Peeling Back the Pixels: Stabilizing Latent Diffusion for Data-Efficient Generation of Pineapple Fields

Manfred Gonzalez Hernandez*
gonzalezhernandez.manfred@gmail.com
Faculty of Engineering and Architecture, Ghent University - imec
Ghent, Belgium

Danny Xie-Li
School of Computing, Costa Rica Institute of Technology
Cartago, Costa Rica

Fabian Fallas Moya
Atlantic Branch, University of Costa Rica
Cartago, Costa Rica

Sebastian Rojas Gonzalez
Faculty of Engineering and Architecture, Ghent University - imec
Ghent, Belgium

Ivo Couckuyt
Faculty of Engineering and Architecture, Ghent University - imec
Ghent, Belgium

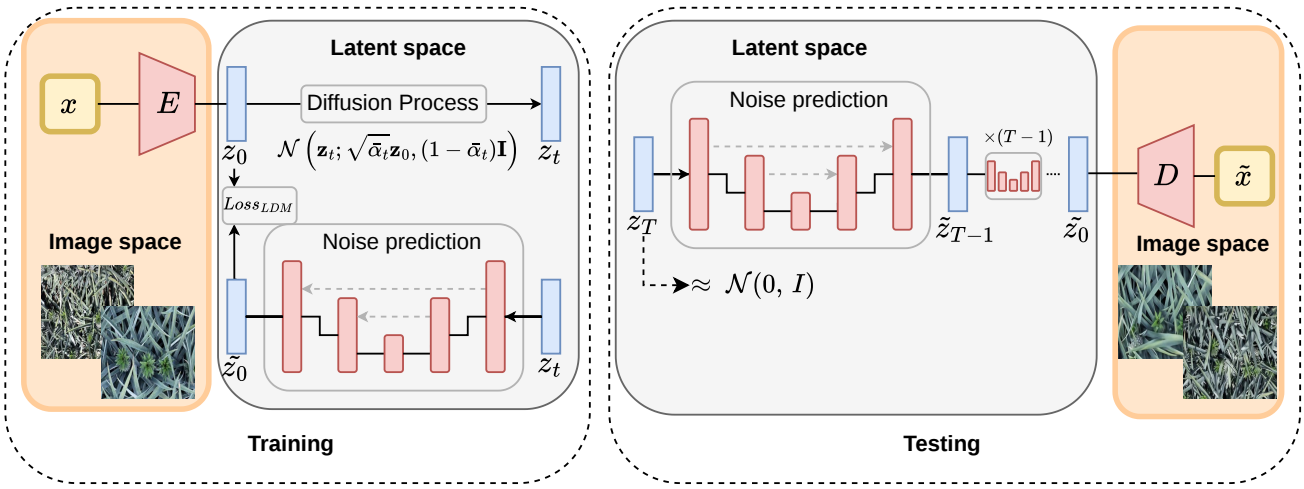


Figure 1: Overview of the Latent Diffusion Framework for synthetic aerial image generation.

Abstract

While Latent Diffusion Models (LDMs) excel at synthetic data generation, they suffer from severe training instability on the small, homogeneous UAV datasets typical of precision agriculture. In this work, we investigate these architectural bottlenecks by training a Denoising Diffusion Probabilistic Model (DDPM) from scratch on constrained pineapple plantation imagery. Our systematic comparison reveals that continuous latent spaces (β-VAE) preserve complex crop structures, whereas discrete architectures (VQ-VAE) fail under scarce data conditions. Additionally, we demonstrate that replacing destabilizing global attention mechanisms with a purely convolutional U-Net backbone ensures smooth convergence in repetitive visual domains. Ultimately, combining this robust architecture with DDIM sampling yields a stable, highly efficient framework for low-data agricultural image synthesis. Code and models will be available at https://github.com/imagine-laboratory/Latent-Diffusion.

CCS Concepts

• **Software and its engineering**; • **Applied computing** → **Agriculture**;

Keywords

Deep Learning, Precision Agriculture, Variational Autoencoder, latent space, Image Generation

This work was supported by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen.

This work is licensed under a Creative Commons Attribution 4.0 International License.
GECCO Companion '26, San Jose, Costa Rica
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2488-6/2026/07
https://doi.org/10.1145/3795101.3814709

ACM Reference Format:
Manfred Gonzalez Hernandez, Danny Xie-Li, Fabian Fallas Moya, Sebastian Rojas Gonzalez, and Ivo Couckuyt. 2026. Peeling Back the Pixels: Stabilizing Latent Diffusion for Data-Efficient Generation of Pineapple Fields. In *Genetic and Evolutionary Computation Conference (GECCO Companion '26), July 13–17, 2026, San Jose, Costa Rica.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3795101.3814709

1 Introduction

At scale, a commercial pineapple plantation is a massive, visually overwhelming expanse of dense and overlapping foliage. Monitoring these millions of sharp-edged plants for microscopic signs of disease or nutritional deficiency requires extreme precision yet remains heavily reliant on slow, manual labor [11]. While Costa Rica and other countries are global leaders in the fresh pineapple export industry [4, 6], the industry's push toward automated crop monitoring via Unmanned Aerial Vehicles (UAVs) and deep learning is bottlenecked by the sheer visual complexity of the crop. Obtaining large-scale, high-resolution aerial datasets that adequately capture the variability of illumination, plant density, and background soil is logistically challenging and very expensive [2].

