Iterative Prompt Refinement for Text-to-Image Generation: A Multi-Metric Evaluation Study

Manjula R[1], Hemalatha S[2]

[1]Lincoln University College, Petaling Jaya,Selangor Darul Ehsan-47301, Malaysia.

[2]Computer science and business systems Panimalar engineering college, Bangalore Trunk road, poonamallee, Chennai Tamil Nadu, India.pincode 600123.

Email: pithemalatha@gmail.com, manjulasaravanaperumal@gmail.com

Abstract: Recent advancements in Generative Artificial Intelligence have significantly accelerated the development of text-to-image generation systems capable of producing high quality and visually realistic images from natural language prompts. Despite these advancements, the quality and accuracy of generated images remain highly dependent on the design of the input prompt, making prompt engineering a critical and challenging task. Users often need to modify prompts to achieve the desired output, which is both time consuming and computationally inefficient. This work proposes an efficient iterative prompt refinement framework based on BLIP (Bootstrapping Language-Image Pretraining) to improve the alignment between textual prompts and generated images. Unlike traditional approaches that rely on Large Language Models (LLMs) for prompt refinement, the proposed system utilizes a vision language model to analyze the generated image and extract a semantic caption. This caption is compared with the original or previously refined prompt to identify missing or mismatched details, enabling the system to progressively generate improved prompts. The iterative feedback loop allows the system to gradually enhance both semantic alignment and visual quality of the generated outputs. The system also achieves better efficiency compared to conventional methods by leveraging semantic feedback directly from generated images.

Keywords: Inverse prompt generation; Vision Language models; CLIP; BLIP; Prompt learning

Introduction

Artificial Intelligence has gain rapid development in recent years, especially in the field of Generative AI. The main purpose of Generative AI is to produce new content such as images, text, music etc. Among these text to image generation models have become very popular because they generate images from natural language prompt. Models like Stable Diffusion, DALL-E and Midjourney have proved that AI can create high quality images using text description only. Use of these technologies have increased significantly in the field of digital art, marketing, gaming, design and creative industries. But the major challenge is prompt engineering [1-2]. The quality of the generated image is highly dependent on the clear and descriptive prompt. Even a single change in wording can lead to generation of a completely different image. Therefore, the user needs to modify the prompt several times to get the desired output. To overcome this, researchers are developing the techniques for automated prompt improvement. Their main aim is to make the system analyze the generated image and improve the prompt in next iterations for better output. But the existing methods use simple keyword addition or rule based modifications which do not capture the semantic difference of the prompt accurately.

Related work

The combination of BLIP and CLIP provides a strong semantic feedback mechanism for prompt refinement. BLIP helps in understanding what has been generated, while CLIP helps in evaluating whether the generated image matches the original prompt. Therefore, BLIP and CLIP complement each other in the proposed framework. BLIP is responsible for image understanding and caption generation, whereas CLIP is responsible for semantic evaluation. This combination enables the system to automatically improve prompts without requiring repeated manual intervention from the user.

Table 1. Comparison of Related work in text to image prompting

REF. No	Methodology	Identification of gaps and Limitations
[3]	· VAE encoder-decoder with denoising U-Net in latent space · Cross-attention for text conditioning via CLIP	· Sensitive to prompt phrasing · No feedback mechanism for quality control
[4]	· Cosine similarity between CLIP image and text embeddings · Reference-free, validated on captioning benchmarks	· Biased toward short, simple prompts · Poor correlation with human preference for complex scenes
[5]	· 18k images annotated with multi-attribute human feedback · Classifier trained to predict text alignment and aesthetics	· High annotation cost limits scalability · Does not study feedback across refinement iterations
[6]	· Benchmarks 7 visual reasoning skills using VQA models · Tested on DALL-E 2, Stable Diffusion, and others	· Skill-specific not generalized to open-domain prompts · VQA models introduce evaluation noise
[7]	· CLIP embeddings for combined semantic and aesthetic scoring · Benchmarked against human ratings on aesthetic datasets	· Correlation with human preference varies by prompt type · No analysis across iterative generation loops
[8]	· Few-shot fine-tuning on 3 to 5 images of a specific subject · Evaluated using DINO and CLIP-I for subject fidelity	· Requires per-subject fine-tuning; not generalizable · No automated prompt refinement or quality feedback loop

