PERSONALIZING TEXT-TO-IMAGE GENERATION WITH VISUAL PROMPTS USING BLIP-2

Jeongwon Lee, Changsun Lee, Joonhyeok Jang *
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon, 34141 South Korea
{gardenlee21, sunny17, jang0727}@kaist.ac.kr

ABSTRACT

In recent years, text-to-image generation has received significant attention as researchers aim to automatically generate realistic images from textual descriptions. Despite the promising results of diffusion models in producing high-quality images, they often struggle to capture the richness and diversity of natural language expressions, making it difficult to generate images that align with user intentions. To tackle this challenge, personalization has been proposed as a potential solution. Personalization involves fine-tuning pre-trained text-to-image generation models using user-provided images that contain specific concepts or subjects, enabling the models to generate images with the desired subject or style. However, existing personalization techniques typically require direct fine-tuning of the text encoder, diffusion model, or both, which can result in high computational costs for incorporating user-provided concepts and potentially compromise the model's knowledge. This paper introduces a novel approach to personalizing a text-to-image model by leveraging a BLIP-2 encoder. We provide the image that contains objects we wish to generate using the Stable Diffusion model as inputs to the BLIP-2 encoder. Then, we use the output queries of the BLIP-2 Q-former as visual prompts to guide the Stable Diffusion model to generate images that capture the visual representations of the input image.

1 INTRODUCTION

The field of text-to-image generation has witnessed estensive resaerch in recent years, aiming to automatically generate realistic images from textual descriptions. Early approaches in this domain employed on a combination of generative adversarial networks (GANs) (Goodfellow et al., 2020) and conditional variational autoencoders (CVAEs) (Sohn et al., 2015) to generate images based on text. However, these models often produced inaccurate of conceptually distant images far from the user's intended meaning, limiting their practical applicability. Diffusion models have emerged as promising alternative for generating high-quality images, allowing for more finer control over the generated output. Nevertheless, these models still struggle to capture the diversity and richness of natural language expressions, making it challenging to generate images that precisely align with user's intentions.

To address this issue, several methods for personalization (Ruiz et al., 2022; Gal et al., 2022; 2023; Chen et al., 2023) have been proposed as a potential solutions. These methods involve training the model on additional data or fine-tuning pre-trained models to adapt to specific tasks, thereby improving their ability to capture user preferences and generate more accurate images. However, these previous methods typically require fine-tuning. For instance, (Ruiz et al., 2022;?) involve fine-tuning the diffusion model and the embedding lookup of the text encoder, respectively. Similiarly, (Gal et al., 2023) enables fast personalization for user-provided concepts, but necessitates pre-training a large-scale dataset containing numerous images from the corresponding class. In (Chen et al., 2023), fine-tuing of the diffusion model is required to obtain an expert model for each user-provided concept. Directly fine-tuning generative models can lead to high costs in terms of memory consumption, computation, and potential degradation of model performance.

*These authors contributed equally.

In this paper, we propose a personalization strategy that does not require model fine-tuning. Intsaed, we leverage a pre-trained diffusion model, including its text encoder, namely BLIP-2 (Li et al., 2023), and a set of training images containing a user-provided concept. Our approach focuses on training BLIP-2, rather than the given diffusion model. Through this personalization process, we can obtain a prompt embedding that encapsulates the provided user-provided concept, enabling the generation of conceptually relevent images.

2 RELATED WORKS

2.1 TEXT-TO-IMAGE GENERATION OR TEXT-GUIDED SYNTHESIS

Early text-to-image models made significant progress by utilizing Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) trained on large paired image-caption datasets. However, training GANs at scale is challenging due to issues like mode collapse. Recently, diffusion models have gained popularity in text-to-image generation. These models can be categorized based on where the diffusion prior is applied, either in the pixel space or in the latent space. Both GLIDE (Nichol et al., 2021) and Imagen (Saharia et al., 2022) employ classifier-free guidance by replacing the label in the class-conditioned diffusion model with text descriptions of the images. The key difference lies in their text encoders. GLIDE trains the text encoder alongside the diffusion prior, while Imagen employs a frozen pre-trained large language model as the text encoder, reducing computational requirements. Stable Diffusion (Rombach et al., 2022) utilizes VQ-GAN for the latent representation of images, enhancing photorealism through an adversarial objective. DALL-E2 (Ramesh et al., 2022) employs a multimodal latent space where image and text embeddings are aligned to generate images reflecting a deeper level of language understanding. However, these models encounter difficulties when generating images based on user-provided concepts, leading to the proposal of various personalization methodologies. We provide an overview of these methods in the following subsection.

