LFQUIAD: Lookup-Free Quantized autoencoder for few-shot Unsupervised Industrial Anomaly Detection via Synthetic Diffusion Inpainting

Shih-Chih Lin
International Intercollegiate Ph.D. Program
National Tsing Hua University
Hsinchu City, Taiwan (R.O.C.)
leolin65@gapp.nthu.edu.tw

Shang-Hong Lai
Department of Computer Science
National Tsing Hua University
Hsinchu City, Taiwan (R.O.C.)
lai@cs.nthu.edu.tw

Abstract

Unsupervised anomaly detection (UAD) is crucial in industrial and medical applications, offering scalable and cost-efficient alternatives to manual inspection by detecting abnormal patterns without requiring labeled anomalies. However, real-world anomalies are often scarce and ambiguous, limiting the effectiveness of conventional methods. We propose ***LFQUIAD****, a novel UAD framework that integrates a quantization-driven autoencoder with a modular* ***Anomaly Generation Module (AGM)****. AGM generates diverse and semantically meaningful synthetic anomalies using prompt-guided, diffusion-based inpainting, providing pixel-level supervision in few-shot scenarios. This enables robust model training without actual anomaly data. At the core of LFQUIAD lies* ***Lookup-Free Quantization (LFQ)****, a codebook-free representation learning method that discretizes features with high precision while improving generalization and robustness. Our method achieves state-of-the-art performance on MVTecAD and VisA benchmarks, excelling in anomaly detection and segmentation under limited data conditions. The plug-and-play nature of AGM also allows seamless integration into other detection pipelines, making LFQUIAD a practical and effective solution for real-world anomaly detection tasks. GitHub – LFQUIAD*

1. Introduction

Unsupervised anomaly detection (UAD) has become indispensable across various industrial and medical applications. By learning to identify patterns that deviate from normal data distributions, UAD enables scalable, cost-effective alternatives to manual inspection, eliminating the need for extensive annotations. However, acquiring real-world anomalous data remains a significant challenge in practice due to its inherent scarcity, ambiguity, and class imbalance. Datasets such as MVTecAD [2] and VisA [31] highlight these issues, with anomalies being vastly outnumbered by defect-free samples.

To overcome the limitations posed by the lack of diverse and labeled anomalies, recent research has explored synthetic anomaly generation as a promising alternative. By augmenting training data with generated anomalous images and Discriminative masks, models can be trained in a more robust and data-efficient manner. This work proposes a synthetic-guided anomaly detection framework that leverages recent advances in vision foundation models and generative modeling. Specifically, we employ Grounding DINO [17] and the Segment Anything Model (SAM) [12] for open-set object detection and Discriminative and integrate Stable Diffusion [21] to generate controllable and realistic anomalies with semantic consistency.

Beyond synthetic data generation, we introduce a novel generative-discriminative architecture designed to effectively localize anomalies with minimal supervision. At the core of our model lies a quantization-based representation learning module, termed **Lookup-Free Quantization (LFQ)** [26], which discretizes feature representations without relying on external codebooks. LFQ enables compact and precise encoding of image features, mitigating the "identical shortcut" problem commonly found in conventional autoencoders. We further incorporate a reconstruction-based discriminator, wherein a tailored autoencoder contrasts input and output to sharpen anomaly localization through residual discrepancies.

Our method is evaluated on standard benchmarks, including MVTecAD [2] and VisA [31], under few-shot settings. Experimental results demonstrate that our approach surpasses existing state-of-the-art methods in detection and Discriminative accuracy and maintains high robustness in low-data regimes. With strong performance and interpretability, our framework presents a viable solution for real-world industrial visual inspection tasks where labeled

anomalies are rare or unavailable. Our key contributions are summarized as follows:

- We introduce a novel synthetic anomaly generation pipeline that produces diverse and realistic anomalous samples with corresponding Discriminative masks, significantly enriching training data for unsupervised anomaly detection.
- We propose a quantization-driven autoencoding architecture based on *Lookup-Free Quantization* (LFQ) [26], which enables compact and expressive representation learning while supporting few-shot anomaly detection scenarios.
- We conduct extensive evaluations on MVTecAD [2] and VisA [31], where our method consistently outperforms state-of-the-art approaches in both anomaly detection and Discriminative, particularly under the few-shot regime.

2. Related Work

2.1. Anomaly Detection

Detecting anomalies in industrial images is critical for localizing defects and identifying abnormal samples [4, 15, 16]. Due to anomalies' inherent rarity and diversity, collecting and annotating real-world anomalous images remains challenging and labor-intensive. As a result, recent research has primarily focused on unsupervised anomaly detection and localization, where models are trained using only normal, defect-free samples.

Two predominant approaches have emerged in this domain: embedding-based and reconstruction-based methods. Embedding-based methods [3, 6, 9, 22] utilize feature representations extracted from pre-trained models to characterize normality. For instance, PatchCore [22] builds a coreset memory bank of normal features and measures deviations at inference. While effective, these methods often suffer from false positives in underrepresented or rare regions that were not sufficiently observed during training.