To address this severe data scarcity, generative artificial intelligence has emerged as a pathway to produce synthetic agricultural imagery. While Generative Adversarial Networks (GANs) have historically been used for data augmentation [12], diffusion-based models — particularly Latent Diffusion Models (LDMs) [13] — recently established a new state-of-the-art. LDMs offer much higher visual fidelity and finer control over the generation process, making them a highly promising alternative for synthesizing realistic UAV imagery to support downstream agronomic analysis [10].

However, there is a fundamental mismatch: most diffusion models are inherently designed for massive, highly diverse datasets, while regional agricultural applications are characterized by small, homogeneous data. This data-constrained regime becomes a critical bottleneck. When working with limited UAV data of homogeneous pineapple fields, we observed that standard diffusion training becomes highly unstable, and the generated images tend to lose essential fine-grained details. A key factor in this failure is the initial latent representation. While continuous Variational Autoencoders (VAEs) [7] provide a compact latent space, they tend to produce overly smooth reconstructions where high-frequency details — such as leaf serrations and plant boundaries — are permanently lost [13]. Conversely, discrete architectures like Vector Quantized VAEs (VQ-VAEs) [16] theoretically preserve sharper textures, but under scarce data conditions, they are highly susceptible to codebook collapse, i.e., a phenomenon where the model only utilizes a small fraction of its discrete embeddings, leading to severe representation bottlenecks [18]. Because any structural information lost during encoding cannot be magically recovered during diffusion generation, systematically comparing these latent settings is essential.

To overcome these specific architectural bottlenecks, this work investigates how to stabilize latent diffusion models trained entirely in low data regimes. Rather than fine-tuning massive, opaque pre-trained models, which often mask the root causes of instability in niche domains, we construct a controlled Denoising Diffusion Probabilistic Model (DDPM) pipeline from scratch [8]. Within this framework, we directly compare the generative performance of continuous (β-VAE) [7] and discrete (VQ-VAE) [16] latent spaces. This controlled setup enables us to isolate how fundamental design choices, such as latent quantization and U-Net attention mechanisms, impact training convergence and the preservation of fine-grained agricultural features. We evaluate our proposed approach using a proprietary dataset of high-resolution UAV imagery from Upala Agrícola, the world's largest exporter of fresh pineapple, located in Costa Rica, providing a rigorous benchmark for synthetic image generation in real-world precision agriculture.

The main contributions are as follows:

- We implement a controlled latent diffusion pipeline from scratch to evaluate generative performance under agricultural data constraints.
- We compare continuous (β-VAE) and discrete (VQ-VAE) latent spaces, revealing how they impact downstream generation in homogeneous domains.
- We demonstrate that standard U-Net attention mechanisms trigger training instability on small datasets, and show that purely convolutional architectures significantly improve convergence.

2 Related Work

The application of deep learning to precision agriculture has rapidly expanded, addressing critical tasks such as crop monitoring, yield estimation, and disease detection. For pineapple cultivation, which features dense planting and complex visual structures, object detection models like YOLO have been widely adopted [2]. Lima et al. [2] demonstrated the efficacy of YOLOv8 combined with Slicing Aided Hyper Inference (SAHI) to automate the detection of pineapple plants in large drone-captured orthomosaics. To capture more comprehensive agronomic features beyond the visible spectrum, Gonzalez-Hernandez et al. [5] explored multispectral drone imaging, showing that early fusion of near-infrared and red-edge channels suggest comparable detection performance. Furthermore, Pereira et al. [11] extended these YOLO-based frameworks to phytosanitary monitoring, successfully detecting visual anomalies such as fusarium wilt, mealybug infestations, and nutritional deficiencies under real-field conditions.

Deploying deep learning models in real-world agricultural scenarios also introduces challenges related to varying drone flight altitudes and dynamic camera motion. Fallas-Moya et al. [4] tackled the issue of object scale discrepancy during model training by proposing scaling augmentation techniques integrated with Feature Pyramid Networks (FPNs), improving generalization across different drone heights. In the domain of continuous video monitoring, Xie-Li et al. [17] introduced PineSORT, a real-time multiple object tracking (MOT) system optimized for drone videos. By incorporating camera motion compensation and a multi-stage association strategy, PineSORT effectively mitigates identity switches caused by the repetitive patterns and homogeneous appearances typical of pineapple fields.

Despite these algorithmic advancements, the performance of downstream monitoring systems—like the aforementioned YOLO architectures and PineSORT trackers—remains heavily bottlenecked by the availability of large, diverse, and well-annotated datasets. These models are effectively "data-starved," struggling to generalize when trained on limited drone flights. To feed these data-hungry algorithms and overcome class imbalances, generative artificial intelligence has emerged as a powerful tool in agriculture. Senthil et al. [1] integrated generative AI with Geographic Information Systems (GIS) and deep learning to simulate crop yield and growth potential, demonstrating that synthetic data can help fill gaps in

remote sensing imagery and promote sustainable precision farming. As comprehensively reviewed by Rahman et al. [12], Generative Adversarial Networks (GANs) are extensively utilized for image augmentation, synthesizing realistic crops to simulate biological variability. However, GAN-based approaches often suffer from training instability and limited diversity (mode collapse), restricting their effectiveness when modeling complex agricultural scenes [9].

