Prompt Optimizer for Text-to-Image Generation: Utilizing Chain-of-Thought Reasoning to Optimize Prompt Design

Anonymous ACL submission

Abstract

Text-to-image generation models have attracted a lot of attention because of their ability to create images from text prompts. However, natural language prompts are often concise and ambiguous, making it difficult to consistently produce high-quality images that meet user expectations. In this work, we investigate the capabilities of large language models in image generation and introduce a method, Prompt Optimizer, which utilizes large language models for prompt augmentation. Using the Pick-a-Pic and CoCo datasets, our experiments employ an improved aesthetic predictor and PickScore as evaluation metrics to evaluate image quality and text-image relevance. Compared to direct generation and other text-to-image prompt generation methods, our method has seen significant improvements in relevance and generation quality.

1 Introduction

In recent years, with the continuous advancement of text-to-image generation models and natural language processing technologies, models such as Stable Diffusion (Rombach et al., 2022) and DALL-E (Ramesh et al., 2022) have demonstrated the ability to generate rich and diverse images from user-provided text prompts. This development not only lowers the cost of artistic creation but also offers large-scale, high-quality datasets for downstream scientific research tasks (e.g., (Kirstain et al., 2023)), particularly in the field of computer vision.

In the image generation process, users first provide text prompts to describe the image they wish to create. They then adjust model hyperparameters and obtain different outputs by modifying random seeds. However, the inherent ambiguity of natural language can make it challenging, especially for novice users, to craft effective prompts that guide the model in generating the desired images. Finding the right keywords often requires extensive trial and error to achieve high-quality results. To address this, previous studies have offered guidelines for text-to-image generation models (Liu and Chilton, 2022), emphasizing the importance of focusing on the main content of the prompts.

Previous studies have introduced various tools to assist users in generating prompts, including automatic prompting techniques (Wang et al., 2023) and systems like PromptMagician (Feng et al., 2023). However, these methods often require users to perform additional steps, which inevitably increases the learning curve and the overall complexity for users.

To assist users in generating more effective prompts, we proposed Prompt Optimizer, a text prompt generation method that aligns with the guidelines for text-to-image generation. This approach focuses on the core content of the prompts. First, a large language model is employed to expand upon content that is either vaguely expressed or omitted when users provide natural language input. Then, the key elements are extracted using a content extraction technique. Our method is inspired by the Chain-of-Thought framework of large language models (Wei et al., 2022). By utilizing a series of intermediate reasoning steps, we leverage the prior knowledge of the model to enhance both the accuracy and effectiveness of the generated prompts. Our contributions are as follows:

- We analyzed the correlation between prompt words and natural language elements required by image generation models, identifying which components of natural language provide the most significant support for image generation tasks.

- We proposed a prompt optimization method for image generation models based on LLMs, enabling the automatic refinement of original prompts, including natural language elements.

Figure 1: Optimization process of the prompt using our method, where missing contents were supplemented.

2 Related Work

Chain-of-thought The original *Chain-of-Thought* approach aims to enhance the performance of tasks related to arithmetic, common sense, and symbolic reasoning (Wei et al., 2022). Building on this, the *Tree of Thoughts* model (Yao et al., 2024) introduces various reasoning pathways and self-evaluation mechanisms, further improving performance in more complex tasks that demand global backtracking. Additionally, *Sparks of Artificial General Intelligence* (Bubeck et al., 2023) (Feng et al., 2024)highlights GPT-4's ability to excel not only in language but also in solving novel, challenging tasks across diverse domains such as mathematics, coding, vision, medicine, law, and psychology—without requiring any specialized prompts. Remarkably, GPT-4's performance in these areas approaches human-level capabilities.(Achiam et al., 2023) This suggests that it is feasible to generate prompts using the GPT-4 model with the Chain-of-Thought approach.