Reconstruction-based approaches [1, 14, 16, 27, 28] aim to recover the original image from corrupted inputs, based on the assumption that anomalies cannot be faithfully reconstructed. Techniques such as DRÆM [27] use synthetic defects during training to force the reconstruction of normal regions while simultaneously segmenting anomalies. However, this assumption does not always hold—some anomalies may be unintentionally reconstructed with high fidelity, diminishing detection accuracy.

Self-supervised learning-based methods attempt to bypass the need for labeled anomalous data by introducing proxy tasks. CutPaste [14] synthesizes anomalies by transplanting image patches but often suffers from discontinuities in appearance. To enhance visual realism, NSA [29] improves this by adopting Poisson image editing [18]. DRÆM [27] further integrates textures from DTD [5] to diversify synthetic anomalies and has shown strong performance. However, it struggles to generalize to structural anomalies, such as partially missing or misplaced components. Ultimately, the success of self-supervised anomaly detection hinges on how well the proxy task mimics real-world defects.

Despite its importance, anomaly synthesis remains an under-explored component in the anomaly detection literature. Most existing methods either generate limited variations or overlook semantic alignment, leading to insufficient diversity or overfitting.

To address these limitations, this work introduces a novel anomaly synthesis framework guided by vision foundation models. Specifically, Grounding DINO [17] is employed as an open-set object detector to identify candidate regions, which are then segmented using the Segment Anything Model (SAM) [12]. These regions are subsequently modified through Stable Diffusion [21], enabling controlled and context-aware anomaly injection. The proposed approach augments training data and enhances model generalization without relying on reconstruction-based assumptions by synthesizing a wide range of anomaly types with spatial precision and semantic diversity.

2.2. Vector Quantization

Sparse representation learning encodes signals using sparse coefficient vectors from a learned dictionary [13], offering benefits such as noise reduction, robustness, and effectiveness in tasks like denoising and super-resolution.

Vector quantization discretizes input features using a codebook and quantization strategy [20, 23, 25, 30], typically by minimizing the mean squared error (MSE) between input vectors and codebook entries. This is a form of discrete representation learning, where outputs often resemble one-hot vectors [26].

Vector Quantized Variational Autoencoder (VQ-VAE) [23, 24] compresses images into discrete latents using an encoder E, decoder D, and codebook C. The loss function is:

$$L = \|x - \hat{x}\|_2^2 + \|\text{sg}(h) - z\|_2^2 + \beta\|h - \text{sg}(z)\|_2^2, \quad (1)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. However, increasing codebook size or reconstruction fidelity does not always improve generation quality.

To address these limitations, we adopt **Lookup-Free Quantization (LFQ)** [26], which deterministically maps encoder outputs to discrete indices without explicit codebook lookup. LFQ enables scalable quantization with large vocabularies, aligning better with token-based modeling and reducing computation overhead.

In our framework, LFQ acts as a compact bottleneck in the autoencoder, enabling efficient and expressive latent to-

