# Semantic Misalignment in Latent Diffusion Models: A Study on Prompt Interpolation and Human Perception

Juyeon Lee

June 4, 2025

**Abstract**

This study investigates whether latent diffusion models (LDMs) can produce conceptually coherent visual transitions when performing linear interpolation between two prompts with large semantic gaps.

We generate 150 interpolated images between two prompts—one describing a futuristic city and the other a magical forest with a cheerful girl—and qualitatively evaluate the visual and conceptual transitions. While the style and lighting evolve smoothly, the model fails to capture the symbolic or narrative progression that humans intuitively expect.

To address this misalignment, we propose Guided Interpolation (inserting explicit intermediate prompts) and Semantic-aware Interpolation (using CLIP-based distances or nonlinear paths) as potential solutions. We briefly discuss how these methods can contribute to interpretable AI, intent-aligned generation systems, and creative design tools.

## 1 Introduction

Latent Diffusion Models have revolutionized text-to-image generation by enabling high-quality synthesis from natural language prompts[1]. However, their interpretability and alignment with human cognition remain open challenges. In this study, we pose a critical question: *Does linear interpolation between two semantically distant prompts produce images that align with human expectations of conceptual transition?*

## 2 Experiment

### 2.1 Prompts

- $P_1$: "A futuristic cityscape, vibrant and detailed"
- $P_2$: "A cheerful young girl with big sparkling eyes, standing in a magical forest, accompanied by a talking animal, soft lighting, Disney-style, whimsical and heartwarming atmosphere"

### 2.2 Setup

We use the Stable Diffusion v1.4 model[1]. Each prompt was tokenized and embedded using the model's tokenizer and text encoder. Then, 150 linearly interpolated embeddings between $P_1$ and $P_2$ were created[5]. The same latent noise vector was used across all steps to isolate the effect of embedding changes. Each embedding was passed through the diffusion pipeline to generate one image. Image resolution was set to 512x512, and inference steps were fixed to 25. The interpolation was performed in batches of 3.

# 3 Results

## 3.1 Positive Observations

1. **Coexistence of visual elements:** Urban structures and neon lights blend gradually with natural textures like trees and magical particles.

2. **Emergence of characters:** Character figures begin to appear amid urban backdrops, forming hybrid visual spaces.

3. **Stylistic continuity:** The transition in lighting and texture is smooth, moving from cold urban hues to warm forest tones.

## 3.2 Negative Observations (Human-Model Misalignment)

1. **Lack of intentional semantic transition:** While the model produces blended images, it fails to exhibit a narrative or symbolic progression that a human viewer would intuitively expect. Humans anticipate a structured, coherent transformation—such as a story where a girl gradually transitions from a futuristic city into a magical forest. For example:

   - The girl appears in a cityscape, then greenery begins to overtake the urban environment.
   - Flowers bloom from the tops of skyscrapers, symbolizing a shift from technology to nature.
   - The city morphs into a fantasy setting, becoming a stage for the character's inner world.

   However, the model instead performs a simple embedding interpolation, producing visuals that are often an arbitrary mixture with no coherent metaphoric or narrative bridge.

2. **Physical/logical inconsistency:** Floating characters or skyscrapers embedded in forests violate spatial realism.

3. **No clear conceptual boundaries:** The model fails to mark clear transitions such as "first appearance of the girl" or "start of the magical setting."