While GANs are used for augmentation, Variational Autoencoders (VAEs) offer a complementary approach by learning compact, interpretable latent spaces from high-dimensional aerial data. Xie-Li et al. [18] investigated the use of VAE variants, including β-VAE and VQ-VAE, for aerial drone imagery of pineapple fields. Their study demonstrated that VAEs can capture relevant plant structures and enable latent-space clustering to separate the crop foreground from complex backgrounds. However, their findings also underscored that vanilla VAEs produce overly smooth reconstructions, whereas VQ-VAEs struggle with capacity limits on this specific dataset. This limitation becomes critical when latent representations are used for downstream generative models.

Consequently, diffusion models have emerged as a robust alternative, offering superior stability and higher image quality under limited data conditions compared to GANs [9]. However, standard diffusion architectures still struggle with agricultural constraints like high intraclass similarity and extreme data scarcity. To overcome this, recent works have had to introduce custom architectural interventions. For example, Zhang et al. [20] demonstrated that standard models fail on the rich textures of grain crops, requiring custom dual-attention U-Net mechanisms to generate high-fidelity images. Similarly, Zhou et al. [22] utilized a few-shot learning framework to combat complex background interference and extreme data scarcity in sunflower disease detection. These studies underscore a critical gap: while highly effective, diffusion architectures and their latent representations must be carefully adapted to handle small, homogeneous datasets without sacrificing fine-grained agronomic details.

Motivated by these challenges, our work builds on these principles by systematically combining structured latent representations with simplified diffusion architectures, aiming to maximize the quality and stability of synthetic image generation under severe data constraints.

3 Methodology

First, we describe how the Variational Autoencoder (VAE) was trained, including the chosen architecture and key reconstruction results. Next, we show effects of the β-VAE [7] and VQ-VAE [16] to explain why we used them in our latent DDPM framework. Finally, we detail the architectural specifications and design decisions that enabled the successful training of U-Net [14] for our image processing task.

3.1 VAE Training and Latent Variations

The latent autoencoder is a critical component of our pipeline, as it reduces the dimensionality of the images processed by our DDPM framework. As noted in prior work [8, 13], training denoising probabilistic models in high-resolution pixel space is computationally expensive. Following Rombach et al. [13], we implement the diffusion process within a compressed latent space. The core architecture used for both our β-VAE and VQ-VAE experiments is detailed in Table 1.

Description	Output ($C\times$ Spatial)
Enc: Conv3, 3 $\rightarrow$ 128; 2$\times$ RB128	$128 \times H \times W$
Enc: $\downarrow$ Conv3, s2; 2$\times$ RB128$\rightarrow$256	$256 \times H/2 \times W/2$
Enc: $\downarrow$ Conv3, s2; 2$\times$ RB256$\rightarrow$512	$512 \times H/4 \times W/4$
Enc: $\downarrow$ Conv3, s2; 3$\times$ RB512; Attn; RB512; GN+SiLU; Conv3, 512 $\rightarrow$ 8; Conv1, 8 $\rightarrow$ 8	$8 \times H/8 \times W/8$
Dec: Conv1, 4 $\rightarrow$ 4; Conv3, 4 $\rightarrow$ 512; RB512; Attn; 3$\times$ RB512	$512 \times H/8 \times W/8$
Dec: $\uparrow$2; Conv3, 512; 3$\times$ RB512	$512 \times H/4 \times W/4$
Dec: $\uparrow$2; Conv3, 512; RB512$\rightarrow$256; 2$\times$ RB256	$256 \times H/2 \times W/2$
Dec: $\uparrow$2; Conv3, 256; RB256$\rightarrow$128; 2$\times$ RB128; GN+SiLU; Conv3, 128 $\rightarrow$ 3	$3 \times H \times W$

Table 1: VAE Architecture. RB: ResBlock, Conv3/1: $3 \times 3/1 \times 1$ **Conv, s2: stride 2, GN: GroupNorm (32 groups).**

Drawing inspiration from Xie et al. [18], we utilized this architecture to compare the trade-offs between continuous and discrete latent representations for aerial imagery. Specifically, we trained two separate models: a β-VAE and a VQ-VAE.

3.1.1 Continuous Representation: β-VAE. In the β-VAE framework, the encoder E outputs the parameters of a Gaussian distribution, μ and $\log \sigma^2$, representing the latent variable z. To ensure high-quality reconstruction while maintaining a regularized latent space, we minimize the following loss function:

$$\mathcal{L}_{\beta\text{-VAE}} = \|x - \hat{x}\|_2^2 + \beta D_{KL}(q_\phi(z|x)\|p(z)) \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence and $p(z) \sim \mathcal{N}(0, I)$. Following the configurations suggested by Xie et al. [18], we set $\beta = 0.001$. In this setup, the final 8-channel output of the encoder is split into 4 channels for the mean and 4 for the log-variance.

3.1.2 Discrete Representation: VQ-VAE. Conversely, the VQ-VAE maps the encoder output $z_e(x)$ to a discrete codebook. We use a codebook $C = \{e_1, \ldots, e_K\}$ of $K = 128$ embeddings with a dimension $D = 64$ (as suggested by Xie et al. [18]). The quantization is performed by finding the nearest neighbor in the codebook:

$$z_q(x) = e_k \quad \text{where} \quad k = \arg\min_j \|z_e(x) - e_j\|_2 \tag{2}$$

The total loss for the VQ-VAE incorporates reconstruction, codebook alignment, and a commitment term to stabilize the encoder:

$$\mathcal{L}_{\text{VQ}} = \|x - \hat{x}\|_2^2 + \|sg[z_e(x)] - e\|_2^2 + \gamma\|z_e(x) - sg[e]\|_2^2 \tag{3}$$

where $sg[\cdot]$ denotes the stop-gradient operator.