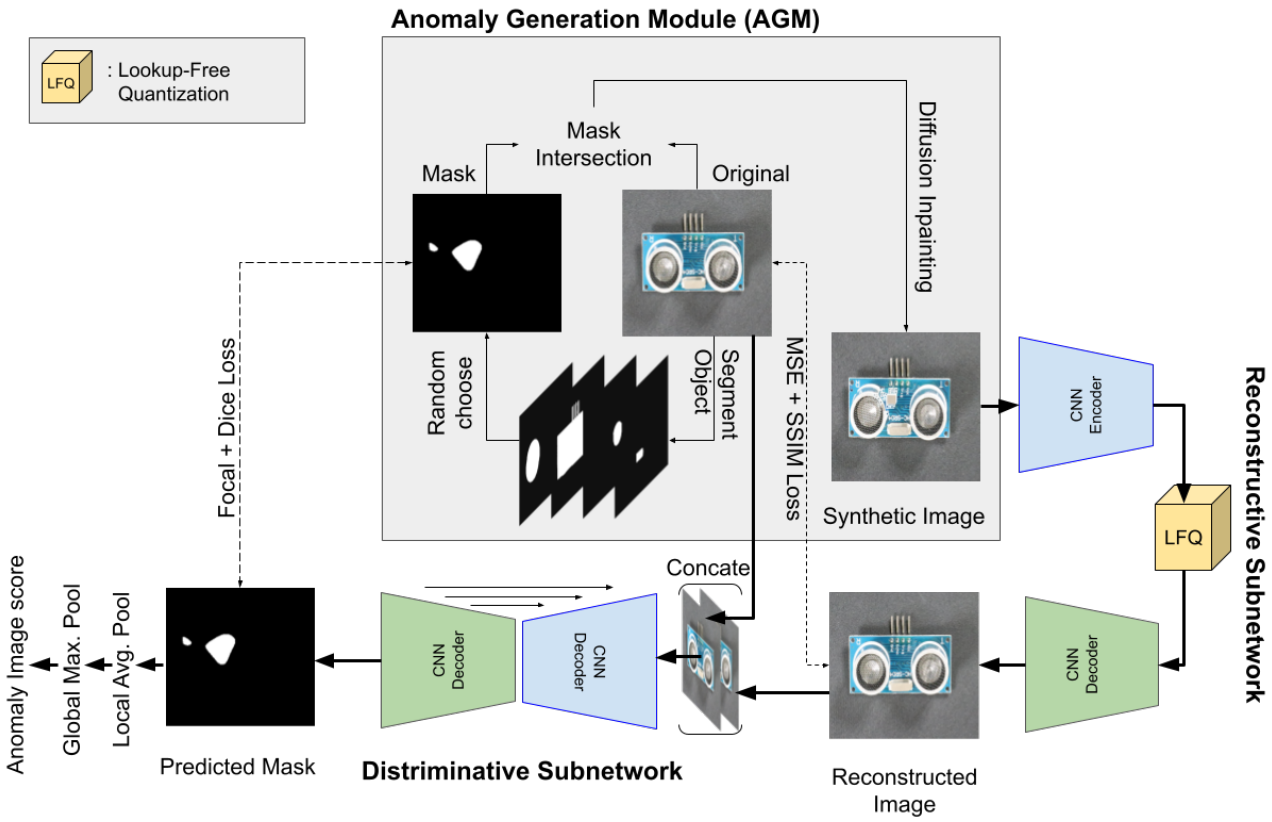

Figure 1. The overall architecture of our proposed framework: LFQUIAD.

kenization of normal patterns. By jointly training the encoder, decoder, and quantizer, LFQ learns meaningful representations without the need for external codebook optimization. Empirically, LFQ improves both reconstruction quality and downstream anomaly detection performance.

3. Proposed Method: LFQUIAD

3.1. Preliminaries

Segment Anything Model (SAM). SAM [12] is a large-scale image segmentation model trained on the SA-1B dataset, offering exceptional generalization capabilities across interactive segmentation tasks. Its promptable interface and versatility have made it a widely adopted foundation model in various visual applications.

Grounding DINO. Grounding DINO [17] integrates natural language descriptions into object detection, enabling open vocabulary and referring-object detection. By leveraging textual prompts, it effectively identifies and localizes arbitrary object categories.

Stable Diffusion. Stable Diffusion [21] is a text-to-image generative model that learns complex distributions of natural images. This work is repurposed to control the inpainting of anomalous regions during anomaly synthesis.

3.2. Synthetic Anomaly Generation Module

To overcome the scarcity of actual anomaly data and support end-to-end unsupervised training, we introduce a high-fidelity anomaly synthesis pipeline based on diffusion-based inpainting guided by prompt conditioning. This module generates diverse and semantically plausible pseudo-anomalies that closely mimic real-world defects while maintaining domain consistency.

Category-Aware Mask Generation. To synthesize meaningful anomalies, we propose a category-aware mask generation framework that adapts to the structural semantics of different image types. Given a normal image $\mathbf{N}$ from datasets such as MVTecAD [2] and VisA [31], our approach differentiates between *texture-centric* and *object-centric* categories.

For *texture-centric* categories (e.g., `carpet`, `grid`), which exhibit repetitive and homogeneous patterns, we simulate localized defects using Perlin noise [27]. This produces soft-edged, randomly shaped masks that mimic realistic irregularities. We apply dilation, Gaussian noise in-

jection, and alpha blending to diversify the synthetic patterns further and introduce local distortions and boundary fuzziness. Additionally, we support randomized inpainting guided by Perlin noise or text-conditioned diffusion, allowing the synthesis of visually diverse and semantically aligned irregular surface anomalies.

In contrast, for *object-centric* categories (e.g., `screw`, `transistor`), where semantic object structure is critical, we utilize the Grounded-SAM pipeline—a combination of Grounding DINO [17] and SAM [12]. Conditioned on prompts such as "a `[class_name]`", the model extracts semantically grounded foreground masks $\mathbf{M}$. For structurally complex objects, this process may yield multiple mask proposals. We randomly sample from these masks during training to guide the inpainting region, ensuring diverse and class-relevant anomaly synthesis.

Prompt-Guided Diffusion Inpainting. To synthesize realistic and semantically coherent anomalies, we adopt a prompt-conditioned inpainting strategy based on the Stable Diffusion Masked DiffEdit pipeline [21], which enables region-specific generation within masked areas. The previously generated masks serve as spatial constraints, guiding the inpainting process to focus exclusively on designated regions.

We employ dual-prompt CLIP-guided [19], conditioning to steer the generation semantics. Specifically, *positive prompts* (e.g., "damaged [class_name]," "scratched [class_name]") are used to encourage the synthesis of plausible visual defects aligned with real-world anomalies, while *negative prompts* (e.g., "perfect [class_name]", "clean surface") function as regularizers to suppress undesired artifacts and maintain fidelity outside the masked region.

During generation, the normal image is first encoded into the latent space using Stable Diffusion's variational autoencoder (VAE). The masked region is then iteratively refined through a denoising process conditioned on the prompt embeddings. Finally, the modified latent representation is decoded into a full-resolution image containing localized, semantically meaningful anomalies. This design ensures that the synthetic anomalies remain visually consistent with the object's category while introducing diversity for robust model training.

