Prompt Matters: Analyzing Prompt Structure in Text-to-Image Generation

Yerin Hong
AIFFEL
ModuLabs
yerinhong0927@gmail.com

***Abstract*—Recent advances in text-to-image models, particularly diffusion-based generators such as Stable Diffusion, have highlighted the significance of prompt design in controlling output quality. Despite widespread interest in prompt engineering, a systematic analysis of how different prompt structures affect generated content remains limited. In this study, we explore the relationship between prompt complexity and semantic alignment of generated images. Using CLIP similarity and latent vector analysis, we evaluate how factors such as prompt length, stylistic keywords, and negative modifiers influence output. Our findings show that longer and stylistically rich prompts generally result in higher CLIP similarity, while latent space analysis reveals meaningful directional shifts corresponding to prompt progression. These results offer empirical insights into prompt effectiveness and provide a foundation for more interpretable prompt design.**

***Index Terms*—Stable Diffusion, Prompt Engineering, Text-to-Image Generation, CLIP, Semantic Similarity**

I. Introduction

Text-to-image generation models are becoming increasingly effective at generating high-resolution images from natural language input. However, their output is highly sensitive to how the prompt text is structured. This dependency has led to a research and practice called *prompt engineering* aimed at systematically designing prompts. In this work, we investigate how specific aspects of prompt structure—such as length, style, and negative phrasing—affect the output of a diffusion-based image generation model.

Despite the emergence of community practices for crafting effective prompts (e.g., adding stylistic modifiers or specifying scene structure), there is limited empirical understanding of how prompt attributes quantitatively affect generation. Key questions remain open: How does prompt length influence semantic alignment? Do style-related keywords enhance or distort the intended meaning? Is there a measurable progression in latent space as prompt semantics evolve?

To address these questions, we conduct a series of controlled experiments using Stable Diffusion. We design prompt series with varying structure—ranging from minimal to stylistically detailed—and generate corresponding images. Using CLIP similarity as a proxy for semantic alignment and latent vector analysis for visual feature evolution, we assess how prompt formulation influences image generation outcomes.

Our work contributes the following:

- A structured prompt dataset designed to isolate length, style, and negativity effects.
- Empirical analysis showing that prompt complexity improves CLIP similarity up to a saturation point.
- Latent trajectory visualizations demonstrating directional shifts in embedding space aligned with prompt modifications.

This study offers insights for both researchers and practitioners seeking interpretable and effective prompt design strategies.

II. Background

Text-to-image generation has rapidly evolved with the introduction of diffusion models, which iteratively denoise random noise conditioned on text prompts. Among them, Stable Diffusion [1] has emerged as a widely adopted open-source model due to its efficiency and quality. It combines a text encoder (usually CLIP) with a denoising UNet and a variational autoencoder (VAE) for image reconstruction.

Prompt engineering refers to the practice of carefully crafting input text to steer the generation toward desired visual outcomes. In practical settings, prompt tuning often involves adding stylistic descriptors (e.g., “cinematic”, “in Van Gogh style”) or specifying object details and environments. Although widely practiced, prompt engineering remains largely empirical, with limited structured analysis on how specific linguistic changes affect generation quality or semantics.

To evaluate generated images, CLIP (Contrastive Language–Image Pretraining) [2] has become a popular metric. CLIP maps images and texts into a shared embedding space, enabling semantic similarity scoring and image retrieval tasks. In addition to CLIP, recent works have explored analyzing latent representations from diffusion models to better understand how prompt structure affects visual outputs. For example, Hertz et al. [3] introduced cross-attention-based prompt editing to control semantic shifts within generated images, revealing directional latent changes triggered by prompt variations.

This study builds on these ideas by systematically varying prompt structure and measuring both semantic alignment (via CLIP) and latent representation movement. Our work focuses on interpreting how prompt composition—such as length, style, and negativity—affects generation behavior.

III. Methodology

To investigate how prompt structure affects image generation in diffusion models, we design a controlled experimental

framework consisting of prompt series construction, image generation, and representation analysis. This section outlines our methodology, including prompt design, generation process, and evaluation metrics.

A. Prompt Series Design

We construct prompt groups centered around 20 different visual concepts (e.g., “a fox”, “a mountain”, “a robot”). Each group contains five prompts of increasing complexity:

1) **Simple**: the base noun phrase (e.g., ”a fox”)
2) **Detailed**: with descriptive adjective (e.g., ”a detailed fox”)
3) **Stylistic**: with emotional or artistic descriptors (e.g., ”a detailed fox in watercolor”)
4) **ArtStyle**: referencing a known artistic style (e.g., ”in Van Gogh style”)
5) **Negative**: with distortive or abstract modifiers (e.g., ”distorted”)

This structure allows us to isolate the effects of prompt length, stylistic keywords, and negative language on generation outputs.

