A Critical Evaluation of Diffusion Models: Exploring Their Limits in Open-World Anomaly Detection

Saksham Madan (Roll no. 24124040)
Kushal Shrivastava (Roll no. 24124025)
Arnav Raghuvanshi (Roll no. 24124007)

Submission date: November 21, 2025

Department of Mathematical Science
Indian Institute of Technology (BHU), Varanasi

Supervisor: Dr. Santwana Mukhopadhyay

Abstract

This report documents an exploratory project that critically evaluates the viability of latent diffusion models, in combination with vision–language embeddings such as CLIP, for open-world industrial anomaly detection on the MVTec AD benchmark. Motivated by the generative power of modern diffusion architectures and the semantic sensitivity of CLIP, we investigate whether this hybrid paradigm can perform identity-preserving reconstructions suitable for pixel–semantic anomaly scoring. The study systematically examines three reconstruction-based pipelines: a small UNet appended to Stable Diffusion latents, prompt-guided Stable Diffusion `img2img`, and LoRA-based fine-tuning aimed at texture specialization. For each approach, we detail the experimental setup, reconstruction behavior, hyperparameter dynamics, and failure modes. We also report practical challenges encountered on constrained compute environments (Kaggle GPU memory limits, instability of large diffusion checkpoints), which shaped the scope of feasible experimentation. Our analysis incorporates mathematical reasoning behind diffusion processes, latent VAE bottlenecks, CLIP contrastive embeddings, and LoRA low-rank adaptation to explain—beyond empirical evidence—why latent diffusion models inherently struggle with the strict geometric and high-frequency fidelity required for industrial anomaly detection. A key contribution of this work is the design of a hybrid anomaly scoring method that integrates pixel-level and semantic cues. Our scoring scheme combines (i) mean squared error between the input and reconstructed image and (ii) CLIP-based cosine similarity deviation to capture semantic/texture-level changes.This composite score enables a richer understanding of anomalies by measuring both low-level reconstruction errors and high-level feature shifts. Despite these methodological advances, our analysis—supported by mathematical reasoning on latent bottlenecks, diffusion denoising objectives, CLIP embeddings, and LoRA low-rank adaptation—shows that latent diffusion models inherently struggle with the strict geometric and fine-texture fidelity required for reliable industrial anomaly detection.

Contents

1 Introduction **2**
- 1.1 Background . . . 2
- 1.2 Problem Statement . . . 2
- 1.3 Aims and Objectives . . . 3
- 1.4 Scope of the Study . . . 3

2 Literature Review **4**
- 2.1 Diffusion models . . . 4
- 2.2 Vision–language models (CLIP) . . . 5

3 Methodology **6**
- 3.1 Overview of Experimental Pipeline . . . 6
- 3.2 Diffusion: Forward and Reverse Processes . . . 6
  - 3.2.1 Forward Diffusion . . . 6
  - 3.2.2 Reverse Denoising . . . 7
- 3.3 CLIP: Contrastive Loss and Semantic Similarity . . . 7
- 3.4 LoRA (Low-Rank Adaptation) . . . 8
- 3.5 Hybrid Anomaly Scoring . . . 8
- 3.6 Dataset(s) Used . . . 8
- 3.7 Research Design . . . 8
- 3.8 Implementation Details . . . 9
- 3.9 Experimental Setup . . . 9
- 3.10 Evaluation Metrics . . . 9

4 Results and Analysis **10**
- 4.1 Summary of experiments . . . 10
- 4.2 Quantitative metrics . . . 11
- 4.3 Why Diffusion-Based Reconstruction Fails . . . 11
  - 4.3.1 Information Bottleneck in Latent Diffusion . . . 11
  - 4.3.2 Diffusion Denoising Is Generative, Not Invertible . . . 11
  - 4.3.3 Effect on Anomalies . . . 11
  - 4.3.4 Contrast With Autoencoders . . . 12
- 4.4 Practical issues encountered . . . 13