Compositional Blending. The final synthetic image $\mathbf{S}$ is composed with a pixel-wise blending scheme given by:

$$\mathbf{S} = \beta(\mathbf{M} \odot \mathbf{N}) + (1 - \beta)(\mathbf{M} \odot \mathbf{A}) + \overline{\mathbf{M}} \odot \mathbf{N},$$

where $\mathbf{M}$ is the binary anomaly mask, $\overline{\mathbf{M}}$ its complement, $\mathbf{A}$ the synthesized anomaly content, and β a blending coefficient. This formulation allows smooth transitions between anomalous and normal regions, increasing realism. $\odot$ denotes element-wise multiplication.

Reconstructive Subnetwork. The reconstructive subnetwork consists of a CNN-based encoder for visual feature extraction, followed by a pre-quantization convolutional layer to align the feature dimensions for Lookup-Free Quantization (LFQ) [26]. After quantization, another convolutional layer ensures compatibility with the decoder. The decoder aims to reconstruct a high-fidelity version of the original image.

Let $\mathbf{I}$ denote the original input image, and $\mathbf{I}_r$ denote the reconstructed image produced by the decoder. The reconstruction objective is to minimize both pixel-wise error and perceptual discrepancy. This is achieved through a combined loss function consisting of Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM).

The reconstruction loss $\mathcal{L}_{\text{rec}}$ is defined as:

$$\mathcal{L}_{\text{rec}} = \alpha_1 \mathcal{L}_{\text{mse}}(\mathbf{I}_r, \mathbf{I}) + \beta_1 \mathcal{L}_{\text{ssim}}(\mathbf{I}_r, \mathbf{I}), \tag{2}$$

where

- $\mathcal{L}_{\text{mse}}(\mathbf{I}_r, \mathbf{I})$ is the Mean Squared Error between $\mathbf{I}_r$ and ground truth $\mathbf{I}$.
- $\mathcal{L}_{\text{ssim}}(\mathbf{I}_r, \mathbf{I})$ is the Structural Similarity Index.
- α_1 and β_1 satisfy $\alpha_1 + \beta_1 = 1$.

3.3. Discriminative Subnetwork

The discriminative subnetwork incorporates a CNN-based autoencoder architecture for pixel-wise anomaly classification. It takes the same input image $\mathbf{I}$ and predicts a binary anomaly map. Let $\mathbf{M}$ denote the ground-truth binary mask, and $\hat{\mathbf{M}}$ be the predicted anomaly map, where each pixel value is either 0 (normal) or 1 (anomaly).

We use a hybrid loss combining Dice Loss and Focal Loss to improve localization accuracy and handle class imbalance. The discriminative loss $\mathcal{L}_{\text{dis}}$ is formulated as:

$$\mathcal{L}_{\text{dis}} = \alpha_2 \mathcal{L}_{\text{dice}}(\hat{\mathbf{M}}, \mathbf{M}) + \beta_2 \mathcal{L}_{\text{focal}}(\hat{\mathbf{M}}, \mathbf{M}), \tag{3}$$

where:

- $\mathcal{L}_{\text{dice}}$ encourages overlap between prediction and ground truth.
- $\mathcal{L}_{\text{focal}}$ addresses class imbalance.
- α_2 and β_2 satisfy $\alpha_2 + \beta_2 = 1$.

3.4. Model: LFQUIAD

This work introduces **LFQUIAD**, a reconstruction-based unsupervised anomaly detection (UAD) framework that synergistically combines a synthetic anomaly generation module with a Lookup-Free Quantized AutoEncoder. LFQUIAD learns quantized latent representations that are both compact and expressive, facilitating adequate distinction between normal and anomalous content.

Integrating LFQ into the latent space promotes discretized and structured feature encoding, enhancing generalization and robustness. In parallel, the synthetic anomaly