Prompt Optimization In PromptMagician(Feng et al., 2023), the authors propose an interactive system to assist users in optimizing input prompts for text-to-image generation models. The system retrieves similar images and prompt words from a large-scale dataset, DiffusionDB(Wang et al., 2022), and identifies important keywords. However, in practical use, this method does not fully simplify the image generation process, as it still requires continuous human-system interaction to optimize the generated results.

RePrompt(Wang et al., 2023) introduces an automatic method that similarly retrieves images and prompt words from DiffusionDB and identifies relevant keywords. However, this approach focuses primarily on the emotional expression within prompts and does not emphasize the key concepts highlighted in the *Design Guidelines for Prompt Engineering in Text-to-Image Generative Models*.

In our approach, we place significant emphasis on automated generation and refinement of keywords to optimize the quality of generated images. Our method automates the identification and enhancement of essential keywords, allowing the prompt construction process to be both efficient and user-friendly.

3 Methodology

We begin by outlining our research content and the overarching methodology that guides our approach. Following this, we present a detailed introduction to our studies and the technical methods used our work.

3.1 Research Content

This study explores the potential of leveraging large language models to optimize and enhance text-to-image generation, addressing key challenges in producing high-quality images based on natural language prompts. Our research encompasses several main areas:

First, we analyze the limitations of current text prompts in image generation, highlighting how overly concise or incomplete prompts can affect the quality, relevance, and consistency of gener-

ated images. This analysis lays the groundwork for developing an optimization framework. Based on these insights, we propose the *Prompt Optimizer* method, which uses the reasoning capabilities of large language models to automatically supplement and refine prompts. This method decomposes the prompt optimization process into multiple subtasks, including key information extraction, missing content completion, and structured prompt generation, to achieve progressive improvement of prompts.

To verify the versatility of Prompt Optimizer, we applied this method across different text-to-image generation models (e.g. Stable Diffusion and DALL-E 2) to examine its adaptability and robustness. These multi-model experiments confirm the method's wide applicability and demonstrate its effectiveness across various architectures. Additionally, we conducted systematic experiments using the Pick-a-Pic dataset, quantifying the impact of Prompt Optimizer on image aesthetics and text-image relevance using evaluation metrics such as Improved Aesthetic Predictor and PickScore.

Finally, we explored the potential applications of Prompt Optimizer in areas such as digital art, content creation, and virtual environment design, and discussed future research directions, including further improvements in adaptability and testing across broader datasets and generation models.

3.2 Methodology

In response to the first question, previous studies have highlighted that the subject and style of the image are the keywords with the most significant impact on the quality of the generated images(Liu and Chilton, 2022). These keywords not only influence the visual characteristics of the output, but also determine whether the generated image accurately reflects the user's intent. However, a potential challenge arises when users, especially those without domain-specific knowledge, attempt to generate images using natural language. Due to the inherent ambiguity and variability of natural language(Beck et al., 2020), the prompts often contain only a subset of the essential keywords, which can result in the omission of critical information necessary for high-quality generation.Previous experiments frequently relied on users to evaluate the quality of the generated images and provide feedback, which was then used to refine and improve the subsequent prompts.

To describe our method, we divided the prompt optimization task into the following tasks:

In the **Task 1**, Our method processes the original prompt in this step, yielding two outputs: the first is the extracted original content, denoted as R'_j, and the second is a vector, η, which indicates whether the original prompt includes the corresponding elements.:

$$(R'_j, \eta) = \mathcal{E}(P_j, IP) \tag{1}$$

In this formula, $\mathcal{E}$ represents the extraction process. The input consists of the j-th original prompt P from the dataset and the task-specific prompt IP. The extraction process yields two outputs: a 4-dimensional vector η, where a value of 0 in any dimension indicates that the corresponding content was not extracted, and the extracted textual content R'_j, representing the refined result.

