

Article

Symmetry-Aware Discrepancy Representation and Collaborative Optimization for Multi-Class Defect Image Generation

Beibei Jia ¹, Haijian Shao ^{1,*} , Dengbiao Jiang ¹, Nian Tao ¹ and Guoquan Yao ²

¹ School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China; 232210104101@stu.just.edu.cn (B.J.); jiangdengb@just.edu.cn (D.J.); 231210702111@stu.just.edu.cn (N.T.)

² Key Laboratory of High Performance Ship Technology of Ministry of Education, Wuhan University of Technology, Wuhan 430063, China; yaogq@whut.edu.cn

* Correspondence: jsj_shj@just.edu.cn

Abstract

Industrial defect image generation is an effective way to alleviate data scarcity and class imbalance in visual inspection. In industrial images, defects usually appear as local asymmetric perturbations on globally regular background structures, which makes defect synthesis dependent on both background consistency and local anomaly fidelity. Existing generative methods still face difficulties when only limited anomalous samples are available, especially in representing fine-grained discrepancies among defect categories, coordinating global and local branches across diffusion stages, and constraining small defect regions and their boundary transitions. To address these issues, this paper develops a symmetry-aware multi-constraint diffusion framework based on the dual-branch architecture of DualAnoDiff. The framework treats multi-class industrial defect generation as a joint optimization problem involving class-conditioned discrepancy representation, diffusion-stage-aware branch coordination, and saliency-guided regional supervision. First, Class-Conditioned Shared-Basis LoRA (CSB-LoRA) models category-specific defect characteristics by combining cross-class shared low-rank bases with class-dependent coefficients, allowing common structural priors and class-specific asymmetric patterns to be represented simultaneously. Second, Temporal Dual-branch Attention Modulation (TDAM) adjusts branch interaction, background information injection, and residual feature fusion according to the denoising stage, so that the generation process can gradually shift from global structure restoration to local defect refinement. Third, Saliency-Guided Reconstruction Loss (SGRL) applies stronger spatial constraints to defect regions and boundary neighborhoods, improving local detail preservation and defect-background continuity. Experiments on the MVTec AD dataset show that the proposed method improves both generation quality and perceptual diversity compared with DualAnoDiff. The average IS increases from 1.93 to 2.07, and IC-LPIPS increases from 0.38 to 0.41. When the generated samples are used for downstream defect segmentation, AP-P improves from 84.5% to 85.7%, and F1-P improves from 78.8% to 79.3%. These results indicate that the generated samples can serve as useful synthetic training data for few-shot and class-imbalanced industrial inspection.

Keywords: industrial defect image generation; multi-class discrepancy modeling; diffusion models; temporal dual-branch attention modulation; saliency-guided reconstruction loss



Academic Editor: Shuang Xu

Received: 9 May 2026

Revised: 1 June 2026

Accepted: 23 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Industrial surface defect detection is a critical component of intelligent manufacturing and quality control, and its performance is directly related to product reliability and

production process stability. In application scenarios such as electronic manufacturing, precision machining, and automated production, detection systems are required not only to accurately identify anomalous regions but also to precisely characterize the location, morphology, and local details of defects, so as to support subsequent tasks such as defect localization, quality assessment, and process optimization. In recent years, the development of deep learning has significantly promoted the advancement of industrial vision technologies, and deep learning-based methods have achieved promising performance in tasks such as defect detection, anomaly localization, and image segmentation [1–3]. However, such methods usually rely on large-scale high-quality annotated data, while in real industrial scenarios, defect samples are difficult to acquire, annotation costs are high, and class distributions are often imbalanced [4], which to some extent restricts the further improvement of model performance.

In practical industrial production processes, defect samples are typically characterized by obvious scarcity. On the one hand, industrial systems generally pursue stable operation and high yield, resulting in a low frequency of anomalous samples and consequently a limited amount of defect data available for training; on the other hand, different categories of defects exhibit significant differences in morphology, scale, and texture patterns, while the fine-grained annotation process is complex and costly, further aggravating the problems of insufficient data and class imbalance. Under such conditions, the feature information of minority defect classes is difficult to be fully learned, the representation of inter-class discrepancies is constrained, and the training process becomes more susceptible to overfitting. Therefore, improving the representation capability of multi-class defect discrepancy features under limited sample conditions has become a key problem that urgently needs to be addressed in the field of industrial defect detection.

To address the issue of insufficient training data, existing studies usually adopt data augmentation or defect image generation strategies to expand the sample scale. Traditional data augmentation methods mainly rely on geometric transformations such as rotation, scaling, and cropping, which can alleviate the problem of insufficient samples to a certain extent but are unable to generate new samples with realistic defect structures. Meanwhile, some methods simulate defects by editing local regions in normal images, such as CutPaste [5]. Although such methods can improve sample diversity, they still suffer from limitations in boundary naturalness, texture consistency, and local realism. With the development of generative models, deep learning-based defect image generation methods have gradually become a research hotspot. Generative Adversarial Networks (GANs) [6] are capable of generating images with high visual quality and have been widely applied in defect synthesis tasks; however, under few-shot conditions, they are prone to problems such as training instability, mode collapse, and insufficient sample diversity. In contrast, diffusion models represented by DDPM [7] and Latent Diffusion Model [8] demonstrate stronger advantages in terms of training stability and image quality. In recent years, they have been gradually introduced into defect image generation tasks and have shown promising potential in improving the realism and diversity of generated results, such as AnomalyDiffusion [9] and DualAnoDiff [10]. From the perspective of structural representation, industrial defects can be regarded as localized asymmetric disturbances embedded in globally consistent background structures. Therefore, industrial defect generation requires not only preserving the structural consistency of background regions but also accurately modeling the asymmetric characteristics of local defect regions. Nevertheless, existing methods still exhibit several limitations in complex industrial scenarios. First, fine-grained discrepancies among different defect categories have not been sufficiently represented, which weakens the modeling of class-specific asymmetric defect patterns. Second, the coordination between global background structures and local defect regions remains insufficient during

the generation process, especially across different diffusion stages. Third, in the absence of explicit region-aware constraints, defect regions and their boundaries may introduce background disturbances or boundary discontinuities, thereby affecting overall structural consistency and local detail quality.

Based on the above analysis, this paper develops a symmetry-aware multi-constraint diffusion framework on the dual-branch architecture of DualAnoDiff. The framework focuses on three practical difficulties in multi-class defect generation: preserving category-specific defect characteristics under shared adaptation, coordinating the global and local branches at different denoising stages, and giving sufficient supervision to small defect regions and their boundary neighborhoods. To handle these difficulties, three modules are integrated into the diffusion process. Class-Conditioned Shared-Basis LoRA (CSB-LoRA) represents different defect categories through class-dependent combinations of shared low-rank bases, so that common structural priors can be preserved while category-specific asymmetric patterns are still distinguished. Temporal Dual-branch Attention Modulation (TDAM) introduces time-dependent regulation into branch interaction, background information injection, and residual feature fusion, allowing the model to place different emphasis on global structure restoration and local anomaly refinement during denoising. Saliency-Guided Reconstruction Loss (SGRL) further assigns stronger reconstruction constraints to defect regions and boundary neighborhoods, which helps preserve local details and maintain smoother defect-background transitions. With these designs, the framework improves class-aware defect representation, local anomaly fidelity, and regional structural consistency under few-shot and class-imbalanced settings.

To validate the effectiveness of the proposed method, experiments are conducted on the MVTec AD dataset with comparisons against multiple representative defect generation methods. The results demonstrate that the proposed framework achieves overall improvements in both generation quality and perceptual diversity, while maintaining competitive performance across structurally complex categories. Compared with DualAnoDiff, the average Inception Score (IS) increases from 1.93 to 2.07, and the average IC-LPIPS [11] increases from 0.38 to 0.41. When the generated samples are further used for downstream defect segmentation, AP-P and F1-P are improved from 84.5% to 85.7% and from 78.8% to 79.3%, respectively, indicating that the generated samples provide effective synthetic training data for industrial defect detection and segmentation. The main contributions of this paper are summarized as follows:

- (1) This work studies multi-class industrial defect image generation from a symmetry-aware collaborative optimization perspective. Instead of treating CSB-LoRA, TDAM, and SGRL as independent modifications, the framework connects class-specific discrepancy representation, diffusion-stage-aware branch coordination, and saliency-guided regional reconstruction within one diffusion-based generation process.
- (2) Based on the dual-branch structure of DualAnoDiff, the method introduces class-conditioned and time-adaptive modeling into defect synthesis. CSB-LoRA uses class-dependent combinations of shared low-rank bases to preserve common structural priors while distinguishing category-specific asymmetric defect patterns. TDAM further uses diffusion time-step information to adjust branch interaction, background information injection, and residual feature fusion across denoising stages.
- (3) SGRL provides region-aware supervision for small defect regions and their boundary neighborhoods. By replacing uniform reconstruction weighting with saliency-guided spatial weighting, it increases the contribution of defect-related regions during training and improves local anomaly fidelity, boundary continuity, and regional structural consistency under few-shot and class-imbalanced conditions.

- (4) Experiments on MVTec AD, supplementary results on VisA, and extended ablation studies show that the framework improves generation quality, perceptual diversity, and downstream defect segmentation performance over DualAnoDiff and other representative methods.

2. Related Work

2.1. Image Generation

Image generation technology aims to learn the distribution characteristics of data and achieve the mapping from latent space to image space, and has become an important research direction in the field of computer vision. Early methods mainly relied on rule-based data augmentation and image editing techniques, which expanded sample distributions through geometric transformations or local region recombination. However, such methods were essentially limited to the reorganization of existing data and were unable to represent complex data variation patterns. With the development of deep learning, data-driven generation methods have gradually become the mainstream. Among them, Generative Adversarial Networks (GANs) [6] progressively approximate the real data distribution through adversarial learning between the generator and discriminator, thereby improving the visual quality of generated images. Based on this framework, related studies have continuously improved model structures and training strategies. For example, DCGAN [12] enhances training stability by introducing convolutional architectures and normalization strategies; Pix2Pix [13] introduces conditional constraints during the generation process to achieve mapping from the input domain to the target domain; CycleGAN [14] achieves cross-domain translation under unpaired data conditions through cycle-consistency constraints; and StyleGAN [15] improves the controllability of generated results in terms of both structure and details through style mapping and layer-wise modulation.

On this basis, recent studies have further developed generation methods based on progressive denoising processes. By constructing a gradual restoration process from noise to images, these methods improve the stability and diversity of generated results. DDPM [7] transforms the complex generation process into a series of continuous steps through progressive noise addition and reverse denoising, thereby enabling high-quality image generation. DDIM [16] further introduces deterministic sampling to improve sampling efficiency while maintaining generation quality. Score-based methods [17] provide a unified perspective for the generation process by learning gradient information of the data distribution and performing sampling with stochastic differential equations. Latent Diffusion Model performs diffusion operations in latent space, effectively reducing computational overhead and enabling high-resolution image generation. Stable Diffusion further combines large-scale data within this framework to achieve high-quality image generation with a certain degree of semantic controllability. Overall, current image generation methods continue to improve in terms of generation quality, structural preservation, and detail representation, while gradually evolving toward higher resolution and higher efficiency.

2.2. Few-Shot Image Generation

In practical applications, due to the high cost of data acquisition and the complexity of the annotation process, few-shot image generation has gradually become a research focus. When training data are extremely limited, models are prone to overfitting, and the generated results often become excessively similar to the training samples, making it difficult to reflect data diversity. Early methods mainly relied on traditional data augmentation strategies, which expanded the sample scale through geometric transformations, color perturbations, or region recombination. However, such methods are still constrained by the original data distribution and are unable to learn new structural characteristics and

complex patterns. To alleviate the above problems, transfer learning mechanisms have gradually been introduced into few-shot image generation tasks. By transferring generative models pretrained on large-scale datasets to the target domain, these methods improve the stability and generalization capability of the models. On this basis, FreezeD [18] effectively alleviates overfitting during the fine-tuning stage by freezing part of the discriminator parameters and incorporating regularization strategies; Defect Transfer GAN [19] exploits transferable defect representations across different product backgrounds, enabling diverse defect synthesis under data-hungry and class-imbalanced industrial inspection scenarios; Few-shot GAN improves generation stability and diversity under extremely limited sample conditions by introducing feature constraints and discriminator regulation; and StyleGAN improves the controllability of generated results in terms of both structure and details through style-space modulation. In terms of further optimization, WeditGAN [20] adjusts the target distribution by learning offset vectors in the latent space, thereby reducing the dependence on large-scale parameter updates; AdAM [21] enhances the adaptability of the model to target data while preserving existing knowledge through low-rank modulation of key convolution kernels.

In recent years, with the development of pretrained generative models, few-shot image generation has further evolved toward parameter-efficient adaptation and semantic controllability. Textual Inversion [22] enables new concepts to be incorporated into the semantic space of pretrained models by optimizing a small number of embedding vectors; DreamBooth [23] maintains semantic consistency across different scenarios through targeted fine-tuning combined with prior constraints; and LoRA [24] performs lightweight modulation of model parameters through low-rank decomposition, reducing training cost while ensuring generation quality and stability. Overall, although existing methods can improve the realism and diversity of generated results under limited data conditions, they still suffer from limitations in complex structure representation and fine-grained detail control. In particular, it remains difficult to simultaneously balance structural integrity and local detail representation, which easily leads to inconsistencies and degraded detail quality in the generated results.

2.3. Defect Image Generation

In industrial defect detection tasks, the scarcity and imbalance of defect samples have a significant impact on model performance, and therefore, image generation-based data augmentation methods have gradually attracted increasing attention. Early methods mainly relied on rule-based data augmentation and region recombination strategies. For example, DeVries et al. [25] simulated defects by introducing random regions into normal images; CutPaste [5] constructed pseudo-defects by copying local regions and pasting them into other locations; and Crop&Paste directly cropped real defects and transferred them onto different backgrounds. Although these methods are simple to implement, they are still limited to the recombination of existing data and are unable to generate novel defect patterns. In addition, they are prone to problems such as boundary discontinuity and texture inconsistency, thereby limiting the realism and diversity of generated results.