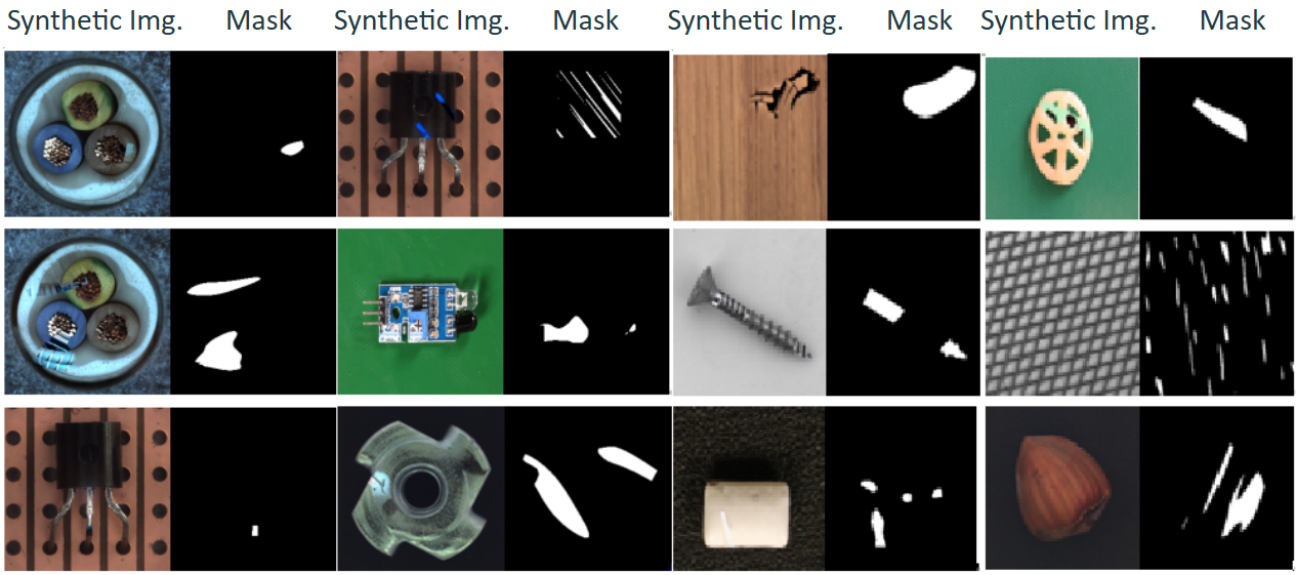

Figure 2. In our training data, we generated realistic synthetic anomalies using a diffusion-based [21] inpainting process, reducing artifacts common in prior methods like CutPaste [14]. This approach ensures coherent and diverse anomalies for robust model training.

generation module augments learning by injecting semantically consistent and spatially diverse defects, improving sensitivity to rare or ambiguous anomalies. Notably, our framework eliminates the need for real-world abnormal samples and requires only a small amount of normal training data. By leveraging high-quality synthetic anomalies, LFQUIAD achieves competitive performance even in few-shot settings, making it highly suitable for practical deployment in data-scarce scenarios.

3.5. Anomaly Scoring and Inference

During inference, LFQUIAD produces both a pixel-wise anomaly segmentation map and an image-level anomaly score. The segmentation output is directly obtained from the discriminative subnetwork's prediction $\hat{\mathbf{M}} \in [0,1]^{H\times W}$, which estimates the probability of each pixel being anomalous.

To derive a global anomaly score for image-level anomaly detection, we compute the mean of the predicted segmentation logits:

$$\mathcal{A}_{\text{img}} = \frac{1}{H \cdot W} \sum_{x=1}^{H} \sum_{y=1}^{W} \hat{\mathbf{M}}_{x,y}, \tag{4}$$

where a higher $\mathcal{A}_{\text{img}}$ indicates greater likelihood of the image containing anomalous regions. This Lookup-Free anomaly score is used to evaluate binary classification metrics such as AUROC at the image level.

Additionally, for fine-grained evaluation, we retain the dense anomaly score map $\hat{\mathbf{M}}$ for pixel-level segmentation analysis. To improve localization sharpness, the predicted score map is optionally refined using Gaussian smoothing post-processing. We also perform bilinear upsampling to restore $\hat{\mathbf{M}}$ to the original input resolution.

Thresholding. For both image-level and pixel-level predictions, a threshold τ is applied to convert soft scores into binary decisions:

$$\hat{\mathbf{M}}_{\text{bin}}(x,y) = \begin{cases} 1, & \text{if } \hat{\mathbf{M}}(x,y) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The threshold τ is selected via a validation set or determined heuristically using the maximum F1 score criterion. In practice, our model is robust across a wide range of τ due to the calibrated nature of synthetic supervision.

4. Experiment

We evaluate **LFQUIAD** under the few-shot anomaly detection (FSAD) setting, where only a small number of normal samples are used for training. To compensate for the absence of real anomalies, our synthetic anomaly generation module produces diverse and semantically coherent pseudo-defects. Experiments are conducted on the MVTecAD[2] and VisA[31] industrial benchmarks, with models trained solely on normal data and evaluated for anomaly detection and segmentation performance.

4.1. Dataset

MVTec AD [2] is a widely adopted benchmark for evaluating visual anomaly detection in industrial settings. It consists of 15 categories, including 10 object classes (e.g.,

bottle, screw) and 5 texture classes (e.g., carpet, grid), each with high-resolution images and pixel-level anomaly annotations. VisA [31] complements this by offering 12 object-centric categories with a broader range of real-world defect types, such as scratches, dents, and contaminations, making it well-suited for testing the generalization and robustness of anomaly detection models in more complex scenarios.

4.2. Evaluation Metrics

Following standard protocols [2], we report AUROC at both image-level (I) and pixel-level (P).

Image-level AUROC evaluates the model's ability to distinguish between normal and anomalous samples. At the same time, pixel-level AUROC assesses spatial localization by comparing the predicted anomaly map $\hat{\mathbf{M}}$ against the ground truth $\mathbf{M}$.

We also include the Per-Region Overlap (PRO) score [3, 7], which averages the region-wise overlap and is suited for evaluating fine-grained segmentation performance.

