Anomagic: Crossmodal Prompt-driven Zero-shot Anomaly Generation

Yuxin Jiang[1], Wei Luo[3], Hui Zhang[2], Qiyu Chen[4], Haiming Yao[3], Weiming Shen*[1], Yunkang Cao*[2]

[1]Huazhong University of Science and Technology, [2]Hunan University, [3]Tsinghua University
[3]Institute of Automation, Chinese Academy of Sciences
yuxinjiang@hust.edu.cn, luow23@mails.tsinghua.edu.cn, zhanghuihby@126.com, chenqiyu2021@ia.ac.cn, yhm22@mails.tsinghua.edu.cn, wshen@ieee.org, caoyunkang@ieee.org,

Abstract

We propose Anomagic, a zero-shot anomaly generation method that produces semantically coherent anomalies without requiring any exemplar anomalies. By unifying both visual and textual cues through a crossmodal prompt encoding scheme, Anomagic leverages rich contextual information to steer an inpainting-based generation pipeline. A subsequent contrastive refinement strategy enforces precise alignment between synthesized anomalies and their masks, thereby bolstering downstream anomaly detection accuracy. To facilitate training, we introduce AnomVerse, a collection of 12,987 anomaly–mask–caption triplets assembled from 13 publicly available datasets, where captions are automatically generated by multimodal large language models using structured visual prompts and template-based textual hints. Extensive experiments demonstrate that Anomagic trained on AnomVerse can synthesize more realistic and varied anomalies than prior methods, yielding superior improvements in downstream anomaly detection. Furthermore, Anomagic can generate anomalies for any normal-category image using user-defined prompts, establishing a versatile foundation model for anomaly generation.

Code — https://github.com/yuxin-jiang/Anomagic

Introduction

Anomalies, characterized as rare or absent patterns in typical conditions (Liu et al. 2024; Cheng et al. 2025), pose significant challenges in domains such as manufacturing, where defects can precipitate critical safety concerns (Roth et al. 2022). To address this, unsupervised anomaly detection has emerged as a pivotal approach, training models exclusively on normal data to identify deviations (Deng and Li 2022; Luo et al. 2025a). Concurrently, efforts to mitigate the scarcity of anomaly samples have led to advancements in anomaly generation, aiming to produce realistic and diverse anomalous instances (Song et al. 2025; Jin et al. 2025).

However, despite their potential, the majority of current anomaly generation methods, including DualAnoDiff (Jin et al. 2025) and AnoGen (Gui et al. 2024), function within a *few-shot anomaly generation* paradigm. This approach depends on a limited set of annotated anomalies from known defect categories. Although these methods can effectively generate visually plausible anomalies within familiar classes, they are unable to generalize to unseen defect types or new object categories. The task of generating realistic anomalies for novel categories without prior examples, known as *zero-shot anomaly generation* (Sun et al. 2025), remains a significant and largely unaddressed challenge.

Initial efforts in zero-shot anomaly generation adopted a cut-and-paste methodology (Zavrtanik et al. 2021; Zhang et al. 2024), overlaying external textures onto normal images. Although these approaches yielded diverse anomalies, their synthetic appearance often undermined practical utility. A notable advancement was achieved with AnomalyAny (Sun et al. 2025), which leverages pre-trained Stable Diffusion (SD) models (Rombach et al. 2022a) guided by textual prompts. By manipulating the attention matrix during the denoising process, AnomalyAny achieves versatility across diverse image categories. However, its attention mechanism (Chefer et al. 2023), limited to a small set of tokens, struggles to capture the semantic richness required for highly diverse anomalies. Furthermore, insights from other image generation domains (Ruiz et al. 2023; Dang et al. 2025) suggest that task-specific fine-tuning can significantly enhance the quality of generated images.

In this work, we present **Anomagic**, an innovative framework for zero-shot anomaly generation that incorporates crossmodal prompts, merging visual and textual semantics to create anomalies that are both highly realistic and adaptable. This method enables guidance through visual inputs, textual inputs, or a combination of both, thereby ensuring versatility across a wide range of scenarios. Furthermore, we introduce a contrastive anomaly mask refinement strategy, which produces precise mask-anomaly pairs, significantly improving upon the coarse masks generated by previous methods. For the training of Anomagic, we develop a crossmodal prompting strategy that utilizes multimodal large language models (MLLMs) to generate accurate captions for anomaly images, informed by both visual and textual data. These captions ensure that the generated anomalies closely match their intended semantic descriptions. By applying this strategy to publicly available datasets such as MVTec AD (Bergmann et al. 2021) and VisA (Zou et al. 2022), we have created **AnomVerse**, which, with 12,987

*corresponding author.

anomaly-mask-caption triplets, stands as the largest dataset of its kind to date.

When trained on AnomVerse using a novel prompt-guided inpainting approach, Anomagic achieves remarkable levels of realism and diversity in the anomalies it generates, while also improving the performance of subsequent anomaly detection tasks. Our experimental findings show that Anomagic not only generates anomalies that accurately reflect the prompts from AnomVerse but also demonstrates strong generalization capabilities, allowing it to handle arbitrary prompts and generate corresponding anomalies for entirely new categories. This positions Anomagic as a foundational model for zero-shot anomaly generation. The key contributions of this work are as follows:

- We introduce Anomagic, a novel framework for zero-shot anomaly generation that utilizes crossmodal prompts to achieve superior realism, supported by a contrastive mask refinement strategy that ensures the precision of anomaly masks.
- We present AnomVerse, an extensive dataset consisting of 12,987 anomaly-mask-caption triplets, developed through a novel crossmodal prompting technique.
- Rigorous experimental validation illustrating that Anomagic outperforms existing methods in generating high-quality anomalies and bolstering anomaly detection performance, while demonstrating strong generalization across diverse prompts, including unimodal, crossmodal, and user-defined inputs.