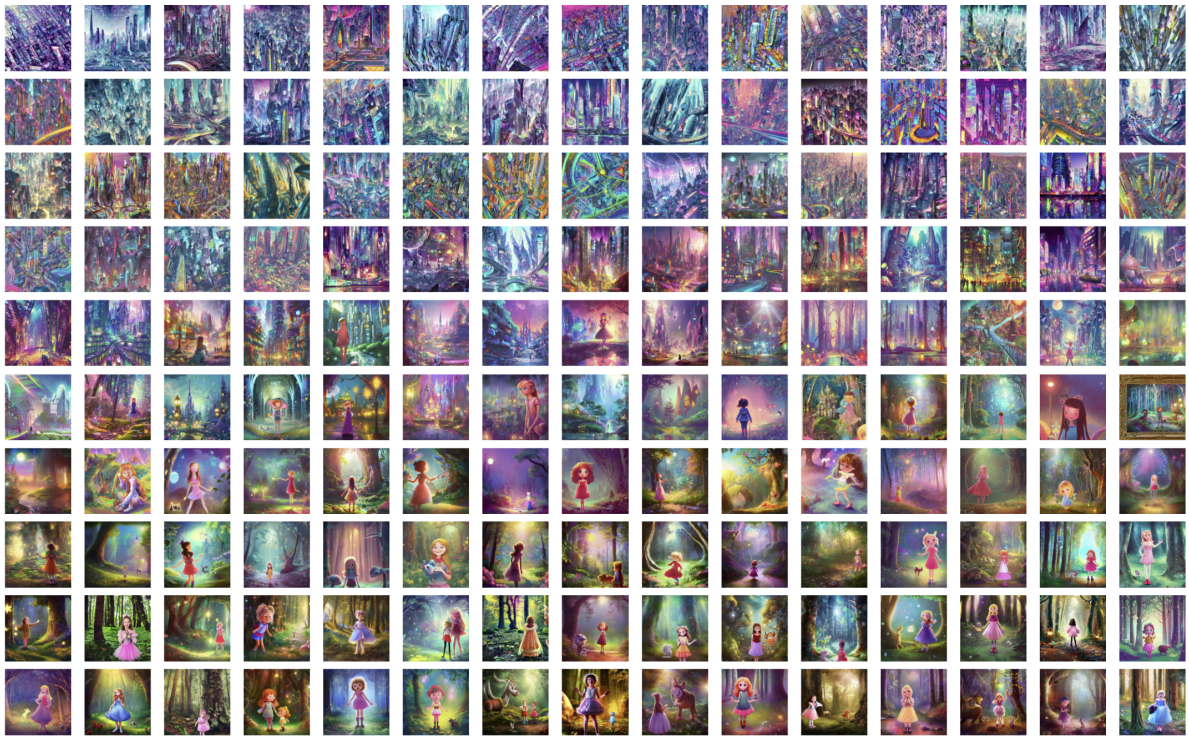

Figure 1: Example of 150 interpolated images between $P_1$ and $P_2$. Visual blending is smooth but lacks semantic or narrative coherence expected by humans.

## 4 Analysis

The model interpolates in embedding space without regard to conceptual structure or symbolic meaning. This results in visually blended but semantically disjointed imagery. The difference arises because humans expect narrative logic and symbolic metaphor, while models rely on vector arithmetic.

## 5 Proposed Solutions

- **Guided Interpolation:** Manually insert semantically meaningful intermediate prompts to guide the transition path. Examples include "A girl standing in a city garden" or "A place where technology and nature coexist." This provides a human-interpretable semantic scaffold for interpolation.
- **Semantic-aware Interpolation:** Replace linear interpolation with semantically guided paths using CLIP-based similarity measures or nonlinear trajectories such as SLERP (spherical linear interpolation)[2]. This helps minimize semantic jumps and follows a more natural curve through concept space.
- **Hybrid Prompt Synthesis:** Utilize large language models (LLMs) to automatically generate intermediate prompts between $P_1$ and $P_2$[3]. For instance, a mid-point prompt like "A dreamlike world where the city slowly transforms into a forest" can smooth transitions.
- **Human-in-the-loop Correction:** Provide user interfaces that allow manual intervention during the interpolation process, especially at points of semantic confusion. Users can insert corrective prompts or adjust interpolation behavior dynamically.

## 6 Applications

- **Visual storytelling:** Generate conceptually coherent image sequences for children's books, pre-visualization, or animatics by incorporating narrative transitions.
- **Creative content tools:** Assist designers and artists in visualizing idea transitions, aiding concept development and brainstorming in fields like fashion, architecture, and entertainment.
- **Interpretable AI research:** Visualize and analyze latent semantic structures in LDMs[4]. Identify conceptual boundaries, transition zones, and overlapping areas in latent space.
- **Cognitive science experiments:** Explore how humans perceive and evaluate semantic transitions in generated images, comparing model-driven interpolation with human expectations.
- **Intent-guided generation systems:** Develop systems that generate images aligned with user-defined conceptual intentions, enhancing customization and alignment in generative workflows.

## 7 Conclusion

This study reveals that although latent diffusion models can produce stylistically smooth transitions, they often diverge from human expectations of conceptual continuity. The misalignment arises due to the lack of semantic structure in interpolation. By introducing guided or semantic-aware methods, future systems can bridge the gap between machine-generated transitions and human cognition.

## References


[1] Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR 2022.

[2] Radford et al. "Learning Transferable Visual Models From Natural Language Supervision," ICML 2021.

[3] Gal et al. "Image Reward Models for Prompt Optimization," NeurIPS 2022.

[4] Kim et al. "Diffusion Concept Transformers for Interpretable Generation," ICLR 2023.

[5] Goyal et al. "Zero-shot Text-to-Image Generation Using Interpretable Latent Directions," CVPR 2023.