Prompt Puzzle: Prompt Fusion for Visual Creativity

Korea University COSE461 Final Project

Seonga Choi
Department Data Science
Team 23
2021320301

Bean Cho
Department Data Science
Team 23
2023320322

Abstract

Text-to-image generation models such as Stable Diffusion have achieved remarkable success in synthesizing high-quality images from natural language prompts. However, user-provided prompts are often vague or underspecified, leading to outputs that lack semantic alignment with user intent. To address this challenge, we propose a novel multi-persona prompt fusion framework that leverages Large Language Models (LLMs) and image-text relevance models to refine and optimize prompts. First, multiple persona-driven rewrites of the input prompt are generated using LLMs, where each persona reflects a distinct stylistic perspective (e.g., engineer, poet, painter, director). For each rewritten prompt, a corresponding image is synthesized, and BLIP-based image-text matching scores are computed to evaluate visual alignment. These scores are normalized and used as fusion weights to combine sentence embeddings across personas. Finally, a fused embedding is decoded into a unified optimized prompt using an LLM. Experimental results on prompts sampled from DiffusionDB demonstrate that our approach improves both semantic alignment (as measured by CLIP scores) and aesthetic quality, while preserving stylistic richness. This work highlights the potential of role-guided prompt fusion as a lightweight and effective strategy for enhancing creativity and consistency in text-to-image generation.

1 Introduction

Text-to-image generation models have recently made significant progress with the advent of large-scale models such as Stable Diffusion, DALL·E 2, and Imagen. These models are capable of producing high-quality images from natural language prompts, and have shown promise in a variety of applications including creative design, digital art, and educational content. However, in real-world scenarios, user-provided prompts are often vague or lacking in visual cues, resulting in outputs that deviate from the intended semantics. Maintaining consistent *semantic alignment* between the prompt and the generated image remains a key challenge.

To address this issue, previous works have leveraged large language models (LLMs) to rewrite the original prompt into a more visually rich and precise description. Additionally, some studies have proposed using multimodal models such as CLIP or BLIP to evaluate and iteratively optimize the alignment between the prompt and the generated image. Nevertheless, these approaches typically rely on a single rewritten prompt and fail to effectively utilize the diversity of multiple candidate prompts.

In this paper, we propose a novel framework inspired by human cognitive processes. We observe that individuals with different perspectives or stylistic tendencies may describe the same scene in notably different ways, leading to diverse visual interpretations. Based on this insight, we design a system in which multiple LLM agents, each assigned a distinct *persona* (e.g., artist, photographer, storyteller),

independently rewrite the input prompt. Images are then generated for each rewritten prompt and evaluated using BLIP-based image-text relevance scores. Instead of selecting a single best candidate, we compute normalized BLIP relevance scores and apply them as fusion weights across all persona prompts.

We introduce a **Prompt Fusion Module** that semantically integrates multiple persona prompts into a single, unified prompt by performing weighted fusion in embedding space, followed by language reconstruction via an LLM. This approach allows the fused prompt to reflect both semantic alignment and stylistic diversity across personas.

Our key contributions are summarized as follows:

- We propose a multi-persona prompt rewriting framework using diverse LLM agents to generate stylistically varied prompts from a single input.
- We introduce a BLIP-based evaluation strategy that computes normalized relevance weights for persona prompts, reflecting their visual alignment.
- We design a *Prompt Fusion Module* that semantically combines multiple persona prompts through weighted embedding fusion and LLM-based reconstruction.
- We conduct both quantitative and qualitative evaluations to demonstrate the effectiveness of our method in improving image-text alignment and prompt expressiveness.

This work presents a new direction for improving prompt quality and semantic alignment in text-to-image generation, contributing to both the reliability and creative potential of generative models.

2 Related Work

2.1 Text-to-Image Generation Models

Text-to-image (T2I) generation models aim to synthesize images that reflect the semantics of an input textual description. Recent advances in diffusion-based models have led to remarkable improvements in generation quality. Notable examples include Stable Diffusion [1], DALL·E [2], and Imagen [3], which are trained on large-scale image-text datasets and capable of capturing complex visual concepts.

