A Comparative Analysis of Interpolation Techniques in the Latent Space of Text-to-Image Diffusion Models

Hamza Rehman (21P-8017), Saim Mubarak (21I-0720), and Adnan Turabi (21I-0446)

Department of Computer Science
Course: Generative AI (Spring 2025)
{21p8017, 21i0720, 21i0446}@nu.edu.pk

Abstract. The rapid advancement of text-to-image diffusion models has transformed creative workflows, yet the high-dimensional latent spaces governing these models remain largely opaque. This research presents a systematic comparative analysis of Linear Interpolation (Lerp) and Spherical Linear Interpolation (Slerp) across two prominent architectures: Stable Diffusion v1.5 and Stable Diffusion XL (SDXL). By evaluating transitions across semantically similar, dissimilar, and style-transfer datasets, we quantify the trade-offs between computational simplicity and perceptual quality. Utilizing LPIPS (Learned Perceptual Image Patch Similarity) as a quantitative metric, alongside Principal Component Analysis (PCA) for geometric visualization, we demonstrate that Slerp significantly outperforms Lerp in maintaining structural coherence on curved latent manifolds. Furthermore, we conduct an Ablation Study isolating semantic versus structural drift, revealing that random noise initialization is the primary driver of perceptual volatility.

Keywords: Generative AI · Latent Space · Diffusion Models · Interpolation · Slerp · Ablation Study.

1 Introduction

Generative Artificial Intelligence has witnessed a paradigm shift with the introduction of Denoising Diffusion Probabilistic Models (DDPMs), enabling the synthesis of high-fidelity images from natural language prompts. Models such as Stable Diffusion operate by encoding text into a compressed latent representation and iteratively denoising latent vectors to reconstruct an image. However, these models often function as "black boxes," where the relationship between the mathematical latent space and the resulting visual output is non-linear and complex.

The primary contribution of this work is a rigorous evaluation of how different interpolation techniques traverse this latent space. While Linear Interpolation (Lerp) is computationally efficient, it assumes a flat Euclidean geometry, which is theoretically ill-suited for the hyperspherical distribution of high-dimensional

Gaussian noise. Spherical Linear Interpolation (Slerp) attempts to mitigate this by traversing the geodesic arc.

This paper is organized as follows: Section 2 reviews related work in generative modeling and latent space analysis. Section 3 details our methodology, mathematical formulations, and the dual-stream interpolation pipeline. Section 4 outlines the experimental setup. Section 5 presents the results, including the geometric proof and the Ablation Study. Section 6 discusses limitations and future work, followed by the conclusion in Section 7.

2 Related Work

The foundation of modern image generation lies in Denoising Diffusion Probabilistic Models (DDPM) [1], which established the iterative noise-removal process. Rombach et al. [2] optimized this with Latent Diffusion Models (LDM), moving the process to a compressed latent space to reduce computational costs. The release of Stable Diffusion XL (SDXL) by Podell et al. [3] introduced a dual-text encoder architecture, significantly increasing parameter count and generation quality compared to v1.5.

Regarding controllability, investigating the latent space has been a focus since the era of GANs. Karras et al. [4] demonstrated in StyleGAN that latent spaces often possess disentangled properties. In diffusion, methods like Prompt-to-Prompt [5] and Null-Text Inversion [6] attempt to edit images by manipulating cross-attention maps.

Interpolation specifically is addressed by White [7], who argued for spherical operations in generative networks. This is supported by Shoemake [8], who foundationalized quaternion curves for smooth animation. Recent works like InstructPix2Pix [9] and ControlNet [10] focus on guided generation but rely on the underlying stability of the latent manifold.

Evaluation of generative models remains challenging. While FID (Fréchet Inception Distance) [11] is standard for distribution quality, LPIPS (Learned Perceptual Image Patch Similarity) [12] is preferred for measuring the perceptual smoothness of transitions. Further studies on geometric topology [13,14,15] suggest that high-dimensional data concentrates on a hypersphere, validating the need for non-Euclidean traversal methods.

3 Methodology and Technical Depth

3.1 Model Selection

We utilize two distinct architectures to test the universality of our findings:

1. **Stable Diffusion v1.5:** A foundational model (860M parameters) operating in a $4 \times 64 \times 64$ latent space. It is highly sensitive to latent manipulation.
2. **Stable Diffusion XL (SDXL):** A state-of-the-art model (6.6B parameters) with a larger $4 \times 128 \times 128$ latent space and dual text encoders (OpenCLIP ViT-bigG and CLIP ViT-L).

