Inducing Rotational Equivariance in a Diffusion Model's VAE through a Dual-Objective Loss Function

Alfred Leong, Xin Qi Liu

June 6, 2025

Abstract

Standard Variational Autoencoders (VAEs) used in latent diffusion models learn powerful, compressed representations of images, but their latent spaces often lack intuitive geometric structure. A transformation such as a rotation applied in the latent space does not typically correspond to a rotation in the decoded pixel space. In this work, we demonstrate that a VAE can be explicitly taught this property. We propose a dual-objective fine-tuning methodology that combines a standard reconstruction loss with a rotation-consistency loss. We fine-tune the VAE from the 'Disty0/SoteMix' Stable Diffusion model on a dataset of anime-style images. Our results show that the fine-tuned VAE successfully learns 180° rotational equivariance, where rotating the latent representation produces a rotated output image. This is quantitatively verified by a near-perfect reconstruction (MSE of 0.0023) after a round-trip transformation, confirming the success of our training objective.

1 Introduction

Variational Autoencoders (VAEs) are a cornerstone of modern generative models, particularly in latent diffusion models like Stable Diffusion [1], where they compress high-resolution images into a manageable latent space for the denoising process. While these latent spaces are highly effective at capturing semantic features, they are not inherently structured to mirror simple geometric transformations in a predictable way. For a standard VAE, applying a 180° rotation to a latent tensor does not result in a 180° rotated image upon decoding, but rather a disorganized artifact (as demonstrated in Figure 2).

This paper explores a method to imbue the VAE's latent space with a specific geometric intuition. We hypothesize that by altering the training objective, a VAE can learn to be **equivariant** with respect to a 180° rotation. This means that performing the rotation operation in the latent space will be equivalent to performing it in the pixel space. We achieve this by fine-tuning a pre-trained VAE with a dual-objective loss function.

2 Methodology

2.1 Dataset Generation

A dataset of 1,000 images was generated using the 'Disty0/SoteMix' model [2], a fine-tune of Stable Diffusion v1.5. A consistent prompt of "1girl" was used to maintain a coherent domain for the VAE fine-tuning. The generation was performed using the `diffusers` library with 50 inference steps.

2.2 Model and Training Architecture

The VAE from the pre-trained 'Disty0/SoteMix model was isolated for fine-tuning. The training was conducted for 5 epochs with a batch size of 3 and a learning rate of 5×10^{-5}, using the AdamW optimizer. All images were resized and center-cropped to a resolution of 256×256 pixels.

2.3 Dual-Objective Loss Function

To enforce rotational equivariance, we designed a loss function, $\mathcal{L}_{\text{total}}$, composed of two Mean Squared Error (MSE) components: a standard reconstruction loss, $\mathcal{L}_{\text{recon}}$, and our proposed rotation-consistency loss, $\mathcal{L}_{\text{rot}}$. The total loss is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{rot}} \tag{1}$$

1. Reconstruction Loss ($\mathcal{L}_{\text{recon}}$): This is the standard VAE objective, ensuring the model can accurately encode and decode an image. Given an input image x, encoder E, and decoder D:

$$\mathcal{L}_{\text{recon}} = \text{MSE}(D(E(x)), x) \tag{2}$$

2. Rotation Loss ($\mathcal{L}_{\text{rot}}$): This novel component teaches the model the desired property. It asserts that decoding a rotated latent vector should yield a rotated image. Given a 180° rotation function, rot_{180}:

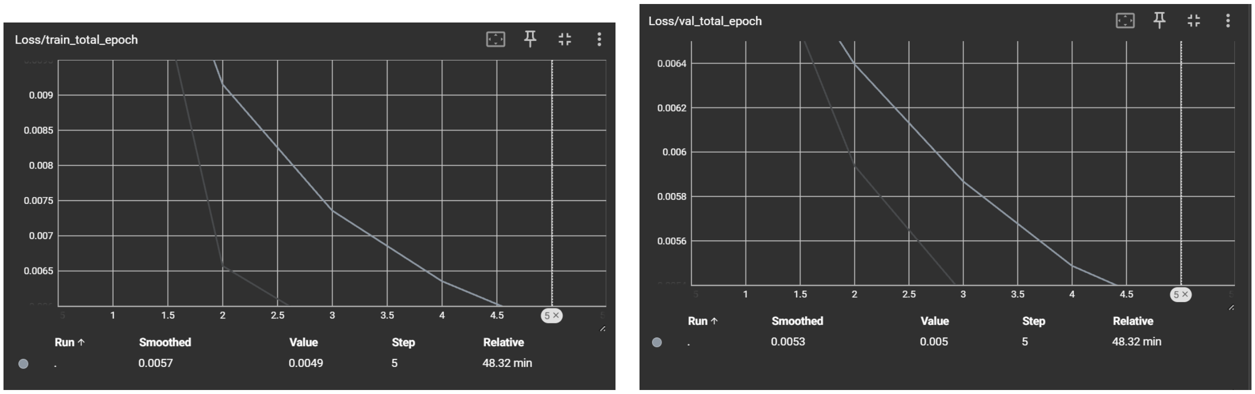
$$\mathcal{L}_{\text{rot}} = \text{MSE}(D(\text{rot}_{180}(E(x))), \text{rot}_{180}(x)) \tag{3}$$

By minimizing both losses simultaneously, the model learns to reconstruct images faithfully while also ensuring its latent space is structurally aligned with our desired 180° rotation.

