CHIMERA: Controllable High-quality Image-Mask Extraction for Reliable Diffusion-Based Anomaly Synthesis

JoungBin Lee[1]*, Hyunkoo Lee[1]*, Jini Yang[1]*, Chaehyun Kim[1], Jung Yi[1], Seok Hwangbo[2], Hyeoncheol Lee[2], Minho Chun[2], Eunjo Jeong[2], Seungryong Kim[1]†

[1]KAIST AI
[2]Samsung Display
{joungbinlee, guu980, jini.yang, kchyun, jungyi9906, seungryong.kim}@kaist.ac.kr
{seok.hwangbo, hayacy.lee, cmh1866.chun, eunjo83.jung}@samsung.com

Abstract

We present **CHIMERA**, a novel framework for generating realistic, generalizable, and prompt-driven industrial anomalies from natural language instructions. Our method addresses two key challenges in text-guided anomaly synthesis: (1) the scarcity of scalable, high-quality paired anomaly data and (2) the difficulty of efficiently adapting large diffusion models to domain-specific tasks without overfitting. To tackle these challenges, we first introduce a *Vision-Language Model (VLM)-guided data curation pipeline* that automatically generates semantically rich and spatially grounded captions from normal images, enabling effective dataset augmentation without manual annotations. Building upon this, we propose a *parameter-efficient fine-tuning strategy* that adapts a pre-trained Diffusion Transformer (Stable Diffusion 3) using lightweight LoRA adapters. By aligning structured prompts with the model's pre-trained language-vision prior and introducing auxiliary attention-based mask supervision, our method prevents overfitting, enhances spatial consistency, and ensures efficient training even with limited data. Extensive experiments show that **CHIMERA** is the first unified framework to achieve controllable, scalable, and generalizable industrial anomaly generation by integrating VLM-guided data curation with efficient diffusion-based training, significantly improving anomaly detection in low-data and unseen scenarios.

Code — https://github.com/cvlab-kaist/CHIMERA

Introduction

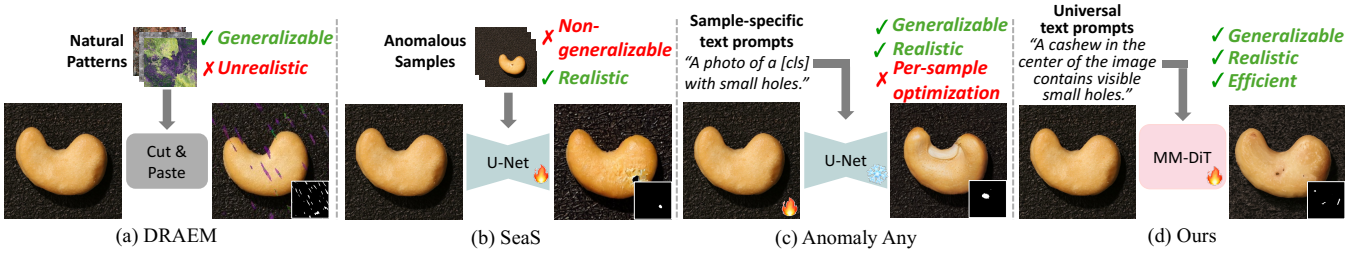
Industrial anomaly inspection, i.e., tasks including defect detection, localization, and so on, is a vital component of manufacturing systems, as it ensures both product quality and reliability (Pang et al. 2021; Duan et al. 2023). To meet the demands of modern industrial environments, it requires not only high-speed inspection but also the capability to detect a wide range of potential defects. For precise detection, it is crucial to accurately distinguish the boundary between the distribution of normal and abnormal data (Bergmann et al. 2020; Roth and et al. 2021; Choe, Lee, and Sim 2023). However, collecting large-scale abnormal data remains a major challenge due to the rarity of real defects and the high cost of data acquisition, which constitutes a fundamental limitation in supervised anomaly inspection. Despite recent advancements, the majority of anomaly detection methods continue to depend on unsupervised learning exclusively with normal data (Fang et al. 2023; Lu et al. 2023; You et al. 2022), and such methods often struggle to generalize to diverse and complex anomaly types (Jin et al. 2025; Sun et al. 2025).

To address data imbalance, a simple yet effective approach is to synthesize additional defective samples. As shown in Fig. 1, some methods augment normal samples by pasting random patterns (Zavrtanik, Kristan, and Skočaj 2021; Zhang, Xu, and Zhou 2024), but these are often applied arbitrarily without regard for object regions, resulting in unrealistic and inconsistent images. Alternatively, generative models have been employed to synthesize visual anomalies, producing more realistic samples. Among them, training-based diffusion models (Yi and Wu 2023; Hu et al. 2024; Jin et al. 2025; Dai et al. 2024) often suffer from a lack of generalizability, as they tend to overfit to semantically meaningless tokens due to insufficient training data, thereby hindering the generation of diverse anomaly types. In contrast, per-sample optimization based methods (Sun et al. 2025) can generate realistic and generalizable anomalies, but they are significantly slow due to iterative optimization for each image.

Given the limitations of existing anomaly synthesis methods, our goal is to develop a generalizable, realistic, and efficient anomaly generation framework that does not overfit to limited training data. Achieving generalizability requires addressing the common issue of diffusion models overfitting to small-scale anomaly datasets, which limits their diversity and scalability.