In addition, we instruct the LLM to process and introduce the extracted content. The response generated by the large model at this stage is not directly included in the final optimized prompt. Instead, the Chain-of-Thought method is employed to leverage the contextual and prior knowledge of the large model. The corresponding formula is as follows:

$$Y = LLM(X, C) = \arg\max_{y \in \mathcal{Y}} P(y \mid X, C) \tag{2}$$

The equation describes how a large language model (LLM) leverages both the input X and the context C to generate the output Y. Specifically, the model seeks to maximize the conditional probability $P(y \mid X, C)$ over all possible outputs $y \in \mathcal{Y}$. Here, X represents the current input, C denotes the contextual information, and $\mathcal{Y}$ is the set of all potential outputs. The final output Y is determined as the one that achieves the highest conditional probability, showcasing the model's ability to utilize prior knowledge and contextual understanding for effective generation.

In the **Task 2**, we instruct the LLM to supplement any potentially missing content based on the previously extracted information. The corresponding formula is as follows:

$$R_j = R'_j + LLM(IP + \eta(P_j)) \tag{3}$$

Among them, R'_j represents the extracted content, while η is a 4-dimensional vector indicating whether each component has been successfully extracted. If any component is missing, the large language model will generate the missing content based on the designed prompt IP, and combine it

with the originally extracted content R'_j to produce the response R_j.

In the **Task 3**, We extract the response R_j generated in the previous step and derive the prompt P_j, which is used for image generation:

$$P_j = \mathcal{E}(R_j, IP) \tag{4}$$

Figure 2 illustrates the process by which our method supplements and optimizes the original prompt to produce a more comprehensive and effective input for image generation. In the initial state, the prompt provided by the user often contains only a brief description of the main subject of the image, lacking details about other important aspects that contribute to the image's overall quality and relevance.

To address these limitations, our approach utilizes a large language model to automatically analyze and enrich the prompt by identifying and incorporating missing elements. Liu et al.(Liu and Chilton, 2022) suggest that when generating an image, it is important to focus on key aspects such as the subject, theme, style, and other relevant details.

Regarding the determination of the impact of each component of the prompt on the quality of the final generated image, we propose the following formula:

$$\mathbf{Q}^* = \arg\max \mathbf{w} Q(S, F, B, A) \tag{5}$$

Among them, Q represents the quality score of the generated image, while S, F, B, and A denote the impact of prompt components related to the subject, features, background, and artistic style, respectively, on the quality of the generated image. The variable $\mathbf{w}$ is a 4-dimensional vector representing the selection strategy for the prompt components. By maximizing Q, we determine whether to retain the corresponding content in the generated prompt. The definition of w is given by the following formula:

$$\mathbf{w} = [w_s, w_f, w_b, w_a], \quad w_i \in \{0, 1\} \tag{6}$$

The value of w_i is either 0 or 1, representing whether this component is retained in the final selection strategy.

After defining the task and the selection strategy, we proceed to the experimental section to validate our approach and determine the specific values of the selection strategy.

4 Experiment

4.1 Setup

Dataset. In this study, we selected the Pick-a-Pic, COCO 2014, and COCO 2017 datasets for experimentation. The Pick-a-Pic dataset is specifically designed to evaluate text-to-image generation models, providing pairs of descriptive text prompts and corresponding images across various categories. Additionally, it includes a scoring system called PickScore, which enables objective comparison of the quality of generated images and their relevance to the provided prompts. The COCO 2014 and COCO 2017 datasets are extensively used in computer vision research, containing over 200,000 diverse images across multiple categories, including people, animals, scenery, and transportation. These datasets offer a wide range of scene and object types, which are essential for improving the generalization ability of image generation models.

Text-to-image model. We chose Stable Diffusion 3.0 (SD 3.0) as the generative model for this study. SD 3.0, an advanced diffusion model, demonstrates strong generative capabilities, producing high-quality and detailed images, particularly in handling complex scenes and fine textures. The model supports multimodal inputs, including text, images, and labels, enabling efficient conditional image generation. This makes it well-suited for tasks such as text-to-image generation. Additionally, its open-source nature allows for customization and optimization according to specific requirements, offering high flexibility and scalability.