After training both models, we evaluate them using a test set which is always unseen. This comparison is particularly important because the quality of the latent representation directly affects the downstream diffusion process: any structural information lost during encoding cannot be recovered during generation. For this reason, selecting the model with the highest reconstruction fidelity was imperative for the success of our latent DDPM framework. In our experiments, the continuous β-VAE consistently provided a

more faithful reconstruction of fine-grained agricultural structures, and was therefore used as the primary latent representation for the diffusion model.

3.2 Latent Diffusion Model Architecture and Training

Rather than operating in high-dimensional pixel space, our diffusion model is trained directly on the latent representations obtained from the selected autoencoder described in Section 3.1. We base our generative pipeline on the Denoising Diffusion Probabilistic Models (DDPM) framework [8] adapted for latent spaces [13]. Instead of exhaustively detailing the standard forward and reverse Markov processes—which gradually perturb a latent $\mathbf{z}_0$ into Gaussian noise $\mathbf{z}_T$ and learn to reverse it, we focus on the specific architectural adaptations required to stabilize this process under the constraints of our small, homogeneous agricultural dataset. The model is trained to predict the added noise ϵ using the standard variational bound objective:

$$L_{LDM}(\theta) := \mathbb{E}_{t,\mathbf{z}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right] \quad (4)$$

where $\epsilon \sim \mathcal{N}(0, I)$ is the true Gaussian noise sampled during the forward process, $\mathbf{z}_0$ is the initial uncorrupted latent, $\bar{\alpha}_t$ dictates the cumulative variance schedule governing the signal-to-noise ratio at timestep t, and $\epsilon_\theta(\cdot)$ is the U-Net model trained to predict this added noise.

A critical requirement for stable diffusion training is ensuring that the target latent space roughly matches the diffusion prior, typically an isotropic Gaussian $\mathcal{N}(0, I)$. Because we evaluate both continuous and discrete latent spaces, the pre-processing of these representations diverges. For the continuous space (β-VAE), as noted by Rombach et al. [13], KL-regularized spaces require explicit variance rescaling. We estimate the empirical standard deviation $\hat{\sigma}$ across all channels and spatial locations using the first training minibatch. All latent activations are then scaled by $\tilde{z} = z/\hat{\sigma}$ prior to forward diffusion, preventing scale mismatches during training. In contrast, for the discrete space (i.e., VQ-VAE), explicit variance scaling is unnecessary. Instead, we extract the pre-quantized latent representations directly from the encoder and absorb the quantization operation entirely into the decoder during the final image reconstruction phase, preserving the discrete topological boundaries during the diffusion process.

Standard LDMs typically incorporate self-attention mechanisms within their U-Net backbones to capture global context [13]; however, in our setting this design choice proved to be detrimental. Around epoch 80, the loss would predictably explode without recovery (see Figure 2). Namely, as highlighted by Dosovitskiy et al. [3], attention is design to capture long-range dependencies across an image; however, pineapple plantations exhibit predominantly local and repetitive pattern. Global attention is designed to capture long-range dependencies and relate distant objects within an image. However, the visual patterns in a pineapple plantation are predominantly local and highly repetitive. In this context, forcing the model to map global relationships across a small, homogeneous dataset introduces unnecessary variance and computational "noise," causing the attention heads to ultimately destabilize the network.

To rectify this, we stripped all attention blocks from the standard U-Net, relying strictly on residual convolutional layers (Table 2). This architectural pruning reduced the model from 865 million to 592 million parameters, successfully eliminating the loss spikes and ensuring smooth convergence (Figure 3b).

Description	Output ($C\times$ Spatial)
Enc: Conv3, 4 → 320; 2× RB320	$320 \times H \times W$
Enc: ↓ Conv3, s2; 2× RB320→640	$640 \times H/2 \times W/2$
Enc: ↓ Conv3, s2; 2× RB640→1280	$1280 \times H/4 \times W/4$
Enc: ↓ Conv3, s2; 2× RB1280	$1280 \times H/8 \times W/8$
Bottleneck: 2× RB1280	$1280 \times H/8 \times W/8$
Dec: 2× RB2560→1280; RB2560→1280; ↑ Conv3	$1280 \times H/4 \times W/4$
Dec: 2× RB2560→1280; RB1920→1280; ↑ Conv3	$1280 \times H/2 \times W/2$
Dec: RB1920→640; RB1280→640; RB960→640; ↑ Conv3	$640 \times H \times W$
Dec: RB960→320; 2× RB640→320	$320 \times H \times W$

Table 2: Compact U-Net Architecture. RB: ResBlock, Conv3: 3×3 Conv, s2: stride 2, ↑ Conv: Upsample + Conv3. Decoder input channels reflect skip connection concatenation.