2.2 PERSONALIZING TEXT-TO-IMAGE MODELS BASED ON DIFFUSION

There is a growing demand for algorithms specialized in generating images related to specific concepts or subjects, beyond algorithms that merely generate images from text or produce general images guided by natural language. As mentioned earlier, several previous attempts have been made to personalize diffusion-based text-to-image models. In DreamBooth (Ruiz et al., 2022), personalization is achieved through naive fine-tuning with a novel autogenous class-specific prior preservation loss, resulting in the generation of realistic images featuring subjects contextualized in various scenes. In the case of (Gal et al., 2022), the embedding lookup in the text encoder is fine-tuned to find an embedding for a pseudo-word, enabling the discovery of the importance of specific words in capturing the user-provided concept. This new embedding in the embedding space is then used to substitute vectors related to tokenized strings. (Gal et al., 2023) introduces an approach called encoder-based domain tuning, where the encoder and diffusion models are trained to transform images of a target concept in a given domain into word embeddings, and the parameters of the text-to-image model are regularized to effectively incorporate additional concepts. However, (Gal et al., 2023) necessitates pre-training the encoder and diffusion model using a large dataset specific to the domain, containing diverse images of the targeted concepts (e.g., ”cat,” ”dog,” etc.). In SuTi (Chen et al., 2023), personalization based on apprenticeship learning is proposed, aiming to personalize a model with multiple concepts. Specifically, expert models are trained with samples of each concept through naive fine-tuning of a given pre-trained model, and an apprentice model is subsequently trained using large-scale generated samples from all expert models. Despite the individual achievements of these previous methods, they typically require either complete or partial pre-training or fine-tuning of the model, leading to high costs in terms of memory consumption, computational requirements, or performance degradation. In contrast, we propose utilizing BLIP-2 (Li et al., 2023) to obtain a representation embedding of a user-provided concept without directly updating the text-to-image diffusion models.

3 METHOD

We aim to guide the text-to-image generation process of Stable Diffusion (Rombach et al., 2022) by providing an additional visual prompt obtained from a frozen BLIP-2 encoder. We employ a simple feed forward neural network to align the visual prompts to the text prompt. The visual prompt is then concatenated to the text prompt token embeddings to guide the model to generate the visual representation captured by the visual prompt in a style that is instructed by the text prompt.

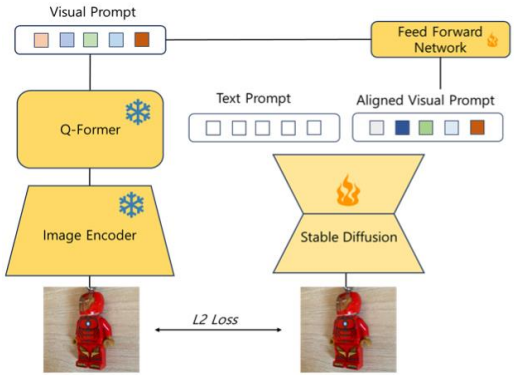

Figure 1: Overview of the proposed architecture. The image encoder and the Q-Former are frozen while the Feed Forward Network and the Stable Diffusion is fine-tuned with the L2 loss

3.1 VISUAL PROMPT ACQUISITION USING BLIP-2

Figure 1 shows the overview of the proposed architecture where the pre-trained image encoder and the multimodal encoder (Q-Former) from BLIP-2 is frozen. Since the pre-trained image encoder is ViT-L/14 (Dosovitskiy et al., 2021) from CLIP (Radford et al., 2021) we assume the image encoder is capable of generating generic image features that aligns well with the CLIP text embeddings that is used as text conditioning for the Stable Diffusion. The Q-Former is pre-trained to generate a query that captures the visual representation of the input image that is the most informative of the provided text. Therefore, we assume that even without finetuning the BLIP-2 encoder architecture, BLIP-2 encoder can generate a visual representation for an unseen image that is capable of guiding the image generation of the Stable Diffusion.