To overcome this, we propose a vision-language model (VLM)-guided data curation pipeline that generates fine-grained captions describing both the semantic category and spatial location of anomalies. These structured prompts provide rich semantic and spatial guidance, facilitating model training by enabling conditional generation of diverse anomalies without requiring extensive manual annotations.

*Equal contribution.
†Co-corresponding authors.



(a) DRAEM (b) SeaS (c) Anomaly Any (d) Ours

Figure 1: Comparison of Anomaly Generation Methods. (a) DRAEM uses cut-and-paste augmentation with natural patterns, offering generalizability but resulting in unrealistic anomalies. (b) Seas leverages Stable Diffusion with anomalous exemplars, producing realistic samples but lacking generalizability. (c) Anomaly Any uses sample-specific prompts, achieving realism and generalizability but requiring costly per-sample optimization. (d) **Ours** introduces universal, structured text prompts to guide Stable Diffusion, enabling anomaly generation that is realistic, generalizable, and efficient.

To ensure generalization from limited data while maintaining efficiency, we employ a lightweight PEFT strategy using LoRA adapters (Hu et al. 2022) and align prompts with the pre-trained diffusion model's linguistic patterns to preserve its language-vision prior and prevent overfitting.

Furthermore, we supervise the image–text cross-attention maps during training to explicitly ground anomalies in correct spatial locations. This cross-attention guidance enhances spatial consistency, stabilizes learning, and reduces overfitting. Consequently, unlike prior approaches that require category-specific models or slow test-time optimization, our method unifies anomaly generation and localization within a single model, enabling fast, controllable, and generalizable anomaly synthesis even for unseen objects and anomaly types.

Our contributions are summarized as follows:

- We propose a generalizable and efficient anomaly synthesis framework that prevents overfitting to limited data by aligning structured prompts with the language-vision prior of pre-trained diffusion models.
- We introduce a VLM-guided data curation pipeline that generates semantically rich and spatially grounded prompts, facilitating effective model training without manual annotations.
- We adopt Stable Diffusion 3 as the generative backbone and apply parameter-efficient fine-tuning (LoRA), enabling high-quality and fast anomaly generation within a unified model.
- We design an auxiliary mask supervision mechanism by decoding attention maps, enhancing spatial consistency, stabilizing training, and improving generation quality.

Related Work

Abnormal Data Augmentation. Due to the rarity and high collection cost of real-world abnormal samples, many prior works focus on synthesizing pseudo-anomalies to enrich the training dataset. Traditional methods often apply handcrafted transformations such as cut-and-paste, noise injection, or texture replacement to normal images in order to simulate defects (Zavrtanik, Kristan, and Skočaj 2021). While these methods are simple and computationally efficient, they frequently fail to capture the semantic plausibility or visual realism of real anomalies, limiting their effectiveness in downstream detection tasks.

Diffusion-based Data Augmentation. Recently, generative models, particularly diffusion models, have emerged as powerful tools for synthesizing realistic images. Diffusion-based anomaly generation methods (Hu et al. 2024; Jin et al. 2025; Dai et al. 2024) aim to generate high-fidelity anomalies by learning the distribution of normal data and guiding the model to generate deviations. Training-based approaches typically rely on special tokens or label conditions, but may overfit to meaningless prompts due to limited training data. Training-free approaches, on the other hand, attempt to edit images directly via attention-based masks, but often suffer from poor spatial consistency between the edited regions and the intended anomalies. For instance, Anomaly Any (Sun et al. 2025) adopts a training-free diffusion strategy guided by attention maps, but lacks precise alignment between the synthesized defects and object semantics.

In this work, we combine the strengths of prompt-driven diffusion and spatially guided supervision to produce diverse and semantically meaningful anomalies.

Abnormal Detection. Anomaly detection aims to identify patterns that deviate from normality, typically under a one-class learning setup where only normal samples are available during training. Recent methods leverage pretrained vision models (Roth and et al. 2022; Cohen and Hoshen 2020) or self-supervised learning (Zavrtanik, Kristan, and Skočaj 2021; Li and et al. 2021) to extract robust representations of normal features. Classical approaches such as PaDiM (Defard and et al. 2021) and PatchCore (Roth and et al. 2022) model the distribution of normal patches, while others like FastFlow (Yu and et al. 2021) utilize flow-based embeddings. Since real-world anomalies are rare, augmenting training with realistic synthetic anomalies (Zavrtanik, Kristan, and Skočaj 2021; Hu et al. 2024) has shown to improve detection and localization. Our method further enhances performance by generating diverse, semantically meaningful anomalies that align well with visual contexts, benefiting downstream tasks across various benchmarks.