Related Works

Anomaly Generation

Few-shot Anomaly Generation. Few-shot anomaly generation methods leverage a small set of abnormal exemplars during training to synthesize novel anomalies of comparable types, thereby enriching the diversity of anomaly categories. Early work such as Defect-GAN (Zhang et al. 2021) and DFMGAN (Duan et al. 2023) adopted Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) for their demonstrated ability to produce high-fidelity images. More recent approaches have transitioned to diffusion-based frameworks: AnomalyDiffusion (Hu et al. 2024) employs text inversion (Gal et al. 2022) to capture anomaly semantics and mask distributions, while Defect-Gen (Yang et al. 2024) introduces a two-stage diffusion process to improve realism. Methods such as AnoGen (Gui et al. 2024) and DefectFill (Song et al. 2025) formulate anomaly synthesis as an inpainting task to guarantee mask consistency, and DualAnoDiff (Jin et al. 2025) further refines generation quality by alternately modeling foreground and background components. SeaS (Dai et al. 2024) extends this paradigm by binding anomaly attributes to distinct prompt tokens, enabling a single model to generate multiple anomaly types. Despite these advances, all of these techniques remain limited to the anomaly types seen during training, restricting their capacity to generalize to unseen defect classes.

Zero-shot Anomaly Generation. Zero-shot anomaly generation aims to produce realistic anomalies for categories not encountered during training, without the need for prior abnormal samples. Initial approaches, such as CutPaste (Li et al. 2021) and DRAEM (Zavrtanik et al. 2021), depended on rudimentary cut-and-paste techniques, superimposing external textures onto normal images. While these methods provided diversity, the generated anomalies frequently lacked authenticity. A notable advancement was made by AnomalyAny (Sun et al. 2025), which leveraged pre-trained SD models (Rombach et al. 2022a) directed by textual prompts, modifying the attention matrix during the denoising process to accommodate a variety of image categories. Nonetheless, its reliance on single-modal prompts compromise its controllability. Our proposed method, Anomagic, addresses these limitations by incorporating crossmodal prompts to enrich semantic understanding and by expediting the generation process.

Conditional Diffusion Models

Crossmodal Conditions. Diffusion models conditioned on multiple modalities have substantially advanced the controllability of generative processes by aligning outputs with diverse guidance, such as textual prompts (Dhariwal and Nichol 2021; Rombach et al. 2022a) and visual cues (Zhang et al. 2023). More recent frameworks—e.g., BLIP-Diffusion (Li, Li, and Hoi 2023) and OmniGen (Xiao et al. 2025)—jointly leverage language and visual embeddings to achieve finer-grained semantic alignment. Unlike earlier anomaly synthesis approaches that rely solely on text-driven control, Anomagic exploits a crossmodal conditioning scheme that seamlessly integrates visual representations of defects with rich textual descriptions to produce highly targeted anomalous examples.

Mask Generation. Accurate pixel-level masks are indispensable for anomaly detection (Gui et al. 2024; Sun et al. 2025). Traditional mask-aware synthesis techniques often depend on heuristic post-processing to extract imperfect region proposals (Nguyen et al. 2023; Qian et al. 2024), which falls short of the pixel-level fidelity required in practice. Approaches such as DefectFill (Song et al. 2025) and AnomalyDiffusion (Hu et al. 2024) constrain generated anomalies to rough mask shapes, while AnomalyAny derives masks from attention maps between text tokens and latent features—yet this tends to yield coarse boundaries. In contrast, Anomagic introduces a contrastive mask refinement module that accuraly derive anomaly masks via discrepancies between input normal image and generated anomalies, yielding pixel-accurate masks suitable for downstream anomaly detection.

Dataset: AnomVerse

Construction Pipeline

We introduce a novel pipeline, as illustrated in Figure 1(a), designed to curate high-quality triplets comprising anomalies, masks, and captions to facilitate the training of zero-shot anomaly generation models. Although anomalies and their associated masks are widely available in public datasets, generating precise and informative captions for anomaly regions poses a significant challenge. To address

this, we employ a crossmodal prompting strategy that leverages MLLMs, specifically the Doubao-Seed-1.6-thinking model[1], to produce detailed captions for the anomalies. To improve the quality and relevance of these captions, we utilize masks to delineate minimal bounding boxes to highlight defective areas. Additionally, we have devised a structured caption template: "The image depicts [general description of the object], with a [type of defect] observed [location description]. The defect is characterized by [detailed description] and exhibits [notable features]." This template ensures consistency and clarity in the caption generation process. By combining visual prompts, in the form of bounding boxes, with this predefined textual template, our method enables the creation of precise anomaly-mask-caption triplets.

Statistics

The AnomVerse dataset integrates data from 13 publicly accessible datasets, including MVTec AD (Bergmann et al. 2021), VisA (Zou et al. 2022), and MANTA (Fan et al. 2025), et al, spanning five distinct domains (Figure 1(b)): industrial (56.5%), textiles (23.6%), consumer goods (8.7%), medicine (5.9%), and electronics (5.3%). Comprising 12,987 anomaly samples that represent 131 defect types, each sample is paired with a structured descriptive caption. In contrast to its predecessor, MMAD (Jiang et al. 2024), which includes 8,366 samples, AnomVerse provides a significantly larger collection, establishing it as a critical resource for advancing research in zero-shot anomaly generation. Further details and examples of AnomVerse are provided in Appendix A.