With the development of deep generative models, data-driven defect generation methods have gradually been applied. Among them, methods based on Generative Adversarial Networks (GANs) have been widely adopted. GANs progressively approximate real data distributions through adversarial training between the generator and discriminator. Based on this framework, SDGAN [26] separately processes background and defect information through a dual-generator architecture to achieve feature fusion between the two types of information; Defect-GAN [27] maintains consistency of generated results under different background conditions by jointly modeling defect generation and restoration processes;

Anomaly-GAN [28] incorporates mask information and anomaly-aware loss to guide the synthesis of detailed anomaly regions, thereby improving the realism and diversity of generated surface anomaly images; and DFMGAN [29] first learns the characteristics of normal samples and then transfers the learned knowledge to defect data, thereby enabling a certain degree of controllable generation. However, under few-shot conditions, these methods still suffer from problems such as training instability, limited diversity of generated results, and unnatural integration between defects and backgrounds.

In recent years, diffusion models have gradually attracted attention in defect image generation tasks due to their stable training process and strong generation capability. Compared with traditional generative methods, diffusion models transform the image generation process into a continuous restoration process through progressive noise addition and reverse denoising, enabling the model to better preserve overall structures and improve detail representation during generation. In terms of specific methods, DDPM achieves high-quality image generation through forward noise addition and reverse denoising processes; DDIM further introduces a deterministic sampling mechanism to improve generation efficiency; and score-based methods provide a unified description of the generation process from the perspective of probability gradients. In studies on industrial defect image generation, to address the problems of training instability and insufficient anomaly region representation under few-shot conditions, related methods have further improved diffusion frameworks. Among them, PTDM [30] progressively optimizes different noise stages through a stage-wise training strategy to improve the stability of the generation process; AnoDDPM [31] introduces denoising diffusion probabilistic models with simplex noise for anomaly detection, using reconstruction discrepancies to support anomaly localization; and different from the above methods, DualAnoDiff constructs a dual-branch structure consisting of global and local branches, enabling the model to simultaneously consider overall structures and local anomalous regions during the generation process, thereby improving the realism and consistency of defect images. Overall, diffusion models have demonstrated promising application potential in defect image generation tasks. However, they still exhibit limitations in multi-class defect discrepancy representation, coordination between global structures and local anomalies, and fine-grained control of key regions. In particular, it remains difficult to achieve a stable balance between structural integrity and local detail representation, which to some extent affects the consistency and detail quality of generated results.

Although current defect image generation methods have improved visual quality and controllability, their performance remains limited in few-shot multi-class scenarios. Differences among defect categories are often not sufficiently separated, which may blur category-specific structures such as cracks, contaminations, scratches, and texture defects. The interaction between global background preservation and local anomaly synthesis is also not fully adaptive during generation, making the generated images vulnerable to structural inconsistency. For small or weak defects, boundary transitions and local textures are still difficult to preserve when the reconstruction objective treats all regions with similar importance. These issues indicate that multi-class discrepancy representation, global–local collaboration, and defect-region supervision should be considered together rather than optimized separately.

Recent studies provide several useful directions for addressing these problems. Parameter-efficient adaptation methods and LoRA-based conditioning variants [24,32] reduce the cost of adapting pretrained generative models, but they mainly target general-domain adaptation or concept personalization, with less emphasis on fine-grained discrepancy modeling among industrial defect categories. Time-step-adaptive diffusion strategies [33] show that different denoising stages contribute differently to structural restoration

and detail refinement. However, most temporal modulation designs are built for a single generation pathway, while industrial defect synthesis needs coordinated modeling of globally regular backgrounds and locally asymmetric anomalies. Region-weighted or mask-guided objectives [28] can strengthen foreground or anomaly-region learning, but hard binary-mask weighting is less suitable for small defects with weak boundaries and subtle texture transitions. Based on these considerations, the proposed framework combines class-conditioned adaptation, diffusion-stage-aware dual-branch modulation, and saliency-guided regional reconstruction for multi-class defect generation. Compared with DualAnoDiff [10], this design further emphasizes class discrepancy modeling, adaptive global–local collaboration, and local defect-region supervision within a unified diffusion framework.

3. Method

As a dual-branch diffusion framework for industrial defect image generation, DualAnoDiff uses a global diffusion branch and a local anomaly diffusion branch to model background structures and defect-specific regions. Its nested prompting and background compensation strategies provide a useful basis for balancing local anomaly synthesis with background preservation under few-shot conditions. When the task is extended to multi-class defect generation, the modeling requirements become more complex. Different defect categories may vary in morphology, texture, scale, and boundary appearance, whereas their normal backgrounds often contain similar regular structures. The model therefore needs to preserve global structural regularity while distinguishing class-specific asymmetric defect patterns and their transitions to surrounding background regions.

Based on the dual-branch architecture of DualAnoDiff, this paper develops a symmetry-aware multi-constraint diffusion framework for multi-class defect image generation. The framework focuses on three aspects that directly affect generation quality in this setting: category-sensitive adaptation for fine-grained defect discrepancies, stage-dependent coordination between global and local branches, and region-aware supervision for small defect areas and boundary neighborhoods. To this end, Class-Conditioned Shared-Basis LoRA (CSB-LoRA), Temporal Dual-branch Attention Modulation (TDAM), and Saliency-Guided Reconstruction Loss (SGRL) are incorporated into the diffusion process. CSB-LoRA provides category-aware discrepancy modeling, TDAM regulates global–local collaboration across denoising stages, and SGRL strengthens reconstruction supervision in defect-related regions. Together, these components form a collaborative optimization path that links category-level representation, diffusion-stage-aware branch modulation, and saliency-guided regional reconstruction.

As shown in Figure 1, CSB-LoRA adopts a class-conditioned parameter-efficient adaptation strategy for multi-class discrepancy modeling. Rather than assigning independent adaptation parameters to each defect category or using one class-agnostic update for all categories, it combines cross-class shared low-rank bases with class-dependent coefficients. This allows the model to retain common structural priors while modulating category-specific asymmetric defect patterns. TDAM adds diffusion time-step information to the dual-branch generation process and adjusts branch interaction, background information injection, and residual feature fusion along the denoising trajectory. In this way, the model can shift its emphasis from global structure restoration to local anomaly refinement as denoising proceeds. SGRL further applies saliency-guided spatial weighting to the reconstruction objective, assigning stronger constraints to defect regions and boundary neighborhoods while maintaining defect–background continuity. These designs improve class discriminability, local anomaly fidelity, and regional structural consistency in few-shot and class-imbalanced industrial defect generation.

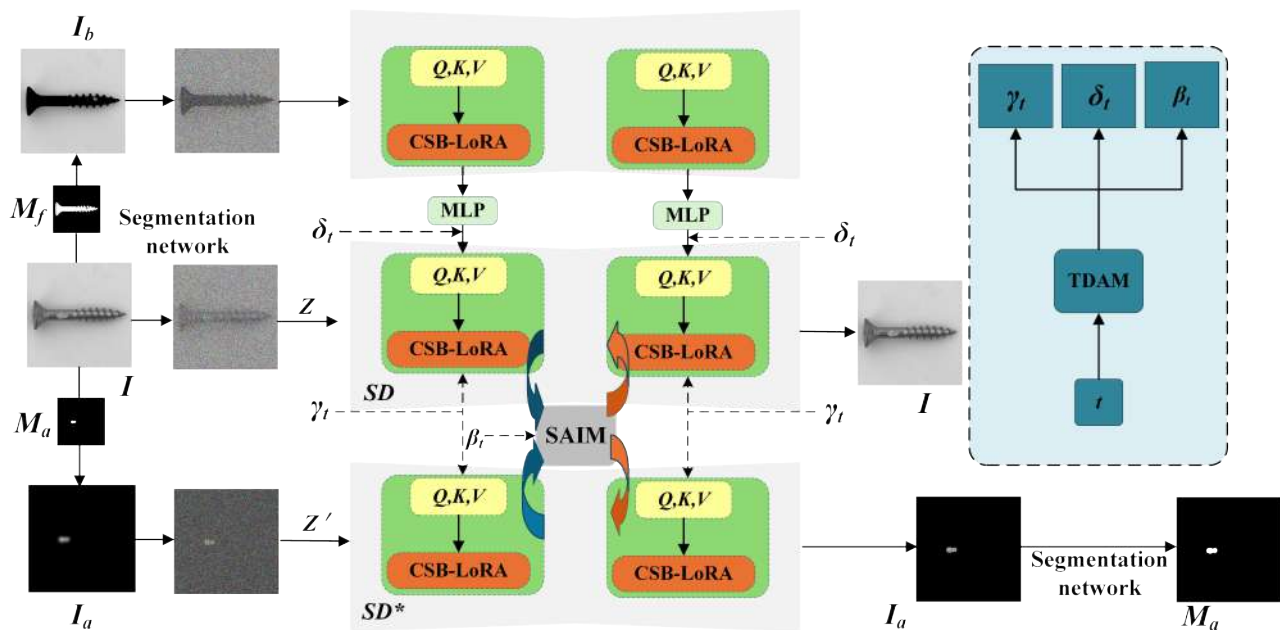


Figure 1. Architecture of the proposed symmetry-aware multi-constraint diffusion framework based on the dual-branch architecture of DualAnoDiff. Here, SD denotes Stable Diffusion, and SD* denotes the adapted Stable Diffusion module in the proposed framework.

3.1. Class-Conditioned Shared-Basis LoRA

DualAnoDiff provides a useful dual-branch basis for industrial defect image generation, but its class-agnostic adaptation is less suited to multi-class defect synthesis. Industrial defects differ markedly in texture, morphology, spatial distribution, and boundary appearance. Crack-type defects usually involve elongated structures and edge continuity, while contamination- or wear-related defects are more closely related to local texture changes and regional shape variations. If all defect categories share the same low-rank update, the model can capture general anomaly patterns, but category-specific cues may be weakened. This may reduce the realism and diversity of generated samples, especially when several defect types are trained within one framework.

LoRA offers an efficient way to adapt pretrained models with a small number of additional parameters, which is useful for few-shot generation. In its conventional form, however, the same low-rank update is usually shared across conditions, making it difficult to represent category-dependent asymmetric defect patterns in multi-class industrial images. To address this problem, we introduce Class-Conditioned Shared-Basis LoRA (CSB-LoRA), as shown in Figure 2. CSB-LoRA learns a set of low-rank bases shared across defect categories and combines them with class-dependent coefficients. This design allows common structural priors to be reused while giving each defect category its own adaptation direction, thereby improving the separation of multi-class defect representations.

In conventional LoRA, a low-rank update matrix ΔW is introduced into a linear layer $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, which can be formulated as:

$$W' = W + \Delta W, \quad \Delta W = BA, \quad B \in \mathbb{R}^{d_{\text{out}} \times r}, \quad A \in \mathbb{R}^{r \times d_{\text{in}}} \quad (1)$$

where $r \ll \min(d_{\text{out}}, d_{\text{in}})$ denotes the rank of LoRA. In DualAnoDiff, the same set of parameters A and B is shared across all defect categories, meaning that different defect types are modulated along a unified direction in the feature space. Although this class-agnostic shared modulation simplifies the model structure, it weakens the ability to capture class-specific discrepancies in the latent space. As a result, it becomes difficult to fully

represent defect characteristics such as texture patterns, spatial distributions, morphological variations, and severity differences across categories.

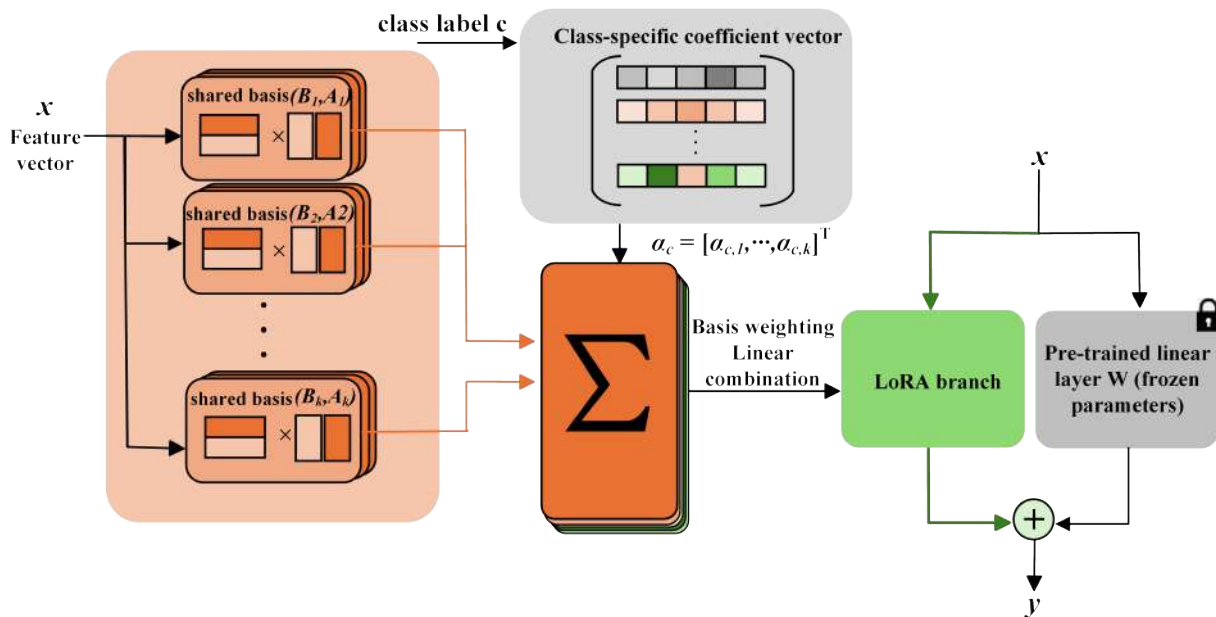


Figure 2. Structure of the proposed CSB-LoRA module.