5 Conclusion and Future Work **14**
- 5.1 Conclusion . . . 14
- 5.2 Future work . . . 14

A Files produced during the project **15**

Chapter 1

Introduction

1.1 Background

Anomaly detection plays a crucial role in industrial inspection systems, where surface defects such as scratches, dents, stains, holes, and texture inconsistencies must be identified with high precision. Traditional reconstruction-based approaches leverage autoencoders, GANs, or flow-based models to learn the distribution of normal samples and highlight deviations as anomalies. In recent years, diffusion models have emerged as powerful generative models capable of producing photorealistic images by iteratively denoising latent representations. Their success across image synthesis, editing, and inpainting has sparked interest in evaluating whether these models can also perform identity-preserving reconstruction. In parallel, vision–language models like CLIP have demonstrated strong semantic understanding, enabling robust similarity measures beyond pixel-level comparisons. The convergence of these two paradigms—generative diffusion and semantic embedding models—motivates an exploratory evaluation of their combined potential for anomaly detection.

(a) Good (b) Color (c) Hole (d) Metal (e) Thread

Figure 1.1: Examples of normal and anomalous samples in MVTec AD (carpet).

1.2 Problem Statement

Open-world anomaly detection introduces the challenge that anomalies during inference may not resemble any anomalies seen (or expected) during training. Industrial datasets such as MVTec AD are particularly difficult because anomalies are often subtle, localized, and embedded within highly repetitive textures. Reconstruction-based anomaly detection relies on the assumption that a model trained on normal samples can accurately regenerate a defect-free version of an input. However, for this assumption to hold, the reconstruction

mechanism must preserve high-frequency spatial details while selectively removing irregularities. This project investigates whether modern *latent diffusion models*, when combined with CLIP-based semantic scoring, satisfy this requirement—or whether architectural and mathematical constraints fundamentally limit their ability to perform identity-preserving restorations.

1.3 Aims and Objectives

The primary aim of this exploratory project is to critically evaluate the use of diffusion-based reconstruction for open-world industrial anomaly detection. Specifically, the study seeks to:

- Examine whether diffusion models (e.g., Stable Diffusion) can reconstruct defect-free images while preserving the original structure and texture.
- Explore the integration of vision–language models (CLIP) into the anomaly-scoring pipeline.
- Test three major reconstruction approaches: a small UNet extension, Stable Diffusion `img2img`, and LoRA-based fine-tuning.
- Analyze reconstruction behavior, hyperparameter sensitivity, and practical implementation constraints.
- Provide a mathematical explanation for the observed limitations of latent diffusion in anomaly detection.
- Compare diffusion-based reconstruction with alternative feature-based baselines.

1.4 Scope of the Study

This project is exploratory in nature and focuses specifically on:

- Unsupervised, reconstruction-driven anomaly detection on the MVTec AD dataset.
- Latent diffusion models (Stable Diffusion family), not pixel-space diffusion.
- CLIP-based semantic scoring, combined with pixel and perceptual losses.
- Qualitative and quantitative evaluation (AUROC, heatmaps) of reconstruction quality.

The study does *not* attempt large-scale retraining of diffusion models from scratch due to computational constraints. It also does not cover advanced industrial anomaly detection methods such as PatchCore or DRAEM in full detail, although they are referenced as stronger baselines for future work.

Chapter 2

Literature Review

2.1 Diffusion models

High-Resolution Image Synthesis with Latent Diffusion Models (Rombach et al.) introduced a major shift in how diffusion models are used for large-scale image generation. Earlier diffusion methods such as DDPM and DDIM operate directly in pixel space, which makes them computationally expensive and slow, especially for high-resolution images. Latent Diffusion simplifies this process by first encoding an image into a much smaller latent representation using a VAE, and then applying the diffusion process in this compressed space. This design dramatically reduces memory footprint and inference time while still retaining the ability to generate visually appealing and high-quality images. The model no longer needs to denoise millions of pixels directly; instead, it works on a compact latent tensor that captures the overall structure and semantic content of the image. Because of this efficiency, latent diffusion models such as Stable Diffusion quickly became popular for tasks like image synthesis, editing, style transfer, and controlled generation using text prompts. However, the key trade-off of this approach is that the VAE compression inherently removes or smooths out fine spatial details. High-frequency textures, subtle surface patterns, and small irregularities are partially lost when images are encoded into the latent space.