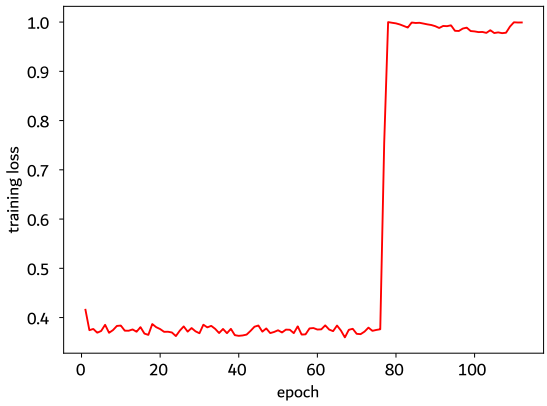
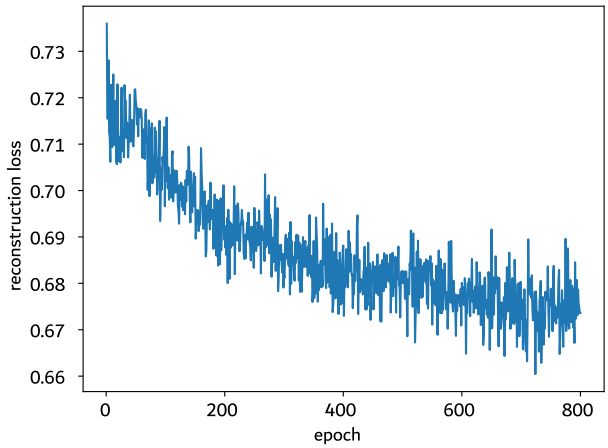


Figure 2: Training loss over epochs with U-Net layers using attention heads.

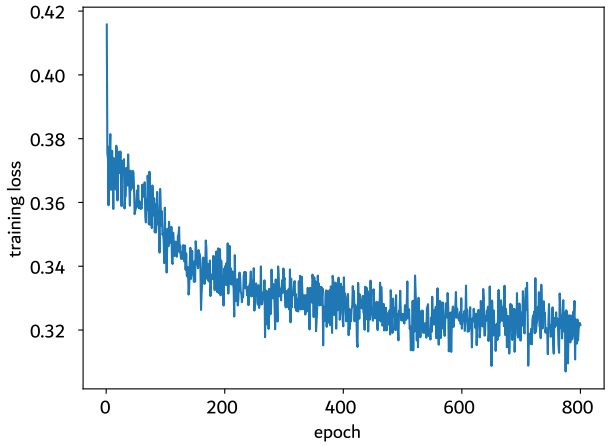
To better understand how well the initial encoded structure (z_0) is preserved during the generative process, we introduce a diagnostic tracking metric. At any randomly sampled timestep t, our U-Net predicts the noise $\epsilon_\theta(\mathbf{z}_t, t)$. Using this prediction, we can estimate the amount of noise injected into z_0 as follows:

$$\hat{\epsilon}_t = \frac{\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(\mathbf{z}_t, t) - \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0}{\sqrt{1-\bar{\alpha}_t}}. \quad (5)$$

This allows us to isolate a "*denoised latent estimate*", defined as $\tilde{z}_t = \mathbf{z}_t - \hat{\epsilon}_t$. During training, we compute the Mean Squared Error (MSE) between this estimate $\tilde{z}_t$ and the true original latent z_0 across timesteps (Figure 3a). While the model optimizes the standard noise prediction loss (Equation 4), this secondary MSE metric acts as a direct, observable proxy for latent degradation, confirming that the



(a) Reconstruction loss of the latent z after denoising



(b) Training loss

Figure 3: Training performance over epochs. Here the U-Net layer were without attention heads.

structural integrity of the crop features remains intact throughout the diffusion schedule.

Finally, at inference time, we replace standard DDPM sampling with the non-Markovian Denoising Diffusion Implicit Models (DDIM) sampler. Utilizing the same trained $\epsilon_\theta(\cdot)$ estimator, DDIM allows us to aggressively skip trajectory steps, yielding high-fidelity pineapple field generation with a 10x to 50x wall-clock speedup.

4 Results

4.1 Dataset

The dataset for this research comprises roughly 5000 images gathered via drone operations at a two-meter altitude [17]. These images were derived from continuous video recordings of pineapple plantations, with individual frames isolated to serve as static data points. This methodology captures high-resolution spatial details and the organic environmental diversity, as shifting light and background noise, typical of real-world agricultural monitoring.

To evaluate the models, the data was partitioned into a 70% training set, a 15% validation set, and a 15% testing set. The findings detailed in Section 4 are derived exclusively from the testing partition, ensuring the data remained entirely unseen by both the VAE and the diffusion model during their respective development phases.

4.2 VAE results

As detailed in Table 3, the continuous $\beta - VAE$ slightly outperforms the discrete VQ-VAE across all evaluated metrics on the test set. Specifically, the $\beta - VAE$ achieves a lower Mean Squared Error (MSE) of 380.14 compared to a MSE of 421.38 for VQ-VAE, alongside slightly higher Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) scores. These metrics indicate that the continuous latent space of the $\beta - VAE$ retains a more mathematically faithful representation of the original aerial imagery prior to any diffusion training.

Table 3: Comparison of reconstruction performance on the test set. VAE refers to β-VAE with $\beta = 0.001$, and VQ-VAE utilizes a codebook of $K = 128$ embeddings with dimension $D = 64$ as suggested by Xie et al. [18].

Method	MSE (↓)	PSNR (dB ↑)	SSIM (↑)
VAE	380.14	22.68	0.85
VQ-VAE	421.38	22.18	0.84

The performance of the autoencoder acts as a critical bottleneck for the downstream Latent Diffusion Model. Because the DDPM operates entirely within this compressed latent space, any structural information lost during the initial encoding step cannot be recovered during the generative reverse process. The higher fidelity of the $\beta - VAE$ suggests that its latent space will provide a richer, more accurate foundation for the diffusion model, potentially leading to higher-quality synthetic image generation. Furthermore, the continuous and heavily regularized nature of the $\beta - VAE$ latent space generally facilitates smoother noise injection and denoising trajectories. This continuous manifold can positively impact the training of the diffusion model by providing a more stable target for the U-Net's noise estimator, whereas discrete quantization boundaries in the VQ-VAE might introduce slight training instabilities if not carefully absorbed into the decoder.