For a given attention linear transformation, such as one of W_q , W_k , W_v , and W_o , CSB-LoRA does not assign an independent set of LoRA parameters to each defect category. Instead, it first learns K groups of shared bases $(A_k, B_k)_{k=1}^K$, where

$$A_k \in \mathbb{R}^{r \times d_{\text{in}}}, \quad B_k \in \mathbb{R}^{d_{\text{out}} \times r}, \quad k = 1, 2, \dots, K \quad (2)$$

For the c -th defect category ($c = 1, 2, \dots, C$), only a class-dependent coefficient vector $\alpha_c = [\alpha_{c,1}, \dots, \alpha_{c,K}]^T \in \mathbb{R}^K$ is learned. This vector performs a weighted linear combination of the shared bases, generating a class-specific LoRA update for the c -th defect category and enabling dedicated feature modulation under the corresponding class condition. In implementation, the class-specific coefficient vector α_c is initialized with small values to avoid assigning excessive preference to any single shared basis at the beginning of training. This initialization provides a stable starting point for shared-basis combination and allows each defect category to gradually learn its own basis weighting during optimization.

$$\Delta W^{(c)} = \sum_{i=1}^K \alpha_{c,i} B_i A_i \quad (3)$$

Finally, the transformation of the linear layer under the class condition c is given by:

$$W^{(c)} = W + \Delta W^{(c)} = W + \sum_{i=1}^K \alpha_{c,i} B_i A_i \quad (4)$$

During forward propagation, given an input feature vector $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$ and the class label c , the output is given by:

$$\mathbf{y}^{(c)} = W^{(c)} \mathbf{x} = \left(W + \sum_{i=1}^K \alpha_{c,i} B_i A_i \right) \mathbf{x} \quad (5)$$

CSB-LoRA offers a middle ground between assigning a separate LoRA module to each category and using a single shared update for all categories. The shared low-rank bases capture structural knowledge that is common across defect categories, so the parameter size does not grow sharply with the number of classes. The class-dependent coefficients then select and combine these bases according to the target defect type, allowing the model to represent category-specific textures, morphologies, and structural variations. This design is well suited to few-shot and class-imbalanced defect generation, where the model needs to reuse shared knowledge while still separating fine-grained category differences.

Within the dual-branch diffusion network, CSB-LoRA is inserted into the key attention modules of both the global diffusion branch and the local anomaly diffusion branch. The resulting class-conditioned modulation allows the two branches to use shared pretrained knowledge while adapting their feature responses to different defect categories. This gives the following TDAM and SGRL modules a more discriminative category-aware representation to operate on during stage-adaptive branch coordination and region-aware reconstruction.

3.2. Temporal Dual-Branch Attention Modulation

CSB-LoRA improves category-aware adaptation, but the dual-branch generation process still needs a mechanism to coordinate global structure modeling and local defect refinement along the denoising trajectory. DualAnoDiff separates the generation process into a global branch and a local anomaly branch, which helps the model attend to background structures and defect regions at the same time. However, the interaction between the two branches and the injection of background information are largely fixed. Such a fixed strategy is not always suitable for diffusion generation, because the role of each branch changes as denoising proceeds. At high-noise stages, the model relies more on global layout and background texture to keep the generated image structurally stable. At later stages, when the noise level is lower, local defect morphology, texture, and boundary details become more important. If the same branch interaction and background injection strength are used throughout the entire denoising process, the model may either introduce unstable local details too early or over-constrain defect regions at later stages. This can affect both the clarity of defect details and the consistency between defects and surrounding backgrounds.

To address this issue, we design Temporal Dual-branch Attention Modulation (TDAM), as shown in Figure 3. TDAM introduces diffusion time-step information into branch interaction, background information injection, and residual feature fusion. Through time-dependent gates, it adjusts the relative contribution of the global branch and the local anomaly branch at different denoising stages. In this way, the model can rely more on global structural cues in early stages and gradually emphasize local anomaly refinement as generation progresses. The implementation details are described below.

(1) Time-Step Embedding and Feature Mapping. Given a diffusion time step t , a sinusoidal positional encoding function is first employed to map it into a high-dimensional time-step embedding vector $E_t \in \mathbb{R}^{d_{\text{emb}}}$. The sinusoidal encoding represents the time index using sine and cosine functions at different frequencies, thereby enhancing the model's ability to capture diffusion stage information.

Based on this, TDAM employs three independent lightweight multilayer perceptrons (MLPs), taking E_t as input, to predict three groups of time-dependent dynamic gating parameters, denoted as γ_t , δ_t , and β_t , respectively. In implementation, the three gating MLPs adopt the same lightweight two-layer structure, which maps the time-step embedding E_t to the corresponding gating parameters through a hidden projection and a nonlinear activation. The time-step embedding dimension is set to $d_{\text{emb}} = 256$. The branch weight-

ing vector γ_t is normalized by softmax so that the global and anomaly branch weights satisfy $\gamma_g + \gamma_a = 1$, with its initial value close to $[0.5, 0.5]$. The residual fusion gate β_t is constrained to $(0, 1)$ by sigmoid activation and optimized jointly with the network. These lightweight designs allow TDAM to introduce stage-adaptive modulation without substantially increasing the parameter scale.

These gating parameters are further used to modulate the response strength of the dual-branch LoRA module, the intensity of reference condition injection, and the fusion magnitude of cross-branch self-attention residuals, respectively. As a result, the key computational pathways within the attention module can be adaptively adjusted according to different diffusion stages.

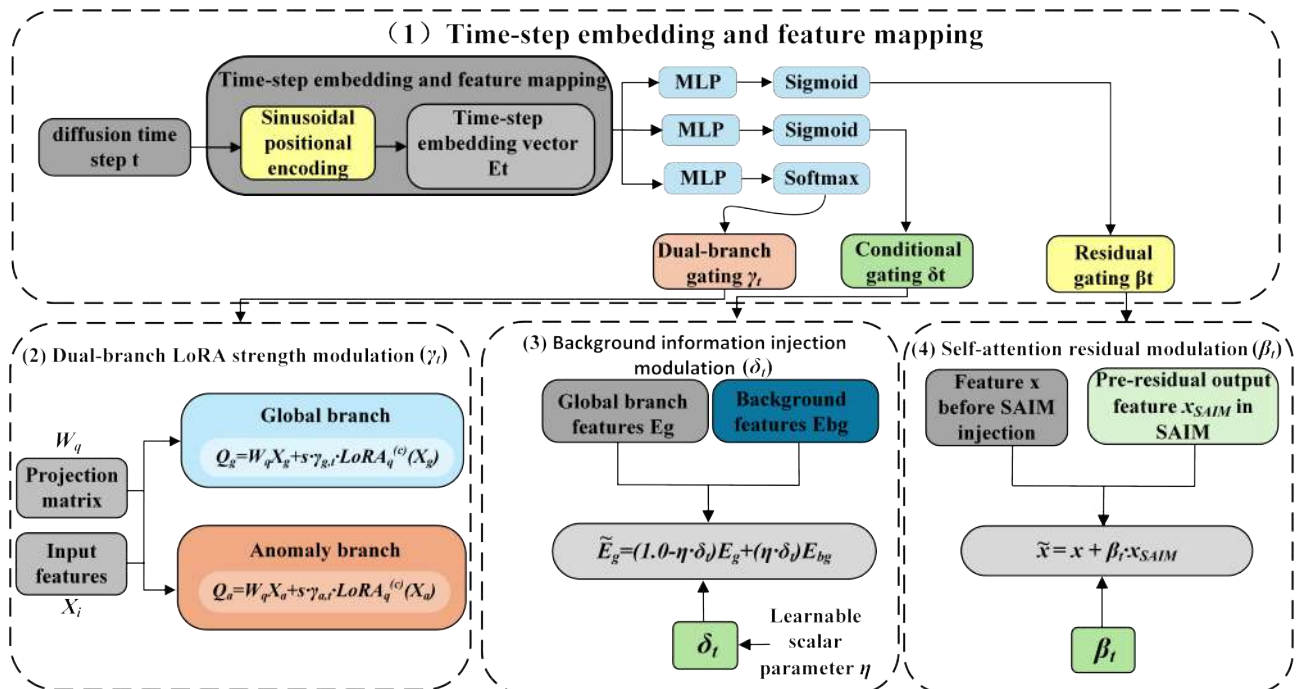


Figure 3. Structure of the proposed TDAM module.

(2) Dual-Branch LoRA Strength Modulation. To dynamically balance the feature extraction capabilities of the global branch and the local anomaly branch across different diffusion stages, TDAM predicts a two-dimensional gating vector $\gamma_t \in \mathbb{R}^2$. After softmax normalization, this vector is decomposed into the global branch gate γ_g and the local anomaly branch gate γ_a , satisfying $\gamma_g + \gamma_a = 1$. This constraint enables a relative allocation of LoRA modulation strength between the two branches.

During the attention projection in each branch, the original Query computation is reformulated as:

$$Q_i = W_q X_i + s \cdot \gamma_{i,t} \cdot \text{LoRA}_q^{(c)}(X_i), \quad i \in \{g, a\} \quad (6)$$

where X_i denotes the input feature of the i -th branch ($i \in \{g, a\}$), W_q is the pretrained Query projection matrix, s is a global scaling factor, and $\text{LoRA}_q^{(c)}(\cdot)$ represents the class-conditioned low-rank feature extracted by CSB-LoRA.

(3) Background Information Injection Modulation. In the background compensation module, the injection of background information plays a crucial role in suppressing artifacts—such as geometric distortion and background contamination—that frequently arise under few-shot conditions. This mechanism is also essential for maintaining structural integrity and background texture consistency. However, if background features are

injected with a fixed intensity throughout the diffusion process, the model fails to adapt to stage-specific generation requirements.

At early diffusion stages, stronger background guidance helps stabilize the global layout and background distribution, thereby mitigating the risk of structural distortion under high-noise levels. In contrast, at later denoising stages, excessive background injection may overly constrain the generation of local defect regions, potentially hindering fine-grained defect modeling. To address this issue, we introduce a time-dependent background gating parameter $\delta_t \in (0, 1)$ to adaptively control the strength of background information injection.

Specifically, the background feature E_{bg} is incorporated into the global branch feature to enhance the modeling of background structure and texture distribution. Before cross-attention computation, the global branch feature E_g is updated as:

$$\tilde{E}_g = (1 - \eta \cdot \delta_t)E_g + \eta \cdot \delta_t E_{bg} \quad (7)$$

where η is a learnable scalar parameter that controls the overall magnitude of background injection. In implementation, this background injection term is initialized close to zero to avoid excessive background interference at the early training stage and is then adaptively updated during training.

(4) Self-Attention Residual Modulation. After independent feature extraction in the dual branches, the model employs a Self-Attention Interaction Module (SAIM) to enable cross-branch information exchange, facilitating the alignment and fusion between global structural information and local anomaly features. However, at high-noise stages, the correspondences established through cross-branch interactions are often unstable. Introducing aggressive cross-branch fusion at this stage may introduce additional artifacts into the denoising process.

To address this issue, we introduce a time-dependent gating parameter $\beta_t \in (0, 1)$, predicted by TDAM, which is applied to the residual connection of the SAIM output to dynamically regulate the strength of cross-branch self-attention fusion. The computation is formulated as:

$$\tilde{\mathbf{x}} = \mathbf{x} + \beta_t \cdot \mathbf{x}_{\text{SAIM}} \quad (8)$$

where $\mathbf{x}$ denotes the feature before SAIM injection, and $\mathbf{x}_{\text{SAIM}}$ represents the output feature of SAIM before the residual connection is applied. Through the residual modulation mechanism, the model adaptively controls the strength of cross-branch information injection across different diffusion stages. Specifically, it suppresses unstable interaction disturbances during high-noise stages, while progressively enhancing feature alignment and information fusion between branches in later denoising stages.

TDAM applies time-dependent modulation to the dual-branch attention pathway, so the interaction between the global branch and the local anomaly branch changes with the denoising stage. As defined in Equations (6)–(8), the three gates γ_t , δ_t , and β_t regulate different parts of the computation. The branch gate γ_t adjusts the LoRA response in each branch, allowing the model to assign different adaptation strengths to global structure restoration and local anomaly refinement. The background injection gate δ_t controls how much background information is introduced into the global branch. This helps stabilize structural regularity at high-noise stages and reduces the risk of excessive background dominance during later denoising. The residual fusion gate β_t controls the strength of cross-branch self-attention residuals, limiting unstable interactions in early stages and increasing feature alignment when local anomaly details become more reliable. With these gates, the dual-branch interaction is no longer fixed throughout the denoising process. The feature evolution follows a coarse-to-fine pattern: global structural cues play a stronger role in

stabilizing the overall layout at early stages, while local anomaly features and cross-branch interactions become more important for defect morphology and boundary refinement at later stages. TDAM therefore provides a stage-adaptive generation process for multi-class industrial defect images.

TDAM also works with CSB-LoRA by linking time-dependent modulation with class-conditioned adaptation. CSB-LoRA determines the category-specific LoRA update through shared low-rank bases and class-dependent coefficients, while TDAM controls how strongly these updates are activated at different denoising stages. When global structure stabilization is dominant, category-related adaptation is kept relatively weak. As denoising proceeds and local anomaly refinement becomes more reliable, the class-specific responses are gradually strengthened. In this process, CSB-LoRA provides category-aware feature responses, and TDAM modulates these responses according to the diffusion stage. This cooperation helps the model combine class-specific discrepancy representation with stage-adaptive global-local collaboration under few-shot conditions.

3.3. Saliency-Guided Reconstruction Loss

CSB-LoRA and TDAM improve category-aware adaptation and stage-dependent branch coordination, but they do not directly solve the imbalance between small defect regions and large background areas in the reconstruction objective. In industrial images, anomalous regions often occupy only a small portion of the image, while most pixels belong to regular background structures. If the diffusion reconstruction loss is applied uniformly over the whole image, the errors from small defects can be overwhelmed by background-dominated regions. As a result, the model may preserve the global structure but still miss fine local textures, edge contours, and boundary transitions that are critical for realistic defect synthesis.