Method: Anomagic

Preliminaries

Latent Diffusion Models. Latent Diffusion Models (LDMs) (Rombach et al. 2022b) constitute a class of diffusion-based generative models (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021; Song et al. 2020b) that operate within a reduced-dimensional latent space, thereby mitigating computational demands. Initially, an encoder E transforms an image $\mathbf{I}$ into a latent representation $\mathbf{z}_0 = E(\mathbf{I})$. During the generative process, this latent representation undergoes a diffusion process involving noise addition and denoising before being reconstructed by a decoder D. Specifically, during the training phase, Gaussian noise $\varepsilon \sim \mathcal{N}(0, \mathbb{I})$ is introduced to $\mathbf{z}_0$ according to a predefined schedule $\{\alpha_t\}_{t=1}^{T}$, resulting in the noisy latent code $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\varepsilon$. A neural network ε_θ is then trained to predict the noise component ε from the noisy latent code $\mathbf{z}_t$ and the timestep t. Through the reverse diffusion process, this noise prediction enables the recovery of a reconstructed latent representation $\hat{\mathbf{z}}_0$. Finally, the image is reconstructed by the decoder D as $\hat{\mathbf{I}} = D(\hat{\mathbf{z}}_0)$. The corresponding loss function is given by:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathbf{z}_0,\varepsilon,t} \left\| \varepsilon - \varepsilon_\theta(\mathbf{z}_t, t) \right\|_2^2 . \tag{1}$$

[1]https://www.volcengine.com/product/doubao

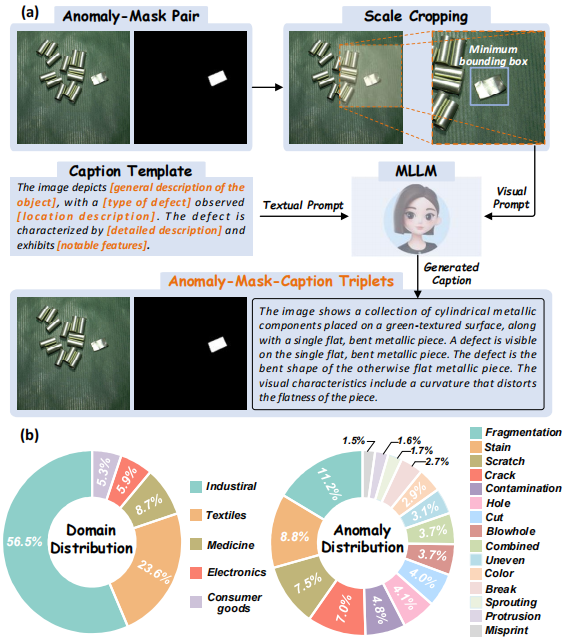

Figure 1: **Overview of AnomVerse.** (a) The construction pipeline for generating anomaly–mask–caption triplets. (b) The distribution across domains and the top 15 most frequent anomaly categories.

Conditional Diffusion Models. To facilitate the generation of images with desired semantic attributes, conditional diffusion models extend LDMs by incorporating supplementary inputs—such as textual descriptions or masks—through cross-attention mechanisms (Rombach et al. 2022a; Zhang et al. 2023). Specifically, a condition embedding $\mathbf{P}$ is integrated into the UNet's latent feature maps at various scales, thereby aligning the denoising process with the provided guidance. This ensures that the synthesized images conform to the specified conditions. The training objective is accordingly adjusted as follows:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathbf{z}_0,\varepsilon,t} \left\| \varepsilon - \varepsilon_\theta(\mathbf{z}_t, t, \mathbf{P}) \right\|_2^2 . \tag{2}$$

Framework Overview

Despite the advancements in conditional LDMs, generating anomalies in industrial contexts presents two primary challenges: (1) formulating precise, defect-specific conditions, and (2) ensuring generated anomalies align with real-world industrial characteristics while maintaining contextual coherence. To address these issues, our proposed framework, Anomagic, as depicted in Figure 2, employs Cross-modal Prompt Encoding (CPE) to derive fine-grained conditions and fine-tunes pre-trained SD models using Low-Rank Adaptation (LoRA) (Hu et al. 2022) in conjunction with an inpainting strategy. During the testing phase, we utilize MLLMs to semantically retrieve the most relevant prompts from external sources, such as AnomVerse, and subsequently apply these conditions to generate anomalies on novel normal images in a zero-shot manner.