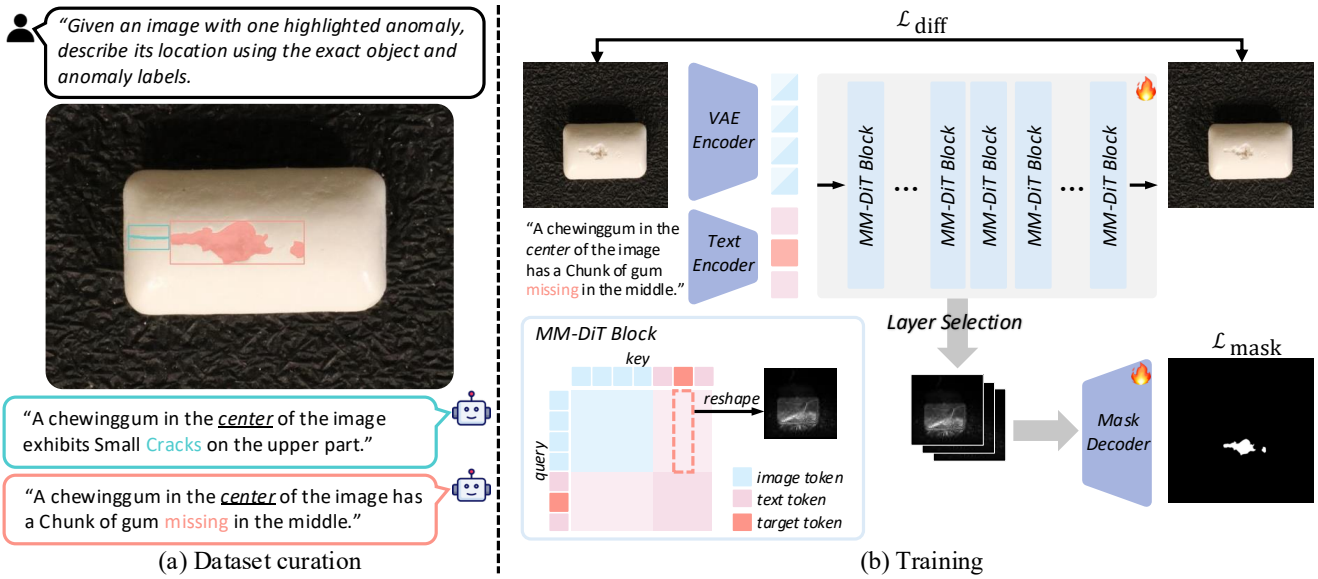


Figure 2: Overall pipeline of CHIMERA. (a) A vision-language model is employed to generate fine-grained captions describing the anomaly category and region. (b) We fine-tune a diffusion transformer using LoRA adapters. During training, a noised image at a randomly sampled timestep is provided along with the corresponding caption. In addition to the standard diffusion objective, we extract image-to-text attention maps corresponding to the anomaly-related text tokens from selected layers. These maps are passed to a mask decoder to predict the anomaly region, which is supervised using a mask loss.

Preliminaries

Rectified Flow Framework

Diffusion models learn to generate images by progressively denoising samples drawn from Gaussian noise (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020). More recently, rectified flow (Liu, Gong, and Liu 2022) simplified this stochastic process by defining a deterministic, linear path between the noise ϵ sampled from $\mathcal{N}(0,1)$ and sample x_0 from image distribution p_0. The forward process for timestep t is defined as:

$$z_t = (1-t)\cdot x_0 + t\cdot \epsilon. \tag{1}$$

The model is trained with a conditional flow matching objective, which aims to predict the velocity $v_t(x)$ between the noise and sample:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,p_t(x)} \left\| (\epsilon - x_0) - v_t(x) \right\|^2. \tag{2}$$

Multi-modal Diffusion Transformer

Architectural advances have significantly propelled diffusion models forward. While early approaches employed U-Net backbones (Rombach et al. 2022), recent diffusion transformers demonstrate superior scalability and image fidelity (Peebles and Xie 2023). In particular, multi-modal diffusion transformer (MM-DiT) variants (Esser et al. 2024; Stability AI 2024; Black Forest Labs 2024) further enhance generation quality through multi-modal attention.

Let $X_I \in \mathbb{R}^{n_I \times d}$ and $X_T \in \mathbb{R}^{n_T \times d}$ denote the image and text features in each layer. These are projected and concatenated into query, key, and value matrices Q, K and $V \in \mathbb{R}^{(n_I+n_T)\times d_K}$ as follows:

$$Q = \mathcal{P}_Q(X_I, X_T) \tag{3}$$
$$K = \mathcal{P}_K(X_I, X_T) \tag{4}$$
$$V = \mathcal{P}_V(X_I, X_T) \tag{5}$$

where $\mathcal{P}_Q$, $\mathcal{P}_K$, and $\mathcal{P}_V$ represent the projection and concatenation operations that map image and text features into the attention space. The multi-modal attention output is then computed as:

$$\text{MM-Attn}(X_I, X_T) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_K}}\right) V. \tag{6}$$

The attention matrix can be partitioned into four regions based on the query-key modality pairings: image-to-image, image-to-text, text-to-image, and text-to-text. Among these, the image-to-text region, which we denote as $A_{i\to t} \in \mathbb{R}^{n_I \times n_T}$, demonstrates strong image-text alignment, resembling the cross-attention behavior observed in U-Net-based architectures (Hertz et al. 2022; Helbling et al. 2025). In this work, we leverage this alignment for anomaly mask prediction during training and also as an anomaly mask for downstream tasks.

Methods

We introduce *CHIMERA*, starting from a data preprocessing pipeline that utilizes vision-language models (VLMs) to automatically generate diverse and semantically rich captions for abnormal images, requiring minimal human supervision. Then, we propose a unified and lightweight training scheme to effectively harness the prior knowledge encoded in diffusion transformers.