Beyond quantitative metrics, a qualitative assessment of the reconstructed images provides further insight into the practical capabilities of the autoencoders. As illustrated in Figure 4, the visual results demonstrate that both models successfully capture the dominant characteristics of the pineapple plantations. When compared to the original data, the reconstructed images look slightly blurry for very fine-grained details (such as sharp soil transitions). However, this minor loss of high-frequency detail is an expected trade-off of the drastic dimensionality reduction. The representation is still highly useful, as the overall quality, macro-structure,

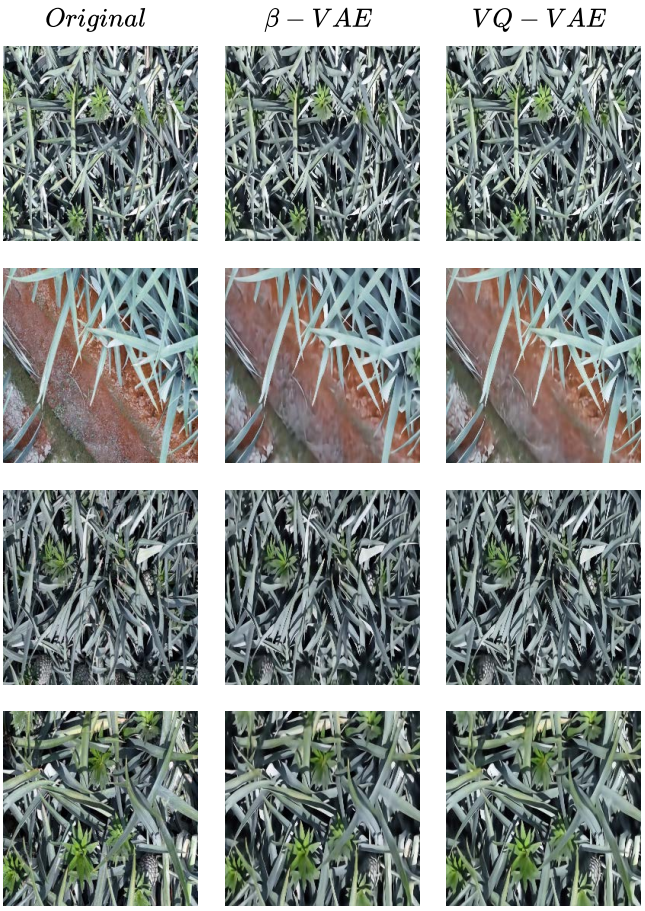


Figure 4: Visual results over the test set of β-VAE with $\beta = 0.001$, and VQ-VAE using a codebook of $K = 128$ embeddings with dimension $D = 64$.

plant density, and lighting conditions look comparable to the original images, ensuring the latent space is robust enough to train a reliable agricultural diffusion model.

4.3 Latent Diffusion Results

To ensure reproducibility, we standardized the core training hyperparameters across our Latent Diffusion experiments. All models were trained using an NVIDIA L40S GPU. Both the discrete (VQ-VAE) and continuous (β-VAE) diffusion models were trained from scratch for 800 epochs. Optimization was performed using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 8. For experimental consistency, all training runs were initialized with a random seed of 42 and executed without additional data augmentation. Furthermore, as noted in our architectural adaptations, the U-Net backbones were trained without global attention mechanisms. Notably, while the VQ-VAE diffusion pipeline maintained a constant learning rate throughout the training process, the β-VAE diffusion model utilized a dynamic learning rate schedule.

We focus on two key aspects: (i) the impact of the latent representation on generative quality, and (ii) the effect of sampling steps on the realism–diversity trade-off.

We begin with a qualitative comparison of the generated images to assess the ability of each latent space to capture the structural properties of pineapple plantations.

As observed in Figure 5a, the diffusion model trained within the continuous β-VAE latent space produces coherent and visually consistent synthetic images that preserve the macro-structure, lighting conditions, and overlapping leaf patterns characteristic of pineapple plantations. The continuous representation proved rich and structured enough for the DDPM to accurately reconstruct the dense macro-structure, lighting, and complex overlapping leaves characteristic of pineapple fields. In contrast, Figure 5b shows that the diffusion model trained on the VQ-VAE latent space struggles to generate coherent agricultural structures. This behavior may be related to the limited utilization of the codebook, where only a small subset of embeddings is effectively used. As a result, the diffusion model receives a less expressive latent representation, which limits its ability to reconstruct consistent spatial patterns. Consequently, we did not evaluate the VQ-VAE at lower DDIM sampling steps, as it failed to synthesize coherent images even at the maximum trajectory of 1000 timesteps. For the successful β-VAE model, Figure 6 illustrates that by leveraging the DDIM sampler, the network remains capable of generating visually convincing images at heavily reduced inference timesteps (50, 100, and 250 steps).



(a) Diffusion model trained with a β-VAE.



(b) Diffusion model trained with a VQ-VAE.

Figure 5: Visual results of synthetic generated images at 1000 sampling steps using different latent architectures.