Together, these metrics (I, P, PRO) provide a comprehensive view of both detection and localization effectiveness.

4.3. Implementation Details

Synthetic Anomaly Generation. We utilize a pipeline comprising Grounding DINO, SAM, and Stable Diffusion to generate synthetic anomalies. We adopt the official implementation of Grounding DINO [17] and SAM [12], provided by Meta AI Research. For object-centric datasets, Grounding DINO is used to detect object bounding boxes conditioned on category prompts (e.g., "a screw"). Detected boxes are passed to SAM to generate high-resolution binary segmentation masks.

SAM is initialized with the pre-trained weights[1] while Grounding DINO uses the[2] checkpoint. Once segmentation masks are obtained, we perform text-guided inpainting using the Stable Diffusion masked diffusion pipeline via the Hugging Face Diffusers API[3].

To synthesize meaningful anomalies, we configure the positive inpainting prompt as:

```
"broken {class_name} with defect"
```

and its inverse (negative prompt) as:

```
"perfect {class_name}"
```

This dual-prompt setup ensures semantic consistency while encouraging localized deviations during inpainting.

[1] sam_vit_h_4b8939.pth
[2] groundingdino_swint_ogc.pth
[3] CompVis/stable-diffusion-v1-4

Training Settings. We train **LFQUIAD** using the AdamW optimizer with a learning rate of 1×10^{-4} for the discriminative subnetwork and 3×10^{-4} for the reconstructive quantized autoencoder. The model is trained for 800 epochs with a batch size of 1. A MultiStepLR scheduler is applied, with decay steps at 80% and 90% of the total training epochs. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24 GB VRAM. Input images are resized to 256×256, and training employs an early stopping mechanism based on validation AUROC, with a patience of 50 epochs.

Quantization Settings. For the Lookup-Free Quantized AutoEncoder, we adopt a codebook size of 2^{16}, an entropy regularization coefficient of 0.02, and a diversity scaling factor $\gamma = 1.0$. Quantization is applied after the encoder via straight-through estimation without codebook lookup. Entropy-aware auxiliary loss is incorporated to encourage high code utilization and reduce code collapse.

5. Experimental Results

We evaluate **LFQUIAD** under the few-shot anomaly detection (FSAD) setting, where only limited normal samples are available. Unlike prior methods that overfit to small datasets or rely on scarce real anomalies, LFQUIAD employs prompt-guided synthetic anomaly generation to produce diverse and semantically meaningful defects with pixel-level supervision—effectively mitigating data scarcity in FSAD.

Our synthesis pipeline, built upon foundation models (SAM, Grounding DINO, and Stable Diffusion), enables high-fidelity anomaly inpainting without requiring manual annotations, providing a rich self-supervised training signal.

Compared to recent unsupervised SoTA methods, LFQUIAD achieves competitive performance across most metrics, and notably surpasses prior approaches in key indicators such as pixel-level segmentation and localization. Ablation studies further highlight the contributions of synthetic supervision and Lookup-Free Quantization (LFQ) in enhancing anomaly representation.

5.1. Anomaly Detection and Localization on MVTecAD and VisA

Few-shot anomaly detection and localization: We compare our LFQUIAD with prior methods designed explicitly for few-shot settings. In Table 1, we illustrate the average experimental results for MVTecAD [2] and ViSA [31]. The experimental results of few-shot anomaly detection and segmentation on the MVTecAD [2] and VisA [31] datasets are summarized in Table 1. Across all shot settings (1, 2, and 4), our proposed LFQUIAD method consistently outperforms prior few-shot SOTA methods in image- and pixel-level AUROC.

		1-shot			2-shot			4-shot		
Dataset	Method	I	P	O	I	P	O	I	P	O
MVTecAD	RegAD [10]	-	-	-	85.7	94.6	-	88.2	95.8	-
	PatchCore [22]	83.4	92.0	79.7	86.3	93.3	82.3	88.8	94.3	84.3
	WinCLIP [11]	93.1	95.2	87.1	94.4	96.0	88.4	95.2	96.2	89.0
	AnomalyGPT [8]	94.1	95.3	-	95.5	95.6	-	96.3	96.2	-
	Ours	**94.9**	**95.7**	**90.4**	**97.9**	**96.9**	**93.1**	**97.8**	**98.2**	**93.5**
VisA	PatchCore	79.9	95.4	80.5	81.6	96.1	82.6	85.3	96.8	84.9
	WinCLIP	83.8	**96.4**	**85.1**	84.6	96.8	**86.2**	87.3	97.2	87.6
	AnomalyGPT	87.4	96.2	-	88.6	96.4	-	**96.6**	96.7	-
	Ours	**88.8**	**96.4**	82.7	**90.4**	**97.0**	85.4	93.3	**97.6**	**92.0**

Table 1. Performance comparison on MVTecAD [2] and VisA [31] under few-shot settings. "I", "P", and "O" refer to image-level AUROC, pixel-level AUROC, and PRO score, respectively. The best performance in each column is highlighted in bold.