In the dual-branch diffusion framework, the global branch mainly maintains background regularity, whereas the local anomaly branch is expected to model defect-related details. Without explicit regional guidance, the two branches may not focus on the same anomalous areas during training, which weakens their cooperation in local defect synthesis. For this reason, we introduce Saliency-Guided Reconstruction Loss (SGRL) by adding saliency-based spatial weights to the dual-branch noise prediction objective. SGRL gives larger weights to defect regions and their boundary neighborhoods, encouraging the model to learn defect textures and transition areas more accurately. Since the same saliency distribution is used for both branches, the optimization of global and local features is better aligned around defect-related regions.

Pixel-level defect annotations are used to construct the saliency weights during training. Each binary defect mask is first projected to the latent resolution and then smoothed to obtain a continuous saliency map. Compared with directly using a hard binary mask, the smoothed map provides dense spatial cues for both defect interiors and boundary neighborhoods. In this way, discrete annotations are converted into continuous reconstruction weights, allowing the diffusion model to emphasize local anomaly structures while keeping the generated defect regions consistent with their surrounding backgrounds.

(1) Defect Saliency Weight Construction. Given a binary defect mask $M \in \{0, 1\}^{H \times W}$ in the pixel space, where 1 indicates defect regions and 0 denotes background regions, M is downsampled to the latent resolution (H', W') using nearest-neighbor interpolation, resulting in $M_{\text{lat}} \in \{0, 1\}^{H' \times W'}$. This interpolation strategy preserves the spatial topology of defect regions and avoids introducing additional class-mixing artifacts during the downsampling process.

To alleviate discontinuities caused by sharp boundaries in the hard mask and guide the model to focus on transition regions around defect boundaries, a Gaussian smoothing operator $\mathcal{G}_\sigma(\cdot)$ is applied to M_{lat} , yielding a soft saliency map:

$$S = \mathcal{G}_\sigma(M_{\text{lat}}), \quad S \in [0, 1]^{H' \times W'} \quad (9)$$

where $\mathcal{G}_\sigma(\cdot)$ denotes a Gaussian-like smoothing operator that softens the hard mask into a continuous saliency map. In implementation, $\mathcal{G}_\sigma(\cdot)$ is approximated by local average smoothing with a 5×5 window. This setting provides smooth transitions around defect boundaries while avoiding excessive diffusion of defect weights into large background regions. The resulting saliency map S takes values close to 1 in defect centers, exhibits smooth transitions near boundaries, and approaches 0 in background regions far from defects.

Based on the saliency map, a pixel-wise weighting map $\Omega \in \mathbb{R}^{H' \times W'}$ is constructed for the local anomaly branch:

$$\Omega = \beta + \lambda \cdot S \quad (10)$$

where $\beta > 0$ is a baseline weight that preserves supervision in background regions, and $\lambda > 0$ is a defect enhancement coefficient that increases the relative importance of defect regions and their neighborhoods. In implementation, β and λ are set to 1.0 and 10.0, respectively. The setting $\beta = 1.0$ maintains the basic reconstruction constraint over non-defect regions, while $\lambda = 10.0$ strengthens the supervision of small defect regions and boundary neighborhoods, where reconstruction errors are more critical for fine-grained defect generation. Larger values of S assign higher weights to reconstruction errors at corresponding locations, guiding the model to prioritize fine-grained detail learning in anomalous regions.

(2) Dual-Branch Saliency-Guided Reconstruction Loss. At a given training step, after the forward diffusion process, the latent representations of the global branch and the local anomaly branch are denoted as $\mathbf{x}_t^{\text{blend}}$ and $\mathbf{x}_t^{\text{fg}}$, respectively. The diffusion model predicts the corresponding noise terms $\hat{\epsilon}_p^{\text{blend}}$ and $\hat{\epsilon}_p^{\text{fg}}$, while the ground-truth noise is denoted as $\epsilon_p^{\text{blend}}$ and ϵ_p^{fg} , where p indexes spatial positions in the latent space.

Based on the saliency weights defined above, the conventional uniform reconstruction loss is reformulated into a weighted form for both branches:

$$\mathcal{L}_{fg} = \frac{1}{\sum_p \tilde{\Omega}_p} \sum_p \tilde{\Omega}_p \left\| \hat{\epsilon}_p^{\text{fg}} - \epsilon_p^{\text{fg}} \right\|_2^2 \quad (11)$$

$$\mathcal{L}_{\text{blend}} = \frac{1}{\sum_p \tilde{\Omega}_p} \sum_p \tilde{\Omega}_p \left\| \hat{\epsilon}_p^{\text{blend}} - \epsilon_p^{\text{blend}} \right\|_2^2 \quad (12)$$

where $\tilde{\Omega}_p$ denotes the normalized saliency-based weighting coefficient derived from Ω at spatial position p . The normalization mitigates loss-scale variations caused by differences in defect sizes, thereby improving training stability and enabling more balanced optimization over regions with varying spatial extents.

Finally, the overall dual-branch reconstruction objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{blend}} + \mathcal{L}_{\text{fg}} \quad (13)$$

Compared with a uniform reconstruction objective, SGRL assigns larger weights to defect regions and their boundary neighborhoods through soft spatial weighting. The model is therefore encouraged to learn fine-grained textures, edge contours, and structural variations in anomalous regions, which improves the visual clarity and realism of generated

defects. Using the same saliency map for both the global branch and the local anomaly branch also guides the two branches toward consistent defect-related regions, making their collaboration more stable during training.

SGRL is particularly useful for small or weak defects, where local details can otherwise be dominated by large background areas. By strengthening reconstruction around defect interiors and transition regions, it improves local detail representation, boundary continuity, and regional structural consistency. The generated defect images can also be paired with their corresponding pixel-level masks, making them usable as annotated synthetic samples for defect segmentation, anomaly localization, quality inspection model training, and few-shot data augmentation. In this way, limited real defect samples can be expanded into task-oriented training data for downstream industrial inspection.

4. Experimental Evaluation and Analysis

To comprehensively evaluate the effectiveness of the proposed method, extensive experiments are conducted on the MVTec AD dataset. Quantitative evaluations are performed from multiple perspectives, including generation quality and diversity. Furthermore, comparative experiments and ablation studies are conducted to systematically validate the contribution of each component in the proposed framework.

4.1. Experimental Setup

4.1.1. MVTec AD Dataset

The MVTec AD dataset is one of the most widely used benchmarks for industrial anomaly detection and defect localization. It consists of 10 object categories and 5 texture categories, with up to 8 defect types per category, with pixel-level defect annotations. The dataset exhibits a significant data imbalance: most categories contain approximately 200–400 normal samples, while each defect type typically includes only 10–25 anomalous samples. This inherent scarcity makes it not only suitable for anomaly localization tasks but also a representative benchmark for defect image generation under few-shot and class-imbalanced conditions.

In this paper, the MVTec AD dataset is adopted as the evaluation benchmark. To comprehensively assess the adaptability of the proposed diffusion-based defect image generation method across diverse object materials, texture structures, and defect patterns, experiments are conducted on all 15 categories under a unified data preprocessing and experimental setting. This setup provides a reliable basis for the subsequent evaluation of generation quality and downstream task performance.

4.1.2. Implementation Details

All diffusion-based experiments are implemented using PyTorch version 2.4 with Python version 3.10 and CUDA version 11.8 on a Windows system, with both training and inference conducted on an NVIDIA RTX 4090 GPU. The model adopts Stable Diffusion v1.5 as the pretrained backbone, and input images are resized to 512×512 .

During training, the Adam optimizer is used with a learning rate of 2×10^{-5} . A batch size of 1 is used, and gradient accumulation steps are set to 1. The total number of iterations is fixed at 5000. The final model checkpoint is adopted for downstream evaluation. The text encoder is jointly fine-tuned with LoRA rank set to 32.

No learning-rate decay schedule or warm-up stage was applied, and the learning rate was kept fixed throughout the 5000 iterations. Early stopping was not adopted because constructing an additional validation split would further reduce the limited defective samples available for few-shot generation. Regularization was mainly provided by the parameter-efficient LoRA adaptation strategy and the saliency-guided reconstruction constraint.

4.1.3. Computational Cost and Complexity Analysis

To provide a clearer assessment of the computational overhead of the proposed method, we further compare the proposed framework with DualAnoDiff in terms of total parameters, trainable parameters, GFLOPs per step, peak GPU memory consumption, and inference time per generated image. All measurements are conducted under the same input resolution, batch size, and hardware configuration. GFLOPs/Step denotes the estimated floating-point operations required for one dual-branch U-Net forward pass within a training step under the same input resolution and batch size and is measured using PyTorch profiler. Peak GPU memory is recorded on the same NVIDIA GeForce RTX 4090 GPU, and the inference time is measured for generating one image under the same inference setting.

As shown in Table 1, compared with DualAnoDiff, the proposed method increases the total number of parameters from 1221.438M to 1263.933M and the trainable parameters from 155.203M to 200.057M. The GFLOPs per step increase only slightly from 2380.829 to 2384.780, indicating that the additional computation introduced by the proposed modules remains limited. The peak GPU memory increases from 12.003 GB to 15.440 GB, and the inference time increases from 11.5 s to 15.2 s per generated image.

Table 1. Computational cost and complexity comparison under the same input resolution and hardware settings.

Method	Total Params	Trainable Params	GFLOPs/Step	Peak Mem. (GB)	Infer. Time (s)
DualAnoDiff	1221.438 M	155.203 M	2380.829	12.003	11.5
Ours	1263.933 M	200.057 M	2384.780	15.440	15.2

More importantly, DualAnoDiff is typically trained or adapted in a category-specific manner, whereas the proposed framework is designed for unified multi-class defect generation. Therefore, the additional overhead should not be interpreted only as an increase in the cost of a single-category generation task. Instead, it mainly comes from supporting class-conditioned representation sharing and unified multi-class discrepancy modeling within one framework. In this regard, the proposed method provides a practical trade-off between computational overhead and unified multi-class generation capability for offline defect data augmentation.

4.2. Evaluation Metrics

4.2.1. IS

Inception Score (IS) is a widely adopted evaluation metric in image generation, which measures the quality of generated images in terms of class discriminability and diversity. It leverages a pretrained Inception network to obtain class probability distributions for generated samples. Specifically, high-quality images are expected to produce conditional label distributions with low entropy, indicating confident predictions, while diverse samples should yield a marginal distribution with high entropy, reflecting a broad coverage of categories.

Based on this formulation, the Inception Score is defined as:

$$IS = \exp\left(\mathbb{E}_{x \sim p_g} [D_{KL}(p(y|x) \parallel p(y))]\right) \quad (14)$$

where $p(y|x)$ denotes the conditional class distribution of a generated image x predicted by the Inception network, $p(y)$ represents the marginal class distribution over all generated images, and $D_{KL}(\cdot)$ denotes the Kullback–Leibler divergence. A higher Inception Score indicates that the generated images exhibit both stronger class discriminability and greater diversity.

It should be noted that IS is not an ideal standalone metric for industrial defect image generation, since it relies on an ImageNet-pretrained classifier and industrial defect images have an obvious domain gap from natural image categories. In this paper, IS is retained mainly because DualAnoDiff and related baseline methods adopt this metric, which enables a fair comparison under the same evaluation protocol. Meanwhile, the evaluation in this study does not rely solely on IS. We further combine IC-LPIPS, qualitative visual comparisons, multi-class discriminability analysis, and downstream U-Net segmentation performance to assess the generated samples. The downstream segmentation results further show that the generated data improve AP-P and F1-P, indicating that the synthetic samples are also beneficial for practical defect segmentation tasks.

4.2.2. IC-LPIPS

Intra-class pairwise LPIPS distance (IC-LPIPS) is adopted to measure the perceptual diversity of generated images. It computes the pairwise distance between generated samples in a deep feature space extracted by a pretrained neural network. Unlike traditional pixel-wise metrics, IC-LPIPS captures differences in texture patterns, local structures, and high-level perceptual characteristics, making it more consistent with human visual perception. Therefore, it serves as an effective indicator for evaluating the diversity of generated samples within the same class.

Let two images be denoted as x and x' . Their feature representations extracted at the l -th layer of a deep network are $f_l(x)$ and $f_l(x')$, respectively. Then, the IC-LPIPS can be expressed as:

$$\text{IC-LPIPS}(x, x') = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{f}_l(x)_{h,w} - \hat{f}_l(x')_{h,w} \right) \right\|_2^2 \quad (15)$$

where $\hat{f}_l(\cdot)$ denotes the normalized feature, w_l represents the channel-wise weighting parameter associated with the l -th layer, and H_l and W_l denote the height and width of the feature map at that layer, respectively. A smaller IC-LPIPS value indicates that the two images are more perceptually similar, whereas a larger value implies greater visual differences.

4.3. Quantitative Analysis of Generation Quality and Diversity

In the defect image generation task, the proposed method is compared with five representative approaches. Crop&Paste is a classical data augmentation technique that synthesizes defects by directly copying and pasting regions. Defect-GAN adopts a damage-recovery mechanism with a hierarchical generator to progressively synthesize defect patterns. DFMGAN leverages pretraining and transfer learning to improve generation quality under limited defect samples. AnomalyDiffusion decouples defect appearance and mask distribution, enabling defect synthesis at specified regions to enhance diversity. DualAnoDiff introduces a dual-branch diffusion framework that jointly generates global defect images and local defect components, while incorporating a background compensation mechanism to improve stability and realism under few-shot conditions.

Table 2 presents a quantitative comparison in terms of Inception Score (IS) and IC-LPIPS, demonstrating the overall effectiveness of the proposed method in balancing generation quality and perceptual diversity under few-shot conditions.

From the overall quantitative results, the proposed method achieves the best average performance in terms of both Inception Score (IS) and IC-LPIPS. Compared with the baseline DualAnoDiff, the average IS is improved from 1.93 to 2.07, while the average IC-LPIPS increases from 0.38 to 0.41. These results indicate that the proposed method

achieves a more favorable balance between visual fidelity and perceptual diversity under few-shot defect generation settings.

Table 2. Comparison of IS and IC-LPIPS on the MVTec AD dataset. The upward arrows indicate that higher values are better, and bold entries highlight the average row label and the best-performing average values.

