Imagine yourself: Tuning-Free Personalized Image Generation

Zecheng He*, Bo Sun*, Felix Juefei-Xu*, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Ning Zhang, Peizhao Zhang, Roshan Sumbaly, Peter Vajda, Animesh Sinha†

GenAI, Meta

Diffusion models have demonstrated remarkable efficacy across various image-to-image tasks. In this research, we introduce Imagine yourself, a state-of-the-art model designed for personalized image generation. Unlike conventional tuning-based personalization techniques, Imagine yourself operates as a tuning-free model, enabling all users to leverage a shared framework without individualized adjustments. Moreover, previous work met challenges balancing identity preservation, following complex prompts and preserving good visual quality, resulting in models having strong copy-paste effect of the reference images. Thus, they can hardly generate images following prompts that require significant changes to the reference image, e.g., changing facial expression, head and body poses, and the diversity of the generated images is low. To address these limitations, our proposed method introduces 1) a new synthetic paired data generation mechanism to encourage image diversity, 2) a fully parallel attention architecture with three text encoders and a fully trainable vision encoder to improve the text faithfulness, and 3) a novel coarse-to-fine multi-stage finetuning methodology that gradually pushes the boundary of visual quality. Our study demonstrates that Imagine yourself surpasses the state-of-the-art personalization model, exhibiting superior capabilities in identity preservation, visual quality, and text alignment. This model establishes a robust foundation for various personalization applications. Human evaluation results validate the model’s SOTA superiority across all aspects (identity preservation, text faithfulness, and visual appeal) compared to the previous personalization models.

Date: July 22, 2024
Correspondence: zechengh@meta.com

1 Introduction

Large scale diffusion models have drawn significant attention. These models, trained on vast amounts of image-text pairs, showcase remarkable semantic understanding capabilities and are able to generate diverse, photo-realistic images based on textual prompts. Due to their unparalleled creative abilities, large-scale diffusion models have found applications across a spectrum of image-to-image tasks beyond the original text-to-image generation, e.g., image editing, image completion, style transfer, and controllable generation.

Personalized image generation techniques have gained significant attention alongside large-scale diffusion models. These methods focus on tailoring image generation to individual preferences or specific user characteristics. By incorporating customization into the generation process, these techniques aim to create images that are more relevant and appealing to the individual user. One line of research tunes a text-to-image model to incorporate the identity (Gal et al., 2022; Ruiz et al., 2023a,b) with a few reference images. However, these methods are not efficient or generalizable as they require a different model to be tuned for each new user.

Recently, another effort has been proposed to obtain personalized diffusion models without subject-specific tuning. This direction of research extracts vision embedding from a reference images and inject it to the diffusion process (Wei et al., 2023; Li et al., 2023; Chen et al., 2023; Ye et al., 2023; Wang et al., 2024; Zhang et al., 2023; Ostashev et al., 2024). While the previous work in this direction can achieve a personalized model
generalizable to all users, it usually comes with a strong over-fitting behavior, i.e., a copy-paste effect, to the reference image. Thus, they can hardly generate images following prompts that require significant changes to the reference image, e.g., change facial expression, head and body poses, hence the diversity of the generated images is low. As a result, these models cannot preserve identity while following complex prompts at the same time.

In this work, we propose Imagine yourself, a state-of-the-art model for personalized image generation without subject-specific fine-tuning. Unlike previous tuning-based personalization techniques which require tuning for each user, Imagine yourself is a tuning-free model where all subjects share a single model. We investigated the key components that lead to the quality improvement of Imagine yourself: **Identity preservation**: Trainable vision encoders with zero conv initialization, and masked vision embedding, **Visual quality**: multi-stage finetune and human-in-the-loop (HITL), **Text alignment**: Synthetic data and parallel attention. Meanwhile, we show that Imagine yourself outperforms the SOTA personalization models (Ye et al., 2023; Wang et al., 2024), with
significant margin in all aspects including identity preservation, visual quality, and text alignment, through large scale human evaluation. In particular, we won +27.8% in text alignment compared to SOTA on complex prompts.