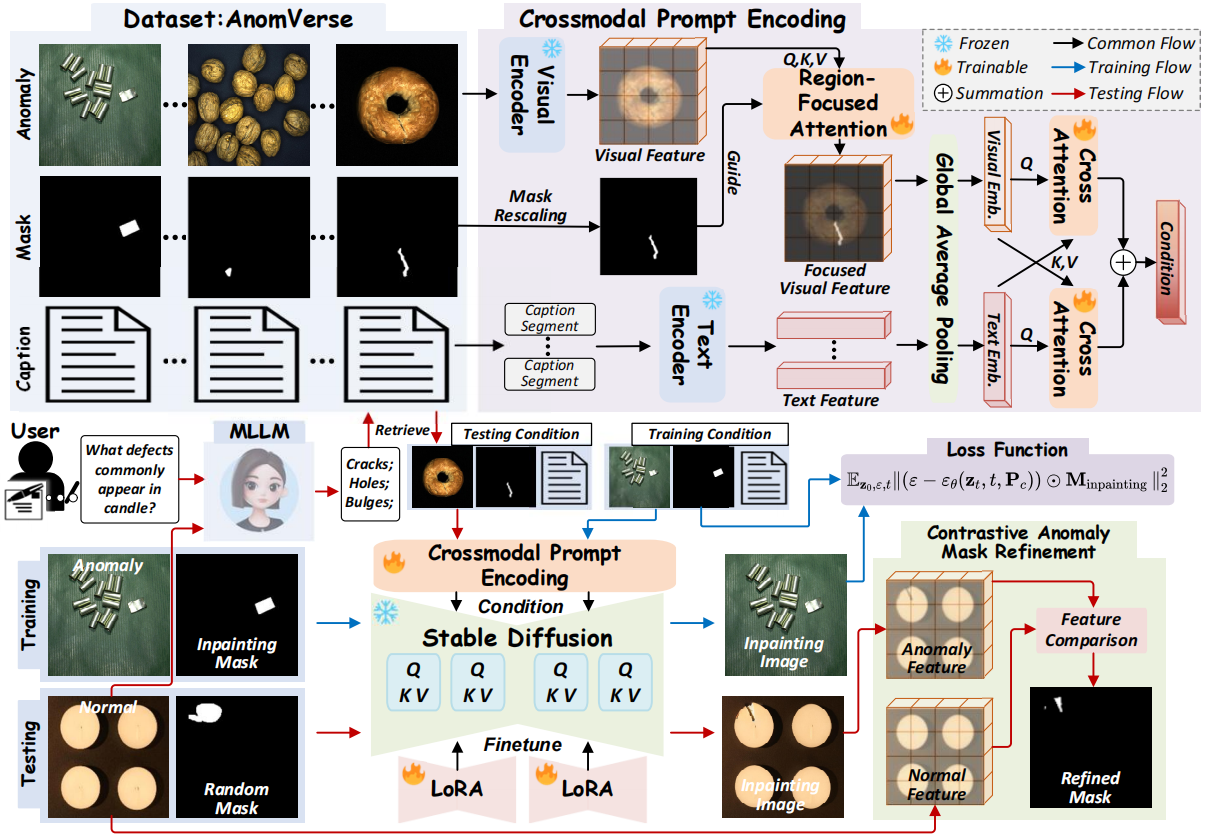

Figure 2: **Overall framework of Anomagic.** Our method employs a crossmodal prompt encoding to extract conditions from selected anomaly–mask–caption triplets in AnomVerse, which are then used to guide the inpainting process. During testing, a contrastive anomaly mask refinement module is introduced to further enhance the accuracy of the predicted anomaly masks.

Crossmodal Prompt Encoding

Both visual and textual modalities possess rich semantic information that can reduce ambiguity while maintaining flexibility. To exploit these characteristics, we introduce the CPE scheme that extracts semantics from crossmodal prompts, including an anomalous image $\mathbf{I}^{\text{ref}}$ as a visual prompt, an anomaly mask $\mathbf{M}^{\text{ref}}$, and corresponding anomaly captions $\mathbf{t}^{\text{ref}}$ as textual prompts.

Region-focused Visual Guidance. To effectively capture visual cues pertinent to anomalies, we leverage a pre-trained CLIP (Radford et al. 2021) image encoder to process the reference image $\mathbf{I}^{\text{ref}}$, yielding a feature map $\mathbf{F}$. However, directly applying the encoder to the entire image may overlook subtle defects due to the dominance of normal object semantics. To counteract this, we introduce a region-focused attention mechanism that isolates anomaly-specific representations by applying the binary anomaly mask $\mathbf{M}^{\text{ref}}$,

$$\mathbf{P}_v = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}} - (1-\mathbf{M}^{\text{ref}})\cdot C\right)\mathbf{V}, \quad (3)$$

where $\mathbf{Q} = \theta_q(\mathbf{F})$, $\mathbf{K} = \theta_k(\mathbf{F})$, and $\mathbf{V} = \theta_v(\mathbf{F})$ are the query, key, and value projections, respectively. The constant C serves as a large scaling factor to attenuate attention weights in masked regions. This approach diminishes the influence of background contexts by assigning negligible attention scores to masked regions.

Fine-grained Textual Semantics. To accommodate verbose technical captions that exceed standard token limits (e.g., CLIP's 77-token constraint (Radford et al. 2021)), we implement a hierarchical encoding strategy for long texts. The input text $\mathbf{t}^{\text{ref}}$ is divided into semantically coherent segments, each of which is encoded by CLIP. These encodings are then aggregated through mean-pooling to form a comprehensive global embedding $\mathbf{P}_t$. This method preserves long-range dependencies and enables precise control over textual prompts.

Crossmodal Prompt Fusion. We further develop a crossmodal fusion architecture to process visual features $\mathbf{P}_v$ and textual embeddings $\mathbf{P}_t$ through modality-specific cross-attention blocks, capturing bidirectional interactions between the modalities. The resulting unified crossmodal semantic feature $\mathbf{P}_c$ serves as the conditioning input for the diffusion process:

$$\mathbf{P}_c = \text{CrossFusion}(\mathbf{P}_v, \mathbf{P}_t), \quad (4)$$

where CrossFusion$(\cdot)$ represents a composite function integrating projection networks for $\mathbf{P}_v$ and $\mathbf{P}_t$ alignment, cross-attention, and fusion operations. Only the attention module within visual guidance and the CrossFusion$(\cdot)$ module are optimized during training, with the pre-trained CLIP model

Algorithm 1: Training process of Anomagic

Input: Reference triplets $(\mathbf{I}^{\text{ref}}, \mathbf{M}^{\text{ref}}, \mathbf{t}^{\text{ref}})$
Output: LoRA weights θ_L and trainable CPE weights θ_{CPE}

