Paint it, BLACK: a Novel Methodology for Prompting

Federico Torrielli[1]

[1]University of Torino, Department of Computer Science, Corso Svizzera 185 - 10149 Torino

Abstract

Latent Diffusion Models have recently emerged as the state-of-the-art approach for synthetic image generation. In the Web context, their adoption may significantly impact the way it is currently approached, from both sides of content generation and exploration. For example, future Web platforms may create alternative and personalised images for individual users or improve the accessibility for users with disabilities. However, due to the nascent stage of this research area, there remains a knowledge gap in effectively utilising these models, which can clutter the digital space with poor-quality AI-generated, thus diminishing the overall perceived impact and the user experience. To address this issue, we propose a novel methodology aimed at generating high-quality prompts with minimal user effort. In particular, we present BLACK (**B**ackground, **L**ighting, **A**menities, **C**ontext, and **K**inesis), a prompt generation model directly designed for achieving high-quality images satisfying a proposed set of five desiderata. We supplement our methodology proposal with illustrative examples, intended to clarify its mechanisms for the reader. As a second contribution, we publicly release a structured resource of prompts along with expected results.

Keywords

Stable Diffusion, Prompt Engineering, Text-to-image Generation, AI Art

1. Introduction and Related Works

Prompting in automatic image generation is a process that involves providing a textual description or instruction as input for image generation [1]. The provided description usually specifies the desired image features, such as objects or scenes to represent, and may include details about the type of images desired, such as realism versus fictionality.

Similarly to web search, where users must efficiently explore a vast information space to retrieve the most useful or relevant information, prompting addresses the problem of orienteering. Orienteering involves exploring search results to determine their relevance to the query, potentially requiring iterative query reformulation.

In the context of automatic image generation, prompting can be viewed as a form of orienteering that helps users navigate the space of possible images that can be generated automatically. However, refining keywords for textual search involves immediate human comparison, whereas exploring an image space can be more challenging, as the links between textual prompts and generated images can be less explicit.

GENERAL'23: GENerative, Explainable and Reasonable Artificial Learning Workshop 2023, held in conjunction with CHITALY 2023

federico.torrielli@unito.it (F. Torrielli)
https://evilscript.eu/ (F. Torrielli)
0000-0001-8037-8828 (F. Torrielli)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3571/short1.pdf

This paper aims to model prompting in the context of image generation, identifying patterns of positive cases and typical pitfalls. However, there are differences between searching within an information space of existing elements and generating through prompting. In the former case, once results are obtained, the user can reformulate based on what has been retrieved. In the latter case, reformulation is based on what has been generated, introducing an additional element of uncontrolled variability. This creates a more significant gap between the threefold space of prompting engineering, expectation, and reformulation.

The widespread adoption of latent diffusion models, such as Stable Diffusion [2], Imagen [3], DALL-E 2 [4], or Midjourney[1], marks the beginning of an era of generative art. AI art not only is more accessible, but it also has a comparable ceiling of mastery compared to traditional "human" art. At the core of this new wave of AI technologies lies a process called **Diffusion** [5] and, of course, natural language through the use of prompts. A prompt in diffusion models is a statement or question that provides context for the model to generate a response or output. The prompt helps the model understand what type of information is expected and allows it to generate a more coherent and relevant response. For example, a prompt for a text-to-image model could be "an image of a dragon [...]", and the model would generate an image of a dragon based on that prompt. Prompts can also be used to control the complexity of the output by providing more or less detailed expectations [6]. In this context, prompt engineering (PE, from now on) is the task of refining a prompt through various techniques to achieve the desired result. Although the literature is still in its early stages regarding the different techniques of prompting, some papers have already distinguished between different parts of the PE task: defining the factual content of the image and the style [7]. While we agree with the analysis of prompts and the division into different parts, there is currently no clear and standard definition of how a prompt for a text-to-image model should be structured, since prompt engineering is still a relatively new field of research. Several papers have already explored prompt permutations, length optimisation, and style changes for older models [8], which are no longer considered as the state of the art. However, in our opinion, it is essential to develop a scientific approach to standardised image generation that can be applied to any text-to-image synthetic model. Such an approach would help ensure reproducibility and facilitate further advancements in the field.