The entire experiment was conducted using the Stable Diffusion 3.0 model. For all generated images, we maintained consistent parameters, including the same random seed, iteration count, and CFG ratio, while generating images at a resolution of 1024x1024. This approach ensures that the prompt is the sole variable in the image generation process.

In parallel, we also selected SD 2.0 and DALL-E models for small-scale experiments to assess the robustness of our method across different generative models.

4.2 Result

For this study, we selected three key evaluation metrics to assess the quality and relevance of the images generated. The first metric, PickScore, is included in the Pick-a-Pic dataset and serves as

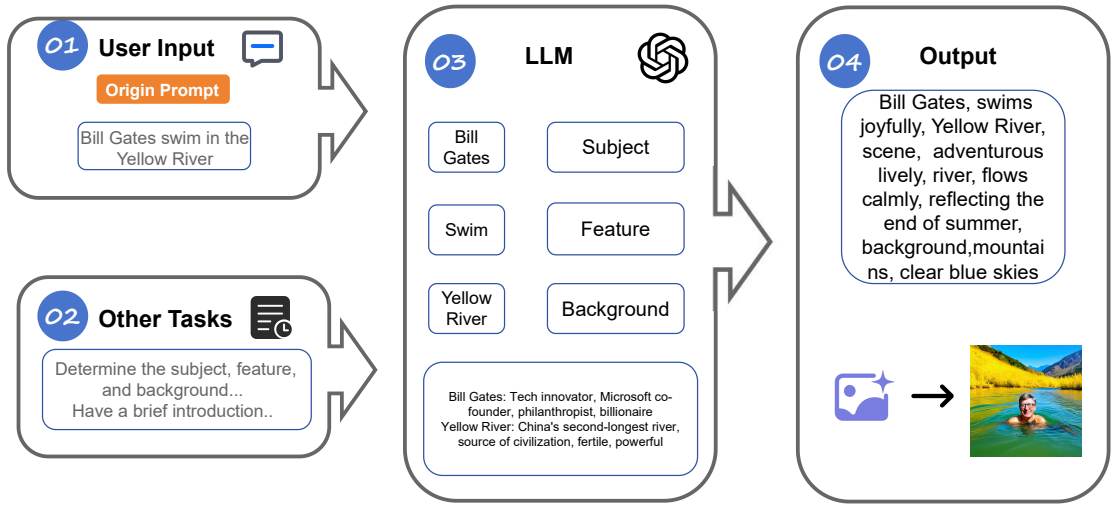

Figure 2: This figure illustrates how we leverage a large language model to optimize the original prompt. The original prompt contains only partial information. The "Introduce" task involves applying the Chain-of-Thought approach, after which the prior knowledge embedded in the large language model is used to supplement and refine the original prompt, enhancing its completeness and relevance.

an objective measure of both image quality and relevance to input prompts. PickScore allows for a standardized assessment of how well the generated images align with the expectations set by the prompt, making it an ideal metric for evaluating prompt-guided image generation models.

The second metric that we used is the improved aesthetic predictor(Dhar et al., 2011), which is specifically designed to evaluate the aesthetic quality of the generated images. Unlike purely objective measures, this predictor focuses on subjective aspects of image quality that are often crucial to human visual perception. By considering these factors, the Improved Aesthetic Predictor provides a nuanced understanding of the aesthetic appeal of generated images, making it a valuable tool in the context of text-to-image generation.

We chose Inception Score(Barratt and Sharma, 2018)as the scoring criterion primarily because it is concise and effective. The IS score evaluates the quality and diversity of generated images by calculating their classification probabilities in a pre-trained Inception network. A higher IS value indicates that the generated image has strong category information and clarity.