Despite these successes, the performance of T2I models [4] is highly sensitive to the input prompt. In particular, vague, underspecified, or ambiguous prompts often lead to misaligned or semantically inconsistent images. As a result, there is growing interest in improving prompt quality through techniques such as rewriting, expansion, or optimization, which can significantly enhance the alignment and expressiveness of generated outputs.

2.2 Persona-Based Prompt Rewriting with LLMs

Large Language Models (LLMs) have recently been leveraged to rewrite user prompts into more detailed, visually grounded descriptions. Prior works have explored prompt engineering techniques [5] and CLIP-guided prompt tuning [6, 7] to enhance image generation quality. However, many approaches rely on a single perspective or style, limiting the diversity and creativity of the rewritten prompts.

In contrast, some emerging studies focus on multi-agent or persona-driven LLMs, where each agent adopts a unique linguistic or stylistic viewpoint. Inspired by the human ability to describe the same scene in different ways, persona-based rewriting enables the generation of multiple prompt candidates that reflect diverse interpretations [8]. While these methods introduce greater diversity, few have explored how to *fuse* such prompts in a way that semantically complements their strengths.

Our work bridges this gap by combining multi-persona prompt rewriting with a novel fusion mechanism. We propose a Prompt Fusion Module that semantically integrates the top-performing prompts, thereby achieving a better balance between prompt diversity and semantic precision in text-to-image generation.

3 Observation

We conducted a comparative evaluation between the baseline prompts and our engineered rewrites using CLIP scores as the primary metric. The results indicate that the performance of the engineered prompts was largely consistent with that of the baseline, showing no significant degradation in generation quality. Importantly, for 27 percent of the evaluated prompts, the engineered versions outperformed the baseline in terms of CLIP score, suggesting that prompt engineering can yield meaningful improvements in image-text alignment. This observation underscores the potential of prompt rewriting as a low-cost yet effective strategy for enhancing the quality of generative outputs, particularly when tailored to specific personas or domains.

In addition to CLIP-based evaluation, we also assessed the aesthetic quality of the generated images using a learned aesthetic scoring function. The analysis revealed that prompts rewritten using the engineer persona achieved higher aesthetic scores than the baseline in approximately 50.6 percent of the cases, while maintaining parity in an additional 1.1 percent These findings suggest that prompt engineering not only preserves semantic alignment with the input prompt (as measured by CLIP) but can also enhance the visual appeal of the generated content. Taken together, our results highlight that persona-guided prompt rewriting holds promise as a lightweight method for improving both factual consistency and aesthetic quality in text-to-image generation.

4 Method

In this section, we describe our BLIP-guided prompt fusion framework designed to generate optimized prompts for text-to-image diffusion models by integrating multiple persona-specific variations. The framework consists of the following core components: (1) prompt collection, (2) relevance-based weighting using BLIP, (3) sentence-level embedding fusion, (4) LLM-based prompt reconstruction, and (5) final image synthesis.

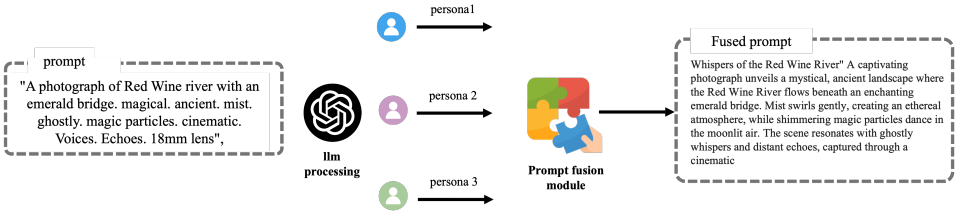

Figure 1: Overview of our BLIP-guided multi-persona prompt fusion framework. An initial user prompt is rewritten by multiple persona agents (e.g., poet, engineer, painter, director) using LLM-based rewriting. Each persona prompt is associated with an image generated via Stable Diffusion, and BLIP scores are computed to measure image-text alignment. These scores are used to compute weighted sentence embeddings, which are fused and decoded by an LLM to generate the final optimized prompt for image synthesis.

4.1 Prompt Collection

Given an original input prompt P_{base}, we construct a set of persona-rewritten prompts $\{P_1, P_2, \ldots, P_N\}$, where each P_i corresponds to a specific stylistic perspective such as "engineer", "poet", or "director". These prompts are generated by prompt engineering techniques or predefined rewriting templates per persona and stored in a structured JSON format. Each sample is thus represented as a set of parallel prompts sharing the same semantic intent but differing in linguistic style.

