Improving Text-to-Image Alignment Using Image Captioning

Hyoryung Kim[1], **Seunghyeon Lee**[2], **Myungjin Lee**[3], **Hyemin Boo**[4]

1 Introduction

Image generation models have brought groundbreaking advancements to the field of artificial intelligence and have demonstrated their applicability across various domains. With these technological advancements, the accuracy with which generated images reflect user intent has emerged as a critical research challenge. Despite extensive research, current image generation models still struggle to accurately reflect complex prompts or relational expressions. While simple prompts are processed well, queries with multiple object attributes or complex relationships face composition generalization issues, leading to omissions or distortions. Additionally, models often suffer from artifact problems, such as unintended elements or misalignment with the prompt, resulting in unnatural images. These challenges hinder models from fully capturing user intent, especially in complex tasks, reducing consistency and accuracy.

1.1 Related Work

Masking to Improve Myung et al. (1) used inpainting to refine erroneous regions in generated images via LLM-based prompt adjustment. However, their approach is limited to well-defined objects and requires user intervention for masking. We propose an automated method that evaluates alignment between the generated image, user prompt, and image captions to determine whether to regenerate the entire image or modify specific objects.

Mitigating Compositional Issues Zarei et al. (2) demonstrated that CLIP's text embedding space is inadequate for capturing complex compositional relationships. To address this, they applied a linear transformation to CLIP's representation space, but emphasized the need for advanced fine-tuning, attention mechanisms, or hybrid models. Building on this insight, our research adopts a hybrid model approach to overcome the limitations of individual models and improve image generation quality.

1.2 Research Objectives

This study aims to improve text-to-image alignment by identifying discrepancies between the user prompt and the generated image caption. Based on these differences, the model determines whether to regenerate the entire image or modify specific regions. Our method automatically detects which parts of the prompt were not accurately reflected and applies targeted masking without manual intervention. The masked areas are refined through inpainting, resulting in images that better align with the user intent while reducing compositional errors and visual artifacts—without requiring model retraining.

2 Methodology

2.1 Data

User Prompt The PartiPrompts dataset is used as user-input prompts for evaluating. This dataset contains a diverse range of prompts across different topics and styles, allowing for an objective assessment of the model's generative performance.

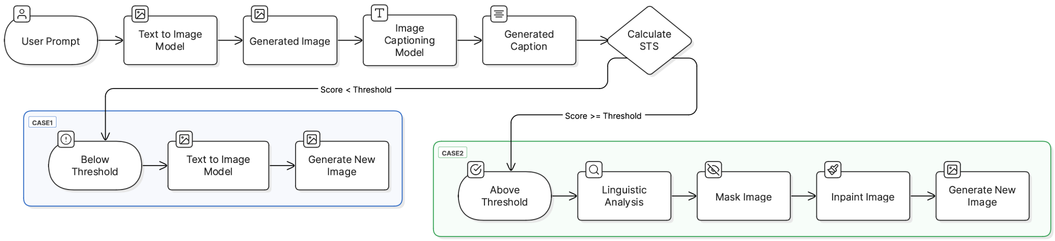

Figure 1: Overview of the proposed pipeline for aligning generated images with user prompts using image captioning and inpainting. See appendix A for pseudo-code explanation.

Image Generation The initial seed value influences image quality, human preference alignment, and text artifacts, resulting in variations in FID scores even with the same prompt (3). To ensure consistency, we adopt Xu et al.'s approach by selecting optimal seeds and using a fixed seed set. Accordingly, we utilize Stable Diffusion XL (SDXL), which generates high resolution and well detailed images based on user prompts while supporting a fixed seed function to maintain consistency.

2.2 Captioning

After generating an image based on the user prompt using SDXL, we leverage one of the latest state of the art model BLIP-2, to generate the caption for the generated image.

2.3 STS score

Then an STS model assesses the alignment between the user prompt and the generated image by computing the similarity score between the image caption and the user prompt. If the score falls below a threshold k (Case1), the image is fully regenerated using a TTI model. Otherwise (Case2), specific objects to be corrected are identified, masked, then refined via inpainting.