Category	Crop&Paste		Defect-GAN		DFMGAN		AnomalyDiffusion		DualAnoDiff		Ours	
	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑
bottle	1.43	0.04	1.39	0.07	1.62	0.12	1.58	0.19	2.17	0.43	2.05	0.39
cable	1.74	0.25	1.70	0.22	1.96	0.25	2.13	0.41	2.15	0.43	2.32	0.46
capsule	1.23	0.05	1.59	0.04	1.59	0.11	1.59	0.21	1.62	0.32	2.14	0.38
carpet	1.17	0.11	1.24	0.12	1.23	0.13	1.16	0.24	1.36	0.29	1.41	0.34
grid	2.00	0.12	2.01	0.12	1.97	0.13	1.96	0.24	2.13	0.29	2.78	0.47
hazelnut	1.74	0.21	1.87	0.19	1.93	0.24	2.13	0.31	1.94	0.35	2.81	0.44
leather	1.47	0.14	2.12	0.14	2.06	0.17	1.94	0.41	1.91	0.35	2.17	0.37
metal_nut	1.56	0.15	1.47	0.30	1.49	0.32	1.96	0.30	1.57	0.32	2.02	0.40
pill	1.49	0.11	1.61	0.10	1.63	0.16	1.61	0.26	1.82	0.38	2.37	0.48
screw	1.12	0.16	1.19	0.12	1.12	0.14	1.28	0.30	1.43	0.36	1.63	0.39
tile	1.83	0.20	2.35	0.22	2.39	0.22	2.54	0.55	2.40	0.50	2.13	0.51
toothbrush	1.30	0.08	1.85	0.03	1.82	0.18	1.68	0.21	2.40	0.48	2.06	0.44
transistor	1.39	0.15	1.47	0.13	1.64	0.25	1.57	0.34	1.71	0.33	1.35	0.32
wood	1.95	0.23	2.19	0.29	2.12	0.35	2.33	0.37	2.24	0.40	1.88	0.40
zipper	1.23	0.11	1.25	0.10	1.29	0.27	1.39	0.25	2.14	0.37	1.91	0.39
Average	1.51	0.14	1.69	0.15	1.72	0.20	1.80	0.32	1.93	0.38	2.07	0.41

Compared with the traditional Crop&Paste strategy, the proposed method obtains consistent improvements in most categories, demonstrating the advantage of diffusion-based generation in representing complex defect morphology and fine-grained structural details. In comparison with Defect-GAN and DFMGAN, the proposed method also achieves more competitive IS and IC-LPIPS values in most categories, suggesting that diffusion-based generative modeling is more effective in capturing diverse defect distributions and local structural variations.

From a category-wise perspective, the proposed method shows particularly pronounced improvements on categories such as grid, hazelnut, pill, cable, and capsule. For instance, in the grid category, the proposed method achieves an IS of 2.78 and an IC-LPIPS of 0.47, clearly outperforming the baseline DualAnoDiff. In the hazelnut category, the proposed method obtains an IS of 2.81 and an IC-LPIPS of 0.44, indicating favorable generation quality and intra-class perceptual diversity. These improvements suggest that the proposed framework enhances the coordination among class-specific defect representation, global structural modeling, and local anomaly refinement, thereby producing visually realistic and structurally coherent defect samples.

Nevertheless, the category-wise results also show that the proposed method does not achieve the highest score in every individual category. In bottle, tile, toothbrush, transistor, wood, and zipper, its performance remains lower than that of some competing methods in certain metrics. Such category-specific performance variations may be attributed to the higher structural complexity, larger intra-class appearance variations, and more challenging local texture distributions of these categories. These factors make the generation process more sensitive to local structural perturbations and may lead to fluctuations in classifier-based confidence or perceptual diversity.

It should also be noted that IS and IC-LPIPS evaluate different aspects of generation performance. IS mainly reflects classifier-based visual quality and semantic separability, whereas IC-LPIPS emphasizes intra-class perceptual diversity. Therefore, although the

proposed method does not consistently achieve the best score for every single category and metric, the average results demonstrate that it provides a more balanced trade-off between generation fidelity and diversity under few-shot conditions.

Overall, by jointly improving class-specific discrepancy representation, global structural modeling, and local anomaly refinement, the proposed method achieves more balanced quantitative performance in terms of both IS and IC-LPIPS. These results further validate its effectiveness for high-quality industrial defect image generation under few-shot and class-imbalanced scenarios.

4.4. Additional Evaluation on the VisA Dataset

To provide preliminary evidence of the cross-dataset adaptability of the proposed framework, a supplementary evaluation is performed on selected categories from the VisA dataset. Compared with MVTec AD, VisA exhibits different object layouts, texture distributions, imaging conditions, and defect appearances, providing a complementary benchmark for evaluating defect generation under distributional variation. In this experiment, three categories are selected, namely PCB4, Macaroni2, and Cashew, to cover representative visual characteristics in VisA. PCB4 corresponds to densely structured electronic components with fine-grained elements such as chips, capacitors, circuits, and solder joints. Macaroni2 represents a multi-instance object scenario with non-fixed spatial layouts and pose variations. Cashew is a single-instance textured object category with a data distribution different from MVTec AD. These categories provide complementary test cases for evaluating the behavior of the proposed framework under different structural and textural conditions.

For consistency with the main quantitative evaluation, Inception Score (IS) and IC-LPIPS are adopted to measure generation quality and perceptual diversity, respectively. Table 3 reports the supplementary comparison between DualAnoDiff and the proposed method on the selected VisA categories.

Table 3. Supplementary comparison of IS and IC-LPIPS on selected VisA categories. The upward arrows indicate that higher values are better, and bold entries highlight the average row label and the best-performing average values.

Category	DFMGAN		AnomalyDiffusion		DualAnoDiff		Ours	
	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑	IS↑	IC-L↑
PCB4	1.21	0.28	1.14	0.25	1.13	0.21	1.19	0.28
Macaroni2	1.47	0.38	1.56	0.38	1.60	0.39	1.72	0.42
Cashew	1.25	0.24	1.27	0.26	1.29	0.24	1.28	0.30
Average	1.31	0.30	1.32	0.30	1.34	0.28	1.40	0.33

As reported in Table 3, the proposed method achieves the best average performance among the compared methods on the selected VisA categories, with an average IS of 1.40 and an average IC-LPIPS of 0.33. These results provide preliminary evidence that the proposed framework maintains competitive generation quality and perceptual diversity under a data distribution different from MVTec AD. Although category-level variations are observed due to differences in object layout, texture distribution, and defect appearance, the average results remain superior to those of the compared methods, indicating the potential adaptability of the proposed framework across different industrial defect datasets.

4.5. Multi-Class Discriminability Analysis

It can also be observed that the proposed method maintains competitive performance across the selected categories. Although category-level variations exist due to differences in object layout, texture distribution, and defect appearance, the average results remain

superior to those of the compared methods. This indicates that the proposed framework retains a favorable balance between generation quality and perceptual diversity under the VisA data distribution. To evaluate the effectiveness of the proposed method in modeling multi-class defect discrepancies, a class discriminability analysis is conducted. Previous evaluations typically rely on Inception Score (IS) and IC-LPIPS to assess the overall quality and perceptual diversity of generated images. However, these metrics primarily reflect visual fidelity and diversity and do not directly measure the discriminability among different defect categories.

Given that the proposed class-conditioned shared-basis LoRA (CSB-LoRA) is specifically designed to enhance multi-class discrepancy representation, further evaluation is conducted from the perspective of class separability.

In the experimental setup, three representative categories with diverse defect types, namely *hazelnut*, *bottle*, and *metal_nut*, are selected for analysis. To evaluate the discriminability of generated samples, a lightweight defect classification network based on ResNet18 pretrained on ImageNet is adopted. The network is fine-tuned on real defect samples using basic data augmentation under supervised learning. During testing, generated images from different methods are used as input, and classification performance is evaluated according to their target categories. If the generated samples can be correctly recognized by a classifier trained on real data, it indicates that the discriminative characteristics of the target categories are well preserved.

The evaluation metrics include classification accuracy (Accuracy) and macro-averaged F1 score (Macro-F1). Specifically, accuracy is defined as the proportion of correctly classified samples:

$$\text{Acc} = \frac{N_{\text{correct}}}{N_{\text{total}}} \quad (16)$$

where N_{correct} and N_{total} denote the number of correctly classified samples and the total number of samples, respectively.

Considering the potential class imbalance across defect categories, the Macro-F1 score is further adopted as a complementary metric. For the c -th defect category, the precision P_c and recall R_c are defined as:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c} \quad (17)$$

Based on this, the F1 score for the c -th class is defined as:

$$F1_c = \frac{2P_c R_c}{P_c + R_c} \quad (18)$$

Finally, the macro-averaged F1 score (Macro-F1) is defined as the average of the F1 scores across all classes:

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (19)$$

where C denotes the total number of defect categories. Compared with classification accuracy, which primarily reflects overall correctness, Macro-F1 provides a more balanced evaluation across different classes. Therefore, it is more suitable for assessing the extent to which generated samples preserve class-discriminative features under class imbalance.

As shown in Table 4, the proposed method achieves the best average performance in terms of both classification accuracy and macro-averaged F1 score across the evaluated categories. On average, it attains an accuracy of 88.3% and a Macro-F1 score of 87.1%, surpassing DualAnoDiff by 1.4% and 2.2%, respectively.

The improvement is particularly pronounced in the hazelnut category, indicating that the proposed method is more effective in preserving class-discriminative features for defect types with subtle inter-class differences and complex local structures. For the bottle and metal_nut categories, the proposed method also maintains consistently better performance gains, demonstrating the robustness of the class-conditioned shared-basis LoRA (CSB-LoRA) in enhancing class recognizability across different object categories.

Compared with Defect-GAN and AnomalyDiffusion, the proposed method exhibits more stable and consistently superior performance across all evaluated categories. This suggests that the proposed framework effectively improves fine-grained multi-class discrepancy representation. In particular, CSB-LoRA enables a better balance between shared representation learning and class-specific feature modeling, thereby enhancing discriminability among different defect categories.

Table 4. Comparison of multi-class discriminability of generated samples by different methods. The bold formatting is retained to highlight the average row label and the best-performing average values.

Category	Metric	Defect-GAN	AnomalyDiffusion	DualAnoDiff	Ours
hazelnut	Acc (%)	79.4	83.8	87.2	89.6
	Macro-F1 (%)	77.1	81.5	85.5	88.6
bottle	Acc (%)	74.8	80.7	85.8	86.3
	Macro-F1 (%)	72.9	78.6	83.1	84.9
metal_nut	Acc (%)	81.2	84.9	87.6	89.1
	Macro-F1 (%)	79.3	83.0	86.1	87.8
Average	Acc (%)	78.5	83.1	86.9	88.3
	Macro-F1 (%)	76.4	81.0	84.9	87.1

4.6. Qualitative Analysis of Defect Image Generation

To further validate the effectiveness of the proposed method in defect image generation, a qualitative comparison is conducted across different approaches. Figure 4 presents representative results on several typical categories from the MVTec AD dataset, including Crop&Paste, Defect-GAN, AnomalyDiffusion, DualAnoDiff, and the proposed method. By comparing the generated results with real defect samples, we analyze the performance of different methods from three perspectives: defect morphology, boundary transition, and background texture consistency.

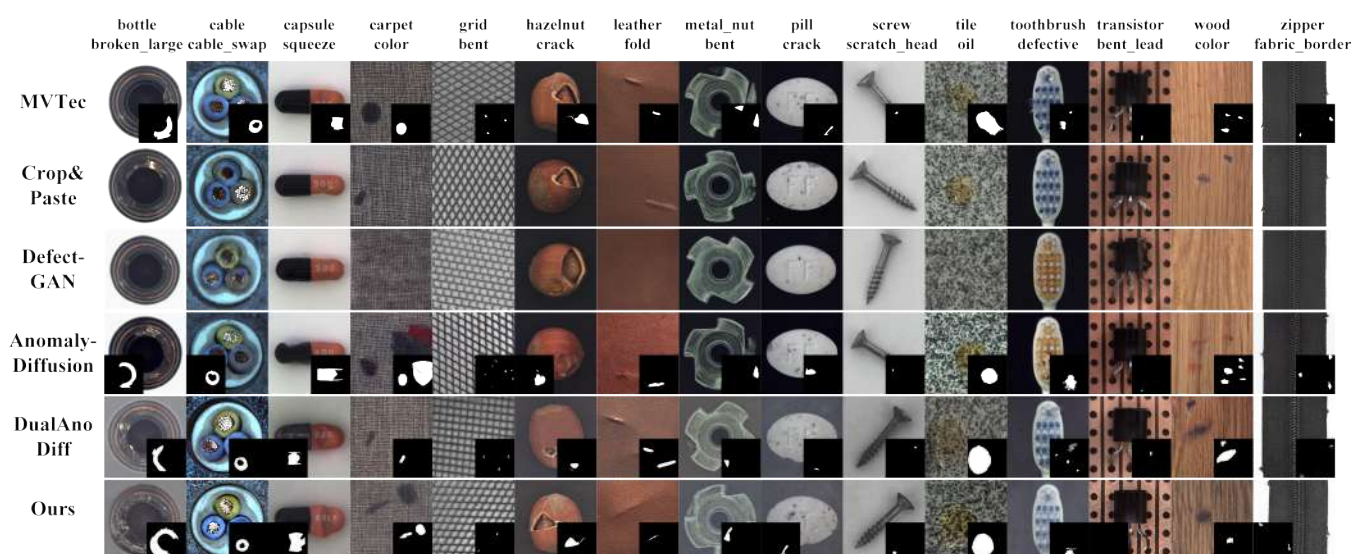


Figure 4. Qualitative comparison of generated results on the MVTec AD dataset.

As shown in Figure 4, Crop&Paste constructs anomalous regions through simple copy–paste operations, which leads to noticeable artificial artifacts across multiple categories. The transition between defect regions and surrounding backgrounds is often unnatural, and this issue becomes more pronounced in categories with complex textures or repetitive structures.

Under limited defect samples, Defect-GAN struggles to capture salient defect characteristics, resulting in generated defects that are less prominent and often too small to be practically useful. AnomalyDiffusion improves the visibility and diversity of anomalous regions; however, it may introduce noticeable distortions in the background structure in some cases. This often disrupts the original texture distribution and local structural consistency, thereby degrading the overall visual coherence of the generated images. DualAnoDiff demonstrates improved coordination between defect regions and backgrounds, producing more natural overall visual results. However, under complex background textures or for small defect regions, it still shows limitations in local anomaly representation and fine-detail recovery.