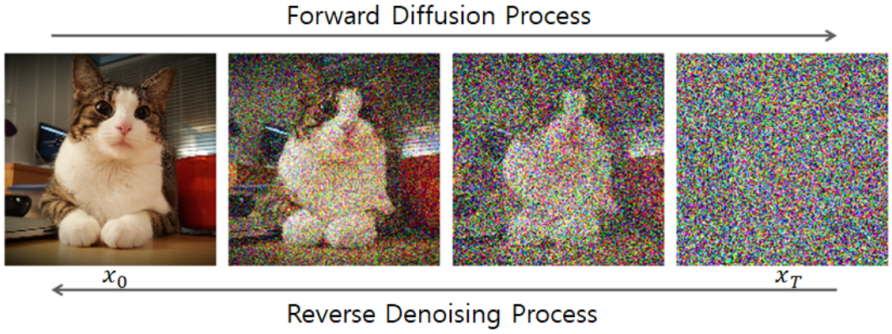

Figure 2.1: Schematic overview of forward noise addition and reverse denoising in diffusion models.

2.2 Vision–language models (CLIP)

CLIP: Learning Transferable Visual Models From Natural Language Supervision (Radford, A., et al.) introduces a powerful framework where image and text encoders are jointly trained using a large-scale contrastive objective. Instead of relying on explicit labels, CLIP learns by matching images with their corresponding textual descriptions across diverse internet-scale data. This training strategy results in a shared embedding space where semantically similar images and texts lie close together, allowing the model to capture both global structure and fine-grained conceptual information.

Because CLIP is sensitive to high-level semantics as well as local visual features, it serves as an effective tool for evaluating differences between two images that may not be apparent from pixel-level metrics alone. This makes it especially valuable in tasks involving subtle texture changes or conceptual deviations, where traditional reconstruction errors (such as MSE) may fail to capture perceptually meaningful differences.

In our work, we leverage this capability by integrating CLIP into a hybrid anomaly scoring method. The score combines pixel-level reconstruction error (MSE) with a semantic distance term computed as:

$$S_{\text{CLIP}} = 1 - \cos\Big(\text{CLIP}(x),\, \text{CLIP}(\hat{x})\Big),$$

where x is the input image and $\hat{x}$ is the reconstructed image.

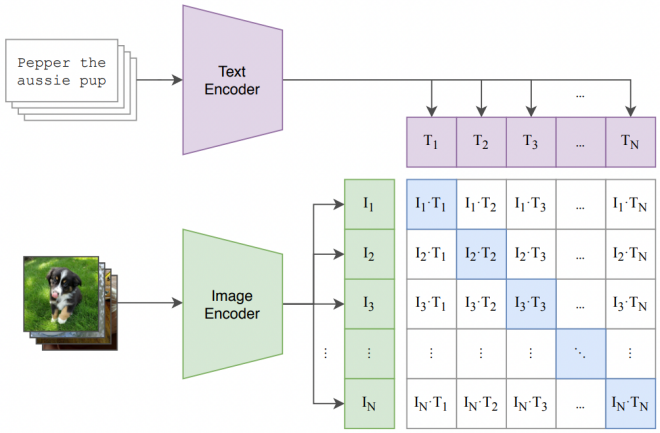

Figure 2.2: CLIP training objective: aligning image and text embeddings.

Chapter 3

Methodology

3.1 Overview of Experimental Pipeline

The methodology explored three major reconstruction-based approaches for anomaly detection using diffusion and vision–language models:

1. **Small UNet on Stable Diffusion latents:** A lightweight UNet was appended after Stable Diffusion's latent representation to attempt refinement and improve texture recovery.
2. **Stable Diffusion `img2img`:** Prompt-guided reconstruction using text descriptions such as "a flawless {category}" and negative prompts for defects. This approach attempted identity-preserving denoising while leveraging generative priors.
3. **LoRA fine-tuning:** Low-rank adaptation was applied to tailor the diffusion model to clean texture distributions using a small number of high-quality normal samples.

For each approach, a hybrid anomaly scoring framework was employed, using both pixel-level (MSE) and semantic (CLIP-based) deviations. Visual heatmaps were generated to interpret the spatial distribution of anomalies.