Dataset Curation

Fine-grained descriptions offer richer conditioning signals for training diffusion models, enhancing both sample quality and semantic alignment (Saharia et al. 2022; Malhi et al. 2025). For diffusion-based anomaly image generation, effective supervision requires detailed information about anomaly types and their locations. To this end, we propose a dataset curation pipeline that decomposes composite anomaly classes and generates fine-grained captions to construct a more informative training set, which is illustrated in Figure 2 (a).

Anomaly Mask Decomposition. To facilitate the diffusion model's capability to comprehend and manipulate fine-grained anomalies through natural language guidance, we decompose the mask into multiple spatially and semantically distinct regions. Even when anomalies belong to the same category, if their spatial distance exceeds a predefined threshold, we treat them as separate instances. This decomposition encourages the model to learn disentangled and localized abnormal features, enabling the model to recognize and control detailed anomaly patterns.

Let input abnormal image and its corresponding anomaly mask be denoted by $I \in \mathbb{R}^{H\times W\times 3}$ and $M \in \{0,1\}^{H\times W}$, respectively. To achieve finer localization, we apply a spatial clustering algorithm (Ester et al. 1996) to group spatially related abnormal regions into coherent clusters. For each anomaly type $c \in \{1,\dots,C\}$, clustering yields N_c spatially disconnected regions for type c, which results in a decomposed binary mask set $\mathcal{M}$:

$$\mathcal{M} = \{M_{c,n} \mid c = 1,\dots,C,\ n = 1,\dots,N_c\} \tag{7}$$

where $M_{c,n}$ corresponds to the n-th spatial cluster of anomaly type c.

Fine-Grained Region-Level Captioning. To extract rich textual descriptions of each abnormal region, we use a vision-language model (VLM) that takes a abnormal image I and a corresponding region mask $M_{c,n}$ as input. The mask is overlaid on the image to visually highlight the spatial extent of the anomaly.

The VLM processes this input to generate a caption $P_{c,n}$ that describes the anomaly type c for each n mask location in detail. We prompt the VLM to produce captions in the structured format: "`The` o `in the center of the image has a` c `in the` $r_{c,n}$" where o, c, and $r_{c,n}$ represent the object, anomaly type, and corresponding region, respectively. The resulting triplet of image I, mask $\mathcal{M}$ and detailed caption set $\mathcal{P} = \{P_{c,n}\}$ is then utilized to train diffusion transformer, enabling it to learn fine-grained anomaly generation.

CHIMERA: Unified Framework for Realistic Anomaly Synthesis

We propose a LoRA-based lightweight fine-tuning scheme for training the MM-DiT model, as illustrated in Figure 2 (b). Unlike category-specific approaches, we train a single unified model across diverse object categories and anomaly types. This strategy promotes robust generalization to unseen objects and anomalies, mitigates overfitting to limited industrial datasets, and eliminates the need for costly per-sample optimization.

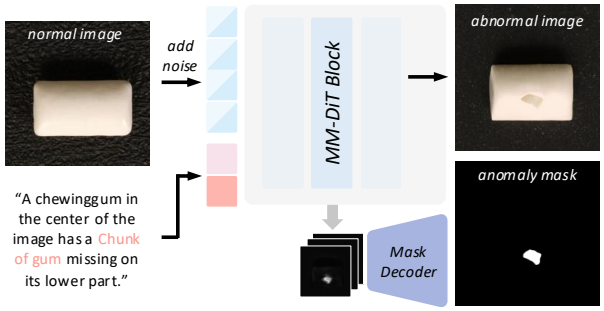


Figure 3: Inference pipeline. Given a normal image and a caption describing the target anomaly, we inject noise at a fixed timestep and guide the denoising process with the caption. Image-to-text attention maps are extracted during generation and decoded into anomaly masks, enabling simultaneous synthesis and localization without post-training or object-specific tokens.

Architecture. To enable realistic and controllable anomaly generation, we adopt a MM-DiT architecture, leveraging its strong generative capacity and semantic understanding (Esser et al. 2024). Given an anomaly image I, anomaly mask $M_{c,n} \in \mathcal{M}$, and corresponding captions $P_{c,n} \in \mathcal{P}$, we encode both the image and text into the latent space. A random timestep t is sampled to add Gaussian noise to the image latent z_t, which is then used as input to the diffusion model.

In parallel, we extract image-to-text attention maps from selected layers $l \in \{start,\cdots,end\}$, denoted as $A^l_{i\to t}$. Let τ is the index of the anomaly-related text token in the image-to-text attention map. We take the τ-th key dimension logits and reshape it to match the spatial dimensions of the input, interpreting it as a soft mask $\hat{m}^l_{c,n} \in \mathbb{R}^{h\times w}$ that indicates the likelihood of each pixel being part of the anomalous region:

$$\hat{m}^l_{c,n} = \text{Reshape}\left(A^l_{i\to t}[:,\tau]\right) \tag{8}$$

where h and w is the reduced spatial resolution of latent image. To produce a high-resolution prediction $\hat{M}_{c,n} \in \mathbb{R}^{H\times W}$, we concatenate all logits in channel dimension and forward through a lightweight mask decoder $\mathcal{D}_\phi$:

$$\hat{M}_{c,n} = \mathcal{D}_\phi\left([\hat{m}^{start}_{c,n},\cdots,\hat{m}^{end}_{c,n}]\right). \tag{9}$$

Training Objective. In addition to the standard diffusion training objective, we introduce an auxiliary supervision loss $\mathcal{L}_{\text{mask}}$ applied to the predicted anomaly mask $\hat{M}_{c,n}$. We define $\mathcal{L}_{\text{mask}}$ using the focal loss (Lin et al. 2017), which encourages the model to attend to fine-grained anomaly regions.