1. Initialization:
Load pretrained SD and CLIP parameters.
Initialize $\theta_L, \theta_{\text{CPE}}$.
2. Main Loop:
foreach *sample* $(\mathbf{I}^{\text{ref}}, \mathbf{M}^{\text{ref}}, \mathbf{t}^{\text{ref}})$ **do**
  (a) Crossmodal Prompt Encoding:
  Compute prompt features $\mathbf{P}_c$ via Equations (3)–(4).
  (b) Input Preparation:
  Dilate $\mathbf{M}^{\text{ref}}$ to get the inpainting mask $\mathbf{M}_{\text{inpainting}}$.
  Mask $\mathbf{I}^{\text{ref}}$ with $\mathbf{M}_{\text{inpainting}}$ to get $\mathbf{I}_{\text{input}}$.
  (c) Compute Loss and Update:
  for $t \sim \mathcal{U}[1, T]$ **do**
    Add noise into the embedding of $\mathbf{I}_{\text{input}}$ to obtain $\mathbf{z}_t$.
    Predict noise with $\varepsilon_\theta(\mathbf{z}_t, t, \mathbf{P}_c)$.
    Compute masked denoising loss $\mathcal{L}'_{\text{LDM}}$ via Eq. (5).
    Backpropagate $\mathcal{L}'_{\text{LDM}}$ to update $\theta_L, \theta_{\text{CPE}}$.
  end
end
repeat
  continue
until *convergence*;

kept frozen to preserve its generalization capacity. The trainable parameters are collectively denoted as θ_{CPE}.

Training via Prompt-guided Inpainting

To achieve precise anomaly generation that adheres to both semantic and spatial constraints, we fine-tune a pre-trained inpainting SD model using LoRA on its cross-attention layers. This efficient fine-tuning approach enables the model to generate defects that align with the crossmodal prompts while respecting the specified spatial regions. In each training iteration, we sample a triplet consisting of $\mathbf{I}^{\text{ref}}$, $\mathbf{M}^{\text{ref}}$, and $\mathbf{t}^{\text{ref}}$ from AnomVerse. Subsequently, $\mathbf{M}^{\text{ref}}$ undergoes dilation to produce an expanded inpainting mask, $\mathbf{M}_{\text{inpainting}}$, within which the pixels of the $\mathbf{I}^{\text{ref}}$ are masked to produce the input image $\mathbf{I}_{\text{input}}$ for the denoising process. During denoising, we compel the diffusion model to inpaint the masked regions with the original anomalies by minimizing the noise prediction error, while ensuring that the non-masked regions follow the original image's denoising trajectory. This is formalized in the modified loss function:

$$\mathcal{L}'_{\text{LDM}} = \mathbb{E}_{\mathbf{z}_0,\varepsilon,t} \left\| (\varepsilon - \varepsilon_\theta(\mathbf{z}_t, t, \mathbf{P}_c)) \odot \mathbf{M}_{\text{inpainting}} \right\|_2^2. \quad (5)$$

The training protocol is detailed in Algorithm 1.

Inference

Prompt-driven Anomaly Generation. Our framework facilitates zero-shot anomaly synthesis by leveraging arbitrary anomaly-mask-caption triplets $(\mathbf{I}^{\text{ref}}, \mathbf{M}^{\text{ref}}, \mathbf{t}^{\text{ref}})$. During the generation phase, we initially extract crossmodal features from the given triplet using the CPE scheme, as detailed in Equation (4), resulting in the crossmodal condition $\mathbf{P}_c$. Subsequently, a coarse anomaly mask, $\mathbf{M}_{\text{inpainting}}$, is randomly sampled. The diffusion model then synthesizes anomalies on the target image $\mathbf{I}$ by updating the latent variables exclusively within the masked regions, thereby preserving the integrity of normal areas. Although our method is capable of accommodating arbitrary prompts, we have implemented an automated pipeline by default to streamline the process. This pipeline selects the most pertinent prompts from AnomVerse, allowing users to simply provide a user query Q, such as "What defects commonly appear in cashews?" Our system then employs an MLLM to generate a semantic response A, for instance, "cracks, holes, bulges, scratches." Then we retrieve semantically aligned anomaly categories from AnomVerse using the MLLM, and select the corresponding anomaly-mask-caption triplets. Additional results regarding the retrieval are elaborated in Appendix B.

Contrastive Anomaly Mask Refinement. Considering that the synthesized anomalies might not entirely fill the initial coarse mask, we introduce a contrastive anomaly mask refinement strategy for better alignment. As illustrated in Figure 2, our inpainting-based generation approach ensures that discrepancies occur only in anomaly regions. Thus, we compute pixel-level differences between input and output images and apply a threshold to obtain refined binary masks, denoted as $\mathbf{M}_{\text{r}}$. To achieve this, we utilize a pre-trained MetaUAS (Gao 2024) model, which is specifically designed to detect pixel-level discrepancies between two images, with a threshold of 0.9. As shown in Figure 3, our refinement step provides better alignment between the anomalies and masks.

Experiments

Experimental Settings

Dataset. Our AnomVerse dataset integrates data from 13 publicly accessible datasets. For evaluation, we predominantly utilize the MVTec AD and VisA datasets, while the remaining 11 datasets are reserved for training purposes. The VisA dataset (Zou et al. 2022) constitutes the primary resource for our experimental investigations, with supplementary results from MVTec AD (Bergmann et al. 2021) detailed in Appendix E.

Implementation details. Our proposed Anomagic is implemented using Diffusers (von Platen et al. 2022), based on Stable Diffusion v1.5 with OpenCLIP ViT-H/14 (Ilharco et al. 2021). We adopt a 20-step DDIM sampler (Song et al. 2020a) to balance generation quality and computational efficiency. For training downstream anomaly detection methods, we generate one anomalous image per normal image.

Metrics. For anomaly generation, we adopt the Inception Score (IS) and the Intra-cluster LPIPS distance (IL) (Hu et al. 2024). For anomaly detection and localization, we report the image-level area under the ROC curve (I-ROC) and maximum F1 score (I-F1), as well as the pixel-level area under the per-region overlap curve (PRO) and maximum F1 score (P-F1) (Luo et al. 2025a).

