Automated Black-box Prompt Engineering for Personalized Text-to-Image Generation

Yutong He[1], **Alexander Robey**[1,2], **Naoki Murata**[3], **Yiding Jiang**[1], **Joshua Williams**[1], **George Pappas**[2], **Hamed Hassani**[2], **Yuki Mitsufuji**[3,4], **Ruslan Salakhutdinov**[1], **J. Zico Kolter**[1,5]
[1]Carnegie Mellon University, [2]University of Pennsylvania, [3]Sony Research,
[4]Sony AI, Sony Group Corporation, [5]Bosch Center for AI

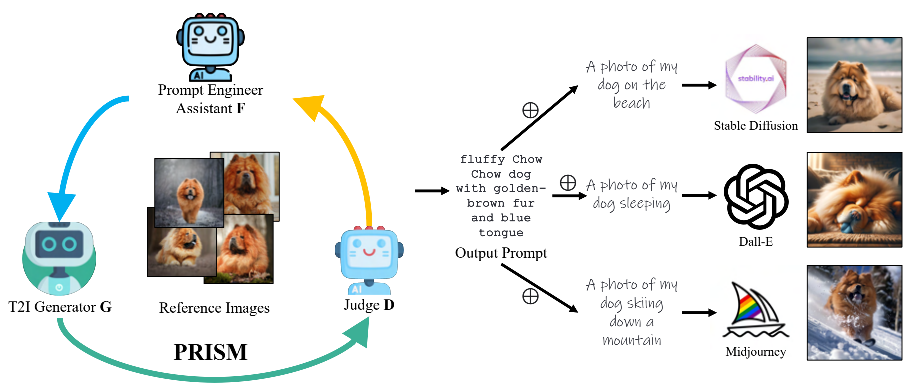

Figure 1: Given a set of reference images, our method, PRISM, is capable of creating human-interpretable and accurate prompts for the desired concept that are also transferable to both open-sourced and closed-sourced text-to-image models.

Recent advances in generative AI (GenAI) for art and image creation have sparked widespread interest, initiating dialogue between artists, content creators, and AI researchers. However, controlling the outputs of generative models remains a persistent challenge, particularly with traditional deep learning techniques. Early attempts, which often centered on particular architectures or tasks, were largely characterized by manually-curated data collection, fine-tuning, or retraining from scratch [1, 2, 3, 4]. These methods are not only resource-intensive but also struggle to generalize across different models and GenAI platforms, limiting their versatility in producing a diverse range of styles and ideas. As a result, despite the promise of these GenAI tools, there is a growing need for more efficient, adaptable, and accessible algorithms that enable better human control over AI-driven creativity.

Today, perhaps the most popular approach for controllable generation is to guide the generation process with a piece of textual information, or *prompt*, that describes the properties of the desired output using text-to-image (T2I) generative models [5, 6]. Through text, T2I models allow users to quickly and easily describe a wide variety of concepts, and users can more efficiently explore the behavior of the model through a myriad of strategies. The predominant method for obtaining such input text is to manually design candidate prompts in an iterative, trial-and-error fashion, a process known as *prompt engineering*, based on what the user (prompt engineer) *believes* will lead to a desirable output. Unfortunately, these practices are often time-consuming, sensitive to different phrasings [7], difficult to scale, and often requiring significant expertise to achieve desired results. Additionally, prompts that work for one platform may not generalize well to others, leading to inefficiencies and making the process mentally exhausting for users.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Motivated by the drawbacks of manual prompt engineering, a recent line of work known as *personalized* or *subject-driven* T2I generation has sought to automate the controllable generation pipeline. Given a collection of reference images that capture specific concepts, such as artistic style or shared objects, personalized T2I algorithms generate images that reflect those concepts. While personalized T2I methods often involve fine-tuning or retraining the underlying T2I model [8, 9, 10], several approaches focus specifically on automating prompt engineering to generate effective prompts. However, these algorithms usually require pre-collected, architecture-specific keywords or white-box, embedding-based optimization [11, 12], leading to non-interpretable prompts [13] and precluding the possibility of directly generating prompts for closed-source T2I models like Midjourney or DALL-E.