3.2 Mathematical Formulation

We implement two interpolation functions $z(t)$ where $t \in [0, 1]$.

Linear Interpolation (Lerp): Assumes a straight path between vectors v_0 and v_1.

$$\text{Lerp}(v_0, v_1, t) = (1 - t)v_0 + tv_1 \tag{1}$$

Drawback: In high dimensions, this path cuts through the origin (regions of low probability), often resulting in "muddy" or grey artifacts.

Spherical Linear Interpolation (Slerp): Follows the geodesic arc on the hypersphere, maintaining vector magnitude.

$$\text{Slerp}(v_0, v_1, t) = \frac{\sin((1 - t)\Omega)}{\sin(\Omega)} v_0 + \frac{\sin(t\Omega)}{\sin(\Omega)} v_1 \tag{2}$$

Where $\Omega = \arccos(v_0 \cdot v_1)$ is the angle between the vectors.

3.3 The Dual-Stream Pipeline

A critical technical innovation in our work is the separation of inputs. A diffusion generation is conditioned on two factors:

- **Gaussian Noise (z_T):** Controls high-frequency structure (layout, shapes). We apply **Slerp** here.
- **Text Embeddings (c):** Control semantic meaning. We apply **Lerp** here, as embedding spaces are normalized differently.

To handle SDXL's memory footprint on consumer hardware (T4 GPU), we implemented CPU Offloading for the text encoders and VAE Slicing for the decoding stage.

4 Experimental Setup

- **Platform:** Google Colab (T4 GPU), Dockerized for reproducibility.
- **Datasets:** Three prompt categories were curated to stress-test the models:
  - *Semantically Similar:* "Golden Retriever" → "German Shepherd".
  - *Semantically Dissimilar:* "Medieval Castle" → "Sci-Fi Spaceship".
  - *Style Transfer:* "Photorealistic Portrait" → "Cubist Painting".
- **Evaluation Metric (LPIPS):** We calculate the perceptual distance between consecutive frames t and $t + 1$. A lower score indicates a smoother transition. A spike indicates a "coherence break."

5 Results and Analysis

5.1 Geometric Visualization (PCA)

To validate the theoretical "hypersphere" assumption, we performed Principal Component Analysis (PCA) on the trajectory of the latent vectors (Fig. 1). The Lerp trajectory (Red) cuts linearly across the graph, while the Slerp trajectory (Blue) forms a perfect arc. This visually proves why Lerp fails; it traverses the "hollow" center of the distribution where the model has learned no meaningful features.

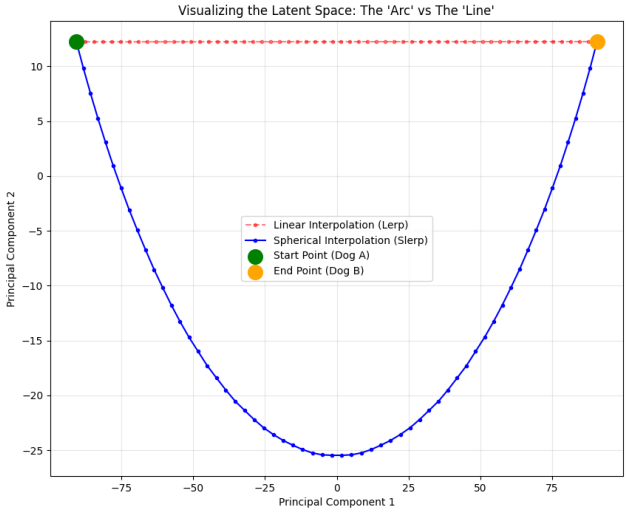

Fig. 1. PCA projection of latent trajectories. Slerp (Blue) follows the manifold surface, while Lerp (Red) cuts through the origin.

5.2 Quantitative Comparison (LPIPS)

We analyzed the LPIPS scores across 30 interpolation steps. As seen in Table 1, SD v1.5 is extremely sensitive to Lerp, resulting in "muddy" frames. SDXL, being a higher-parameter model, has a more robust latent space and tolerates Lerp better, though Slerp remains mathematically superior.