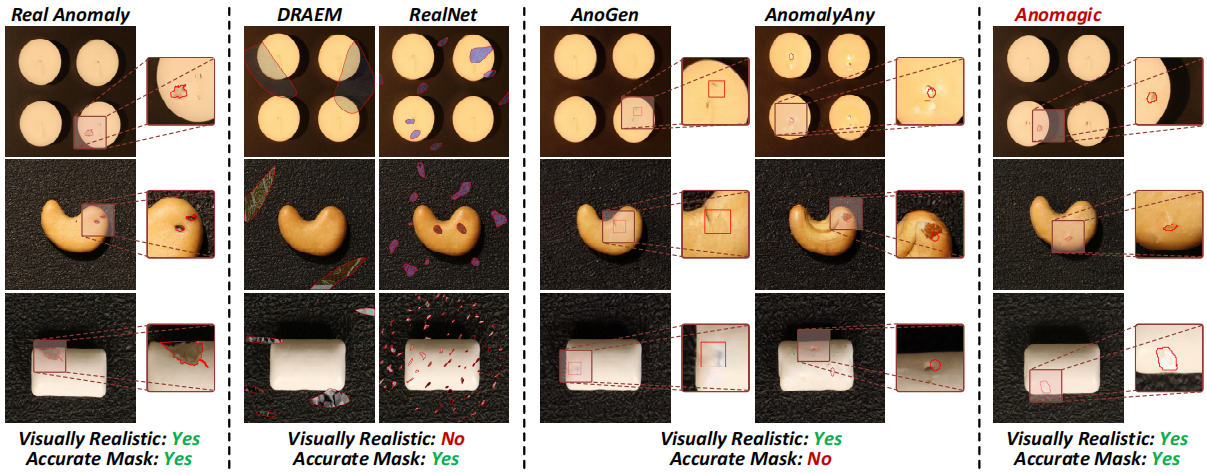

Figure 3: **Qualitative comparison of anomaly generation performance.** Anomalies are highlighted with red circles. Unlike existing zero-shot methods (DRAEM and RealNet) and few-shot methods (AnoGen), Anomagic uniquely achieves both visually realistic anomaly synthesis and accurate anomaly mask generation.

Comparison in Anomaly Generation

Baselines. We primarily compare our method against existing zero-shot anomaly generation techniques, namely DRAEM (Zavrtanik et al. 2021), RealNet (Zhang et al. 2024), and AnomalyAny (Sun et al. 2025). Additionally, we benchmark our approach against AnoGen (Gui et al. 2024), a state-of-the-art (SOTA) few-shot anomaly generation method trained using one image per defect type per category from VisA.

Quantitative Results. As illustrated in Table 1, our method surpasses all zero-shot anomaly generation methods in IS and IL, exceeding even AnoGen, which requires real anomalies for training. These results indicate that Anomagic excels in both the realism and diversity of generated anomalies.

Cate.	AnoGen	DRAEM	RealNet	AnoAny	**Anomagic**
pcb1	1.48/0.26	1.35/0.24	1.38/0.23	**1.64/0.10**	1.58/0.33
pcb2	**1.79/0.45**	1.33/0.43	1.32/0.40	1.36/0.25	1.70/0.42
pcb3	1.70/0.32	1.53/0.28	1.46/0.27	1.62/0.18	**2.00/0.28**
pcb4	1.46/0.42	1.36/0.39	1.34/0.38	1.46/0.38	**1.62/0.41**
maca.1	**2.64/0.41**	1.87/0.37	1.97/0.37	1.81/0.39	2.07/0.39
maca.2	**2.73/0.49**	2.42/0.47	2.40/0.48	2.59/0.38	2.55/0.49
caps.	1.69/0.61	1.55/0.60	1.52/0.60	1.69/0.48	**1.73/0.60**
cand.	**2.61/0.24**	2.49/0.22	2.60/0.23	1.80/0.12	2.55/0.24
cash.	2.28/0.42	2.04/0.40	2.01/0.39	1.91/0.45	**2.41/0.44**
chew.	2.30/0.51	2.17/0.47	2.12/0.46	2.48/0.49	**2.68/0.48**
fryum	1.96/0.33	1.91/0.30	1.88/0.29	1.95/0.23	**2.08/0.29**
pipe.	2.51/0.26	2.23/0.34	2.31/0.34	**3.04/0.36**	2.94/0.36
mean	2.10/0.39	1.85/0.37	1.86/0.37	1.94/0.33	**2.16/0.39**

Table 1: **Comparison of IS/IL on VisA dataset.** Best results are in **bold**.

Qualitative Comparisons. Figure 3 demonstrates that our method produces anomalous samples with greater realism compared to cut-paste-based approaches, while also outperforming AnoGen and AnomalyAny in both generation quality and mask accuracy. Note that precise anomaly masks are critical for downstream anomaly detection training. Unlike AnomalyAny, which requires approximately 3 minutes to generate a single anomalous image, our end-to-end synthesis pipeline—including anomaly generation and mask refinement—achieves an average runtime of only 1.2 seconds per sample with a resolution of 512×512 on a single NVIDIA A100 Tensor Core GPU. Appendix D and E further highlights the versatility of our approach across various domains, including medical imaging and web-crawled image data.

Data Distribution Comparison. Figure 4 visualizes the feature distributions of normal samples, real anomalies, and synthesized anomalies using t-SNE with a pre-trained ResNet50 (He et al. 2016). As shown, the anomalies generated by Anomagic exhibit a distribution closely aligned with that of real anomalies, outperforming the few-shot method AnoGen, suggesting that Anomagic is capable of synthesizing anomalies that resemble real-world defect patterns.