In this work, we present a novel methodology for image generation that builds upon previous literature and recent studies on generative models. Our approach leverages precise tokenisation, it minimises the use of irrelevant tokens, and it incorporates small, incremental changes to produce consistent outputs. We think that this approach might become one possible standard practice for prompt engineering, involving factual descriptions, style changes, etc. within a structured pipeline we named **BLACK** (**B**ackground, **L**ighting, **A**menities, **C**ontext, and **K**inesis). Our methodology draws upon the latest advancements in the field and provides a systematic framework for generating images. Furthermore, we created a structured lexical resource of prompts following the proposed pipeline and examples of expected results[2].

[1] https://www.midjourney.com/

[2] https://docs.google.com/spreadsheets/d/1xer8JvcsVYEVsRMhksJD--JaAXzZuJ6QJjyT3b5bpJc

2. Exploring Latent Space on a Prompt

This proposal aims at placing some fundamental bricks towards the creation of bases and guidelines to systematically generate a large number of different images. Our newly introduced method, called *BLACK*, relies on a fixed token structure and a permutation of single or multiple terms to achieve results ranging from simple and straightforward to complex and intricate. The prompting phase of the proposal derives from the current literature on text-to-image generation and can be divided into two parts: the factual description (Subject + Verb + Object) and the style embellishment.

A working template that is commonly utilised for this prompt is "`A/The [subject] [verb] [support tokens] [object]`" where the verb and the support tokens (e.g. conjunctions) can be optionally included or excluded [9]. These groups of tokens are comma separated and can be weighted differently. The use of templates is justified because tests have shown that maintaining focus on keywords is more impactful than rephrasings during the prompt engineering process [8].

The definition of a style is another fundamental aspect of image generation. A style encompasses both the content and the form of an artistic piece: it determines how the content is represented, illustrated and conveyed [10]. Since style depends heavily on the weights of the model, there are checkpoints[3] that are either tailored to a particular style or general enough to handle different tasks. Synthetic style is defined in one or few words that are unique to artists, art movements and everything that is radically recognisable: everything that has influential aesthetic can be a style (e.g. *cyberpunk*) [9]. An example of a working prompting template is the following: `[factual prompt] in the style of [style name]`. When discussing style, it is optimal to be concise and direct, and this approach should always be considered. In all observed cases, using single-term style names has been found to be more impactful than using style definitions.

3. BLACK: a prompting methodology

In this paper, we introduce a novel method called *BLACK* for text-to-image prompting techniques, which aims to improve the generation of synthetic images using tools such as Stable Diffusion. The name *BLACK* is an acronym for its five key components: *Background*, *Lighting*, *Amenities*, *Context*, and *Kinesis*.

Background represents the environment where the subject is located, while *Lighting* groups light, shadows, and everything in between. *Amenities* are model-specific tokens essential for highly-detailed generation. *Context* tokens work alongside style, and *Kinesis* tokens are used for movement, motion blur, and dramatic representation of reality.

When using tools like Stable Diffusion, prompts are employed to add features to an image, with the model navigating the latent space and collecting features that are direct mappings of the prompts. Negative prompts [11] can be utilised to instruct the model on what to avoid while

[3] In Stable Diffusion, checkpoints are pre-trained weights or models that can generate images of a specific genre or style. These checkpoints are created by training the model on a dataset and can be used to generate images of a particular genre or style.

navigating the latent space. This can be particularly useful for constraining any generative model, as specifying that unwanted colours, elements, or styles should be avoided or reduced can help eliminate undesirable variations and steer the output towards the desired themes.

3.1. Background

The background context provided in a prompt can encompass the environment surrounding a character or object. This context can be minimally specified with a simple phrase like `'flat black background'` or more intricately described in vivid detail. However, more lengthy background descriptions are discouraged, as they can inadvertently 'bleed' into and influence the generated subject. For instance, a prompt specifying `'pink-gold arabic era very intricate city background'` may result in the generated subject adopting pink or gold colours approximately 20% of the time, as the background context exerts an unintended effect on the subject's attributes.

3.2. Lighting

Lighting and shadows are crucial components of high-quality art generation. The choice of appropriate lighting depends on the background context and subject matter. As discussed in the Experiments section that follows, lighting selection holds a greater impact than may be intuitively evident. For instance, pairing a dark, ominous background with `'eerie lights'` or a cheerful, natural subject with `'cinematic lighting'` or `'soft lighting'` can significantly influence the tone, interpretation, and perception of the generated art, as will be demonstrated.