3.2 Diffusion: Forward and Reverse Processes

3.2.1 Forward Diffusion

The forward diffusion (noise-adding) process is a Markov chain transforming a clean sample $\mathbf{x}_0$ into a noisy latent $\mathbf{x}_T$:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\Big(\mathbf{x}_t;\ \sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\ \beta_t\mathbf{I}\Big)\,,$$

with variance schedule $\{\beta_t\}_{t=1}^{T}$. A closed-form expression allows sampling directly from $\mathbf{x}_0$:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

3.2.2 Reverse Denoising

The reverse process learns to denoise $\mathbf{x}_t$:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\Big[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\Big].$$

Latent diffusion replaces $\mathbf{x}$ with VAE latents $\mathbf{z} = \text{Enc}(\mathbf{x})$ to improve training efficiency at the cost of spatial precision.

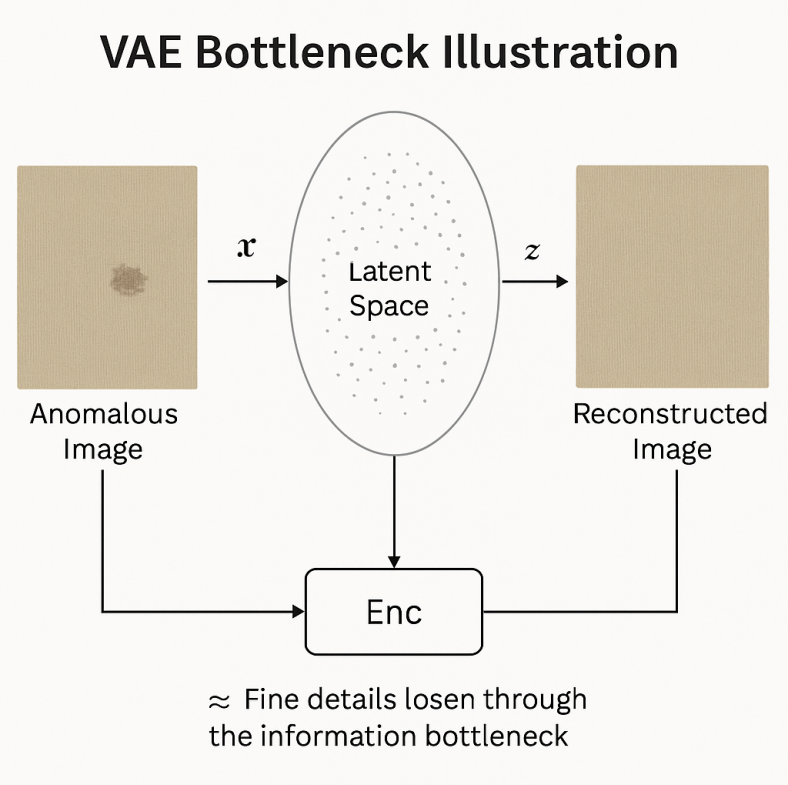

Figure 3.1: Latent diffusion architecture showing VAE encoder bottleneck before denoising UNet.

3.3 CLIP: Contrastive Loss and Semantic Similarity

CLIP jointly trains an image encoder $f(x)$ and a text encoder $g(t)$ using a contrastive objective:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(f(x_i), g(t_i))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(f(x_i), g(t_j))/\tau)},$$

where $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$ and τ is a learned temperature. In our pipeline CLIP provides a semantic deviation signal:

$$S_{\text{CLIP}}(x, \hat{x}) = 1 - \text{sim}(f(x), f(\hat{x})).$$

This captures changes in texture, structure, and semantics beyond raw pixel differences.

3.4 LoRA (Low-Rank Adaptation)

LoRA modifies a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ by learning a low-rank update:

$$W = W_0 + \Delta W, \qquad \Delta W = BA,$$

with $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ and $r \ll k, d$. Only A and B are trained, keeping W_0 frozen:

$$\min_{A,B} \ \mathbb{E}_{x\sim\mathcal{D}}\, \mathcal{L}_{\text{task}}(W_0 + BA; \theta_{\text{other}}).$$

LoRA excels at learning stylistic or conceptual adjustments, but it **cannot modify** the latent VAE bottleneck, meaning fine industrial textures lost during encoding cannot be recovered—even with fine-tuning. Our experiments confirmed such distortions.