On MVTecAD, LFQUIAD improves upon AnomalyGPT [8] in image-level AUROC by +0.8%, +2.4%, and +1.5% in the 1-shot, 2-shot, and 4-shot settings, respectively. Similarly, for pixel-level AUROC, our method achieves gains of +0.4%, +1.3%, and +2.0%, demonstrating enhanced spatial anomaly localization capabilities.

For the VisA dataset, LFQUIAD maintains competitive performance. It outperforms AnomalyGPT [8] in image-level AUROC by +1.4% and +1.8% in the 1-shot and 2-shot settings while showing a slight drop (-2.7%) in the 4-shot case. In pixel-level AUROC, LFQUIAD achieves marginal improvements of +0.2%, +0.6%, and +0.9% across the three shot scenarios.

These results highlight the effectiveness of our synthetic anomaly generation module and quantized representation learning, which together enable LFQUIAD to deliver robust performance under low-data regimes while achieving or exceeding state-of-the-art results in both detection and segmentation tasks.

5.2. Qualitative comparisons

Figure 3 showcases a qualitative comparison of anomaly localization results on representative samples from the MVTecAD and VisA datasets. From left to right, each row displays the original test image, its reconstructed output, the corresponding anomaly heatmap, and the pixel-level ground truth annotation.

Our proposed model, incorporating the Lookup-Free Quantization (LFQ) mechanism, demonstrates significantly improved localization precision and boundary sharpness compared to existing methods. By enforcing a discretized latent space through LFQ, the model imposes more substantial representational constraints on the autoencoder, making it inherently more difficult to reconstruct semantically inconsistent regions. This mitigates the common issue of over-adaptation in reconstruction-based models, where synthetic anomalies are often unintentionally reconstructed with high fidelity.

Table 2. Effect of our Anomaly Generation Module (AGM) on anomaly detection performance. AGM enhances both image-level (I) and pixel-level (P) AUROC across texture and object categories by generating semantically aligned, diffusion-based synthetic anomalies.

Model	Texture		Object	
	I	P	I	P
DRÆM	99.3	98.2	96.8	96.0
+Our AGM	**99.4**	**98.5**	**97.7**	**96.9**

The predicted heatmaps exhibit better spatial alignment with ground-truth anomaly regions, particularly in challenging cases with fine-grained textures or small object-level defects. These results qualitatively confirm the effectiveness of combining prompt-guided synthetic supervision and quantized representation learning, enhancing the model's sensitivity to subtle deviations while preserving structural consistency.

5.3. Ablation

Table 2 presents the performance improvements obtained by integrating our Anomaly Generation Module (AGM) into the DRÆM [27] framework under the full-data setting. The original DRÆM synthesizes irregular anomalies without incorporating semantic awareness, which often leads to the placement of anomalies in background regions—especially in object-centric images—this lack of contextual grounding results in suboptimal supervision signals during training. In contrast, our approach explicitly distinguishes between *texture-centric* and *object-*

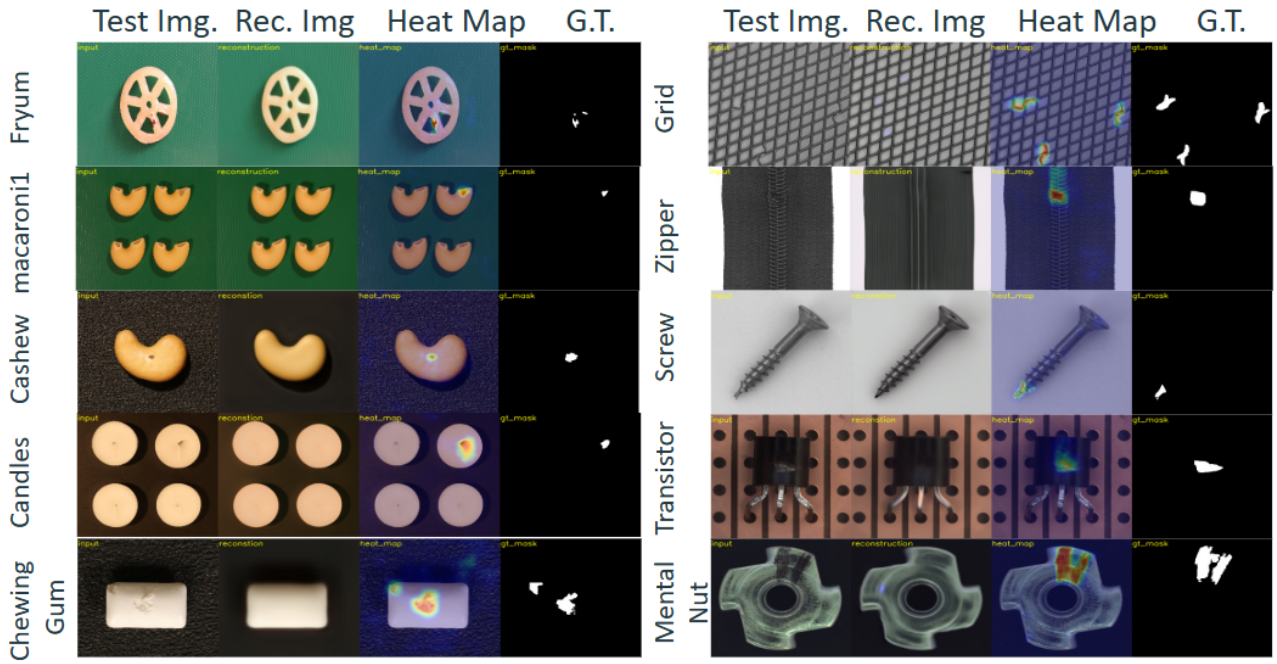

Figure 3. Our LFQUIAD model demonstrates superior reconstruction quality and anomaly localization accuracy. From left to right: testing data, reconstructed images, heatmap, and ground truth. Best viewed by zooming in.