Despite the advancements, a key challenge in text-to-image generation remains prompt engineering is given in Table1. The wording and structure of prompts significantly influence the generated output. Users often need to modify the prompt iteratively to achieve the desired results. Motivated by this challenge, this work explores and iterative prompt refinement framework that

integrates text-to-image generation, image captioning and large language models to automatically improve prompts and enhance alignment between the generated images and textual descriptions[3-8].

This work proposed a iterative prompt refinement system which combines multiple AI models. First, the image is generated using Diffusion model. Then BLIP image captioning model analyzes the image and tells what is actually generated in that image. After that the Large Language Model compares the original prompt and the BLIP caption and generates an improved prompt which describes the missing concepts.

This process is repeated for multiple iterations which helps the system to gradually improve the prompt and generate better images[9]. To evaluate the performance of system, two metrics were used:

- CLIP Score – Measures the semantic similarity of image and prompt.
- Aesthetic Score – Estimates the visual quality of the image.

Here the intelligent refinement approach were used to a baseline method in which only generic keywords are added. Experimental results show that iterative prompt refinement improves both image quality and prompt alignment.

CLIP has become an important model in modern text-to-image research because it provides a way to evaluate how closely an image matches its textual prompt[10]. Unlike traditional image quality metrics that only measure realism, CLIP focuses on semantic alignment between text and image. It maps both text and images into the same embedding space and calculates their similarity. Because of this property, CLIP is widely used in text-to-image systems for prompt ranking, prompt optimization, image retrieval, and image-text matching[11]. In many recent diffusion-based studies, CLIP Score is used as an objective metric to compare generated images with their corresponding prompts. Higher CLIP scores generally indicate that the generated image is more semantically aligned with the intended prompt.

Similarly, in proposed work BLIP (Bootstrapping Language-Image Pretraining) which focuses on vision language understanding tasks such as captioning images, visual question and answering. BLIP Models can generate descriptive captions for images, allowing systems to analyze generated images and compare them with the intended textual prompts.

BLIP is especially useful in iterative prompt refinement systems because it acts as an automatic observer of the generated image. After an image is generated, BLIP can describe what visual content is actually present in the image. This allows the system to compare the BLIP-generated caption with the original prompt and identify missing concepts, attributes, or scene details. For example, if the original prompt contains concepts such as ‘red bird’ and ‘sunset’, but the generated image only contains a bird on a branch, BLIP can help identify that the color and lighting details are missing. These missing concepts can then be added back into the prompt for the next iteration[12-13].Recent research has also explored large-scale multimodal datasets for training vision-language models[14-15]. These datasets have helped the training of large multimodal models capable of learning complex relationships between images and textual descriptions.

Conclusions

This work proposed an iterative prompt refinement framework which is designed to improve the alignment between prompt given to the text-to-image generation system and the generated image. System integrates multiple AI models like Stable Diffusion for image generation, BLIP for image

captioning, LLM for prompt refinement. CLIP and Aesthetic scoring model for image evaluation. The proposed approach follows a feedback-based mechanism in which generated images are analyzed and the prompts are automatically refined. This refinement process helps in identifying the difference between the original prompt and the image generated. The system uses multiple iterations to improve the prompts gradually to generate better images. The experimental results shows that LLM-based refinement method performs better than keyword-based refinement method. Evaluation metrics such as CLIP score and aesthetic score shows how iterative prompt refinement improves the image quality and prompt alignment. This result demonstrates that vision models, language models and evaluation techniques can be combined to create an automated prompt optimization framework. Proposed framework improves the text to image generation workflow as well as reduce the effort of manual prompt engineering. Overall, this work shows how multimodal AI can be integrated to iterative feedback loop to improve the performance and reliability significantly for generative image generations. Overall, the proposed framework provides a robust, scalable, and practical solution for automated prompt optimization in text-to-image generation systems, making the process more efficient, accessible, and user-friendly.