The overall training objective combines the diffusion and mask losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \alpha\cdot\mathcal{L}_{\text{mask}}, \tag{10}$$

where α is a weighting coefficient that balances the localization supervision against the generation objective.

Category	NAS		RealNet		DFMGAN		AnomalyAny		AnoDiff		DualAnoDiff		SeaS		CHIMERA	
	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑
bottle	1.35	0.11	1.58	0.16	1.62	0.12	1.73	0.17	1.58	0.19	2.17	0.43	1.78	0.21	2.15	0.13
cable	1.55	0.38	1.65	0.41	1.96	0.25	2.06	0.41	2.13	0.41	2.15	0.43	2.09	0.42	2.94	0.32
capsule	1.38	0.17	1.62	0.18	1.59	0.11	2.16	0.23	1.59	0.21	1.62	0.32	1.56	0.26	1.62	0.22
carpet	1.03	0.22	1.03	0.25	1.23	0.13	1.10	0.34	1.16	0.24	1.36	0.29	1.13	0.25	1.50	0.28
grid	2.26	0.36	2.24	0.37	1.97	0.13	2.31	0.38	2.04	0.44	2.13	0.42	2.43	0.44	1.63	0.06
hazelnut	2.04	0.31	2.21	0.31	1.93	0.24	2.55	0.32	2.13	0.31	1.94	0.35	1.87	0.31	2.23	0.24
leather	1.07	0.26	1.67	0.36	2.06	0.17	2.26	0.41	1.94	0.41	1.91	0.35	2.03	0.40	3.61	0.36
metal nut	1.66	0.23	1.67	0.24	1.49	0.32	1.82	0.27	1.96	0.30	1.57	0.32	1.64	0.31	1.62	0.28
pill	1.34	0.24	1.45	0.26	1.63	0.16	2.91	0.30	1.61	0.26	1.82	0.38	1.62	0.33	2.41	0.23
screw	1.20	0.31	1.20	0.31	1.12	0.14	1.33	0.32	1.28	0.30	1.43	0.36	1.52	0.31	1.64	0.30
tile	1.31	0.37	1.57	0.48	2.39	0.22	2.66	0.53	2.54	0.55	2.40	0.50	2.60	0.50	5.52	0.47
toothbrush	1.18	0.17	1.19	0.18	1.82	0.18	1.64	0.22	1.68	0.21	2.40	0.48	1.96	0.25	1.85	0.24
transistor	1.34	0.22	1.49	0.30	1.64	0.25	1.66	0.28	1.57	0.34	1.71	0.33	1.51	0.34	1.77	0.26
wood	1.30	0.32	2.22	0.41	2.12	0.35	1.93	0.41	2.33	0.37	2.24	0.40	2.77	0.46	2.52	0.32
zipper	1.52	0.22	1.88	0.24	1.29	0.27	2.14	0.33	1.39	0.25	2.14	0.37	1.63	0.30	3.01	0.32
Average	1.44	0.26	1.64	0.30	1.72	0.20	2.02	0.33	1.80	0.32	1.93	0.38	1.88	0.34	2.40	0.27

Table 1: Comparison on IS and IC-LPIPS on MVTec dataset with various anomaly generation methods.

Inference. We present our inference pipeline in Figure 3 Given a normal image I and a caption P that specifies the desired anomaly type and region, we adopt an SDEdit-style inference strategy (Meng et al. 2021) to synthesize an abnormal image. This approach preserves the structure of the original image while injecting fine-grained anomalies, resulting in high-fidelity abnormal samples.

We introduce noise at a fixed timestep $t = t_{\text{inf}}$, and perform denoising to obtain the abnormal image. Simultaneously, we extract image-to-text attention maps as done during training, and decode them into an anomaly mask using the trained mask decoder. Notably, our method does not require post-training adaptation or object-specific tokens, demonstrating its unified and generalizable nature across diverse object categories and anomaly types.

Experiments

Experimental Settings

Dataset. We conduct extensive experiments on MVTec AD (Bergmann and et al. 2019), which contains 5,354 images across 15 categories, and on VisA (Zou et al. 2022), which includes 10,821 images from 12 categories with pixel-level annotations.

We use only one-third of the abnormal data from each dataset for training. Specifically, we follow the previous anomaly generation setup (Jin et al. 2025) by selecting 466 abnormal samples from the test set and 1,000 normal samples from the training set of MVTec AD. Similarly, to emphasize generalizability under an unseen dataset setting, we sample 396 abnormal samples from the test set and 1,000 normal images from VisA. In total, 2,862 training images are used to train our unified model.

For evaluation, we follow the protocol of AnomalyDiffusion (Hu et al. 2024) and use the normal test set from MVTec AD along with the remaining two-thirds of its abnormal test samples.