The first step of the experiment is to identify the specific keywords that need to be emphasized when using a large language model for prompt completion. For example, consider the formula we provided earlier:

$$\mathbf{Q}^* = \arg\max \mathbf{w} Q(S, F, B, A) \tag{7}$$

In the above equation, to determine the content that should be included in the optimal prompt, we need to systematically remove the subject, features, background, and artistic style information from the prompt. This allows us to test the impact of each content component on the quality of the generated image. To achieve this, we conducted a small-scale experiment, selecting 500 random prompts from the Pickapic dataset for comparative analysis. The results are shown in Table 1:

As can be seen in the table 1, the different prompt words included in the prompt have a noticeable impact on the generated results. Based on our scoring criteria, we define the quality of the image generation results from three perspectives: the correlation with the original prompt, represented by Pickscore; the overall quality of the image generation, represented by the average values of IPA and IS; and the stability of image generation, represented by the variance of the IPA and IS scores.

From this column of results, it can be seen that in the Pickscore (representing the first column),

Table 1: This table represents the impact of different contents in the prompt on the quality of image generation, where S, F, B, and A represent the subject, features, background, and artistic style, respectively.

Datasets	**Methods**	**Scores**					
		Pickscore	IPA Ave	IS Ave	IPA Var	IS Var	Q Value
PickaPic	Origin	/	15.1924	10.4584	103.2728	32.7592	/
	F+B+A	0.4759	19.2023	10.5711	121.9197	35.0822	6.65
	S+B+A	0.4758	19.2949	10.5845	117.5262	31.1142	7.14
	S+F+A	0.4845	18.2441	9.8058	125.6750	31.8859	5.89
	S+F+B	0.4296	18.8345	11.4459	93.2812	36.5210	7.26
	S+F+B+A	0.4887	18.5619	10.3748	131.6659	32.7690	5.99

prompts without background-related content show a relative decrease of about 10% in relevance compared to prompts with other content. However, the stability of the generated content has improved significantly. Specifically, the variance of IPA in the S+F+B+A prompt, which includes all content, decreased from 131.66 to 93.28 after removing background-related content, representing a reduction of approximately 29%.

To comprehensively consider these factors, we propose a combined evaluation metric for image generation quality.

$$Q = P\left(IPA + IS - \alpha\left(Var(IPA) + Var(IS)\right)\right) \tag{8}$$

This formula allows us to evaluate the optimal generation method by considering the prompt relevance, image quality, and stability of the generation process.

Here, we aim to balance the quality and stability of image generation. To achieve this, we set the weight α to 0.1, which allows a rough balance between these parameters of the same order of magnitude, ensuring that the weight values of quality and stability are appropriately balanced in the formula.

The last column of the table 1 represents the result of the calculation Q of the formula. From the value Q, we can determine the comprehensive evaluation of the impact of each relevant content in the prompt on the generated image. After considering the quality, stability, and relevance of the generation, we have concluded that the method of automatically completing prompts by removing artistic style-related content, namely the S+F+B prompts, is the most effective.

The results of our experiments on the three datasets, PickaPic, COCO2014, and COCO2017, are shown in Table 2:

The following results in table 2 are worth noting:

Using the previously defined formula, we comprehensively evaluate the correlation and aesthetic quality of the generated images, with the results presented in the final column, Q. Since Pickscore requires two images and a prompt as input parameters, it is not possible to obtain a rating when using the original image and prompt as inputs. Therefore, no Pickscore score is available for the original prompt, and correspondingly, the Q value is also unavailable for the original prompt, as it requires a Pickscore for correlation evaluation. Regarding the evaluation metric Q, our method significantly outperforms traditional model training methods, such as PromptMagician, across three different datasets. The difference between the two ranges from 6.36 to 7.43 on the PickaPic dataset used for image generation, representing a 16% improvement. The effect is even more pronounced on the COCO2014 and COCO2017 datasets, with the mean Q increasing from 2.58 to 7.92.

In Figure 3, we ran several experiments with a single prompt, and the results show that while some of the images generated by the Prompt Optimizer still had negative IAP scores, the overall area of the score was still better than the original method.