3.5 Hybrid Anomaly Scoring

The final anomaly score combines pixel, semantic and optionally perceptual metrics:

$$S_{\text{hybrid}}(x) = \alpha \cdot \text{MSE}(x, \hat{x}) \ + \ (1-\alpha) \cdot \Big(1 - \text{sim}_{\text{CLIP}}(x, \hat{x})\Big).$$

This hybrid design compensates for shortcomings of pixel differences on textured surfaces and enhances CLIP's semantic sensitivity. This scoring method is a key contribution of the project.

3.6 Dataset(s) Used

The study primarily uses the **MVTec Anomaly Detection (MVTec-AD)** dataset, a widely adopted benchmark for industrial anomaly detection. It covers 15 object and texture categories such as *carpet*, *tile*, *wood*, *bottle*, each containing:

- multiple high-quality **normal (defect-free)** training samples,
- diverse **anomalous** test samples with defects such as stains, scratches, holes, misaligned textures.

This dataset was selected due to:

- its industrial relevance,
- high-resolution texture patterns (challenging for latent diffusion),
- compatibility with reconstruction-based anomaly detection evaluation.

3.7 Research Design

The overall research design is a **comparative exploratory study** involving:

1. Analyzing multiple reconstruction pipelines using diffusion models.
2. Comparing reconstruction fidelity against identity-preservation requirements.
3. Evaluating anomaly localization through hybrid scoring and heatmaps.
4. Interpreting failure modes using mathematical and architectural reasoning.

As a baseline, a CLIP patch-based feature model was implemented for comparison.

3.8 Implementation Details

Software and Frameworks

- Python, PyTorch
- HuggingFace Diffusers (Stable Diffusion, DDIM scheduler)
- OpenAI CLIP
- Kaggle GPU environment for training/inference

Model Architecture Summary

- Stable Diffusion v1.5 latent UNet + VAE encoder/decoder.
- Small UNet refinement module (experimentally attached).
- LoRA modules injected into diffusion attention layers.
- CLIP ViT-B/32 for semantic scoring.

3.9 Experimental Setup

Diffusion Hyperparameters

- **Strength**: 0.1–0.7 Controls deviation from input image during `img2img`.
- **Guidance scale**: 3.5–9 Controls adherence to textual prompt.
- **Inference steps**: 20–50 Higher steps improve realism but not necessarily identity preservation.

Scoring Hyperparameters

- α: 0.3–0.9 Relative weighting of pixel vs semantic differences.

3.10 Evaluation Metrics

The following metrics were used:

- **AUROC (Area Under ROC Curve)** Standard metric for binary anomaly detection (normal vs anomalous).
- **Pixel-wise heatmaps** Visual inspection of anomaly localization quality.
- **Reconstruction quality indicators** MSE, SSIM.
- **CLIP cosine similarity** Semantic deviation measure.

Chapter 4

Results and Analysis

4.1 Summary of experiments

We summarize the high-level outcome of each pipeline attempt:

- **Small UNet on SD latents:** produced blurry reconstructions; model could not recover high-frequency textures lost by the original latent encoder.
- **Stable Diffusion `img2img`:** at low strength the reconstruction preserved the input (defects remain); at high strength defects were removed but texture identity changed substantially (new pattern/hallucination). Intermediate values did not yield a reliable tradeoff.
- **LoRA fine-tuning:** reduced some defects but introduced color/tone drift and hallucinated pattern shifts (overfitting to a slightly different style), consistent with the theoretical limitation that LoRA cannot modify the encoder/decoder bottleneck.