In addition, we employ an internal dataset consisting of mobile device panel images, comprising 1,238 normal images and 238 abnormal images. Among them, 1,000 normal images are used for training, while the remaining 238 normal images and 238 abnormal images are used for evaluation. Note that evaluation metrics and comparisons with baseline methods are conducted on MVTec AD and internal dataset, following prior works.

Metric. For generation evaluation, we employ Inception Score (IS) (Salimans et al. 2016) to assess the quality of generated images and Intra-cluster pairwise LPIPS distance (IC-LPIPS) (Zhang et al. 2018) to evaluate sample diversity.

For anomaly detection, we use Area Under the Receiver Operating Characteristic (AUROC), Average Precision (AP), the F1-max score, and Intersection over Union (IoU). Additionally, we report the Per-Region-Overlap (PRO) (Bergmann et al. 2020) metric to evaluate region-level detection performance.

Implementation Details. For anomaly generation, we deploy the Stable Diffusion 3 model (AI 2024). Unlike previous works that train a separate model for each anomaly type (Hu et al. 2024; Jin et al. 2025; Dai et al. 2024), we fine-tune a single unified model that jointly handles multiple object categories and anomaly types, generating 1,000 anomalous images with corresponding paired masks for downstream anomaly detection tasks.

During inference, we inject noise at a fixed timestep ($t = 0.4T$), following an SDEdit-style generation strategy (Meng et al. 2021), enabling the model to preserve the structural content of the original normal image while generating fine-grained anomalies guided by textual prompts. This deterministic path ensures better controllability and se-

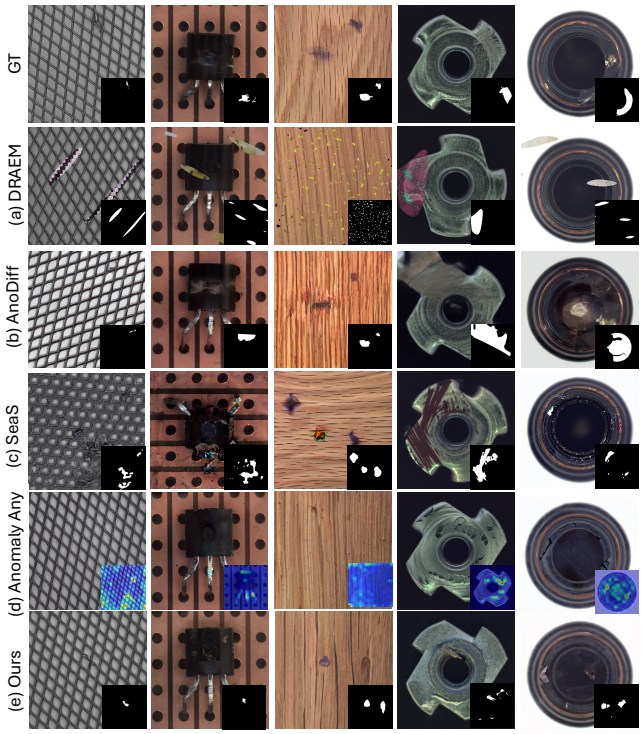


Figure 4: Qualitative comparisons between existing anomaly generation methods on the MVTec AD dataset. CHIMERA synthesizes fine-grained and realistic anomalies that align with object structures, while previous methods often generate unrealistic or spatially inconsistent defects.

mantic consistency in the generated outputs.

For training, we employ a lightweight fine-tuning scheme using LoRA (Hu et al. 2021) with a rank of 8, allowing efficient adaptation of the diffusion transformer while minimizing the number of trainable parameters. The model is optimized using a combination of standard diffusion loss (Ho, Jain, and Abbeel 2020) and a focal mask loss (Lin et al. 2017) to enhance anomaly localization capabilities. We train the model for 30,000 iterations with a batch size of 1 and gradient accumulation steps of 2, effectively simulating a batch size of 2. All experiments are conducted on 4 NVIDIA A6000 GPUs.

Comparison in Anomaly Generation

Baseline. We compare our method against a broad set of anomaly generation approaches, including NAS (Schlüter et al. 2022), RealNet (Zhang, Xu, and Zhou 2024), DFM-GAN (Duan et al. 2023), Anomaly Any (Sun et al. 2025), AnoDiff (Hu et al. 2024), DualAnoDiff (Jin et al. 2025), and SeaS (Dai et al. 2024).

Evaluation results for anomaly generation. As shown in Table 1, our method surpasses existing state-of-the-art approaches in IS scores. By adopting an inversion-based strategy that utilizes normal data typically available in real-world industrial scenarios, we generate high-quality and realistic object images. Unlike previous methods that under-utilize the semantic conditioning capabilities of diffusion models, we fully leverage the prior knowledge of diffusion transformer model by constructing prompts that explicitly describe both the anomaly type and its spatial location. This enables effective generation with only a few anomaly samples, allowing the model to align closely with the target anomaly distribution. With these semantically rich prompts and minimal tuning via lightweight LoRA, our approach achieves high fidelity to anomaly types, domain consistency, and diversity in shape and location, while preserving the strong prior of the pre-trained diffusion transformer.