Building on these observations, the proposed method further improves defect morphology, boundary transition, and background consistency. For various defect types such as cracks, scratches, and contamination, the generated results better align with their real visual characteristics, exhibiting clearer contours and more natural local textures and boundary details. Moreover, under complex textured backgrounds, the proposed method is able to highlight anomalous regions while preserving the underlying structural and textural distribution, resulting in outputs that are closer to real defect samples and exhibit improved realism and visual coherence.

Furthermore, as illustrated in Figures 5–7 for the hazelnut, bottle, and metal_nut categories, the proposed method produces defect regions with higher realism and visual plausibility. Compared with existing methods, the generated defects exhibit improved consistency with real samples in terms of structural morphology, local texture patterns, and boundary continuity, indicating more accurate modeling of category-specific visual characteristics.

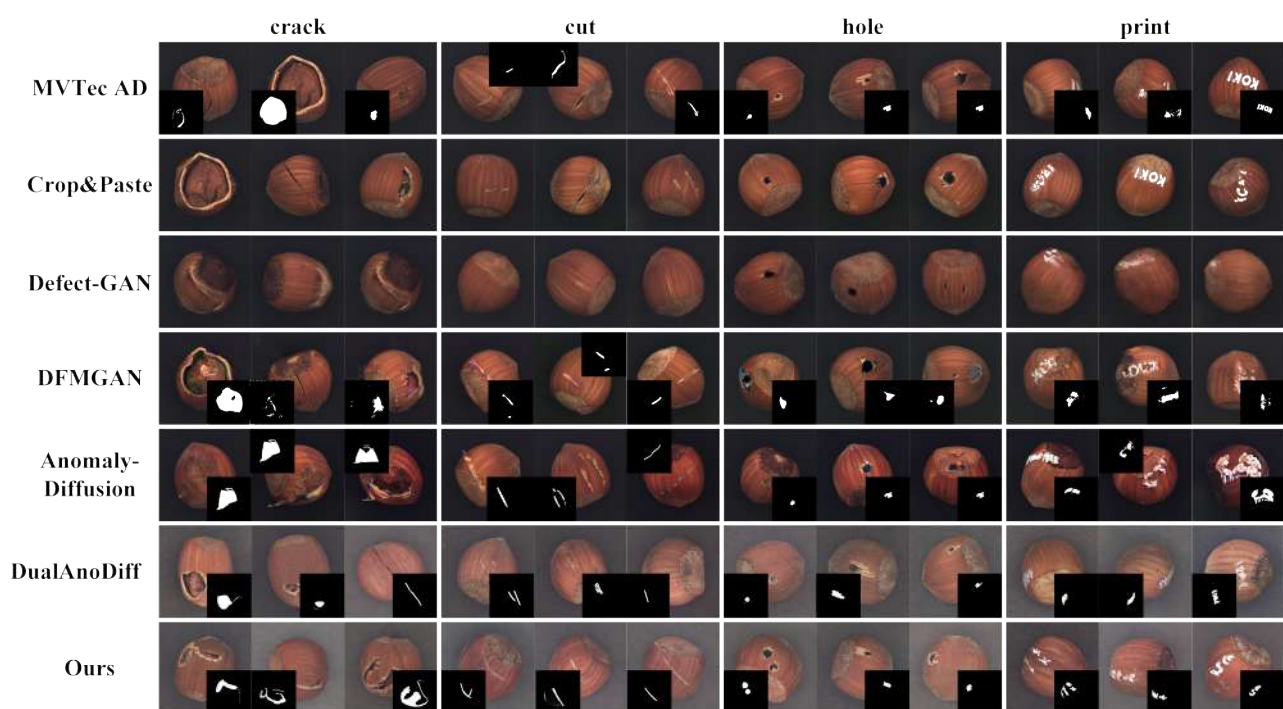


Figure 5. Qualitative comparison of generated defect images in the hazelnut category.

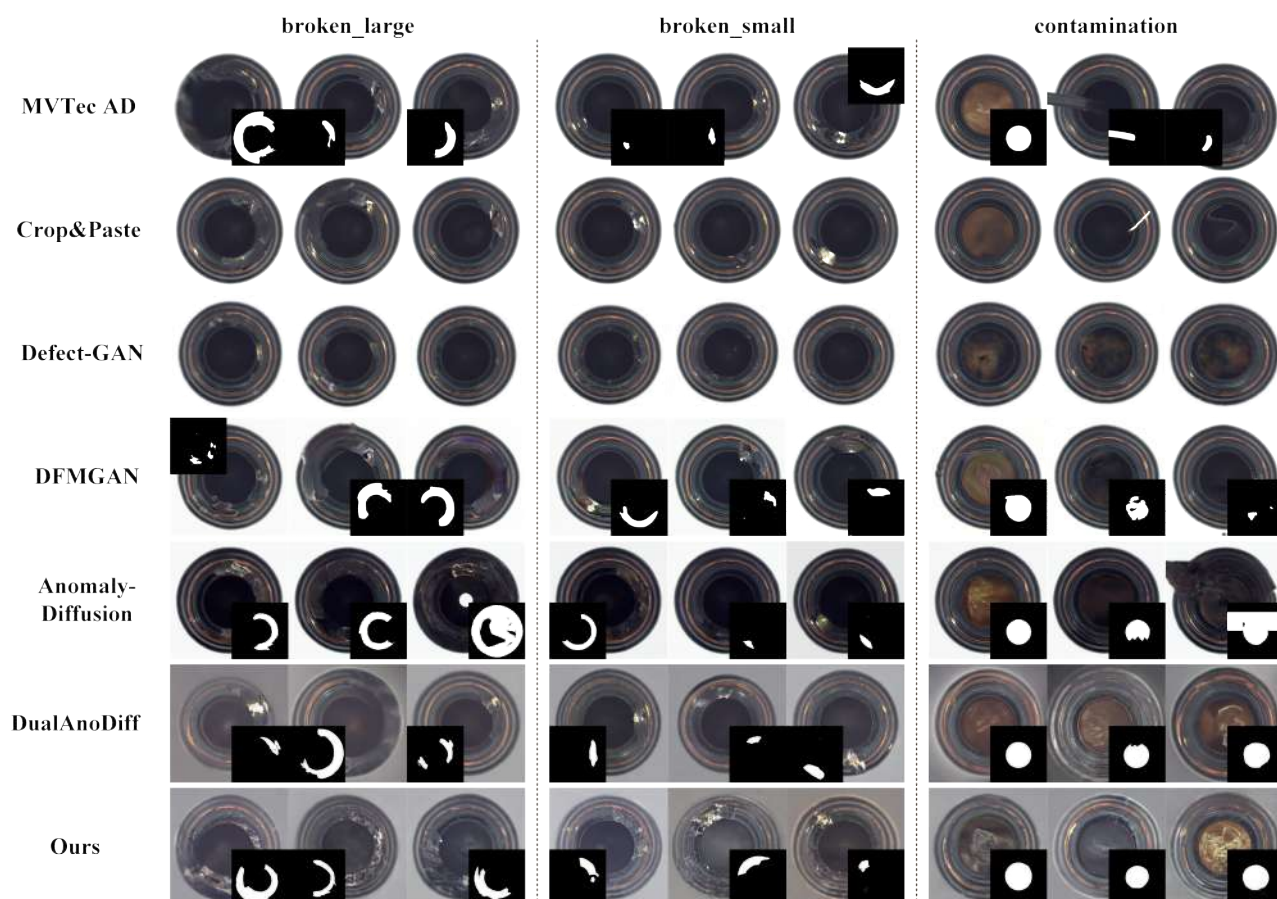


Figure 6. Qualitative comparison of generated defect images in the bottle category.

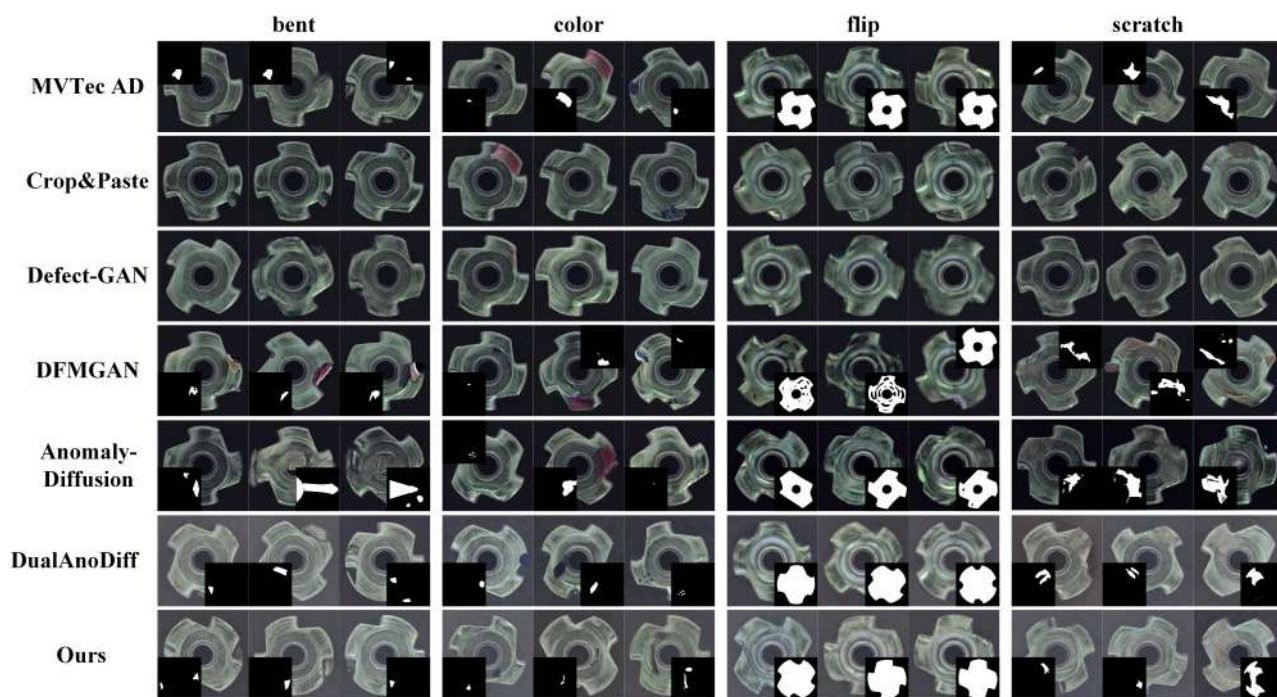


Figure 7. Qualitative comparison of generated defect images in the *metal_nut* category.

Meanwhile, the proposed method improves the representation of anomalous regions while maintaining overall background consistency. This helps alleviate visual artifacts such

as boundary discontinuities and texture inconsistencies, resulting in improved perceptual coherence of generated defect images.

Figure 8 further presents localized visual comparisons of representative defect regions and their surrounding background contexts, covering bottle-broken_large, bottle-contamination, hazelnut-crack, and metal_nut-scratch. These examples are used to inspect local defect morphology, boundary transitions, and contextual texture consistency.

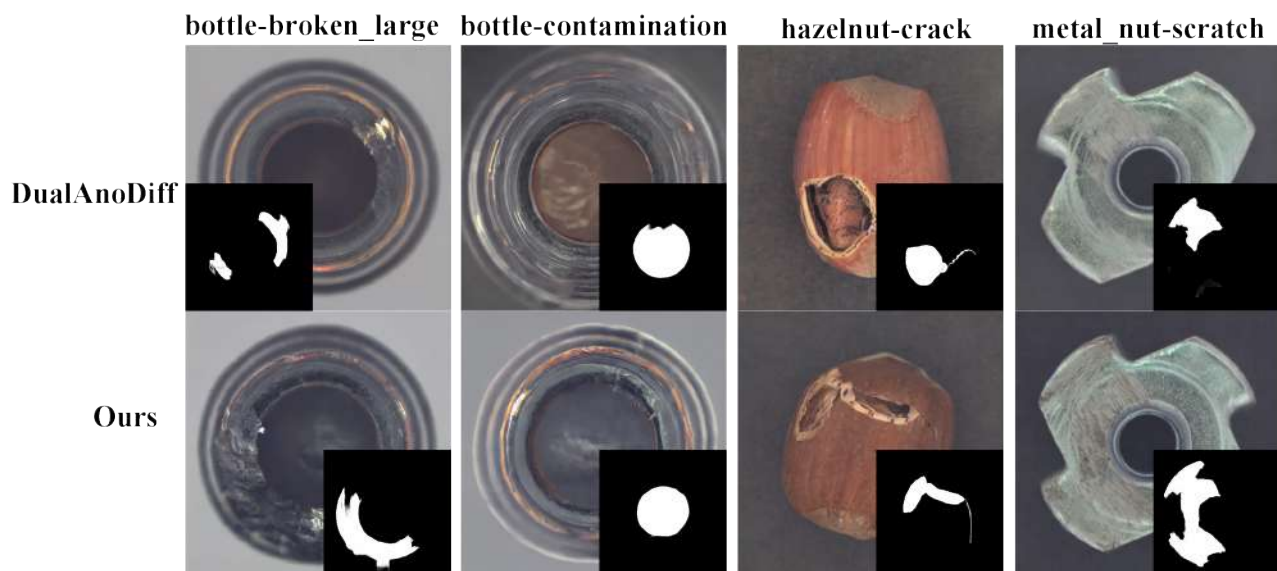


Figure 8. Localized visual comparison of representative defect regions and surrounding background contexts, including bottle-broken_large, bottle-contamination, hazelnut-crack, and metal_nut-scratch. The enlarged regions highlight defect morphology, boundary transitions, and contextual texture consistency.