2.4 Case 1: Prompt Engineering and Case 2: Text-Based Object Modification

In Case1, TTI model receives the user request, the image caption, and the original image to regenerate an image that better aligns with user intent by incorporating textual differences. Specifically, the prompt is structured as: *"The user prompt was X, but the generated image was captioned as Y. Considering the difference between X and Y, regenerate the image to align with the user prompt."* This approach guides the model to produce images that more accurately reflect user intent.

In Case 2, the model identifies and masks target objects using both textual and visual information. The two proposed approaches explore different methods for locating objects to be modified, with the most effective strategy to be selected through further experimentation. Approach 1 (Attribute-Based Dictionary) employs word tokenization, POS tagging, and synonym filtering to extract key objects based on descriptive attributes. Approach 2 (Relationship-Based Graph) leverages graph-based learning inspired by SHINE (6), using tokenization, POS tagging, and NER to analyze sentence structure and identify differences. Once the target objects are identified, the SAM (7) model masks them, and Stable Diffusion performs inpainting to generate context-aware updates based on the user prompt. See Appendix B for details.

These processes not only automate image correction but also enhance semantic consistency and visual coherence—without the need for additional model training.

3 Evaluation

To evaluate the improvement of images regenerated using Case 1 or Case 2, we employ CLIP-score and FID metrics. These measures assess performance gains in terms of text alignment and image quality compared to the original image.

References

[1] Myung, Jiyoon, & Park, Jihyeon . "Inpaint Biases: A Pathway to Accurate and Unbiased Image Generation." arXiv preprint arXiv:2405.18762, 2024.

[2] Zarei, Arman & Rezaei, Keivan & Basu, Samyadeep & Saberi, Mehrdad & Moayeri, Mazda & Kattakinda, Priyatham & Feizi, Soheil. *Understanding and Mitigating Compositional Issues in Text-to-Image Generative Models*. arXiv preprint arXiv:2406.07844, 2024.

[3] Xu, Katherine, Zhang, Lingzhi, & Shi, Jianbo. "Good Seed Makes a Good Crop: Discovering Secret Seeds in Text-to-Image Diffusion Models." arXiv preprint arXiv:2405.14828, 2024.

[4] Kim, Yunji, & Lee, Jiyoung, & Kim, Jin-Hwa, & Ha, Jung-Woo, & Zhu, Jun-Yan, "Dense text-to-image generation with attention modulation" Proceedings of the IEEE/CVF International Conference on Computer Vision, (7701–7711), 2023.

[5] Han, Xu & Jin, Linghao & Liu, Xiaofeng & Liang, Paul Pu "Progressive compositionality in text-to-image generative models." arXiv preprint arXiv:2410.16719, 2024.

[6] Wang, Yaqing, Wang, Song, Yao, Quanming, & Dou, Dejing. "Hierarchical Heterogeneous Graph Representation Learning for Short Text Classification." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 8203–8214, 2020. arXiv preprint arXiv:2004.12085.

[7] Kirillov, Alexander, Mintun, Eric, Ravi, Nikhila, Mao, Hanzi, Rolland, Chloe, Xiao, Tete, Whitehead, Spencer, Berg, Alexander C., Lo, Wan-Yen, Dollar, Piotr, & Girshick, Ross. "Segment Anything." arXiv preprint arXiv:2304.02643, 2023.

Appendix

A Pseudo Code

This appendix provides the pseudo-code for our proposed method, which refines image generation by evaluating alignment between the generated image and the user prompt. The algorithm determines whether to fully regenerate the image or apply inpainting to specific objects based on similarity scores.

Algorithm 1 Improving Text-to-Image Allignment Using Image Captioning

```
$user\_prompt \leftarrow$ get_user_prompt(PartiPrompts)
$seed \leftarrow$ select_best_seed()
$generated\_image \leftarrow$ generate_image($user\_prompt, seed$)
$image\_caption \leftarrow$ generate_caption($generated\_image, BLIP$-2)
$similarity\_score \leftarrow$ compute_STS($user\_prompt, image\_caption$)
if $similarity\_score < k$ then
    Case 1: Image scene mismatch → Regenerate full image using prompt engineering
    $regenerated\_image \leftarrow$ regenerate_image($generated\_image, prompt, seed$)
else
    Case 2: Partial mismatch → Identify incorrect objects and inpaint
    $incorrect\_objects \leftarrow$ detect_incorrect_objects($user\_prompt, image\_caption$)
    $masked\_image \leftarrow$ apply_masking($generated\_image, incorrect\_objects, SAM$)
    $regenerated\_image \leftarrow$ inpaint($masked\_image, stable\_diffusion$)
end if
$original\_score \leftarrow$ evaluate_image($generated\_image$, metrics={CLIP, FID})
$improved\_score \leftarrow$ evaluate_image($regenerated\_image$, metrics={CLIP, FID})
```