Metrics →	Mobile Device Panel Dataset					
Methods ↓	IS ↑	IC-L ↑	AUC-I ↑	AP-I ↑	AUC-P ↑	AP-P ↑
DRAEM	4.93	0.40	77.4	77.7	52.9	7.0
AnomalyAny	2.78	0.38	80.7	84.1	38.1	7.4
CHIMERA	2.87	0.32	84.1	88.1	57.2	12.5

Table 2: Comparisons of generation quality and 1-shot anomaly detection performance across different anomaly generation methods. Generation quality is evaluated by Inception Score (IS) and IC-LPIPS, while downstream detection performance is assessed on the mobile device panel dataset using image-level and pixel-level metrics.

Comparison in Anomaly Detection

To evaluate the effectiveness of our method, we conduct experiments under the challenging one-shot anomaly detection setting, where only a single normal image is provided and no anomalous samples are available. Following the procedure described in (Jin et al. 2025), we generate 1,000 anomalous samples conditioned on the given normal image, each accompanied by a spatially aligned anomaly mask.

As shown in Table 3, although our training does not specialize in any particular object or anomaly type, the proposed unified model still delivers competitive results across diverse anomaly patterns. As illustrated in Figure 4, our method produces fine-grained and spatially coherent anomalies that are better aligned with object structure and the described anomaly type, even when generated from only a single normal example.

This visual consistency is attributed to the use of text captions that encode both spatial location and anomaly semantics, enabling the model to synthesize diverse and interpretable anomalies with precise localization.

Anomaly Generation and Detection on Unseen data

We further assess the generalization capability of our method on unseen data. Specifically, we evaluate it on the internal 1000 test set, which consists of high-resolution images of mobile device panel images with scratches as the anomaly type.

Our model successfully generates realistic and diverse anomalies on this previously unseen object and anomaly type, as shown in Table 2, demonstrating strong generalization in terms of visual quality. As shown in Figure 5, the

Metrics →	MVTec AD						
Methods ↓	AUC-I ↑	AP-I ↑	F_1-I ↑	AUC-P↑	AP-P↑	F_1-P ↑	PRO ↑
Per-Class Training Method							
AnoDiff	99.2	99.7	98.7	99.1	81.4	76.3	94.0
DualAnoDiff	–	98.9	–	99.1	84.5	78.8	–
SeaS	98.53	99.41	97.40	96.87	76.19	72.77	92.23
Generalized Method							
DRAEM	94.6	97.0	94.4	92.2	54.1	53.1	83.1
CHIMERA	94.61	96.36	92.80	90.89	53.11	53.75	80.46

Table 3: Comparison of 1-shot anomaly detection performance across different anomaly generation methods. We compare our approach with per-class training–based diffusion models as well as generalized methods on the MVTecAD dataset.

Method	IS ↑	IC-L ↑	AUC-P↑	F_1-P ↑	PRO ↑
CHIMERA	1.50	0.12	99.0	75.0	96.1
Model Structure Ablations					
(a) LoRA only	1.14	0.11	74.4	14.4	47.6
(b) + Mask Loss	1.30	0.20	96.8	64.9	90.4
(c) + Decoder	1.20	0.13	96.9	71.2	90.4
(d) + Layer Selection	1.18	0.14	96.6	65.8	88.6
Data Pre-processing Ablations					
(e) + 'sks' Token	1.11	0.12	68.2	4.5	37.4
(f) + w/o Location	1.20	0.11	92.0	54.1	81.6

Table 4: Ablation study on model structure and data preprocessing. All metrics follow a higher-is-better convention.

generated results show higher visual fidelity compared to AnomalyAny (Sun et al. 2025), AnoDiff (Hu et al. 2024) and DRAEM (Zavrtanik, Kristan, and Skočaj 2021).

Unlike training-free methods such as AnomalyAny, our approach captures both spatial irregularities and semantic context, enabling high-quality generation without requiring test-time optimization.

Ablation Study

We evaluate the effectiveness of each component in our method through an ablation study on the anomaly detection task, as summarized in Table 4. Starting with (a), simple LoRA fine-tuning produces images that are almost entirely normal, indicating that the model fails to synthesize meaningful anomalies without additional mechanisms. In (b), directly applying mask loss to the attention map leads to unstable and unrealistic results. In (c), incorporating a mask decoder that modulates the mask signal results in more plausible and localized anomalies. In (d), selecting an appropriate attention layer improves anomaly localization by better aligning with the anomaly type.

We further investigate the impact of data pre-processing through two additional ablation studies. In (e), replacing our carefully crafted anomaly-type tokens with a generic placeholder (e.g., "a vfx with sks.") significantly degrades performance, as it removes the semantically meaningful cues embedded in the diffusion prior. In (f), omitting spatial location and detailed descriptors from the prompt (e.g., "A wood has scratch.") hinders the model's ability to localize the intended anomaly regions, resulting in less accurate mask predictions. These findings underscore the importance of incorporating both semantically rich anomaly descriptions and spatially grounded language to fully exploit the text-conditioned prior knowledge of the diffusion transformer.