Compared with DualAnoDiff, the proposed method produces more faithful local defect structures and more natural transitions between anomalous regions and adjacent background textures. In the bottle-broken_large and bottle-contamination cases, the generated anomalies are more consistent with the circular bottle-bottom geometry and surrounding texture patterns. In the hazelnut-crack case, the generated defect conforms more coherently to the natural surface structure while maintaining the continuity of neighboring texture regions. In the metal_nut-scratch case, the generated scratch better follows the metallic surface pattern and geometric contour. These localized observations provide qualitative evidence that the proposed framework enhances the coordination between global structural modeling and local anomaly refinement, thereby improving local visual coherence across different materials and defect types.

4.7. Ablation Study

To further validate the effectiveness of each proposed component in industrial defect image generation, ablation studies are conducted on representative categories of the MVTec AD dataset. The original DualAnoDiff is adopted as the baseline model. Under the same training settings, Class-Conditioned Shared-Basis LoRA (CSB-LoRA), Temporal Dual-branch Attention Modulation (TDAM), and Saliency-Guided Reconstruction Loss (SGRL) are progressively incorporated to evaluate their individual and combined contributions.

To avoid limiting the ablation analysis to a single category, the experiments are extended to an additional category with distinct visual characteristics. Specifically, hazelnut is used as the original object category with natural surface textures, while grid is selected as an additional texture category characterized by regular high-frequency patterns and small

defect disturbances. This setting enables the effectiveness of the proposed components to be examined under both object-surface and regular-texture scenarios.

The performance of different component combinations is quantitatively compared using Inception Score (IS) and IC-LPIPS. IS reflects the visual quality and discriminability of generated images, while IC-LPIPS evaluates the perceptual diversity among generated samples within the same category. The ablation results on hazelnut and grid are summarized in Tables 5 and 6, respectively.

As shown in Tables 5 and 6, the proposed components lead to overall performance gains over the baseline across the two representative categories. For hazelnut, the complete model improves IS and IC-LPIPS from 1.94 and 0.35 to 2.81 and 0.44, respectively. For grid, the corresponding metrics are improved from 2.13 and 0.29 to 2.78 and 0.47. These results indicate that the proposed components are effective not only for object-surface defects but also for regular high-frequency texture patterns.

The ablation results further show that introducing individual components generally improves the overall performance, while the relative gains vary across different modules and module combinations. In most cases, combining two modules provides better performance than using a single module alone, while integrating all three modules achieves the best overall results in both representative categories. The variations among different configurations suggest that the three components contribute from complementary perspectives rather than through simple additive effects.

Table 5. Ablation study results on the MVTec AD dataset (hazelnut category). The upward arrows indicate that higher values are better. The checkmark indicates that the corresponding module is included. Bold values highlight the best-performing results under the corresponding metrics.

CSB-LoRA	TDAM	SGRL	IS↑	IC-LPIPS↑
—	—	—	1.94	0.35
✓	—	—	2.18	0.38
—	✓	—	2.11	0.37
—	—	✓	2.07	0.36
✓	✓	—	2.43	0.41
✓	—	✓	2.32	0.39
—	✓	✓	2.36	0.40
✓	✓	✓	2.81	0.44

Table 6. Ablation study results on the MVTec AD dataset (grid category). The upward arrows indicate that higher values are better. The checkmark indicates that the corresponding module is included. Bold values highlight the best-performing results under the corresponding metrics.

CSB-LoRA	TDAM	SGRL	IS↑	IC-LPIPS↑
—	—	—	2.13	0.29
✓	—	—	2.37	0.36
—	✓	—	2.28	0.35
—	—	✓	2.25	0.33
✓	✓	—	2.57	0.42
✓	—	✓	2.43	0.38
—	✓	✓	2.48	0.39
✓	✓	✓	2.78	0.47

Regarding the individual contributions of the proposed components, CSB-LoRA generally produces the most evident improvement among the individual components in both categories, suggesting its importance in category-specific discrepancy modeling through

class-conditioned shared-basis adaptation. TDAM also yields stable improvements by adaptively coordinating global structural preservation and local anomaly refinement across denoising stages. Although SGRL brings relatively moderate gains when used independently, its consistent improvements suggest that region-aware reconstruction supervision contributes to defect-region refinement and boundary-related structural consistency.

The results of dual-component combinations further demonstrate the complementarity among CSB-LoRA, TDAM, and SGRL. In particular, the combination of CSB-LoRA and TDAM consistently achieves stronger performance than either component alone in both categories, suggesting that category-specific discrepancy modeling and diffusion-stage-aware global–local coordination reinforce each other during generation. When all three components are integrated, the best results are achieved in both categories, confirming the combined effectiveness of the proposed modules. Overall, the ablation results provide additional evidence for the effectiveness and complementarity of CSB-LoRA, TDAM, and SGRL under different defect patterns.

In addition to the module-level ablation, we further examine the influence of the shared basis number K in CSB-LoRA on the hazelnut category. As shown in Table 7, increasing K from 2 to 3 improves IS from 2.27 to 2.81 and IC-LPIPS from 0.38 to 0.44, indicating that a moderate number of shared bases benefits class-specific discrepancy representation. When K is further increased to 4, both metrics decrease, suggesting that excessive basis components may introduce parameter redundancy and unstable adaptation under few-shot conditions. Therefore, $K = 3$ is adopted in the main experiments to balance representation capacity, parameter efficiency, and generation stability.

Table 7. Sensitivity analysis of the shared basis number K in CSB-LoRA using the hazelnut category. The upward arrows indicate that higher values are better. Bold entries highlight the best-performing setting and results.

K	IS $\uparrow$	IC-LPIPS $\uparrow$
2	2.27	0.38
3	2.81	0.44
4	2.46	0.41

4.8. Visualization Analysis of TDAM Gating Parameters

To further examine whether the proposed Temporal Dual-branch Attention Modulation (TDAM) learns stage-adaptive modulation patterns during the denoising process, we visualize the temporal responses of its gating parameters after training on the hazelnut category of MVTec AD. Specifically, we record the global-branch gate γ_g , the anomaly-branch gate γ_a , the background information injection gate δ_t , and the SAIM residual fusion gate β_t at different denoising time steps. For each denoising step, the gate values are averaged over multiple generated samples and the TDAM-equipped attention layers in the U-Net, thereby obtaining an overall temporal response trend. In Figure 9, the horizontal axis denotes the denoising progress sorted according to the timestep order, where the left side corresponds to the early high-noise stage and the right side corresponds to the late low-noise stage.

As shown in Figure 9, the four gating parameters are not fixed throughout the denoising process but exhibit distinct temporal variations. In the early high-noise stage, γ_g maintains a relatively high response, while δ_t also remains at a comparatively high level. This indicates that the model tends to rely more on global structural information and background constraints at the beginning of generation, which helps stabilize the overall layout and reduce structural distortion under strong noise perturbations. As denoising proceeds, γ_g gradually decreases, whereas γ_a increases and surpasses γ_g in the later stage,

suggesting that the generation focus progressively shifts from global structure recovery to local anomaly detail refinement.

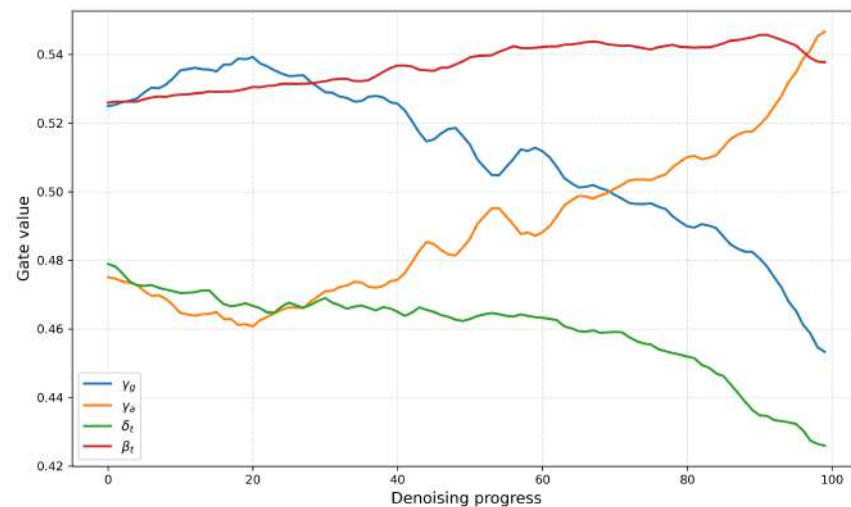


Figure 9. Visualization of TDAM gating parameters during the denoising process.

Meanwhile, the decreasing trend of δ_t suggests that the background constraint is gradually relaxed in the later denoising stage, providing greater flexibility for defect texture synthesis, boundary transition modeling, and local morphology generation. In addition, β_t maintains a relatively high and stable response throughout the denoising process, with a slight increase in the middle-to-late stage. This trend indicates that cross-branch residual fusion continues to contribute to the collaborative modeling of global structural information and local anomaly features.

Overall, these observations provide direct evidence that TDAM dynamically adjusts global structure modeling, background information injection, anomaly-region modulation, and cross-branch fusion across different denoising stages. The learned gating trends are generally consistent with the coarse-to-fine generation process of diffusion models, further supporting the stage-adaptive behavior and interpretability of the proposed TDAM.

4.9. Downstream Task Evaluation Using Generated Data

4.9.1. Generated Data Construction and Downstream Evaluation Setup

In accordance with the experimental protocol of MVTec AD, defective samples within each category were partitioned into training and testing subsets at an approximate ratio of 1:2. The training subset, accounting for roughly one third of the defective samples, was used to train the defect image generation model and to construct the augmented training set for downstream segmentation. The remaining testing subset, accounting for approximately two thirds of the defective samples, was reserved exclusively for downstream testing and evaluation. Given the limited number of defective samples in certain categories, no additional validation subset was established to avoid further reducing the already constrained few-shot training data. All categories followed the same data preprocessing, training, and generation protocols to ensure experimental consistency.

During the synthetic data construction stage, defect images were generated independently for each object category using the corresponding defective training subset. Specifically, for each defect type within each object category, 1000 synthetic defect images were generated and added to the training data of the downstream segmentation model. Consequently, the total number of generated samples for each object category depended on the number of defect types contained in that category. The original testing samples were excluded from both the generation and augmentation procedures and were used only

for testing, ensuring a clear separation between training and evaluation data. Detailed counts of original training samples, generated samples, and original testing samples for each category are reported in Table 8, providing a comprehensive overview of the dataset scale used in the downstream evaluation.

For the downstream task setup, the classical U-Net is adopted as the baseline model to evaluate the effectiveness of the generated defect data. The network follows an encoder–decoder architecture, where multi-scale semantic features are progressively extracted through downsampling and spatial resolution is restored via upsampling. Meanwhile, skip connections are introduced to effectively fuse low-level detailed information with high-level semantic representations, enabling the model to capture both global structural context and fine-grained details.

Benefiting from its fully convolutional design and parameter efficiency, U-Net serves as a stable and widely adopted baseline under limited training data. Therefore, it is well suited as an evaluation model for assessing the quality and utility of the generated defect data.

Table 8. Statistics of training and testing samples for each category on the MVTec AD dataset.

Category	Training Samples		Testing Samples Original
	Original	Generated	
bottle	22	3000	41
cable	34	8000	58
capsule	38	5000	71
carpet	32	5000	57
grid	20	5000	37
hazelnut	24	4000	46
leather	33	5000	59
metal_nut	33	4000	60
pill	49	7000	92
screw	41	5000	78
tile	29	5000	55
toothbrush	10	1000	20
transistor	16	4000	24
wood	22	5000	38
zipper	43	7000	76

4.9.2. Downstream Task Results and Analysis

Considering that industrial defect analysis requires not only identifying the presence of anomalies but also accurately localizing abnormal regions, the performance is evaluated from both pixel-level anomaly localization and image-level anomaly detection perspectives. Let denote the pixel-level anomaly score map, and denote the corresponding ground-truth defect mask. The pixel-level and image-level evaluation metrics are defined as follows:

(1) Pixel-level AUROC (AUC-P). The pixel-level Area Under the Receiver Operating Characteristic Curve (AUROC), denoted as AUC-P, is employed to measure the separability between anomalous and normal pixels. It is defined as the area under the ROC curve, which characterizes the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) under varying decision thresholds.

The value of AUC-P ranges from 0 to 1, where a higher value indicates stronger capability to distinguish defect regions from normal regions at the pixel level. Formally, AUC-P can be expressed as:

$$\text{AUC-P} = \int_0^1 \text{TPR}(t) d\text{FPR}(t) \quad (20)$$

The TPR and FPR are defined as:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \quad (21)$$

where TP , FP , FN , and TN denote the numbers of true positives, false positives, false negatives, and true negatives, respectively.

(2) Pixel-level Average Precision (AP-P). The pixel-level Average Precision (AP-P) is defined as the area under the precision–recall (PR) curve and is used to evaluate pixel-wise detection performance under varying threshold values. Compared with AUC-P, AP-P is generally more sensitive to class imbalance and is more suitable for scenarios with low anomaly prevalence.

Formally, AP-P is defined as:

$$\text{AP-P} = \sum_n (R_n - R_{n-1}) P_n \quad (22)$$

where P_n and R_n denote the precision and recall at the n -th threshold, respectively.

(3) Pixel-level F1 Score (F1-P). The pixel-level F1 score (F1-P) measures the balance between precision and recall under a fixed threshold and is defined as the harmonic mean of precision and recall:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \quad (23)$$

$$\text{F1-P} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (24)$$

A higher F1-P value indicates a better trade-off between false positives and false negatives.

(4) Image-level Average Precision (AP-I). The image-level Average Precision (AP-I) is used to evaluate anomaly detection performance at the image level. Specifically, the pixel-wise anomaly score map is aggregated to obtain an image-level anomaly score. In this work, the maximum value of the pixel-wise score map is adopted:

$$s = \max_{x,y} S(x, y) \quad (25)$$

Based on these scores, the image-level precision–recall curve is constructed, and the area under the curve is used as AP-I.

Tables 9 and 10 present a comprehensive quantitative evaluation of pixel-level anomaly localization and image-level anomaly detection across 15 industrial categories using U-Net under different data augmentation strategies. Overall, the proposed method achieves the highest average AP-P and F1-P values among the compared methods, which are more sensitive to localization accuracy under class-imbalanced conditions. Specifically, AP-P and F1-P reach 85.7% and 79.3%, respectively. Compared with DFMGAN, the proposed method improves AP-P and F1-P by 23.0% and 17.2% on average and consistently outperforms AnomalyDiffusion and DualAnoDiff in AP-P.