3.3. Amenities

Despite its name, the 'amenities' component of a prompt is critical to generating high-quality outputs. 'Amenities' refers to unique keywords that enhance the final product, and for that reason Oppenlaender calls them *'Quality Boosters'* [9]. These keywords are drawn from the training dataset and reflect its content. For example, prompts for models trained on the LAION dataset[4] or DeviantArt[5] data might include fundamental terms like `'8k, best quality, masterpiece'`. Prompts targeting 3D-rendered art might instead include phrases like `'trending on CGSociety, extremely detailed CG unity wallpaper'`. In general, selecting amenities relevant to the target style or dataset is essential to successful generation.

3.4. Context

The 'context' refers to a set of categories of tokens that include color palettes, level of detail, setting, and complementary style tokens (e.g. `'vibrant'`, `'muted'`, etc.). These tokens encode information such as time of day, whether the art is an illustration, 3D rendering or photograph, camera angles, and other relevant details. Additionally, context tokens may represent season (e.g. `'winter'`), time period (e.g. `'medieval'`), subject matter (e.g. `'landscape'`), and mood (e.g. `'melancholy'`).

[4]https://laion.ai/blog/laion-5b/
[5]https://www.deviantart.com/

3.5. Kinesis

The 'kinesis' component refers to tokens that encode motion and pose attributes, such as movement (e.g. `'walking'`, `'running'`, `'jumping'`) or stance (e.g. `'sitting'`, `'crouching'`, `'lying down'`). These tokens are often used in conjunction with context tokens to specify the nature of the movement, such as `'sprinting through a forest'` or `'motion blur'`. Including kinesis tokens can help generate dynamic art with a sense of action or energy. The specific kinesis tokens selected will depend on the desired style and subject matter. For example, prompts for sports photography may include additional athletic motions like `'kicking'`, `'throwing'`, or `'swinging'`, while prompts for portraits could include more subtle poses like `'tilting head'` or `'crossing arms'`.

4. Experiments

In this section, we demonstrate the ability of our methodology to generate art that is faithful to the prompt. To achieve this, we utilise a trained checkpoint called **DreamShaper** [12], which is capable of generating art in styles ranging from realistic to fantastical. To ensure consistent results, we fix the random seed for each generation, then only modifying individual tokens in the prompt.