3 Results

3.1 Training Performance

The model was trained successfully for 5 epochs. The training and validation losses, shown in Figure 1, demonstrate a consistent downward trend, indicating that the model was effectively learning from the dual-objective loss without overfitting.



Run ↑	Smoothed	Value	Step	Relative
.	0.0057	0.0049	5	48.32 min

Run ↑	Smoothed	Value	Step	Relative
.	0.0053	0.005	5	48.32 min

Figure 1: Total training loss (left) and validation loss (right) per epoch.

3.2 Qualitative Analysis

A standard VAE's behavior is shown in Figure 2. When the latent representation is flipped 180°, the decoded output is incoherent.

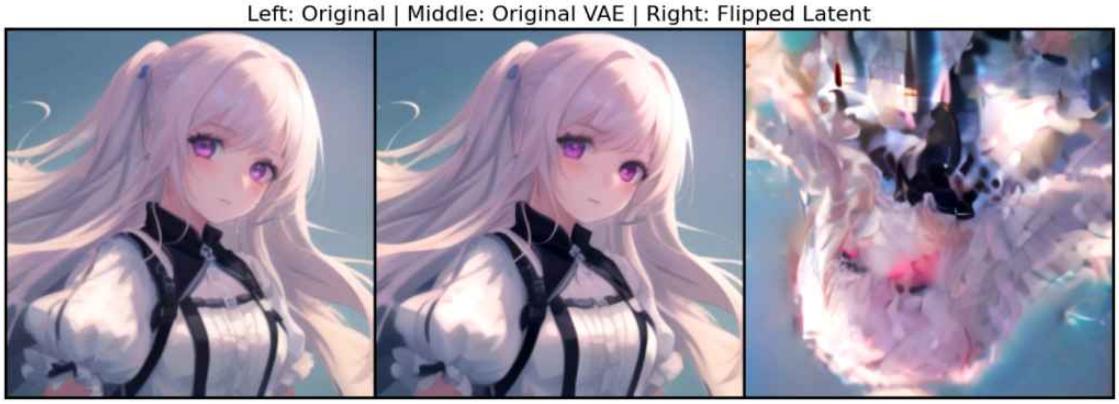

Left: Original | Middle: Original VAE | Right: Flipped Latent

Figure 2: A standard VAE attempting to decode a flipped latent vector results in a scrambled image.

In contrast, our fine-tuned model demonstrates clear rotational equivariance, as shown in Figure 3. Panel 3 shows the output of decoding a flipped latent vector, which now correctly corresponds to a 180° rotation of the original image. Subsequently, rotating Panel 3 by 180° in pixel space (Panel 4) recovers the original reconstruction almost perfectly.

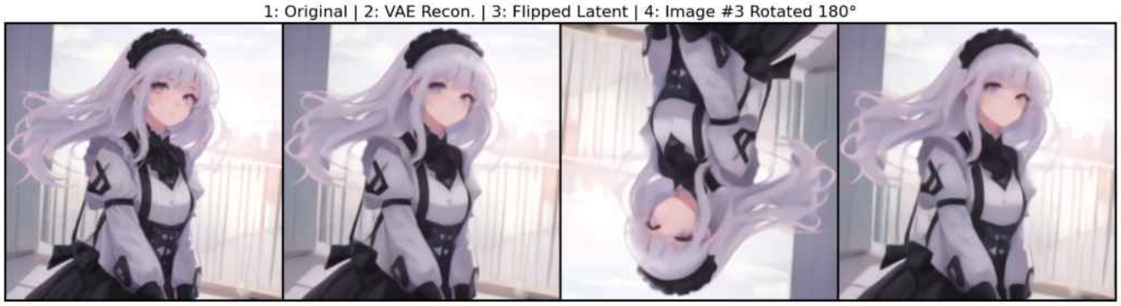

1: Original | 2: VAE Recon. | 3: Flipped Latent | 4: Image #3 Rotated 180°

Figure 3: Our fine-tuned model. (1) Original Image, (2) Standard Reconstruction, (3) Reconstruction from Flipped Latent, (4) Image #3 Rotated back 180°.

3.3 Quantitative Verification

To numerically validate the success of our method, we calculated the Mean Squared Error between the original input image (Figure 3, Panel 1) and the final round-trip image (Figure 3, Panel 4). The resulting MSE was **0.0023**. This exceptionally low value confirms that the process of rotating in the latent space and then rotating back in the pixel space is a near-perfect identity function, proving that the model has learned the intended geometric property.

4 Conclusion

In this work, we have successfully demonstrated that a VAE from a latent diffusion model can be fine-tuned to be equivariant to 180° rotations. By introducing a rotation-consistency term to the loss function, we were able to structure the VAE's latent space in an intuitive and predictable way. This shows that targeted training objectives can imbue deep generative models with specific, human-understandable properties, opening up potential avenues for more controllable and interpretable image generation. Future work could explore inducing equivariance for other transformations, such as 90° rotations or horizontal flips.

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[2] Disty0. (2023). *SoteMix*. Hugging Face. Retrieved from `https://huggingface.co/Disty0/SoteMix`