In contrast, AUC-P and AP-I, which evaluate pixel-level separability and image-level detection performance, reach consistently high scores for most methods, with most average values exceeding 98.8% and only marginal differences across methods. This suggests that AUC-P and AP-I are close to saturation in this downstream evaluation and are therefore less sensitive to fine-grained differences among different generation-based augmentation strategies. Accordingly, the following analysis mainly focuses on AP-P and F1-P, which provide more discriminative evaluation under stricter pixel-level localization and segmentation criteria.

The proposed method achieves average AP-P and F1-P values of 85.7% and 79.3%, respectively. Compared with DualAnoDiff, AP-P increases from 84.5% to 85.7%, and F1-P increases from 78.8% to 79.3%. Although these improvements are relatively modest, they still indicate that the generated samples provide useful supplementary supervision for downstream segmentation under stricter pixel-level evaluation criteria. Compared with DFMGAN, the proposed method improves AP-P and F1-P by 23.0% and 17.2% on average, suggesting that diffusion-based defect generation with consistent annotation constraints is more effective than GAN-based augmentation in this setting.

Further category-wise analysis shows that the performance gains mainly stem from improved sensitivity to subtle defects and robustness under complex background patterns. For categories with relatively salient defects, such as bottle, hazelnut, and metal nut, all methods achieve high performance across metrics, with image-level AP-I reaching or approaching 100%. However, for more challenging categories with small defect regions that are easily overwhelmed by background textures, such as screw and grid, DFMGAN performs poorly, with AP-P values of only 2.2% and 14.3%, respectively. In contrast, the proposed method achieves higher AP-P values of 63.5% and 65.5% for these categories. Similarly, for the capsule category, it achieves 77.1% in AP-P and 71.5% in F1-P, outperforming both AnomalyDiffusion and DualAnoDiff. These results suggest that, under a well-designed data augmentation strategy and consistent annotation constraints, the proposed diffusion-based approach can provide useful additional supervision for defect segmentation, especially in fine-grained defect scenarios.

Although the proposed method improves the average AP-P and F1-P, the category-wise downstream results are not uniformly improved across all defect objects. In particular, toothbrush, transistor, and wood show lower AP-P or F1-P values than DualAnoDiff in Tables 9 and 10. This suggests that the benefit of generated samples depends on the structural characteristics and defect distributions of each category. For categories with slender structures, complex components, or highly variable textures, synthetic samples may introduce additional appearance variations that do not always translate into better segmentation performance. Therefore, the downstream results should be interpreted as an overall improvement in average task performance rather than uniform gains for every individual category.

Table 9. Quantitative evaluation results of anomaly detection (AUC-P, AP-P). Bold entries highlight the average row label.

Defect Object	DFMGAN		AnomalyDiffusion		DualAnoDiff		Ours	
	AUC-P	AP-P	AUC-P	AP-P	AUC-P	AP-P	AUC-P	AP-P
bottle	98.9	90.2	99.4	94.1	99.5	93.4	99.4	93.1
cable	97.2	81.0	99.2	90.8	97.5	82.6	98.4	85.5
capsule	79.2	26.0	98.8	57.2	99.5	73.2	99.5	77.1
carpet	90.6	33.4	98.6	81.2	99.4	89.1	99.6	92.5
grid	75.2	14.3	98.3	52.9	98.5	57.2	99.2	65.5
hazelnut	99.7	95.2	99.8	96.5	99.8	97.7	99.8	98.2
leather	98.5	68.7	99.8	79.6	99.9	88.8	99.8	92.3
metal_nut	99.3	98.1	99.8	98.7	99.6	98.0	99.9	98.5
pill	81.2	67.8	99.8	97.0	99.6	95.8	99.9	97.3
screw	58.8	2.2	97.0	51.8	98.1	57.1	98.6	63.5
tile	99.5	97.1	99.2	93.9	99.7	97.1	99.5	95.7
toothbrush	96.4	75.9	99.2	76.5	98.2	68.3	97.1	67.4
transistor	96.2	81.2	99.3	92.6	98.0	86.7	95.6	83.1
wood	95.3	70.7	98.9	84.6	99.4	91.6	95.6	86.7
zipper	92.9	65.6	99.4	86.0	99.6	90.7	99.7	88.7
Average	90.0	62.7	99.1	81.4	99.1	84.5	98.8	85.7

Table 10. Quantitative evaluation results of anomaly detection (F1-P, AP-I). Bold entries highlight the average row label.

Defect Object	DFMGAN		AnomalyDiffusion		DualAnoDiff		Ours	
	F1-P	AP-I	F1-P	AP-I	F1-P	AP-I	F1-P	AP-I
bottle	83.9	99.8	87.3	99.9	85.7	100.0	84.5	99.8
cable	75.4	97.8	83.5	100.0	76.9	98.3	79.4	98.7
capsule	35.0	98.5	59.8	99.9	67.0	99.2	71.5	100.0
carpet	38.1	98.5	74.6	98.8	80.2	99.9	84.1	100.0
grid	20.5	90.4	54.6	99.5	54.9	99.7	58.2	99.7
hazelnut	89.5	100.0	90.6	99.9	92.8	100.0	95.1	100.0
leather	66.7	100.0	71.0	100.0	78.8	100.0	76.5	99.7
metal_nut	94.5	99.8	94.0	100.0	93.0	99.9	95.2	100.0
pill	72.6	91.7	90.8	99.6	89.2	99.0	94.5	99.6
screw	5.3	64.7	50.9	97.9	56.1	95.0	60.1	98.3
tile	91.6	100.0	86.2	100.0	91.0	100.0	89.4	99.8
toothbrush	72.6	100.0	73.4	100.0	68.6	99.7	65.1	99.4
transistor	77.0	92.5	85.7	100.0	79.6	93.7	75.5	91.7
wood	65.8	99.4	74.5	99.4	83.8	99.9	79.5	98.1
zipper	64.9	99.9	79.2	100.0	82.7	100.0	80.2	100.0
Average	62.1	94.8	76.3	99.7	78.8	98.9	79.3	99.0

5. Conclusions

This paper develops a symmetry-aware multi-constraint diffusion framework based on the dual-branch architecture of DualAnoDiff for industrial defect image generation under data scarcity and class imbalance. The framework introduces three coordinated components for multi-class defect synthesis. CSB-LoRA models category-specific defect discrepancies by combining shared low-rank bases with class-dependent coefficients, allowing the model to reuse common structural priors while distinguishing different anomaly patterns. TDAM adjusts the interaction between the global branch and the local anomaly branch across denoising stages, so that global structure preservation and local defect refinement can be handled with different emphasis during generation. SGRL further uses pixel-level defect cues to construct region-aware supervision, giving more attention to small anomalous regions and boundary neighborhoods. With these components, the generated samples better preserve global background regularity while capturing category-related defect structures and local details.

Experiments on MVTec AD show improvements in average generation quality and perceptual diversity under few-shot and class-imbalanced settings. When the generated samples are added to downstream training data, the average defect segmentation performance is also improved, especially for defect-region recognition and localization. Some category-level fluctuations still appear in structurally complex or texture-sensitive classes, but the supplementary experiments on selected VisA categories suggest that the method can maintain competitive performance under a different industrial defect distribution. These results show that the generated samples can enrich limited defect data and support downstream industrial inspection tasks. Since the generated images can be paired with corresponding defect masks, they can also be organized as annotated synthetic data for defect segmentation, anomaly localization, quality inspection model training, and few-shot data augmentation.

Several issues still need further study. Although the MVTec AD experiments and the selected VisA results provide preliminary evidence of cross-dataset adaptability, more diverse industrial datasets and imaging conditions are needed to evaluate practical robustness. The additional adaptation and modulation modules also increase the training and sampling cost, so improving computational efficiency remains an important direction. For

categories with complex structures or large intra-class variations, more precise modeling of defect boundaries and surrounding structural context may further improve generation stability. Future work will focus on broader industrial validation, more efficient diffusion sampling, and stronger boundary-aware modeling for highly variable defect categories.

Author Contributions: Conceptualization, B.J. and H.S.; methodology, B.J. and H.S.; software, D.J.; validation, N.T. and G.Y.; formal analysis, B.J.; investigation, B.J. and H.S.; resources, B.J.; data curation, H.S.; writing—original draft preparation, B.J. and H.S.; writing—review and editing, B.J. and H.S.; visualization, D.J., N.T. and G.Y.; supervision, D.J., N.T. and G.Y.; project administration, B.J.; funding acquisition, B.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Open Fund of Key Laboratory of High Performance Ship Technology (Wuhan University of Technology), Ministry of Education, grant number gxnc25112203.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9584–9592. <https://doi.org/10.1109/CVPR.2019.00982>.
2. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *Pattern Recognition. ICPR International Workshops and Challenges*; Springer: Cham, Switzerland, 2021; pp. 475–489. https://doi.org/10.1007/978-3-030-68799-1_35.
3. 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), Montreal, QC, Canada, 10–17 October 2021; pp. 8330–8339.
4. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. <https://doi.org/10.1186/s40537-019-0197-0>.
5. 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 (CVPR), Virtual, 19–25 June 2021; pp. 9664–9674.
6. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems; Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
7. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; pp. 6840–6851. <https://doi.org/10.48550/arXiv.2006.11239>.
8. 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 (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695. <https://doi.org/10.1109/CVPR52688.2022.01042>.
9. 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 (AAAI), Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 8526–8534. <https://doi.org/10.1609/aaai.v38i8.28696>.
10. Jin, Y.; Peng, J.; He, Q.; Hu, T.; Wu, J.; Chen, H.; Wang, H.; Zhu, W.; Chi, M.; Liu, J.; Wang, Y. Dual-Interrelated Diffusion Model for Few-Shot Anomaly Image Generation. *arXiv* **2024**, arXiv:2408.13509. <https://doi.org/10.48550/arXiv.2408.13509>.
11. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595. <https://doi.org/10.1109/CVPR.2018.00068>.
12. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016. <https://doi.org/10.48550/arXiv.1511.06434>.
13. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134. <https://doi.org/10.1109/CVPR.2017.632>.

14. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2232. <https://doi.org/10.1109/ICCV.2017.244>.
15. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410. <https://doi.org/10.1109/CVPR.2019.00453>.
16. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021. <https://doi.org/10.48550/arXiv.2010.02502>.
17. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021. <https://doi.org/10.48550/arXiv.2011.13456>.
18. Mo, S.; Cho, M.; Shin, J. Freeze the discriminator: A simple baseline for fine-tuning GANs. *arXiv* **2020**, arXiv:2002.10964. <https://doi.org/10.48550/arXiv.2002.10964>.
19. Wang, R.; Hoppe, S.; Monari, E.; Huber, M.F. Defect Transfer GAN: Diverse Defect Synthesis for Data Augmentation. *arXiv* **2023**, arXiv:2302.08366, .
20. Duan, Y.; Niu, L.; Hong, Y.; Zhang, L. WeditGAN: Few-Shot Image Generation via Latent Space Relocation. *arXiv* **2023**, arXiv:2305.06671. <https://doi.org/10.48550/arXiv.2305.06671>.
21. Zhao, Y.; Chandrasegaran, K.; Abdollahzadeh, M.; Cheung, N.-M. Few-shot Image Generation via Adaptation-Aware Kernel Modulation. In Proceedings of the Thirty-Sixth Annual Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022. <https://doi.org/10.48550/arXiv.2210.16559>.
22. 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. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. <https://doi.org/10.48550/arXiv.2208.01618>.
23. Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv* **2022**, arXiv:2208.12242. <https://doi.org/10.48550/arXiv.2208.12242>.
24. 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 Proceedings of the International Conference on Learning Representations (ICLR), Online, 25–29 April 2022. <https://doi.org/10.48550/arXiv.2106.09685>.
25. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with Cutout. *arXiv* **2017**, arXiv:1708.04552. <https://doi.org/10.48550/arXiv.1708.04552>.
26. Niu, S.; Li, B.; Wang, X.; Lin, H. Defect Image Sample Generation With GAN for Improving Defect Recognition. *IEEE Trans. Autom. Sci. Eng.* **2020**, *17*, 1611–1622.
27. Zhang, G.; Cui, K.; Hung, T.-Y.; Lu, S. Defect-GAN: High-Fidelity Defect Synthesis for Automated Defect Inspection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), online, 5–9 January 2021; pp. 2524–2534.
28. Liu, R.; Liu, W.; Zheng, Z.; Wang, L.; Mao, L.; Qiu, Q.; Ling, G. Anomaly-GAN: A Data Augmentation Method for Train Surface Anomaly Detection. *Expert Syst. Appl.* **2023**, *228*, 120284.
29. Duan, Y.; Hong, Y.; Niu, L.; Zhang, L. Few-Shot Defect Image Generation via Defect-Aware Feature Manipulation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 571–578.
30. Wang, Y.; Zhou, Z.; Tan, X.; Pan, Y.; Yuan, J.; Qiu, Z.; Liu, C. Unveiling the Potential of Progressive Training Diffusion Model for Defect Image Generation and Recognition in Industrial Processes. *Neurocomputing* **2024**, *592*, 127837.
31. Wyatt, J.; Leach, A.; Schmon, S.M.; Willcocks, C.G. AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models Using Simplex Noise. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 650–656.
32. Choi, J.Y.; Park, J.R.; Park, I.; Cho, J.; No, A.; Ryu, E.K. Simple Drop-in LoRA Conditioning on Attention Layers Will Improve Your Diffusion Model. *arXiv* **2024**, arXiv:2405.03958. <https://doi.org/10.48550/arXiv.2405.03958>.
33. Tang, S.; Jin, Z.; Zhang, Y.; Lu, J.; Li, H.; Yang, J. A Timestep-Adaptive-Diffusion-Model-Oriented Unsupervised Detection Method for Fabric Surface Defects. *Processes* **2023**, *11*, 2615. <https://doi.org/10.3390/pr11092615>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.