Regarding the quality of image generation, it can be observed from the IPA dataset that our generation method significantly outperforms the original dataset. The improvement in aesthetic predictor scores for the three datasets were 23%, 18%, and 19%, respectively. Although the PromptMagician generation method achieves higher IPA scores, its average Pickscore is only 43% of our method's average. This suggests that PromptMagician is constrained by the Diffusion DB dataset used for training, and after optimizing the original prompt, the resulting content may deviate considerably from the original prompt.This trend is particularly evident

Table 2: This table compares the performance of our method with other prompt generation methods and raw prompts across multiple datasets.

Datasets	**Methods**	**Scores**					
		Pickscore	IPA Ave	IS Ave	IPA Var	IS Var	Q Value
PickaPic	Origin	/	15.1924	10.4584	103.2728	32.7592	/
	PromptMagician	0.2618	28.5341	9.7634	111.3139	26.6703	6.36
	PromptOptimizer	0.4296	18.8345	11.4459	93.2812	36.5210	7.43
CoCo2014	Origin	/	9.5477	18.9431	49.2601	88.7491	/
	PromptMagician	0.1736	25.2156	15.1639	91.8183	77.3831	2.28
	PromptOptimizer	0.4517	11.2882	19.1551	44.3437	88.3283	7.72
CoCo2017	Origin	/	10.4565	18.1300	46.2994	77.1154	/
	PromptMagician	0.1395	24.4401	14.0787	94.0516	76.2918	2.78
	PromptOptimizer	0.4434	12.4577	18.1022	45.4382	75.2305	8.13

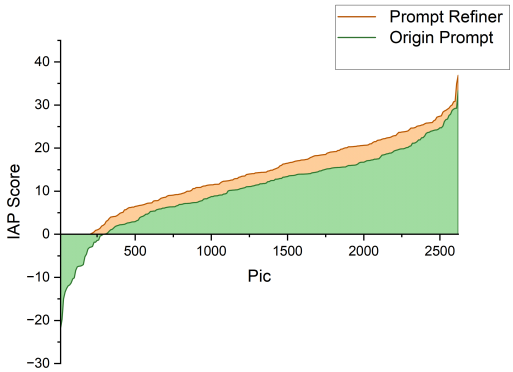

Figure 3: Comparison of Improved Aesthetic Predictor scores between the two methods, with all images ranked in ascending order of score.

across different datasets. Although the datasets we selected provide both images and annotated text, there is a fundamental difference between the PickaPic, COCO2014, and COCO2017 datasets.

The underlying reasons for this phenomenon may include several factors. First, the elevated average aesthetic score suggests an improvement in the perceived quality of individual images; however, the substantial increase in variance signals a loss in generation stability. This instability is likely due to the addition of generic or imprecise style prompts that may not perfectly align with the content and context of every original prompt. Such prompts, while enhancing aesthetic appeal in some cases, may conflict with the intended subject or theme, leading to unpredictable outcomes.

Additionally, the decline in PickScore points to a potential reduction in correlation between the generated images and their original prompts. This reduction may be a result of "priority interference," where the focus on stylistic elements overshadows or dilutes the content-specific guidance intended by the user. Since these artistic style prompts are automatically generated by the large language model—rather than being tailored by the user—their influence can sometimes dominate over the original content, creating a dissonance between the image and prompt. As a result, although the generated image might be visually appealing, it may not accurately reflect the user's original intention, leading to a lower PickScore.

Another contributing factor could be the inherent subjectivity associated with artistic styles. Artistic styles often encompass broad interpretations, making it challenging for the model to apply a style uniformly across different content types without specific guidelines. When the model applies a certain style automatically, it may accentuate particular visual aspects that do not necessarily align with the user's intended message, further impacting prompt-image alignment. This subjective application of style could explain the fluctuation in aesthetic scores and the decreased consistency in image generation.

In summary, while artistic style prompts can enrich the aesthetic appeal of generated images, they introduce challenges regarding consistency and prompt fidelity. The high aesthetic scores accompanied by increased variance and decreased PickScore highlight a trade-off between achieving visual quality and maintaining prompt alignment. These results suggest that when using automatically generated artistic styles, careful consideration is needed to balance aesthetic enhancement with the retention of core prompt content, ensuring that

Table 3: This table compares the performance of our method with other prompt generation methods and raw prompts across multiple datasets.

