Personalized Text-to-Image Generation with Auto-Regressive Models

Kaiyue Sun[1] **Xian Liu**[2] **Yao Teng**[1] **Xihui Liu**[1]
[1]The University of Hong Kong [2]The Chinese University of Hong Kong

Code: https://github.com/KaiyueSun98/T2I-Personalization-with-AR

Abstract

Personalized image synthesis has emerged as a pivotal application in text-to-image generation, enabling the creation of images featuring specific subjects in diverse contexts. While diffusion models have dominated this domain, auto-regressive models, with their unified architecture for text and image modeling, remain underexplored for personalized image generation. This paper investigates the potential of optimizing auto-regressive models for personalized image synthesis, leveraging their inherent multimodal capabilities to perform this task. We propose a two-stage training strategy that combines optimization of text embeddings and fine-tuning of transformer layers. Our experiments on the auto-regressive model demonstrate that this method achieves comparable subject fidelity and prompt following to the leading diffusion-based personalization methods. The results highlight the effectiveness of auto-regressive models in personalized image generation, offering a new direction for future research in this area.

1. Introduction

The rapid advancement of text-to-image generation models has revolutionized the field of computer vision, enabling the creation of highly realistic and diverse images from textual descriptions. Among the various applications of these models, personalized image synthesis—generating images of specific subjects in new contexts—has garnered significant attention. This capability is particularly valuable for applications in digital art, advertising, and virtual reality, where the ability to seamlessly integrate personalized content into diverse scenes is crucial.

While diffusion models have been at the forefront of personalized image generation, auto-regressive models, which employ a unified architecture for text and image modeling, have not been extensively explored for this task. Auto-regressive models [5, 6, 9–11] have demonstrated remarkable success in text-to-image generation by predicting image tokens sequentially. However, their potential for personalized image synthesis remains largely untapped. This paper aims to investigate the adaptation of auto-regressive models for personalized image generation.

We propose a novel two-stage training strategy that firstly optimizes text embeddings and then fine-tunes transformer layers together. Our experiments on the Lumina-mGPT 7B model [5] show that this approach outperforms existing optimization-based techniques like Textual Inversion [3] and shows comparable performance with DreamBooth [8] in terms of subject fidelity and prompt following. The results underscore the potential of auto-regressive models in personalized image generation and pave the way for future research in this domain.

This work explores the potential of auto-regressive models for personalized image synthesis, adapting them to meet the specific demands of text-to-image generation. Our findings suggest that auto-regressive models, when properly optimized, can achieve competitive performance in personalized image generation, offering a promising alternative to diffusion-based approaches.

2. Preliminaries

2.1. Personalizing Text-to-Image Models via Optimization

Textual Inversion. Textual Inversion [3] proposes a personalization method by creating a new "pseudo-word" (e.g., S_*) within the text embedding space of a text-to-image diffusion model. Using just 3-5 images of a specific subject provided by the user, this method optimizes the embedding vector corresponding to the pseudo-word to represent that subject. This word can then be used to compose natural language prompts, such as "a S_* on the beach", to generate personalized images in novel contexts.

DreamBooth. Instead of a "pseudo-word", DreamBooth [8] opts to optimize a unique identifier "[V]" that precedes the subject class (for example, "a [V] cat/dog/toy on the beach"). This approach helps to link the prior knowledge of the class with the subject, thereby reducing training time. However, using the class name can lead to a grad-

ual loss of the model's broader semantic knowledge during fine-tuning, a phenomenon known as language drift. To address this issue, a class-specific prior preservation loss is introduced to retain the model's ability to generate diverse instances of the class.

These optimization-based approaches are implemented and proved to be effective on text-to-image diffusion models. They can effectively perform various personalization tasks, including subject recontextualization, text-guided view synthesis, and artistic rendering.

In this paper, we explore the adaptation of these optimization-based personalization techniques to auto-regressive models and offer insights into the finetuning of auto-regressive models.

2.2. Text-to-Image Generation via Next-Token Prediction

Auto-regressive text-to-image models generate images in three steps. First, a tokenizer converts the input text into a sequence of discrete tokens, which are transformed into vector embeddings. These text embeddings, denoted as c, are then fed into an auto-regressive transformer that outputs logits l_t. The logits are converted into probabilities where the next image token x_t is sampled. The newly sampled token is concatenated with the preceding tokens to predict the subsequent token. Finally, an image decoder translates the complete sequence of tokens $x = (x_1, x_2, \ldots, x_T)$ into image pixels.