References

1. Mohd Hafizul Afifi Abdullah, et al., "Systematic Literature Review of Information Extraction From Textual Data: Recent Methods, Applications, Trends, and Challenges" IEEE Access, 10535 – 10562, 2023. 10.1109/ACCESS.2023.3240898
2. Hao Ma, et al., "CLIP-Flow: Decoding images encoded in CLIP space", Computational Visual Media , vol 10,pp 1157 – 1168, ,2024. 10.1007/s41095-023-0375-z
3. Robin Rombach, et al., "High-Resolution Image Synthesis with Latent Diffusion Models", Computer Vision and Pattern Recognition, pp.10684–10695, 2022. https://doi.org/10.48550/arXiv.2112.10752
4. Hessel, J, et al.," CLIPScore: A Reference-Free Evaluation Metric for Image Captioning", EMNLP pp. 7514–7528, 2021. https://doi.org/10.48550/arXiv.2104.08718
5. Liang, V, et al.," Rich Human Feedback for Text-to-Image Generation", IEEE-CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 10.1109/CVPR52733.2024.01835
6. Cho, J, et al., "DALL-Eval: Probing Reasoning Skills of T2I Generation Models", Computer Vision and Pattern Recognition, 2023. https://doi.org/10.48550/arXiv.2202.04053
7. Wang, J, et al., "Exploring CLIP for Assessing the Look and Feel of Images", The Thirty-Seventh AAAI Conference on Artificial Intelligence, Vol. 37, 2023. https://doi.org/10.1609/aaai.v37i2.25353
8. Ruiz.N, et al., "DreamBooth: Fine Tuning T2I Diffusion Models for Subject-Driven Generation". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 17-24, 2023. 10.1109/CVPR52729.2023.02155
9. Gaojie Wu, et al., "Adaptive Weight Generator for Multi-Task Image Recognition by Task Grouping Prompt", IEEE Transactions on Multimedia, vol 26, pp 9906 – 9919, 26, 2024. 10.1109/TMM.2024.3390152
10. Yuanbo Wen, et al., "All-in-One Weather-Degraded Image Restoration Via Adaptive Degradation-Aware Self-Prompting Model", IEEE TRANSACTIONS ON MULTIMEDIA, vol 27, pp 3343 – 3355, 2025. 10.1109/TMM.2025.3535316
11. Atsushi Takada, et al., "Example-Based Conditioning for Text-to-Image Generative Models", IEEE Access, vol 12, pp 162191 – 162203, 2024. 10.1109/ACCESS.2024.3486055

12. Yu-Ming Shang, et al., “Multi-Modal Relation Extraction Enhanced by Out-of-Text Visual Information”, IEEE Transactions on Audio, Speech and Language Processing, vol 33, pp 3428 – 3442, 2025. 10.1109/TASLPRO.2025.3589892
13. Shaofeng Ming, et al., “PromptTrace: A Fine-Grained Prompt Stealing Attack via CLIP-Guided Beam Search for Text-to-Image Models”, Symmetry, vol *18*, 2026 https://doi.org/10.3390/sym18010161
14. William Brach, et al., “The Effectiveness of Large Language Models in Transforming Unstructured Text to Standardized Formats”, IEEE Access, vol13, pp 91808 – 91825, 2025 10.1109/ACCESS.2025.3573030
15. Yiming Lei, et al., “Prompt learning in computer vision: a survey”, Frontiers of Information Technology & Electronic Engineering, vol 25, pp 42–63, 2024. https://doi.org/10.1631/FITEE.2300389