These qualitative observations are supported by the quantitative metrics detailed in Table 4, which further highlight the impact of the latent representation on diffusion performance. The limitations of the VQ-VAE configuration are also reflected in the quantitative metrics, where Precision and Recall drop to 0.00, accompanied by a high Fréchet Distance (FD) of 210.51, indicating poor alignment with the real data distribution.



Figure 6: Sampling results of synthetic generated images at sampling steps 50, 100, 250 using the diffusion model trained with the β-VAE space.

Conversely, the β-VAE configuration reveals an interesting generative dynamic. The model achieves its absolute best overall fidelity at 250 sampling steps, indicating an optimal trade-off point with the lowest Fréchet Distance (78.02) and Kernel Distance (0.057). Varying the number of sampling steps reveals a clear trade-off between realism and diversity. At 100 steps, the model achieves its highest Precision (0.46), indicating that the generated samples are highly realistic and closely match the true data distribution. However, reducing the sampling trajectory further to 50 steps maximizes the Recall (0.33), suggesting that the model produces a wider, more diverse variety of field textures, albeit at the cost of a slight drop in strict visual fidelity.

Overall, these results highlight the importance of selecting an appropriate latent representation for diffusion models in low-data regimes, where continuous latent spaces provide a more stable and expressive foundation for generative modeling.

5 Conclusion

Using application of UAV imagery of pineapple fields, we addressed the challenge of training Latent Diffusion Models on small, homogeneous agricultural datasets to highlight the limitations of standard diffusion architectures, which are typically designed for large-scale and highly diverse datasets. By building a novel DDPM pipeline from scratch, we isolated the architectural vulnerabilities causing standard models to fail. Our comparative analysis of latent spaces revealed that continuous representations (β-VAE) provide a robust, structured foundation that allows the diffusion model to successfully synthesize complex crop macro-structures. Conversely, discrete architectures (VQ-VAE) proved highly susceptible to codebook collapse in this homogeneous domain, starving the downstream diffusion process of necessary spatial features and resulting in poor generative performance. Furthermore, we demonstrated that global attention mechanisms within the standard U-Net backbone can destabilize training when exposed to localized, repetitive crop patterns. Pruning these attention blocks in favor of a purely convolutional architecture eliminated loss explosions and ensured smooth convergence. Ultimately, combining a simplified convolutional architecture with continuous latent representations and DDIM sampling results in a stable and efficient framework capable of generating realistic agricultural imagery. These findings emphasize the importance of adapting generative models to the constraints of real-world agricultural data, where data availability and variability are inherently limited.

6 Future Work

This work opens several directions for improving generative modeling in agricultural settings. First, we plan to explore enhancements to discrete latent representations, particularly addressing the codebook collapse observed in VQ-VAE. Techniques such as codebook re-initialization, exponential moving average updates, or entropy regularization could improve embedding usage and increase the expressiveness of discrete latent spaces in low-data regimes. Additionally, we aim to investigate alternative attention mechanisms better suited for homogeneous environments, such as localized or parameter-efficient attention, which may capture relevant context without introducing the instability observed with global attention.

Another important direction is the evaluation of synthetic data in downstream tasks. While this work focuses on generative quality, future experiments will assess how the generated images impact applications such as crop detection, counting, and disease identification. We also plan to extend this framework to more diverse agricultural scenarios, including multispectral imagery and varying acquisition conditions, to evaluate the generalization of latent diffusion models beyond controlled UAV settings.

Although our architectural adaptations were manually derived through systematic experimentation, this design process can naturally be framed as a global optimization problem. Given the complex nature of these search spaces, evolutionary Neural Architecture Search (NAS) offers a promising alternative to exhaustive manual tuning. Recent literature demonstrates that evolutionary NAS can efficiently optimize diffusion architectures for specific domain and computational constraints [15, 19, 21]. Exploring evolutionary NAS to automate the discovery of stable, data-efficient latent diffusion

Table 4: Evaluation of generative performance across models and DDIM sampling steps. Metrics were extracted using the Inception network. Fréchet Distance (FD) and Kernel Distance (KD) are lower-is-better, while Precision (Prec), Recall (Rec), Density (Dens), Coverage (Cov), and Authenticity Percentage (Auth) are higher-is-better.

Metric	β-VAE (1000)	VQ-VAE (1000)	β-VAE (250)	β-VAE (100)	β-VAE (50)
FD (↓)	79.96	210.51	**78.02**	81.77	84.45
KD (↓)	0.058	0.257	**0.057**	0.060	0.063
Prec (↑)	0.45	0.00	0.44	**0.46**	0.39
Rec (↑)	0.20	0.00	0.30	0.25	**0.33**
Dens (↑)	0.16	0.00	**0.16**	0.16	0.15
Cov (↑)	0.23	0.00	**0.24**	0.20	0.19
Auth (↑)	94.90	100.00	93.67	94.29	96.12

models remains a compelling direction for future research, bridging applied deep learning with the core metaheuristic methodologies.

References

[1] Senthil G. A, S.U. Suganthi, L. Prinslin, R. Selvi, and R. Prabha. 2025. Generative AI in Agri: Sustainability in Smart Precision Farming Yield Prediction Mapping System Based on GIS Using Deep Learning and GPS. *Procedia Computer Science* 252 (2025), 365–380. doi:10.1016/j.procs.2024.12.038 4th International Conference on Evolutionary Computing and Mobile Sustainable Networks.