4.2 BLIP-based Image-Text Relevance Scoring

To assess the visual-linguistic alignment of each persona prompt, we employ a pretrained BLIP [**?**] image-text matching model. For each prompt P_i, we generate a corresponding image I_i using a Stable Diffusion model [**?**]. The BLIP model then computes a similarity score $s_i = \text{BLIP}(I_i, P_i)$

0.9

Minimalist:
Rewrite the prompt as a **minimalist**. Keep only **core objects and actions**. Remove adjectives, styles, and vague modifiers. Focus on **factual essentials**, while preserving the scene's structure. Stay **under 70 tokens**.

Director:
Rewrite the prompt as a **film director**.
Describe the scene using cinematic language: include **composition**, **camera angle**, **lighting**, and **color tone**. Mention **lens type** if relevant. Use **clear, visual terms** that help the model frame the scene like a shot in a movie. Keep all key visual elements and stay **under 70 tokens**.

Figure 2: Prompt rewriting instructions for **Minimalist** and **Director** personas. The Minimalist retains only core objects and factual essentials, minimizing adjectives and vague modifiers. The Director employs cinematic language by describing composition, camera angle, lighting, color tone, and lens type to create a visually rich scene.

0.9

Painter:
Rewrite the prompt as a **visual artist**. Keep all key objects and describe them using **clear words** about **color, lighting, material**, and **layout**. Follow the order: structure → details → style. Use **visual language**, and write a **descriptive sentence under 70 tokens**.

Engineer:
Rewrite the prompt as an **engineer**. Use **technical terms** about **structure, material, shape**, and **function**. Preserve all original elements. Avoid emotion or style. Output must be **concise, literal**, and **under 70 tokens**.

Poet:
Rewrite the prompt as a poet. Express the same scene using symbolic, emotional, and sensory language. Keep all major visual components, but describe them with feeling and metaphor. Write one evocative sentence under 70 tokens.

Figure 3: Prompt rewriting instructions for **Painter**, **Engineer**, and **Poet** personas. The Painter emphasizes visual details such as color, lighting, material, and layout in a descriptive manner. The Engineer utilizes technical and functional descriptions while avoiding emotion or style. The Poet employs symbolic, emotional, and metaphorical language to evoke vivid imagery.

Figure 4: Persona-specific prompt rewriting guidelines used to generate diverse variants of the input prompts.

indicating how well the image matches the prompt. These scores are normalized using a softmax function:

$$w_i = \frac{\exp(s_i)}{\sum_{j=1}^{N} \exp(s_j)}, \tag{1}$$

where w_i denotes the weight assigned to persona i. This weighting mechanism ensures that more visually aligned prompts contribute more to the fused representation.

4.3 Sentence Embedding Fusion

We represent each persona prompt P_i as a dense vector $\mathbf{e}_i \in \mathbb{R}^d$ using a pretrained Sentence-BERT model [**?**]. The final fused embedding is computed as a weighted sum of individual embeddings:

$$\mathbf{e}_{fused} = \sum_{i=1}^{N} w_i \cdot \mathbf{e}_i. \tag{2}$$

This fusion strategy allows us to integrate multiple stylistic perspectives into a single semantic representation that reflects both diversity and visual relevance.

4.4 LLM-based Prompt Reconstruction

Directly decoding the fused embedding into natural language is nontrivial due to the absence of a universal inverse mapping. Instead, we reconstruct the final fused prompt using a Large Language Model (LLM), specifically GPT-4o. The model is prompted with the list of persona prompts and instructed to generate a single optimized prompt that synthesizes their stylistic strengths while preserving coherence and suitability for diffusion-based image generation.

```
System: You are a prompt engineer for image generation. Merge the
following persona prompts into one optimized prompt that is visually
rich, coherent, and stylistically balanced.
```

This LLM-based decoding ensures that the final prompt is fluently expressed and aligned with the input variations, functioning as a semantic "centroid" of the personas.

4.5 Image Generation and Output Storage

The reconstructed prompt P_{fused} is fed into the same Stable Diffusion model to generate the final image I_{fused}. The image, together with its associated fused prompt, embedding vector, and weighting information, is stored for downstream analysis and qualitative evaluation. Results are serialized in JSON format and saved alongside the generated images.