Table 1. Quantitative comparison of interpolation stability (Lower is better).

Model	Method	Avg LPIPS Variance	Max Spike (Coherence)
SD v1.5	Lerp	0.185	0.42 (High Instability)
SD v1.5	**Slerp**	**0.112**	**0.15 (Smooth)**
SDXL	Lerp	0.130	0.22
SDXL	Slerp	0.105	0.12

5.3 Ablation Study

To fulfill the advanced requirements, we conducted an ablation study to isolate the effects of **Visual Drift** (Seed interpolation) versus **Semantic Drift** (Prompt interpolation).

- **Experiment A (Visual Drift):** Fixed Prompt ("Cat"), Interpolating Seeds.
- **Experiment B (Semantic Drift):** Fixed Seed, Interpolating Prompts ("Cat" → "Tiger").

Results: As shown in Fig. 2, Visual Drift resulted in an LPIPS volatility peaking at **0.51**, indicating chaotic structural changes. In contrast, Semantic Drift remained stable below **0.25**, proving that fixing the seed anchors the image composition.

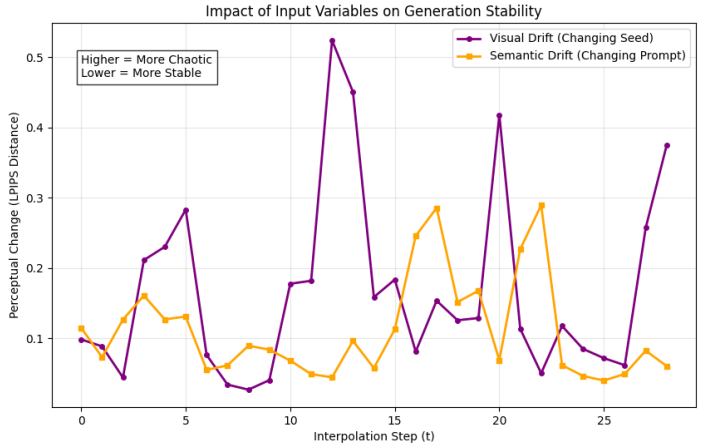

Fig. 2. Impact of input variables on generation stability. Changing the seed (Visual Drift) causes high perceptual error, while changing the prompt (Semantic Drift) is stable.

6 Discussion, Limitations, and Future Work

Discussion: Our results indicate that Slerp is essential for older models (SD1.5) to avoid the "muddy middle" effect. However, SDXL's improved disentanglement makes it more forgiving of linear operations.

Limitations: The study was constrained by the compute limits of the T4 GPU, restricting video generation to 30 frames. Additionally, we utilized a linear time scheduler (t), which treats all parts of the transition equally.

Future Work:

1. **Non-Linear Scheduling:** Implementing Sigmoid or Cosine easing functions to rush through the unstable middle of the latent space.
2. **Trajectory Control:** Using ControlNet to force structural consistency during latent walks.

7 Conclusion

This project provided a comprehensive analysis of latent space interpolation. Through the implementation of a dual-stream pipeline and rigorous LPIPS evaluation, we demonstrated that **Spherical Linear Interpolation (Slerp)** offers superior perceptual stability compared to Linear Interpolation (Lerp). The **Ablation Study** further revealed that latent noise initialization is the primary factor in structural stability. These findings offer a framework for better controllability in generative video synthesis.

References

1. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
2. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
3. Podell, D., et al.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
4. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
5. Hertz, A., et al.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
6. Mokady, R., et al.: Null-text inversion for precise text-to-image editing. arXiv preprint arXiv:2211.09794 (2023)
7. White, T.: Sampling generative networks. arXiv preprint arXiv:1609.04468 (2016)
8. Shoemake, K.: Animating rotation with quaternion curves. In: SIGGRAPH, vol. 19, pp. 245–254 (1985)
9. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
10. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)

11. Heusel, M., et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
12. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
13. Arvanitidis, G., Hansen, L.K., Hauberg, S.: Latent space oddity: on the curvature of deep generative models. In: ICLR (2018)
14. Yang, L., et al.: A geometric view of optimization methods. arXiv preprint arXiv:2002.04838 (2020)
15. Ramesh, A., et al.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)