centric categories and separates foreground from background. Leveraging segmentation masks generated by vision foundation models (e.g., SAM and Grounding DINO), our pipeline applies prompt-guided diffusion-based inpainting to inject anomalies into semantically meaningful regions. We use randomized inpainting guided by Perlin noise or text-conditioned diffusion for texture categories, simulating irregular surface defects. For object categories, we constrain the inpainting within foreground regions, resulting in visually realistic and structurally consistent anomalies.

Quantitatively, AGM improves object-category AUROC by **+0.9%** (I: 96.8 $\rightarrow$ 97.7, P: 96.0 $\rightarrow$ 96.9), and texture-category AUROC by **+0.1%** (I: 99.3 $\rightarrow$ 99.4) and **+0.3%** (P: 98.2 $\rightarrow$ 98.5), demonstrating its effectiveness in boosting robustness and localization accuracy under few-shot constraints.

6. Conclusion

In this paper, we proposed the Lookup-Free Quantized-based Unsupervised Industrial Anomaly Detection (LFQUIAD) method, demonstrating significant advancement in unsupervised anomaly detection. By leveraging innovative techniques such as Lookup-Free Quantization (LFQ) and our proposed Anomaly Generation Module (AGM), LFQUIAD achieves remarkable precision in anomaly localization and segmentation tasks.

The AGM is critical in bridging the data scarcity gap by generating diverse, semantically aligned, high-fidelity synthetic anomalies, particularly under few-shot settings. This enriched supervision significantly boosts the model's robustness and generalization. Moreover, AGM is designed as modular, allowing it to be easily integrated into a wide range of models that benefit from synthetic supervision. Its plug-and-play nature promotes flexibility and scalability, making it applicable to broader anomaly detection frameworks beyond LFQUIAD. Our experiments demonstrate superior results across various datasets, highlighting the efficacy of LFQUIAD in real-world industrial scenarios. Furthermore, our model maintains strong performance even with limited training data, outperforming existing methods. These findings underscore the potential of LFQUIAD—with its synergy of synthetic anomaly generation module and quantized encoding—to deliver cost-effective, accurate solutions for industrial anomaly detection.

Acknowledgements

This work was supported in part by the National Science and Technology Council, Taiwan, under grant NSTC 113-2634-F-007 002.

References

[1] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect seg-

mentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018. 2

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1, 2, 3, 5, 6, 7

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 2, 6

[4] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024. 2

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 2

[6] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pages 475–489. Springer, 2021. 2

[7] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, 2022. 6

[8] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024. 7

[9] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 98–107, 2022. 2

[10] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pages 303–319. Springer, 2022. 7

[11] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 7

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2, 3, 4, 6

[13] Adrian Łańcucki, Jan Chorowski, Guillaume Sanchez, Ricard Marxer, Nanxin Chen, Hans JGA Dolfing, Sameer Khurana, Tanel Alumäe, and Antoine Laurent. Robust training of vector quantized bottleneck models. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 2

[14] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 2, 5

[15] Shih-Chih Lin and Shang-Hong Lai. Squad: Scalar quantized representation learning for unsupervised anomaly detection and localization. In *Computer Vision – ECCV 2024, Part IV*. Springer, 2024. 2

[16] Shih-Chih Lin, Ho-Weng Lee, Yu-Hsuan Hsieh, Cheng-Yu Ho, and Shang-Hong Lai. Masked attention convnext unet with multi-synthesis dynamic weighting for anomaly detection and localization. In *Proceedings of the 34th British Machine Vision Conference*, 2023. 2

[17] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2, 3, 4, 6

[18] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, page 313–318, New York, NY, USA, 2003. Association for Computing Machinery. 2

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[20] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 4, 5

[22] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 2, 7

[23] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[24] Xin Wang, Shinji Takaki, Junichi Yamagishi, Simon King, and Keiichi Tokuda. A vector quantized variational autoencoder (vq-vae) autoregressive neural f_0 model for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:157–170, 2019. 2

[25] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 2

[26] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 1, 2, 4

[27] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 2, 3, 7

[28] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr – a dual subspace re-projection network for surface anomaly detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, page 539–554, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[29] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023. 2

[30] Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric vqgan for stablediffusion. *arXiv preprint arXiv:2306.04632*, 2023. 2

[31] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 1, 2, 3, 5, 6, 7