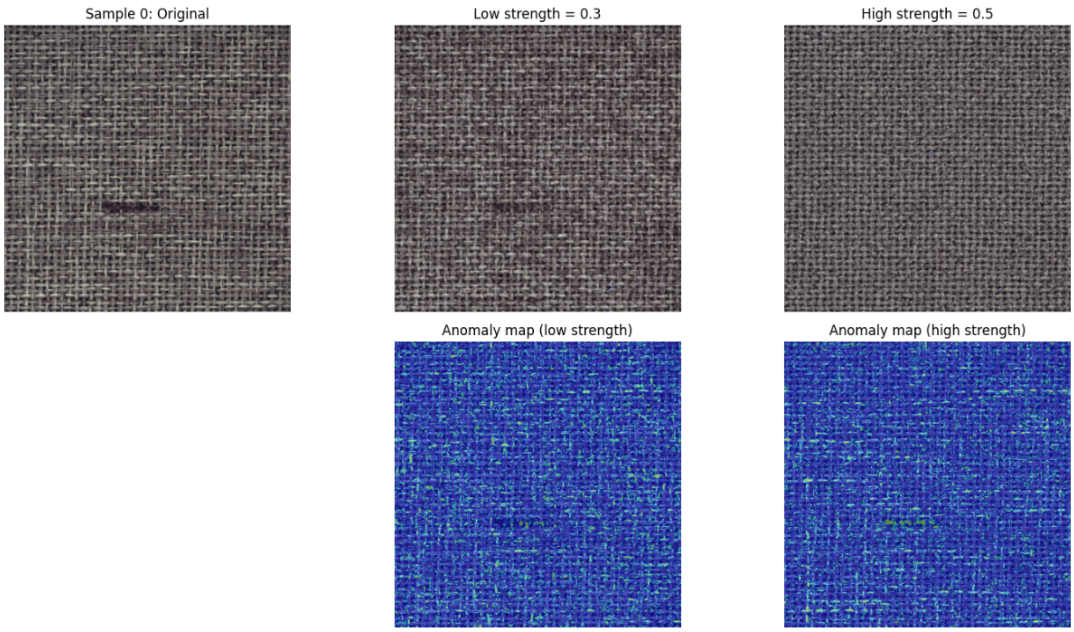

Figure 4.1: Effect of diffusion strength: low strength preserves defects, high strength removes both defects and original texture.

4.2 Quantitative metrics

We evaluated AUROC over MVTec categories using the hybrid score. While AUROC values sometimes looked acceptable, qualitative inspection showed that good AUROC came from the model changing texture globally, which increased pixel and semantic distances in a way that still correlated with defect presence. Thus numeric AUROC can be misleading when reconstructions are not identity-preserving.

4.3 Why Diffusion-Based Reconstruction Fails

This section presents a simplified mathematical and intuitive explanation of the fundamental limitations that prevent latent diffusion models from performing reliable reconstruction-based anomaly detection in open-world industrial settings.

4.3.1 Information Bottleneck in Latent Diffusion

Stable Diffusion relies on a VAE encoder $\text{Enc}(x)$ that compresses an image x into a much smaller latent z. This mapping is inherently lossy:

$$x \approx \text{Dec}(\text{Enc}(x)),$$

and the reconstruction error contains the high-frequency textures and fine surface details that are crucial for detecting small anomalies. Because anomalies are typically subtle and localized, they often fall below the resolution capacity of the latent space and are discarded during encoding.

Intuition: The encoder keeps only the "broad texture" of a surface. Small scratches, stains, holes, or fiber irregularities get smoothed out before diffusion even begins.

4.3.2 Diffusion Denoising Is Generative, Not Invertible

The diffusion reverse process does not invert the input image. It samples from a distribution of plausible clean images:

$$\hat{x} \sim p_\theta(x_0 \mid z_{\text{init}}),$$

meaning that the output is a realistic example of the category (e.g., "a carpet"), not the same carpet with its defects removed. Small changes in noise or guidance cause different reconstructions, showing that the mapping is one-to-many and not identity preserving.

Intuition: Diffusion models answer: "What does a typical defect-free carpet look like?" But anomaly detection needs: "What does **this exact carpet** look like without defects?"

4.3.3 Effect on Anomalies

Let an anomalous image be $x_{\text{anom}} = x_{\text{norm}} + \delta$, where δ is a tiny defect. After VAE compression:

$$\text{Enc}(x_{\text{anom}}) \approx \text{Enc}(x_{\text{norm}}).$$

Thus the diffusion model receives almost identical latents for normal and anomalous samples and generates similar outputs. As a result:

- At low noise strength, defects are preserved.

- At high noise strength, the model removes the defect *and* the original texture, producing a new artificial pattern.

Intuition: The anomaly signal is smaller than the model's reconstruction error. Diffusion either keeps the defect or erases the whole identity along with it.