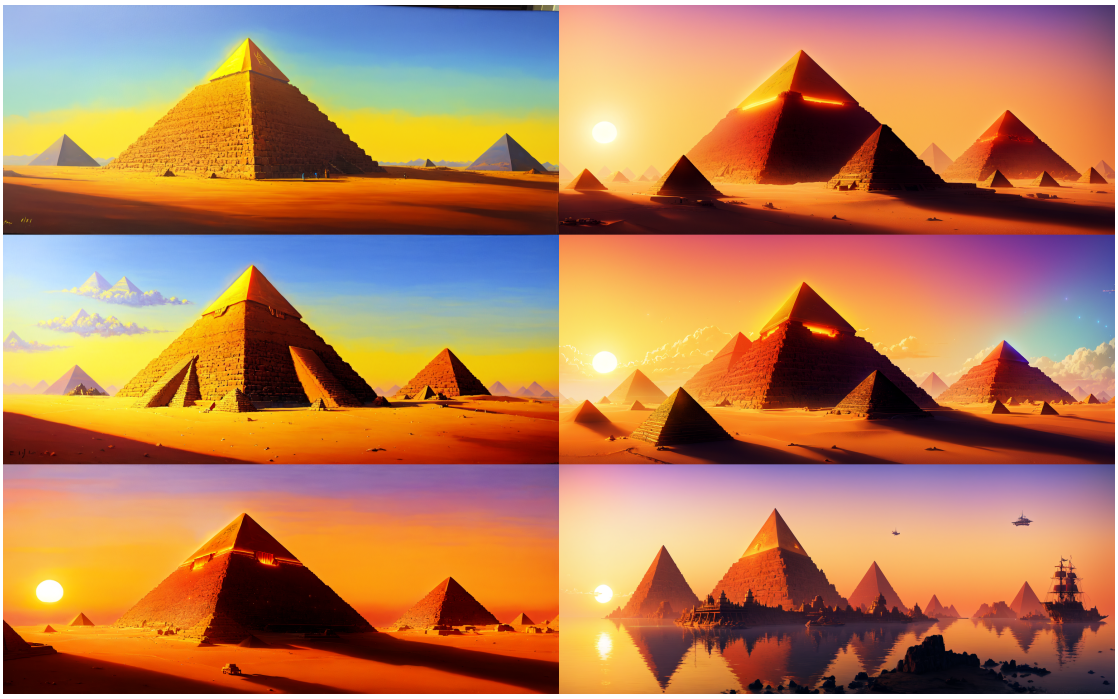

Figure 1: Iterations of our BLACK method. On the left, from top to bottom: (1) first generation, (2) with **b**ackground and (3) with **l**ighting. On the right, from top to bottom: (4) with **a**menities, (5) with **c**ontext and finally (6) **k**inesis.

4.1. Image Synthesis and Parameters

In generating high-quality images, the selection of appropriate parameters is crucial. In our experiment, we employed the DPM++ Keras solver [13] to sample 20 steps on images with dimensions ranging from 512×512 to 512×768 pixels. The CFG scale was set to 9, with outputs upscaled by 1.8× using a latent space upscaler that involved 20 steps and a denoising strength of 0.6. The initial seed is 526430048.

When determining parameters, several fundamental decisions must be made, starting with the choice of solver. Based on our findings, DPM++ appears to be the fastest, yielding satisfactory results in just under 25 steps. It is widely recognised that diffusion models have a significant limitation in terms of generation speed. This issue is expected to be addressed by Consistency Models in the near future [14]. These models promise a streamlined, few-step process for generating high-quality images and support a one-step generation without the requirement of adversarial training.

We advocate for a generate-then-upscale approach as opposed to generating images directly at a higher resolution. This strategy helps circumvent issues such as visual clutter, errors, and residual noise, thus enhancing the overall quality of the generated images.

4.2. A scenario: *Pyramid Song*

The top-left image in Figure 1[6] shows the initial generated image from our model. As can be seen, the output image exhibits some flaws: the colours are unrealistic, the contrast is overly exaggerated, and the background appears neither like a tundra nor a desert but somewhere in between. However, as we will demonstrate in this section, our proposed model and training methodology are capable of generating more natural and compelling images than this initial attempt.

4.2.1. *B*LACK

The mid-left image[7] in Figure 1 shows the results of providing our model with additional context about the background scenario. Specifying that the image should depict a desert landscape leads to notable improvements, including more realistic colours and clearer shadows. There are more details in the pyramid and in the sky. Overall, this example demonstrates how giving our model more contextual guidance can significantly enhance the quality and coherence of the generated images.

4.2.2. B*L*ACK

The bottom-left image[8] in Figure 1 demonstrates continued improvements in the quality of the generated images as we provide more context and feedback to our model. The lighting and shadows are more realistic, and additional details have been added, including shining elements atop the pyramid and smaller surrounding objects. The sunset appears more realistic and casts

[6]Initial prompt: *a oil painting of a alien pyramid.*
[7]Prompt (with Background): *a oil painting of a alien pyramid, ancient egypt background.*
[8]Prompt (with Lighting): *a oil painting of a alien pyramid, ancient egypt background, cinematic lighting, sunset.*

a 'bloom' effect onto the dunes. Each iteration reveals how our model can produce increasingly compelling results when given more guidance about the desired style, content, and context of the images.

4.2.3. BL*A*CK

By incorporating amenities tokens in the top-right image[9] of Figure 1, we achieve further improvements in image quality. The bloom effect from the previous image is stronger, and the smaller surrounding pyramids are moved into the foreground, giving the primary pyramid subject greater prominence. The ground is detailed and more striking, and additional nuances are visible on the pyramid itself. Overall, the image appears smoother and with less noise than the previous iteration, demonstrating how continued refinements to our training methodology can steadily enhance the results.

4.2.4. BLA*C*K

The mid-right image[10] in Figure 1 demonstrates some unexpected changes that emerge from our model as it continues to be instructed. There are now clouds, while the gradient has become more prominent and some stars are visible. The shadows are longer and smoother, and the sky takes on an oil painting-like quality rather than photo-realism.