Training objective. During training, the auto-regressive transformer models the conditional probability $p\left(x_t \mid x_1, x_2, \ldots, x_{t-1}, c\right)$ of the sequential tokens using the standard next-token prediction objective. We denote $x_{1\sim t-1} = \{x_1, x_2, \ldots, x_{t-1}\}$, the model predicts the next token $x_t \in V$, where V denotes the vocabulary. The loss function f for a single prediction can be written as follows:

$$L(\theta) = f\left(y_t, p_\theta\left(x_t \mid x_{1\sim t-1}, c\right)\right), \tag{1}$$

$$p_\theta\left(x_t \mid x_{1\sim t-1}, c\right) = \text{Softmax}(l_t), \tag{2}$$

where $L(\theta)$ is the loss, parameterized by the model parameters θ and loss function f. In image generation, we predict the tokens from the image split of the total vocabulary. y_t represents the target label of the next token, which is derived by tokenizing the ground-truth image associated with the input text. f is cross-entropy loss.

3. Method

Personalizing a text-to-image diffusion model generally involves two strategies. The first strategy is to associate a unique text embedding with the subject. This text embedding can either represent the subject as a whole or serve as an adjective describing the subject class. However, because the number of parameters for a text embedding is limited, personalized images often struggle to capture all the essential features of the subject. To effectively embed the subject's appearance in the model, fine-tuning of the model parameters is usually required. Figure 1 shows the overview of our fine-tuning strategy.

In this section, we present our method for personalizing an auto-regressive model and explain the rationale behind our choices.

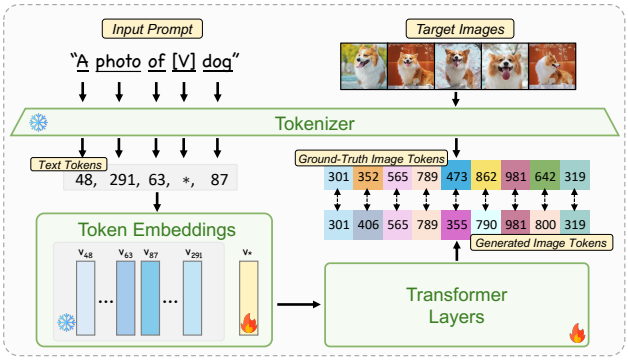

Figure 1. **Overview of Fine-tuning.** We fine-tune a text-to-image auto-regressive model using 3-5 input images, each paired with a text prompt that includes a unique identifier and the subject class name (e.g., "A photo of [V] dog"). The process involves two stages: first, we fine-tune the text embedding for the identifier [V], and second, we additionally fine-tune the transformer layers to enhance the model's performance.

3.1. Optimizing Text Embeddings

We generally follow the DreamBooth [8] approach to optimize a text embedding for a specific subject. We introduce a placeholder word [V], to represent the unique identifier of the new subject we wish to learn. The input text that includes the identifier [V] and the subject class name is then converted to tokens. We replace the embedding associated with the token for [V] with a new randomly initialized embedding, denoted as v_*. With a small set of reference images (e.g. 3-5) of the subject in various backgrounds or poses, we optimize v_* based on the cross-entropy loss defined in Equation 1. For the input text, we use the templates provided by Textual Inversion [3], which contain neural context such as "A photo of [V] [class_name]", "A rendition of [V] [class_name]". Our optimization goal can thus be defined as follows:

$$v_* = \arg\min_v f\left(y_t, p_\theta\left(x_t \mid x_{1\sim t-1}, c\right)\right), \forall t \tag{3}$$

It is expected to encourage embedding v_* to learn the common features in the reference images while discard elements that are unique to each image, such as the background.

3.2. Fine-tuning Transformer Layers

We conduct experiments using the Lumina-mGPT 7B model [5]. We have observed that the generated images fail to accurately replicate the reference subject if we optimize the text embeddings only. Additionally, when optimizing the text embeddings on a single data point, the model does not overfit; instead, after a slight decrease in the loss, it stabilizes around a specific level. Given the limited capacity of text embeddings, fine-tuning the auto-regressive transformer becomes necessary to effectively implant the subject into the model's output domain.

Two-stage training. DreamBooth [8] fine-tunes the layers conditioned on the text embeddings and the diffusion UNet simultaneously. In our experiments, we find that when fine-tuning the text embeddings and transformer layers together, the text embeddings cannot get fully trained. If we revert to the original transformer layers during inference, the text embeddings alone fail to convey any meaningful content.