B. Image Generation

All images are generated using Stable Diffusion v1.4 via the HuggingFace `diffusers` library. For consistency, we generate one image per prompt using a fixed random seed. This ensures that observed changes are due to prompt structure, not stochastic noise.

C. Evaluation Metrics

We evaluate the generated images using two complementary approaches:

- **CLIP Similarity**: We compute the cosine similarity between the prompt text and generated image embeddings using OpenAI’s CLIP ViT-B/32 model. This serves as a proxy for semantic alignment.
- **Latent Representation Analysis**: We extract the final latent vectors from the diffusion model and project them into 2D using PCA. This allows us to visualize directional changes in latent space across prompt stages.

Prompt-level features such as token count, stylistic keyword presence, and negative modifiers are also logged to support quantitative analysis.

IV. Results and Analysis

In this section, we analyze the impact of prompt structure on the behavior of text-to-image generation using Stable Diffusion. Specifically, we evaluate how changes in prompt length, style-related keywords, and progressive prompt refinements affect CLIP similarity and latent vector space representations.

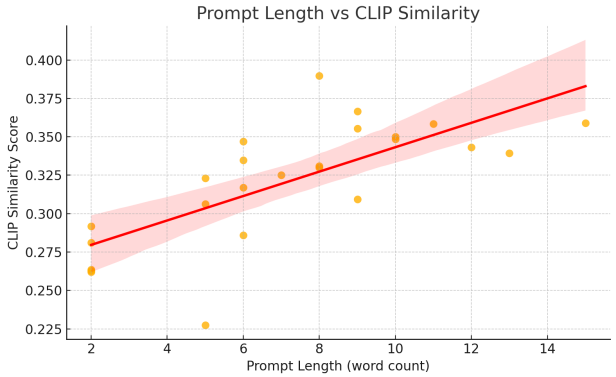

Fig. 1. Scatter plot showing the correlation between prompt length (in words) and CLIP similarity score. A positive trend is observed, with diminishing gains beyond 5-6 words.

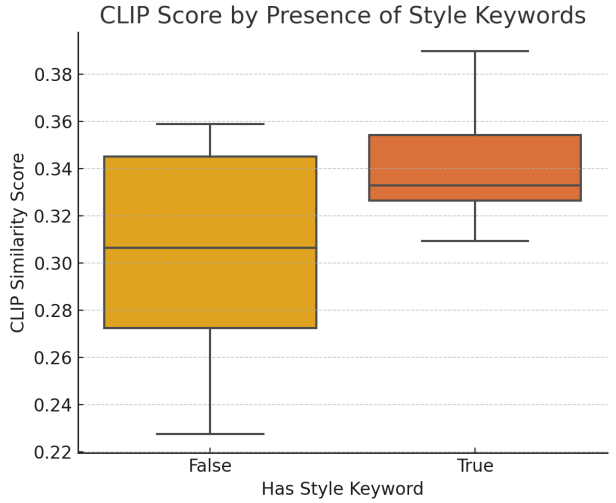

Fig. 2. Boxplot comparing CLIP similarity scores for prompts with and without stylistic keywords. Style-enhanced prompts yield higher alignment.

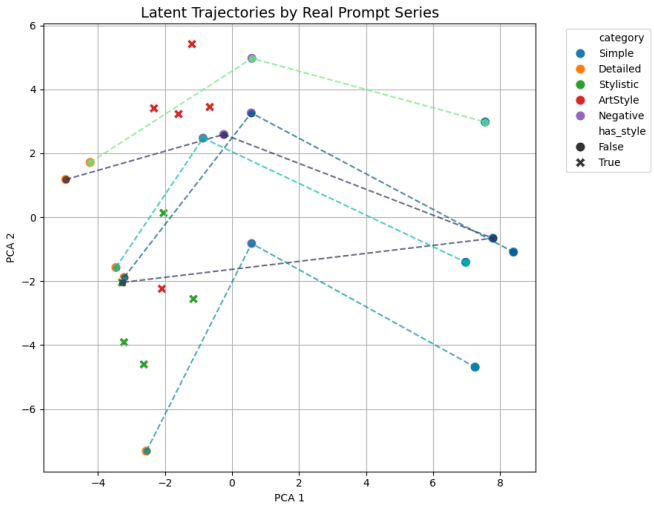

Fig. 3. Latent vector trajectories for multiple prompt groups projected in PCA space. Arrows indicate the semantic progression across prompt stages.