4.2.5. BLAC*K*

The bottom-right image[11] in Figure 1 shows the results of incorporating 'kinesis' tokens into our model prompting. To demonstrate how our model can represent dynamic objects, we included a pirate ship token part, which produces some ships in a now present lake in the final image. With the addition of the kinesis and pirate ship tokens, our image-prompting process using the BLACK methodology is now complete. As shown, this process is capable of generating diverse, compelling images with unique content, styles, and implied motion.

5. Unresolved Questions

This research serves as an exploratory foundation aimed at developing and assessing standards in prompt engineering for generative AI. Ensuring comparability and reproducibility in LDMs presents ongoing challenges. Notably, even minor modifications to prompts can yield varying outcomes. In this section, we address potential questions from readers.

[9] Prompt (with Amenities): *a oil painting of a alien pyramid, ancient egypt background, cinematic lighting, sunset, perfect, Beeple, concept art, fantasy art, trending on ArtStation, trending on CGSociety, Intricate, High Detail, photorealistic painting art by midjourney and greg rutkowski, 8k.*

[10] Prompt (with Context): *a oil painting of a alien pyramid, ancient egypt background, cinematic lighting, sunset, perfect, Beeple, concept art, fantasy art, trending on ArtStation, trending on CGSociety, Intricate, High Detail, photorealistic painting art by midjourney and greg rutkowski, 8k, summer, 4k landscape, wallpaper.*

[11] Prompt (with Kinesis): *a oil painting of a alien pyramid, ancient egypt background, cinematic lighting, sunset, perfect, Beeple, concept art, fantasy art, trending on ArtStation, trending on CGSociety, Intricate, High Detail, photorealistic painting art by midjourney and greg rutkowski, 8k, summer, 4k landscape, wallpaper, (pirate ship in the background), (flying pirate ship), dynamic.*

1. **Does the order of prompts significantly influence the results?** Preliminary evidence suggests it does, although the exact impact is not yet quantifiable. Tokens positioned at the beginning of a prompt tend to receive more attention [15]. However, the importance of individual words is often outweighed by their contextual relationships.
2. **How does placing *amenities* before *background* in a prompt affect the outcome?** This question is intrinsically linked to the previous one. Our experiments indicate that altering a sequence of tokens (rather than just one) can significantly impact the result. However, a precise qualitative evaluation remains elusive. Future studies will delve deeper into this aspect.
3. **Is BLACK compatible with all image generators?** Yes, provided they support prompting and are trained on internet data.

6. Conclusion

In conclusion, our study presents a straightforward and replicable token-based prompting methodology, known as BLACK, that caters to both novice and advanced users. We have developed a comprehensive example to elucidate the thought process underlying our approach. To assist readers in crafting prompts using the BLACK method, we have compiled a spreadsheet featuring over 1000 unique words, each assigned to a corresponding BLACK category such as subject or style tokens. Additionally, we have incorporated a random generator and sample prompts within the document to further facilitate the prompt creation process. All the work is already available on our repository[12].

References

[1] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 10674–10685.

[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding, ArXiv abs/2205.11487 (2022).

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, ArXiv abs/2204.06125 (2022).

[5] J. N. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, ArXiv abs/1503.03585 (2015).

[6] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, T. Scialom, Augmented

[12] https://github.com/federicotorrielli/BLACK

language models: a survey, 2023. URL: https://arxiv.org/abs/2302.07842. doi:10.48550/ARXIV.2302.07842.

[7] S. Witteveen, M. Andrews, Investigating prompt engineering in diffusion models, ArXiv abs/2211.15462 (2022).

[8] V. Liu, L. B. Chilton, Design guidelines for prompt engineering text-to-image generative models, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (2021).

[9] J. Oppenlaender, A taxonomy of prompt modifiers for text-to-image generation, 2022. arXiv:2204.13988.

[10] S. Ross, Style in art, The Oxford handbook of aesthetics (2003) 228–244.

[11] Negprompt, 2023. URL: https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt.

[12] Lykon, Dreamshaper (revision 51416b0), 2023. URL: https://huggingface.co/Lykon/DreamShaper. doi:10.57967/hf/0297.

[13] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu, Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models, ArXiv abs/2211.01095 (2022).

[14] Y. Song, P. Dhariwal, M. Chen, I. Sutskever, Consistency models, ArXiv abs/2303.01469 (2023).

[15] Deane, Does prompt order matter in stable diffusion? - prompt phantom, 2023. URL: https://promptphantom.com/does-prompt-order-matter-in-stable-diffusion/.