Implementation Details. All sentence embeddings are computed using the `all-mpnet-base-v2` variant of Sentence-BERT. The BLIP score is computed using a pretrained image-text matching model, and image synthesis is performed via a Stable Diffusion v1.5 pipeline. LLM-based decoding is conducted using the GPT-4o API with a temperature of 0.7 and a maximum token limit of 70.

4.6 Data

We construct our dataset based on **DiffusionDB** [9], a large-scale text-to-image prompt dataset containing over two million prompts and their corresponding images generated by Stable Diffusion. From this dataset, we randomly sample 1000 prompts that serve as the *base prompts* for our experiments.

To study stylistic variation in prompt formulation, we augment each base prompt with a set of **persona-specific rewrites**. Specifically, for each base prompt, we generate four persona variants corresponding to *poet*, *engineer*, *painter*, and *director*, where each persona reflects a distinct linguistic and stylistic perspective. These persona prompts are synthetically generated using a Large Language Model (GPT-4o), which is instructed to rewrite each base prompt following the conventions and domain knowledge of the respective persona.

For each persona prompt, we generate a corresponding image using a Stable Diffusion v1.5 model, resulting in a total of 5,000 images (including the base prompts). These images and their associated prompts serve as input for our prompt fusion and evaluation pipeline.

Formally, for a given base prompt P_{base}, we produce a set of $N = 4$ stylistically distinct persona prompts $\{P_1, P_2, P_3, P_4\}$. These serve as input to our BLIP-guided fusion module, which computes a fused embedding and reconstructs a final optimized prompt P_{fused} for downstream image generation.

This dataset enables a systematic investigation of prompt-level fusion across multiple stylistic perspectives, supporting our broader objective of generating visually grounded, stylistically balanced prompts that improve both semantic coherence and creative diversity in diffusion-based image generation.

5 Analysis

We conduct quantitative analysis to evaluate the effectiveness of our BLIP-guided prompt fusion framework. In particular, we compare the BLIP image-text matching scores between the fused prompts and the original base prompts to assess whether fusion improves semantic alignment.

Table

Method	Fused Mean	Original Mean	Fused > Original Ratio
All Prompt Fusion	0.343	0.321	56.37
Top-K Prompt Fusion	0.359	0.322	63.58

Table 1: Performance Comparison of Fusion Methods

summarizes the results across two fusion strategies: *All Prompt Fusion* (using all persona rewrites) and *Top-3 Prompt Fusion* (fusing the top 3 persona prompts with the highest BLIP scores). In both cases, we observe that the fused prompts consistently achieve higher BLIP scores than the original prompts. For the All Prompt Fusion setting, the fused prompts yield a mean score of 0.3435 compared to 0.3210 for the original, while for Top-3 Prompt Fusion, the mean score further improves to 0.3595.

Furthermore, the proportion of cases where fusion outperforms the original prompt is substantial. Specifically, in 56.37% of cases under All Prompt Fusion and 63.58% under Top-3 Prompt Fusion, the fused prompt achieves a higher BLIP score than the original. This demonstrates that even a simple persona-based fusion mechanism can lead to meaningful improvements in image-text relevance.

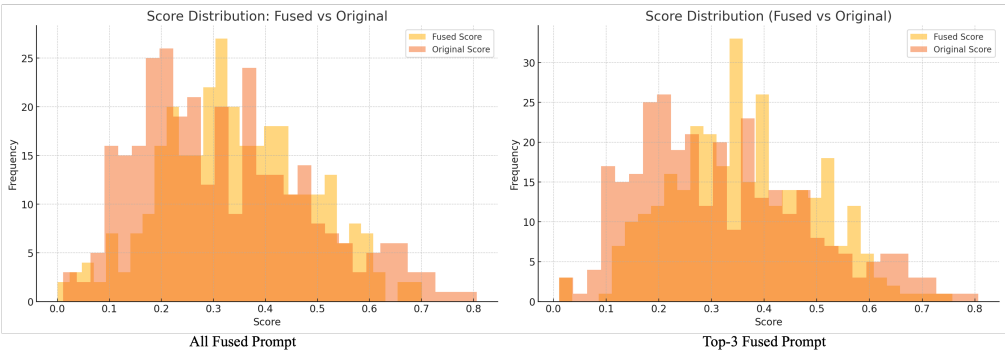

Figure 5: Score distribution comparison between fused prompts and original prompts. The left plot shows the results for All Prompt Fusion, while the right plot shows the results for Top-3 Prompt Fusion. In both cases, fused prompts generally achieve higher BLIP scores, with Top-3 fusion demonstrating a clearer shift toward higher scores. This indicates that selecting and integrating highly aligned persona prompts leads to better semantic alignment in text-to-image generation.