B Detailed Methodology of Case 2

This appendix describes the process applied when the STS score between the generated image and the image caption exceeds the threshold k. The two approaches presented below are among the proposed methods for identifying objects to be modified in Case 2. Through further experimentation, the most effective methodology for performance improvement will be selected. Additionally, this appendix includes a description of the subsequent inpainting process.

B.1 Approach 1: Attribute-Based Dictionary

We propose a methodology to identify and mask target objects by utilizing word tokenization, POS tagging, and synonym filtering. First, each sentence is tokenized into words, and differences between tokens are identified. Then, a POS tagging model is applied to classify the differing words. If the differing word is an adjective or adverb, we determine the noun or key entity it modifies using a dictionary-based approach and store the identified noun or entity as an object. To ensure meaningful differentiation, a synonym filtering mechanism is employed to exclude words with minimal semantic differences from the object list, allowing only significant variations to be retained. By integrating these steps, our approach improves the identification of masking targets through POS tagging, leading to more accurate and context-aware masking applications.

B.2 Approach 2: Relationship-Based Graph

Inspired by the SHINE (6) paper, which proposes a method for constructing and learning a hierarchical heterogeneous graph that considers sentence-level similarity using GNNs, we aim to adopt a similar approach. Using tokenization, POS tagging, and Named Entity Recognition (NER), we generate a graph structure and dynamically establish relationships between sentences to learn their similarities. This dynamic graph structure enables improved learning performance even with a small number of labeled samples. We seek to leverage the key features and advantages of such graph-based approaches by representing user prompts and image captions as graphs. By enhancing the expressiveness of each sentence, this method facilitates effective comparison, allowing us to identify differences and extract the corresponding objects.

B.3 Inpaint

Image masking is the process of concealing or removing specific objects or regions within an image. By leveraging a pre-selected object detection model, relevant objects can first be identified, and then SAM (7) can be used to automatically detect and mask those objects within the image. The masked image is fed into the Stable Diffusion Inpainting model, along with the user's original prompt as a text prompt. The model then performs inpainting on the masked area, filling it in according to the user's desired content.

C Validity of Study

This appendix provides theoretical foundations relevant to our approach, drawing insights from prior studies that have demonstrated effective techniques for improving image generation quality and compositional understanding. By leveraging these methodologies, our framework aims to enhance text-to-image alignment while addressing challenges related to compositional generalization and visual artifacts.

DenseDiffusion (3) demonstrated that modifying intermediate attention maps in pretrained Text-to-Image models can effectively enhance image quality and text fidelity. This study underscores the feasibility of optimizing image generation without altering the underlying model weights or requiring additional retraining.

Progressive Compositionality (4) proposed a method for quantitatively evaluating text-image similarity during data generation, enhancing the composition of object interactions and spatial relationships. This approach enables the model to distinguish between semantically similar yet structurally different sentences, such as "A woman chases a dog" and "A yellow dog chases a woman," thereby improving its understanding of object relationships and ultimately enhancing text-to-image generation performance. Similarly, in Case 2, our method aims to improve image generation by enabling the model to better comprehend object interactions and spatial relationships in sentences through a novel approach. By refining this understanding, our framework seeks to generate images that more accurately reflect user intent while addressing compositional generalization challenges.

Building upon these insights, the proposed framework is inspired by DenseDiffusion, which demonstrated that image generation performance can be improved without additional retraining. Similarly, the Progressive Compositionality study highlights the importance of text-to-image similarity in guiding models to better understand compositional relationships. Drawing from these ideas, our research develops a novel approach that enhances compositional generalization and text-image alignment. By leveraging these principles, our framework aims to construct a more robust image generation pipeline that accurately reflects user intent while mitigating challenges related to compositional generalization and unintended visual artifacts.