointly synthesizing anomalies and their host products. Furthermore, a gated self-attention module is introduced to facilitate robust few-shot adaptation. Extensive experiments demonstrate that GroundingAnomaly generates high-quality anomaly images with accurately aligned masks, achieving state-of-the-art performance. In future work, we will explore zero-shot anomaly synthesis and investigate more powerful generative models to achieve improved generation capacity.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Chongqing under Grant CSTB2023NSCQ-MSX0095.

References

1. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9584–9592 (2019). https://doi.org/10.1109/CVPR.2019.00982
2. Cao, Y., Xu, X., Zhang, J., Cheng, Y., Huang, X., Pang, G., Shen, W.: A survey on visual anomaly detection: Challenge, approach, and prospect. arXiv preprint arXiv:2401.16402 (2024)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chen, Q., Luo, H., Lv, C., Zhang, Z.: A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In: European Conference on Computer Vision. pp. 37–54. Springer (2024)
5. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9737–9746 (2022)
6. Duan, Y., Hong, Y., Niu, L., Zhang, L.: Few-shot defect image generation via defect-aware feature manipulation. In: AAAI (2023)
7. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
9. Gui, G., Gao, B.B., Liu, J., Wang, C., Wu, Y.: Few-shot anomaly-driven generation for anomaly classification and segmentation. In: European Conference on Computer Vision. pp. 210–226. Springer (2024)
10. Guo, J., Lu, S., Fan, L., Li, Z., Di, D., Song, Y., Zhang, W., Zhu, W., Yan, H., Chen, F., et al.: One dinomaly2 detect them all: A unified framework for full-spectrum unsupervised anomaly detection. arXiv preprint arXiv:2510.17611 (2025)
11. He, H., Bai, Y., Zhang, J., He, Q., Chen, H., Gan, Z., Wang, C., Li, X., Tian, G., Xie, L.: Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. Advances in Neural Information Processing Systems **37**, 71162–71187 (2024)

12. Heckler-Kram, L., Neudeck, J.H., Scheler, U., König, R., Steger, C.: The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. arXiv preprint arXiv:2503.21622 (2025)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
15. Hu, T., Zhang, J., Yi, R., Du, Y., Chen, X., Liu, L., Wang, Y., Wang, C.: Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
16. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19606–19616 (June 2023)
17. Jin, Y., Peng, J., He, Q., Hu, T., Wu, J., Chen, H., Wang, H., Zhu, W., Chi, M., Liu, J., et al.: Dual-interrelated diffusion model for few-shot anomaly image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30420–30429 (2025)
18. Jocher, G.: ultralytics/yolov5: v3.1 - bug fixes and performance improvements. https://github.com/ultralytics/yolov5 (Oct 2020). `https://doi.org/10.5281/zenodo.4154370`
19. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020)
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
21. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9664–9674 (2021)
22. Li, X., Zhang, Z., Tan, X., Chen, C., Qu, Y., Xie, Y., Ma, L.: Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16838–16848 (2024)
23. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
24. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
25. Lu, R., Wu, Y., Tian, L., Wang, D., Chen, B., Liu, X., Hu, R.: Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 8487–8500. Curran Associates, Inc. (2023)
26. Luo, W., Cao, Y., Yao, H., Zhang, X., Lou, J., Cheng, Y., Shen, W., Yu, W.: Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 9974–9983 (June 2025)
27. Ma, H., Yang, G., Zhao, D., Ji, Y., Zuo, W.: Remp-ad: Retrieval-enhanced multimodal prompt fusion for few-shot industrial visual anomaly detection. In: Proceed-

ings of the IEEE/CVF International Conference on Computer Vision. pp. 20425–20434 (2025)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
31. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14318–14328 (June 2022)
32. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
33. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
34. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
35. You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., Le, X.: A unified model for multi-class anomaly detection. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 4571–4584. Curran Associates, Inc. (2022)
36. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International journal of computer vision **129**(11), 3051–3068 (2021)
37. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8330–8339 (October 2021)
38. Zhang, J., Chen, X., Wang, Y., Wang, C., Liu, Y., Li, X., Yang, M.H., Tao, D.: Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. arXiv preprint arXiv:2312.07495 (2023)
39. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
40. Zhewei, D., Shilei, Z., Haotian, L., Xurui, L., Feng, X., Yu, Z.: Seas: Few-shot industrial anomaly image generation with separation and sharing fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
41. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In: The Twelfth International Conference on Learning Representations (2023)
42. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European conference on computer vision. pp. 392–408. Springer (2022)