3.2 VISUAL-TEXT PROMPT ALIGNMENT

Simply appending the output queries of the frozen BLIP-2 encoder as visual prompts to the text prompt leads to poor performance in image generation as the text encoder of the Stable Diffusion has never learnt to encode a query that contains a visual representation. As shown in Figure 1, we introduce a Feed Forward Network that consists of two linear layers where each linear layer is followed by a GELU (Hendrycks & Gimpel, 2023) activation. The feed forward network is trained to approximate the appropriate alignment of the visual prompt from the BLIP-2 encoder to the text embeddings of the Stable Diffusion.

The visual prompt are fed through the feed forward network and projected to match the dimension of the text prompt token embeddings of the Stable Diffusion. Then the final conditioning prompt to the Stable Diffusion is given as "[text prompt embeddings] [projected visual prompt]" where the [projected visual prompt] conveys the representation of the desired object from the input image that we wish reconstruct in the style suggested by the [text prompt embeddings].

3.3 TRAINING OBJECTIVE

We aim to fine-tune the Stable Diffusion to generate images that heavily captures the visual representation from the visual prompt. We take a naïve approach of training the Stable Diffusion to reconstruct the same input image as its output. While optimizing the Stable Diffusion with the re-

construction loss (L2 loss) we assume we may obtain images that contain the objects capturing the visual representations of the visual prompt in a style suggested by the text prompt.

4 RESULTS

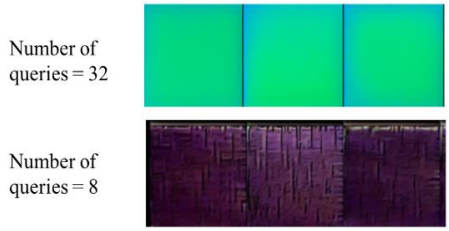

Figure 2: Generated samples when fine-tuned with the input image shown in Figure 1 with the text prompt "a red brick wall that looks like". The samples in the top and the bottom row are generated with 32 and 8 queries for the visual prompt, respectively. Having many number of queries for the visual prompt overwhelms the text embedding. Only the bottom row images are shown to generate images that reflect "a red brick"

As shown in Figure 2, the model is not yet able to reconstruct the object within the image and have also failed to understand the prompt. We fine-tuned the Stable Diffusion for two varying number of queries of the visual prompt. Originally BLIP-2 outputs 32 queries of dimension 768. The text prompt embedding for the Stable Diffusion is fixed to maximum sequence length of 77 with dimension of 1280. As shown in Figure 2, using all the 32 queries overwhelms the text prompt hindering the model to be conditioned on the text prompt. When we only use 8 queries for the visual prompt, the Stable Diffusion can then be conditioned on the text prompt. How to integrate the visual prompt as a condition for the Stable Diffusion along with the text prompt requires further investigation in training objectives and model architecture.

5 CONCLUSION

In this paper, we propose a deep learning-based approach to personalize a text-to-image generation model using BLIP-2. By leveraging the power of BLIP-2, we aim to obtain a meaningful representation of user-provided concepts, enabling the generation of diverse images showcasing those concepts from various perspectives. We anticipate that improved results can be achieved through further optimization of the Q-former with a paired dataset consisting of an image and a single and unique text that describes the image. Learning the visual representation that is informative of only a single text would allow the Stable Diffusion to easily learn the mapping between a certain text and the image it must reconstruct. We expect that a Q-former trained in such a manner can generate queries that effectively capture and represent previously unseen concepts, leading to the generation of high-quality images. We also expect employing LoRA to fine-tune Stable Diffusion model will prevent the model from generating trivial solutions.

REFERENCES

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, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

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. *arXiv preprint arXiv:2208.12242*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.