To address these shortcomings, we propose ***P**rompt **R**efinement and **I**terative **S**ampling **M**echanism* (PRISM), a new automated prompt engineering algorithm for personalized T2I generation. A key observation is that prompt engineers repeatedly update their "belief" of what makes an effective prompt based on the difference between their desired results and the generated images from previous iterations. Inspired by jailbreaking attacks on large language models (LLMs) [14], we design an algorithm that operates with only limited human input, is capable of generating human interpretable and editable prompts, makes minimal assumptions about the T2I generative model, and generalizes across different T2I platforms, including popular black-box models such as DALL-E and Midjourney.

Given a set of reference images, our method first generates an initial prompt and its corresponding image using a multimodal LLM and a T2I model. We then obtain a score indicating the visual similarity of the generated image and the reference image via another multimodal LLM. Leveraging LLMs' in-context learning abilities [15, 16, 17], we instruct the LLM to update the candidate prompt distribution based on the previously generated prompt, images, and the evaluation scores. This processing is then repeated for a predetermined number of iterations. In the end, PRISM outputs the best-performing prompt by re-evaluating the top prompts generated from this process.

Through experiments, we demonstrate that PRISM consistently generates accurate, human-interpretable prompts for personalized T2I generation and direct image inversion (Figure **??**), outperforming all baselines. With detailed experiments in the appendix, our method also shows significantly better generalizability and transferability as we achieve the best performance in almost all metrics when experimenting with closed-source models in comparison to existing methods. Finally, we show that because of the interpretability provided by our method, the prompts produced by PRISM are also easily editable (Figure 2), enabling a wide range of creativity possibilities in real life.

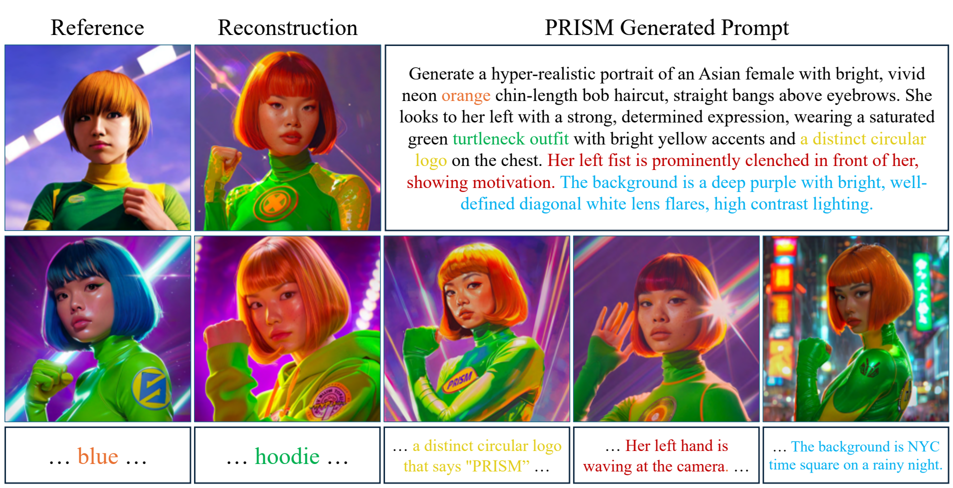

Figure 2: With accurate, human-interpretable direct image inversion from PRISM, artists can adjust specific attributes of a reference image with the rest of the scene intact on platforms like Midjourney.

Acknowledgements

This work was supported in part of the ONR grant N000142312368. We also thank Sony AI for sponsoring part of this work. A.R., H.H., and G.P. are supported by the NSF Institute for CORE Emerging Methods in Data Science (EnCORE). A.R. is also supported by an ASSET Amazon AWS Trustworthy AI Fellowship. Y.J. is supported by Google PhD Fellowship.