4.3.4 Contrast With Autoencoders

Autoencoders used in anomaly detection explicitly optimize:

$$\|x - \mathrm{AE}(x)\|^2 \quad \text{(identity-preserving reconstruction)}.$$

Diffusion models instead optimize a generative denoising objective. Their residual error consists of VAE loss plus stochastic sampling variance, which is often much larger than the anomaly itself. Therefore the difference $x - \hat{x}$ is dominated by generative drift rather than true defect structure.

Short Intuition: Why Diffusion Fails for Open-World Anomaly Detection

Open-world anomaly detection requires:

- retaining exact object identity,
- preserving fine textures,
- adapting to unseen, unpredictable anomalies.

Latent diffusion provides:

- lossy encoding that removes tiny defects,
- stochastic sampling that changes textures,
- generation of typical examples rather than faithful reconstructions.

Thus, diffusion models are excellent texture generators but fundamentally misaligned with identity-dependent anomaly detection.

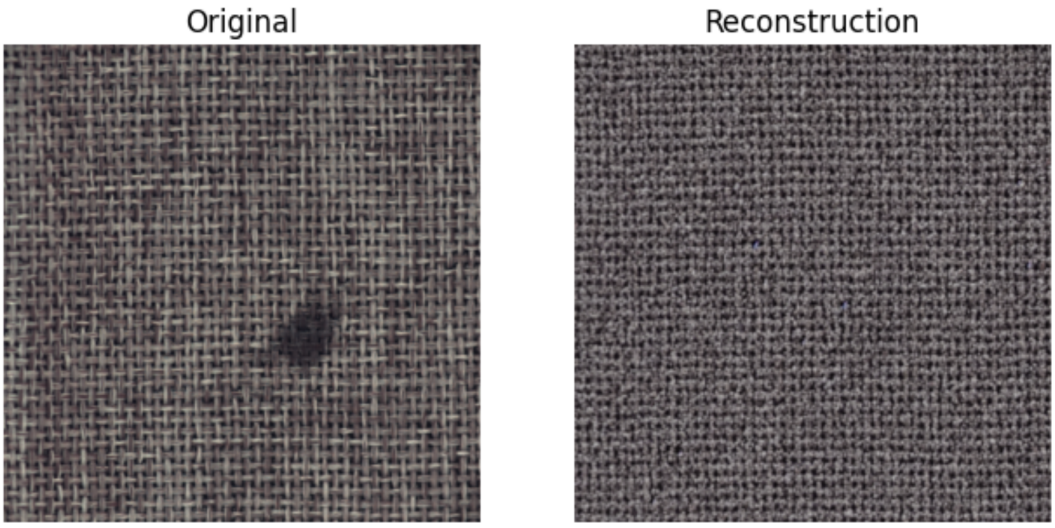

Figure 4.2: Example showing identity drift: input image vs SD reconstruction.

4.4 Practical issues encountered

We documented engineering challenges during experiments:

- GPU session instability and VRAM OOMs on Kaggle when running large SD models.
- Attention-slicing and fp16 help but do not eliminate memory constraints.
- DDIM/latent inversion is approximate; exact inversion would require model changes.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Our exploratory study provides a clear, evidence-backed conclusion:

Latent diffusion models (Stable Diffusion family) are not well-suited for reconstruction-based industrial anomaly detection that requires identity-preserving, pixel-accurate restorations.

The main reasons are the VAE compression bottleneck, stochastic generative sampling, and the model's objective mismatch: diffusion models optimize for visual realism and sample diversity rather than deterministic inversion.

5.2 Future work

Possible directions if additional time/resources become available:

1. Implement and benchmark PatchCore / WinCLIP baselines on the same train/test splits used here.
2. Explore pixel-space DDPMs trained specifically with a reconstruction objective on industrial textures (expensive but potentially informative).
3. Develop hybrid pipelines that combine CLIP patch embeddings with discriminative models for better localization.
4. Explore invertible architectures (normalizing flows) tailored to high-frequency textures.

Appendix A

Files produced during the project

- Project report.
- Slide deck summarizing pipelines, hyperparameters and visuals.
- Experiment code and Kaggle-friendly scripts.
- Everything is uploaded on GitHub: github.com/s4kr3d-w0r1d/exploratory_project