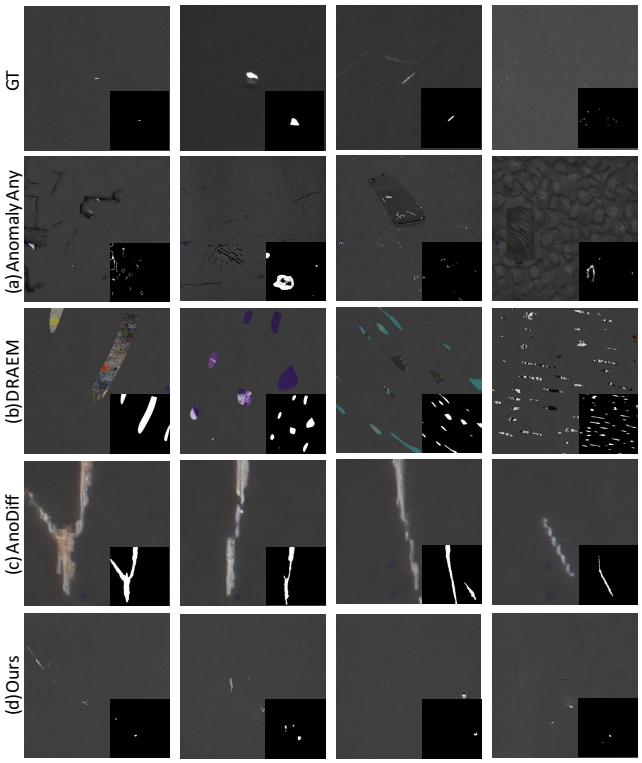


Figure 5: Qualitative comparisons between existing anomaly generation methods on unseen abnormal type. Our method generates fine-grained and realistic anomalies, while previous methods produce unrealistic defects.

Conclusion

We present a unified anomaly generation framework, CHIMERA, which for the first time leverages a multimodal diffusion transformer to synthesize diverse and realistic anomalies across multiple object categories. By leveraging vision-language models for detailed prompt construction and employing a lightweight LoRA-based fine-tuning strategy with auxiliary mask supervision, our method achieves high-quality anomaly synthesis and accurate localization. Unlike existing approaches that often overfit to specific anomaly types, our model is universal on a wide range of objects and anomaly types—allowing it to generalize effectively to unseen scenarios without additional calibration. Experimental results demonstrate our method is on par with other baselines in both image generation quality and downstream anomaly detection tasks.

Acknowledgements

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592, RS-202502214479, RS-2024-00457882, RS-2025-25441838, RS-2025-25441838, RS-2025-02214479, RS-2025-02217259), the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068, RS-2023-00222280, RS-2023-00266509), and National Research Foundation of Korea (RS-2024-00346597).

References

AI, S. 2024. Stable Diffusion 3. *https://stability.ai*. Accessed: 2025-07.

Bergmann, P.; and et al. 2019. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. *CVPR*, 9592–9600.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4183–4192.

Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

Choe, J.; Lee, J.; and Sim, J. 2023. GLASS: Global to Local Attention for Self-Supervised Anomaly Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20502–20511.

Cohen, N.; and Hoshen, Y. 2020. Sub-image anomaly detection with deep pyramid correspondences. In *CVPR*.

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*.

Defard, T.; and et al. 2021. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *ICCV*.

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.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, 226–231.

Fang, Z.; Wang, X.; Li, H.; Liu, J.; Hu, Q.; and Xiao, J. 2023. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17481–17490.

Helbling, A.; Meral, T. H. S.; Hoover, B.; Yanardag, P.; and Chau, D. H. 2025. Conceptattention: Diffusion transformers learn highly interpretable features. *arXiv preprint arXiv:2502.04320*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

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, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

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.

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.; and et al. 2021. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In *CVPR*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Lu, R.; Wu, Y.; Tian, L.; Wang, D.; Chen, B.; Liu, X.; and Hu, R. 2023. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36: 8487–8500.

Malhi, I.; Dutta, P.; Talius, E.; Ma, S.; Driscoll, B.; Holden, K.; Pruthi, G.; and Narayanaswamy, A. 2025. Preserving Product Fidelity in Large Scale Image Recontextualization with Diffusion Models. *arXiv preprint arXiv:2503.08729*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Pang, G.; Shen, C.; Cao, L.; and Hengel, A. V. D. 2021. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2): 1–38.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. 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.; and et al. 2021. PatchCore: Towards Efficient Anomaly Detection and Localization on Image-Level and Pixel-Level. In *NeurIPS*.

Roth, K.; and et al. 2022. Towards Total Recall in Industrial Anomaly Detection. In *CVPR*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2234–2242.

Schlüter, H. M.; Tan, J.; Hou, B.; and Kainz, B. 2022. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, 474–489. Springer.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Stability AI. 2024. Stable Diffusion 3.5. https://github.com/Stability-AI/sd3.5.

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.

Yi, J.; and Wu, J. 2023. Generating Diverse Industrial Defects with Diffusion Models for Few-Shot Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; and Le, X. 2022. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35: 4571–4584.

Yu, Y.; and et al. 2021. FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows. In *arXiv preprint arXiv:2111.07677*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 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, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595.

Zhang, X.; Xu, M.; and Zhou, X. 2024. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16699–16708.

Zou, X.; Liu, Z.; Wang, Y.; Ji, S.; Wu, L.; Wang, F.; and Loy, C. C. 2022. Spot the Difference: Learning Visually Similar Anomalies for Industrial Anomaly Detection. In *European Conference on Computer Vision (ECCV)*, 274–291. Springer.