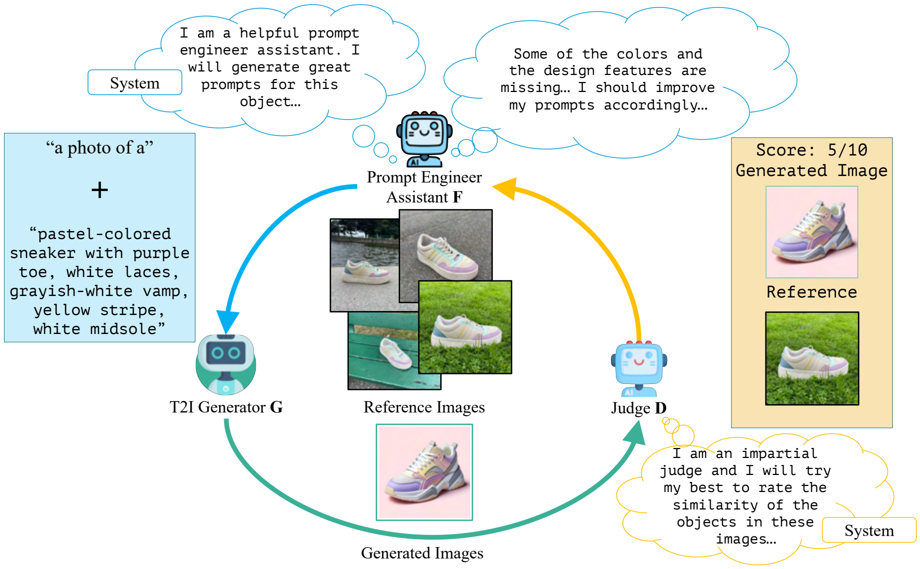

Figure 3: An illustration of PRISM. The label "System" indicates the system prompts setups for the multimodal LLMs.

Figure 4: Qualitative results for personalized T2I generation on DreamBooth dataset.

Figure 5: Qualitative results for personalized style T2I generation on Wikiart dataset.

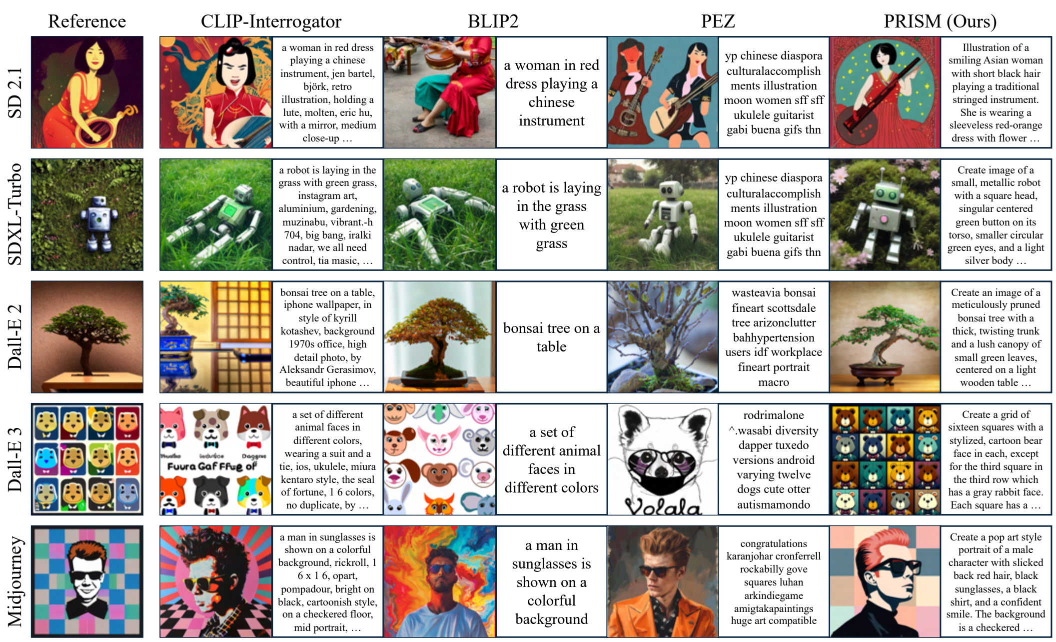

Figure 6: Image inversion results for different methods on different T2I models.

References

[1] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[2] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. 2017.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. pages 10684–10695, 2022.

[6] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. Featured Certification.

[7] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2300–2344. Association for Computational Linguistics, 2022.

[8] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. pages 22500–22510, 2023.

[9] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. 36:30286–30305, 2023.

[10] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning, 2023.

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. 2023.

[12] Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. *arXiv preprint arXiv:2312.12416*, 2023.

[13] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. volume 36, pages 51008–51025, 2023.

[14] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

[15] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4222–4235. Association for Computational Linguistics, 2020.

[16] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. 2023.

[17] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. 2024.