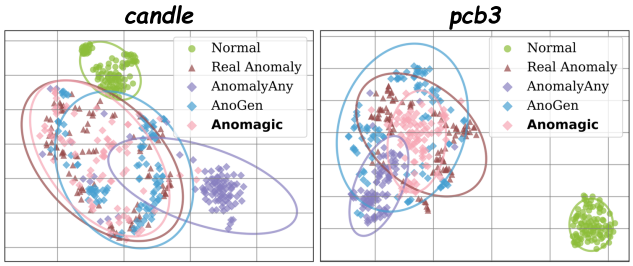

Figure 4: **t-SNE visualization of marginal group distributions for "candle" and "pcb3" objects from VisA.**

Comparison in Anomaly Detection

Evaluation Paradigm: Unlike existing methods (Hu et al. 2024; Jin et al. 2025) that typically employ UNet for supervised training without comparing to SOTA anomaly detection techniques, our approach focuses on enhancing a recent

SOTA method, INP-Former++ (Luo et al. 2025b), by integrating advanced anomaly generation techniques. More details about this paradigm are presented in Appendix C.

Quantitative Comparison. As shown in Table 2, our generated samples enhance anomaly detection performance in INP-Former++ on VisA, and even surpass AnoGen that requires real anomalies for training, with higher I-F1/PRO/P-F1 (96.77%/95.92%/54.00% vs 96.55%/95.62%/52.61%).

Method	I-ROC	I-F1	PRO	P-F1
PatchCore	95.10	94.10	91.20	44.70
RD4AD	96.00	94.30	70.90	42.60
Dinomaly	98.90	96.20	95.30	48.60
AnoGen	**99.09**	96.55	95.62	52.61
DRAEM	99.03	96.58	95.59	51.94
RealNet	99.03	96.75	95.70	52.87
AnoAny	99.01	96.48	95.57	50.76
Anomagic	99.08	**96.77**	**95.92**	**54.00**

Table 2: **Comparison of anomaly detection performance on VisA**. Below presents INP-Former++ augmented with selected anomaly generation methods. Best results are in **bold**.

Ablation study

Overall Ablation. As shown in Table 3, incorporating CPE and LoRA significantly improves both the quality of anomaly generation and the performance of downstream anomaly detection, yielding a 2.06% increase in P-F1 score over the baseline DRAEM.

Module		Metric					
CPE	LoRA	IS	IL	I-ROC	I-F1	PRO	P-F1
		1.85	0.375	99.03	96.58	95.59	51.94
✓		**2.16**	**0.394**	99.04	96.71	95.88	53.87
✓	✓	**2.16**	**0.394**	**99.07**	**96.77**	**95.92**	**54.00**

Table 3: **Ablation Study on CPE and LoRA.** The method excludes both CPE and LoRA modules corresponds to DRAEM. **Bold** values indicate the best performance.

Unimodal and Crossmodal Prompt Evaluation. To assess the adaptability of Anomagic in generating anomalies, we conducted experiments on VisA using both unimodal prompts—namely Anomagic-Text and Anomagic-Visual—and crossmodal prompts (Anomagic-Cross). As shown in Figure 5, Anomagic reliably produces high-fidelity anomalies even when provided with single-modality prompts. Moreover, anomalies synthesized by Anomagic-Cross display enhanced realism, closely mirroring the texture, structure, and contextual coherence of genuine anomalous instances. These findings underscore the method's flexibility, allowing users to employ either unimodal or crossmodal prompts according to their specific requirements.

Anomaly Generation under User-Defined Prompts. Figure 6 illustrates our method's capability to generate realistic anomalies using arbitrary user-specified prompts, including those derived from real anomalies distinct from our training data. For example, although our model was not trained on VisA, it effectively produces convincing anomalies based on diverse prompts inspired by VisA. Furthermore, our method can also effectively generate realistic anomalies even with novel web-sourced prompts, as detailed in the Appendix E. This positions it as a foundational framework for adaptable anomaly generation across diverse, user-defined prompts.

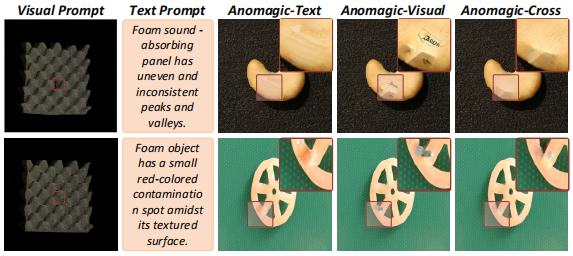

Figure 5: **Visualization of anomaly generation results under unimodal and crossmodal prompts.** Our method effectively synthesizes realistic anomalies in both settings, with Anomagic-Cross producing notably superior results.

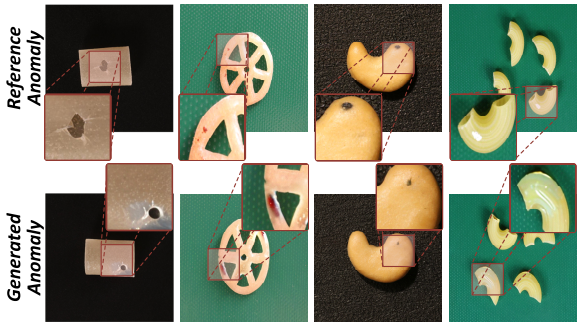

Figure 6: **Visualization of anomaly generation results on VisA using unseen prompts from the same category.** Noting that Anomagic is untrained on VisA, it can still generate realistic anomalies aligned with the provided prompts.