To address this issue, we devise a two-stage training strategy. In the first stage, we fully optimize the text embeddings, and in the second stage, we fine-tune the transformer layers to maximize the subject fidelity. This two-stage approach is mutually beneficial: the first stage stabilizes the training and reduces the effort needed in the second stage, while the second stage compensates for any defects from the first stage due to its inherent limitations.

4. Experiments

4.1. Dataset and Evaluation

We evaluate our model's personalization capability on Dreambench [8], which provides a dataset consisting of 30 subjects, each with 4-6 images. These subjects are divided into two groups: 21 objects and 9 live subjects/pets. Each subject is tested on 25 prompts, which include scenarios such as re-contextualization, accessorization, and property modification. Their purpose is to assess whether the key features of the subject can be preserved under different semantic modifications while the generated image adheres to the prompt. Following Dreambench [8], we employ DINO [1] and CLIP-I [7] to assess subject fidelity, and CLIP-T [7] to measure the prompt following. For evaluation, we generate images using a fixed Classifier-free Guidance of 4.0 and an image top-k value of 2000.

4.2. Quantitative Results

Table 1 presents the evaluation results of various models on Dreambench [8]. By fine-tuning the auto-regressive model of Lumina-mGPT [5] using our method, it outperforms Textual Inversion [3], Re-Imagen [2], and zero-shot BLIP-Diffusion [4] in both subject fidelity (Dino and CLIP-I) and prompt following (CLIP-T). Additionally, it achieves comparable results to stable diffusion-based DreamBooth [8] and fine-tuned BLIP-Diffusion [4] in DINO. Notably, our method achieves the highest CLIP-T among all the models listed. These findings demonstrate that auto-regressive models can be fine-tuned to incorporate new concepts without compromising their original generation capabilities.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Images	0.774	0.885	N/A
Textual Inversion [3]	0.569	0.780	0.255
Re-Imagen [2]	0.600	0.740	0.270
DreamBooth (Stable Diffusion) [8]	0.668	0.803	0.305
DreamBooth (Imagen) [8]	**0.696**	**0.812**	0.306
BLIP-Diffusion (zero-shot) [4]	0.594	0.779	0.300
BLIP-Diffusion (fine-tune) [4]	0.670	0.805	0.302
Ours (Lumina-mGPT [5])	0.671	0.785	**0.314**

Table 1. **Quantitative results comparison on Dreambench [8].** We show subject fidelity (DINO, CLIP-I) and prompt following (CLIP-T) scores across different models. For all three metrics, the scores range from 0 to 1, where a higher score indicates better performance. The bold values highlight the highest score achieved.

4.3. Qualitative Results

In Figure 2 and Figure 3, we present qualitative generation results of our model. The re-contextualization examples demonstrate the model's ability to accurately reproduce the subject's appearance while merging it into the new backgrounds. Furthermore, the model can accurately modify the color and shape properties of the subject, even in challenging cases such as "cube-shaped". This indicates that the model not only learns new concepts, but also effectively decomposes and recomposes them with its prior knowledge. In accessorization examples, the model can seamlessly integrate subjects with outfits, demonstrating its ability to understand the structure and meaning of the subject rather than merely replicating its appearance. These results validate the model's strong ability to follow prompts and maintain high subject fidelity, as reflected in the quantitative evaluation.

5. Conclusion

In this paper, we demonstrate the potential of auto-regressive models for personalized image synthesis through a two-stage training strategy, first optimizing text embedding and then fine-tuning transformer. Our approach achieves comparable subject fidelity and prompt following to the state-of-the-art stable diffusion-based methods such as DreamBooth [8]. However, auto-regressive models are slow, taking minutes to generate images, and the fine-tuning also requires 15-20 minutes, limiting real-time applicability. Additionally, the ability to create personalized images raises ethical concerns, such as misuse for misleading content, a challenge common to all generative models. Future work should focus on improving efficiency, addressing ethical risks, and ensuring responsible advancements in personalized generative technologies.

Figure 2. Qualitative results of personalizing objects, categorized by generative capabilities.

Figure 3. Qualitative results of personalizing animals, categorized by generative capabilities.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[2] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 3

[3] 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. 1, 2, 3

[4] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. 2023. 3

[5] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 1, 3

[6] OpenAI. Dalle-2, 2023. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[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. In *CVPR*, 2023. 1, 2, 3

[9] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 1

[10] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[11] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1