Datasets	**Methods**	**Scores**					
		Pickscore	IPA Ave	IS Ave	IPA Var	IS Var	Q Value
PickaPic	Origin	/	10.1653	10.0045	105.9577	33.2099	/
	PromptMagician	0.2360	21.2009	8.5963	105.2533	28.1316	/
	PromptOptimizer	0.3837	14.9215	11.0820	83.3579	33.4278	/
CoCo2014	Origin	/	8.5712	17.9607	31.9323	88.0926	/
	PromptMagician	0.1499	18.3210	12.4703	76.6914	68.1638	/
	PromptOptimizer	0.4455	10.7920	17.2218	35.7865	85.0113	/

the generated image remains true to the user's original intention.

4.3 Robustness Experiment

To further verify the robustness of the Prompt Optimizer, we conducted small-scale experiments on DALL-E 2. The experimental results are shown in Table 3. It can be seen from the results that although there are differences in the image generation mechanisms of these models, the Prompt Optimizer can always improve the generation quality of each model. The IAP scores have been significantly improved compared with the original prompts in both generation methods using two datasets, and the IS scores also have certain improvements on the Pickapic dataset. This multi-model adaptability demonstrates that the optimization strategy of the Prompt Optimizer does not rely on any specific generation model but can be effective across various architectures.

Prompt Optimizer demonstrates strong performance in terms of generation consistency and stability. We conducted variance analysis on the images generated by each experimental group to assess fluctuations in generation quality. The results indicate that the variance in image scores is significantly reduced following optimization with Prompt Optimizer, resulting in more stable outputs across various generation tasks. This implies that Prompt Optimizer not only optimizes image quality but also enhances the stability of the generation results, ensuring more consistent outcomes when generating images from the same prompt multiple times.

Although our method performs well in terms of scoring, it still sometimes achieves lower scores, which may be due to the following reasons. Due to constraints in training data, no current method can fully determine whether a generated image A is definitively superior to image B. Image detection algorithms can compare image quality by analyzing key factors such as distortion and gradient.

5 Conclusion

This paper proposes a new method Prompt Optimizer that leverages the reasoning capabilities of large language models to optimize user prompts. This includes extracting the information provided by the user and supplementing any missing parts of the prompt based on relevance. We randomly selected a subset of the Pick-a-Pic dataset to evaluate our method, using two metrics: PickScore and the Improved Aesthetic Predictor, to assess both effectiveness and stability. The results demonstrate that our method can effectively and consistently optimize the original prompts provided by users, while also validating the reasoning abilities of large language models in the context of image generation. This provides new insights into the application of large language models in the field of computer vision.

6 Limitation

Due to the limited availability of diverse datasets and standardized evaluation metrics in the field of image generation, verifying the robustness of our method across various datasets remains challenging. The existing datasets often lack the breadth and variety necessary to test how well the method generalizes to different types of prompts, content, and styles. Although the subset we used for evaluation is derived from the original dataset's training set, there is a potential bias introduced by the specific image generation techniques and stylistic choices within this dataset. Such biases may impact the generalizability of the evaluation metrics, limiting our ability to conclusively measure the robustness of our method across unseen data distributions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shane Barratt and Rishi Sharma. 2018. A note on the inception score. *arXiv preprint arXiv:1801.01973*.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*, pages 1657–1664. IEEE.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2024. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36.

Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2023. Promptmagician: Interactive prompt engineering for text-to-image creation. *IEEE Transactions on Visualization and Computer Graphics*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 36652–36663. Curran Associates, Inc.

Vivian Liu and Lydia B. Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23, New York, NY, USA. Association for Computing Machinery.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Yunlong Wang, Shuyuan Shen, and Brian Y Lim. 2023. Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–29.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.