[2] Matheus Arroyo de Lima, Thiago Mantovani Martins, Vitor Matheus Soares Siqueira De Almeida, Rafael Buosi Jardim, Victor Hugo Zani, and Luiza Hoehl Loureiro Alves Barbosa. 2024. Automated Detection of Pineapple Plants in Drone-Captured Aerial Imagery for Precision Agriculture. *Anais do XIX Workshop de Visão Computacional (WVC 2024)* (2024). https://api.semanticscholar.org/CorpusID:277375708

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* abs/2010.11929 (2020). https://api.semanticscholar.org/CorpusID:225039882

[4] Fabian Fallas-Moya, Manfred Gonzalez-Hernandez, and Amir Sadovnik. 2021. Size Does Matter: Overcoming Limitations during Training when using a Feature Pyramid Network. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 1553–1560. doi:10.1109/ICMLA52953.2021.00249

[5] Manfred Gonzalez-Hernandez, Fabian Fallas-Moya, Werner Rodriguez-Montero, Danny Xie-Li, Bryan Roman-Solano, Francini Corrales-Garro, Amir Sadovnik, and Hairong Qi. 2023. Evaluation of the Influence of Multispectral Imaging for Object Detection in Pineapple Crops. In *2023 IEEE 5th International Conference on BioInspired Processing (BIP)*. 1–6. doi:10.1109/BIP60195.2023.10379335

[6] Phillip Gorris and Maynor Alberto Vargas Vargas. 2025. Agrochemicals in Costa Rica's pineapple industry: A review of environmental and human health impacts. *Revista Académica Arjé* 8, 1 (Jun. 2025), 1–26. doi:10.47633/29p6qn27

[7] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.. In *ICLR (Poster)*. OpenReview.net.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6840–6851.

[9] Xing Hu, Haodong Chen, Qianqian Duan, and Dawei Zhang. 2025. A Comprehensive Review of Diffusion Models in Smart Agriculture: Progress, Applications, and Challenges. arXiv:2507.18376 [cs.LG] https://arxiv.org/abs/2507.18376

[10] Federico Pallottino, Simona Violino, Simone Figorilli, Catello Pane, Jacopo Aguzzi, Giacomo Colle, Eugenio Nerio Nemmi, Alessandro Montaghi, Damianos Chatzievangelou, Francesca Antonucci, Lavinia Moscovini, Alessandro Mei, Corrado Costa, and Luciano Ortenzi. 2025. Applications and perspectives of Generative Artificial Intelligence in agriculture. *Computers and Electronics in Agriculture* 230 (2025), 109919. doi:10.1016/j.compag.2025.109919

[11] Elainne Ladislau Ferreira Pereira, Romário Da Costa Silva, Flávio Rafael Trindade Moura, Carlos Renato Lisboa Francês, and Diego Lisboa Cardoso. 2026. Pineapples Health Detection Using Deep Learning Models. *IEEE Access* 14 (2026), 19302–19319. doi:10.1109/ACCESS.2026.3660135

[12] Zahid Ur Rahman, Mohd Shahrimie Mohd Asaari, Haidi Ibrahim, Intan Sorfina Zainal Abidin, and Mohamad Khairi Ishak. 2024. Generative Adversarial Networks (GANs) for Image Augmentation in Farming: A Review. *IEEE Access* 12 (2024), 179912–179943. doi:10.1109/ACCESS.2024.3505989

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.

[15] Zeyu Tang. 2026. DE-ENAS:Diffusion-Enhanced Evolutionary Neural Architecture Search via Generative Mutations. (2026).

[16] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6309–6318.

[17] Danny Xie-Li and Fabian Fallas-Moya. 2025. PineSORT: A Simple Online Real-Time Tracking Framework for Drone Videos in Agriculture. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 65–74. doi:10.1109/CVPRW67362.2025.00012

[18] Danny Xie-Li, Manfred Gonzalez-Hernandez, Blaž Meden, and Fabian Fallas-Moya. 2025. Learning Compact Representations of Agricultural Fields: A Study of Variational Autoencoders Variants for Aerial Drone Imagery.

[19] Yu Xue, Chunxiao Jiao, Yong Zhang, Ali Wagdy Mohamed, Romany Fouad Mansour, and Ferrante Neri. 2026. Lightweight Diffusion Models Based on Multi-Objective Evolutionary Neural Architecture Search. *International Journal of Neural Systems* (2026). doi:10.1142/S0129065725500595

[20] Shuohua Zhang, Lei Liu, Guorun Li, Yuefeng Du, Xiuheng Wu, Zhenghe Song, and Xiaoyu Li. 2025. Diffusion model-based image generative method for quality monitoring of direct grain harvesting. *Computers and Electronics in Agriculture* 233 (2025), 110130. doi:10.1016/j.compag.2025.110130

[21] Bingye Zhou, Caiyang Yu, and Chenwei Tang. 2025. Evolution Meets Diffusion: Efficient Neural Architecture Generation. (2025). https://arxiv.org/abs/2504.17827

[22] Huachen Zhou, Weixia Li, Pei Li, Yifei Xu, Lin Zhang, Xingyu Zhou, Zihan Zhao, Enqi Li, and Chunli Lv. 2025. A Novel Few-Shot Learning Framework Based on Diffusion Models for High-Accuracy Sunflower Disease Detection and Classification. *Plants* 14, 3 (2025). doi:10.3390/plants14030339

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009