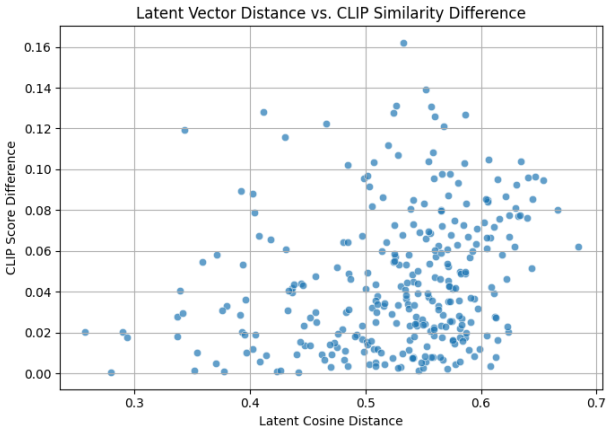

Fig. 4. Scatter plot of latent cosine distance vs. CLIP similarity difference. Weak correlation suggests latent proximity does not imply semantic similarity.

A. Prompt Length and CLIP Similarity

We observe a positive correlation between the length of a prompt (measured in word count) and the CLIP similarity between the prompt and the generated image. As shown in Fig. 1, shorter prompts such as *"a fox"* yield lower CLIP scores, whereas prompts with detailed and stylistic descriptors tend to increase the alignment. This trend saturates beyond a certain length, suggesting diminishing returns from excessive detail.

B. Style Keywords and CLIP Similarity

We analyze the effect of including style-related words (e.g., "cinematic," "in watercolor," "in Van Gogh style") on CLIP similarity. Prompts containing stylistic keywords show consistently higher CLIP scores compared to those without (Fig. 2). This indicates that stylistic prompts are more likely to yield visually distinctive results that align semantically with the input text, possibly due to stronger prior associations in the training dataset.

C. Latent Trajectories of Prompt Series

To investigate how semantic transformations in prompts affect the model's latent space, we projected the latent vectors into 2D using PCA. For each prompt group (e.g., G1: "a fox" series), we observed the trajectory of latent vectors across five stages: Simple $\rightarrow$ Detailed $\rightarrow$ Stylistic $\rightarrow$ ArtStyle $\rightarrow$ Negative.

Fig. 3 shows that some groups exhibit a coherent trajectory in a particular direction, while others show abrupt shifts, particularly at the transition to ArtStyle or Negative prompts. These transitions may indicate that the addition of stylistic or distorted keywords pushes the latent representation into a different semantic region.

D. Latent Distance and CLIP Similarity Difference

We computed the cosine distance between latent vectors and compared it to the absolute difference in CLIP similarity between corresponding prompts. While some degree of correlation was expected, our findings indicate no strong or consistent relationship (Fig. 4). This suggests that proximity in latent space does not reliably reflect semantic similarity from the CLIP perspective.

E. Summary of Findings

Our experiments suggest the following:

- Prompt length has a positive but saturating effect on CLIP similarity.
- Style-related keywords boost semantic alignment as measured by CLIP.
- Latent trajectories vary by group; meaningful progression is observed in some cases.
- Latent vector distances do not consistently align with CLIP-based semantic distances.

These insights provide guidance for prompt engineering in generative models and open avenues for deeper latent space interpretability.

V. Conclusion

In this paper, we investigated how different aspects of prompt structure affect the behavior of diffusion-based text-to-image generation. Using Stable Diffusion as our backbone model, we designed a controlled prompt dataset covering variations in length, stylistic descriptors, and negative modifiers.

Our quantitative results show that longer prompts generally lead to higher CLIP similarity scores, though with diminishing returns after a certain point. Prompts containing stylistic keywords also tend to increase semantic alignment as measured by CLIP. These findings validate the common intuition that more expressive or descriptive prompts yield images that better align with their intended meaning.

In addition, we explored how prompt modifications influence latent representations. Through PCA projection and trajectory visualization, we observed that semantically related prompts form progressive shifts in latent space, although the direction and consistency vary by concept and modifier type. The use of negative or distortive words did not exhibit clear trends in either CLIP similarity or latent displacement.

Future directions include incorporating human preference ratings, leveraging automatic captioning models, and exploring fine-tuning strategies to better reflect prompt semantics. Additionally, concept-level editing and removal techniques [4] may be employed to further investigate how specific elements of a prompt can be isolated, suppressed, or disentangled within the generation process.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh *et al.*, "Learning transferable visual models from natural language supervision," *ICML*, 2021.

[3] A. Hertz, R. Mokady, G. Tevet *et al.*, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[4] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 2427–2437.