Interestingly, the Top-3 Prompt Fusion strategy yields the best performance, indicating that selectively integrating only the most visually aligned persona prompts results in better semantic consistency. This suggests that while prompt diversity introduces useful variation, not all rewrites contribute equally, and carefully weighting the contributions based on relevance leads to superior fused prompts.

Overall, these findings support the conclusion that BLIP-guided multi-persona prompt fusion is an effective technique for enhancing the alignment between prompts and generated images in text-to-image generation tasks.

6 Conclusion

In this work, we proposed a BLIP-guided multi-persona prompt fusion framework for enhancing semantic alignment and creativity in text-to-image generation. By leveraging persona-based rewriting with Large Language Models and weighting the diverse prompts using image-text relevance scores from BLIP, our system was able to generate fused prompts that demonstrate improved image alignment compared to the original base prompts. Quantitative evaluations show that prompt fusion consistently yields higher BLIP scores, with additional gains when selectively integrating highly aligned persona rewrites. These results indicate that prompt fusion is an effective and lightweight approach for improving text-to-image generation quality.

A key insight from our study is that generating persona prompts with strong and distinctive stylistic characteristics is crucial for effective fusion. Diverse and well-differentiated rewrites allow the fusion mechanism to capture complementary visual cues, ultimately leading to richer and more coherent synthesized images.

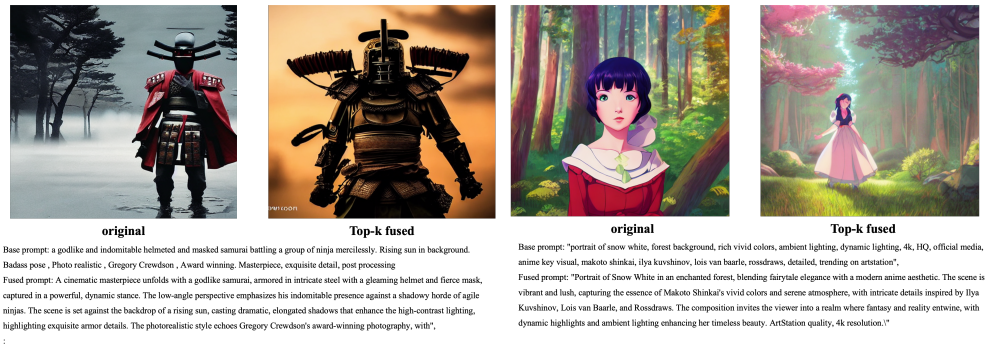

Figure 6: Qualitative examples of images generated using persona-guided prompt rewriting. The left two images depict samurai-style scenes generated from persona-rewritten prompts, while the right two images show anime-style portraits in natural settings. The diverse visual outputs demonstrate how persona rewrites introduce stylistic variations while maintaining semantic consistency with the original prompt.

However, our work also has several limitations. First, the overall dataset size used for training and evaluation is relatively small, which may limit the generalizability of our findings. Second, the current fusion strategy relies on a simple softmax-based weighting scheme that may not fully capture complex interactions between personas. In future work, more sophisticated fusion strategies such as learned weighting mechanisms or attention-based fusion could be explored. Additionally, expanding the scale of training data and incorporating more diverse personas may further improve the robustness and flexibility of the proposed framework.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[2] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[5] Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. Editing knowledge representation of language model via rephrased prefix prompts, 2024.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[7] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation, 2022.

[8] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language, 2022.

[9] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models, 2023.

A Appendix: Team Contributions

Seonga Choi: Responsible for prompt fusion implementation, writing the Method to Analysis sections of the paper, and contributing to prompt design.

Bean Cho: Responsible for prompt dataset creation, image generation, and writing the paper sections from Introduction to Observation. Also contributed to prompt design.