Conclusion

We have presented Anomagic, a crossmodal prompt-driven, zero-shot anomaly synthesis method, together with AnomVerse, a dataset containing 12,987 anomaly–mask–caption triplets. Our prompt-guided inpainting mechanism produces high-fidelity, diverse anomalies that are precisely aligned with their masks. When integrated into INP-Former++, these synthetic anomalies lead to SOTA detection performance. We further demonstrate Anomagic's flexibility in generating anomalies across a wide range of categories using both unimodal and crossmodal prompts, establishing it as a foundational tool for anomaly generation.

Future Work: We plan to augment Anomagic and AnomVerse with fine-grained control over anomaly attributes, such as geometric and material properties, to enable more customizable and targeted anomaly synthesis. To further validate the contrastive anomaly mask refinement strategy,

we will compare boundary accuracy between refined and ground-truth masks, while assessing its robustness across a wider range of anomalies, including low-contrast cases.

Acknowledgments

This work was supported by Fundamental Research Funds for the Central Universities (HUST: 2021GCRC058) and was part by the HPC Platform of Huazhong University of Science and Technology where the computation is completed.

References

Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; and Steger, C. 2021. The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *International Journal of Computer Vision*, 129(4): 1038–1059.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4): 1–10.

Cheng, Y.; Cao, Y.; Yao, H.; Luo, W.; Jiang, C.; Zhang, H.; and Shen, W. 2025. A Comprehensive Survey for Real-World Industrial Defect Detection: Challenges, Approaches, and Prospects. arXiv:2507.13378.

Dai, Z.; Zeng, S.; Liu, H.; Li, X.; Xue, F.; and Zhou, Y. 2024. SeaS: few-shot industrial anomaly image generation with separation and sharing fine-tuning. *arXiv preprint arXiv:2410.14987*.

Dang, M.; Singh, A.; Zhou, L.; Ermon, S.; and Song, J. 2025. Personalized Preference Fine-tuning of Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8020–8030.

Deng, H.; and Li, X. 2022. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9737–9746. Doi: 10.1109/CVPR52688.2022.00951.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Duan, Y.; Hong, Y.; Niu, L.; and Zhang, L. 2023. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 571–578.

Fan, L.; Fan, D.; Hu, Z.; Ding, Y.; Di, D.; Yi, K.; Pagnucco, M.; and Song, Y. 2025. Manta: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25518–25527.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gao, B.-B. 2024. MetaUAS: Universal Anomaly Segmentation with One-Prompt Meta-Learning. *Advances in Neural Information Processing Systems*, 37: 39812–39836.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gui, G.; Gao, B.-B.; Liu, J.; Wang, C.; and Wu, Y. 2024. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *European Conference on Computer Vision*, 210–226. Milan, Italy, September 29-October 4, 2024: Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

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 *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 8526–8534.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP. GitHub repository.

Jiang, X.; Li, J.; Deng, H.; Liu, Y.; Gao, B.-B.; Zhou, Y.; Li, J.; Wang, C.; and Zheng, F. 2024. MMAD: The First-Ever Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection. In *International Conference on Learning Representations*.

Jin, Y.; Peng, J.; He, Q.; Hu, T.; Wu, J.; Chen, H.; Wang, H.; Zhu, W.; Chi, M.; Liu, J.; et al. 2025. Dual-Interrelated Diffusion Model for Few-Shot Anomaly Image Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30420–30429.

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, D.; Li, J.; and Hoi, S. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36: 30146–30166.

Liu, J.; Xie, G.; Wang, J.; Li, S.; Wang, C.; Zheng, F.; and Jin, Y. 2024. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1): 104–135.

Luo, W.; Cao, Y.; Yao, H.; Zhang, X.; Lou, J.; Cheng, Y.; Shen, W.; and Yu, W. 2025a. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9974–9983.

Luo, W.; Yao, H.; Cao, Y.; Chen, Q.; Gao, A.; Shen, W.; Zhang, W.; and Yu, W. 2025b. INP-Former++: Advancing Universal Anomaly Detection via Intrinsic Normal Prototypes and Residual Learning. *arXiv preprint arXiv:2506.03660*.

Nguyen, Q.; Vu, T.; Tran, A.; and Nguyen, K. 2023. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36: 76872–76892.

Qian, H.; Chen, Y.; Lou, S.; Shahbaz Khan, F.; Jin, X.; and Fan, D.-P. 2024. Maskfactory: Towards high-quality synthetic data generation for dichotomous image segmentation. *Advances in Neural Information Processing Systems*, 37: 66455–66478.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328. Doi: 10.1109/CVPR52688.2022.01392.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Song, J.; Park, D.; Baek, K.; Lee, S.; Choi, J.; Kim, E.; and Yoon, S. 2025. DefectFill: Realistic Defect Generation with Inpainting Diffusion Model for Visual Inspection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18718–18727.

Song, J.; et al. 2020a. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sun, H.; Cao, Y.; Dong, H.; and Fink, O. 2025. Unseen Visual Anomaly Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25508–25517.

von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; Nair, D.; Paul, S.; Berman, W.; Xu, Y.; Liu, S.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models. GitHub repository.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; and Liu, Z. 2025. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13294–13304.

Yang, S.; Chen, Z.; Chen, P.; Fang, X.; Liang, Y.; Liu, S.; and Chen, Y. 2024. Defect spectrum: a granular look of large-scale defect datasets with rich semantics. In *European Conference on Computer Vision*, 187–203. Springer.

Zavrtanik, V.; et al. 2021. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.

Zhang, G.; Cui, K.; Hung, T.-Y.; and Lu, S. 2021. Defect-GAN: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2524–2534.

Zhang, L.; et al. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, X.; et al. 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*, 16699–16708.

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.