Our contributions can be summarized as follows:

1. We propose Imagine yourself, an innovative state-of-the-art model for personalized image generation. The proposed model can take any reference image as input for customized image generation and does not need tuning for each subject.

2. Imagine yourself incorporates new components and shows significant improvements over the existing models: a new synthetic paired data generation mechanism to encourage image diversity, a fully parallel attention with three text encoders and a fully trainable vision encoder architecture to improve the text faithfulness, and a novel coarse-to-fine multi-stage finetuning methodology that gradually pushes the boundary of visual quality.

3. We provide comprehensive qualitative and quantitative evaluation results compared to the state-of-the-art models. We provide human annotations on thousands of test examples as a golden standard to demonstrate the superior performance of Imagine yourself in all aspects, including identity preservation, prompt alignment, and visual appeal.

Table 1 Quantitative evaluation results of Imagine yourself vs. the SOTA adapter-based model and the SOTA control-based model under head-to-head human evaluation setting.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Head-to-head (win rate)</th>
<th>SOTA adapter-based model</th>
<th>Imagine yourself</th>
<th>Tie</th>
<th>SOTA control-based model</th>
<th>Imagine yourself</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt Alignment</td>
<td>1.2%</td>
<td>46.3%</td>
<td>52.6%</td>
<td>1.6%</td>
<td>32.4%</td>
<td>66.0%</td>
<td></td>
</tr>
<tr>
<td>Identity Preservation</td>
<td>15.1%*</td>
<td>3.2%</td>
<td>81.7%</td>
<td>3.8%</td>
<td>5.8%</td>
<td>90.7%</td>
<td></td>
</tr>
<tr>
<td>Visual Appeal</td>
<td>11.5%</td>
<td>31.6%</td>
<td>57.0%</td>
<td>3.3%</td>
<td>4.2%</td>
<td>92.5%</td>
<td></td>
</tr>
</tbody>
</table>

2 Related Work

2.1 Text-to-Image Diffusion Models

Text-to-image diffusion models represent a cutting-edge paradigm in the domain of deep learning, captivating researchers with their capacity to translate textual descriptions into vibrant visual representations. At their core, these models operate through an iterative refinement process, wherein an initial noise vector is progressively denoised based on text prompts, ultimately yielding the desired image output. A common practice is first translating images to a latent space and denoising in that space. Stable Diffusion \cite{rombach2022high} and its variants SDXL \cite{podell2023stablediffusion}, Stable Diffusion Turbo \cite{sauer2023turbo}, and Stable Diffusion-3 \cite{esser2024 Dezeo} follow this path by increasing the model size, distilling a large model to fewer denoise steps, and leveraging new transformer architectures, respectively.

2.2 Tuning-based Personalization Models

Diffusion-based personalized image generation has indeed garnered increased attention in recent times. This approach involves leveraging diffusion models to generate high-quality and personalized images based on a given input or set of inputs. Technically two streams of personalization models have been proposed. One stream of research tunes a text-to-image model to incorporate the identity. Textual Inversion \cite{gal2022adapting} finetunes a special text tokens for the new identity. DreamBooth \cite{ruiz2023dreambooth} leverages a few images from the same person as reference and a special text token to represent the identity. To accelerate the finetuning process, LoRA \cite{hu2021lora} only tunes a light-weight low-rank adapter rather than the whole diffusion model. HyperDreambooth \cite{ruiz2023hyperdreambooth} further predicts the initial weights of LoRA

* We observed the SOTA control-based model is better in identity preservation than Imagine yourself, due to its hard copy-pasting of the reference image at the center of the image, resulting in unnatural images despite the high identity metric.
from the reference images. However, a major drawback of tuning-based personalization model is that the finetuned model becomes specific for the corresponding identity, and cannot be generalized to new identities. Furthermore, tuning for each user is costly and introduces a long waiting time.

### 2.3 Tuning-free Personalization Models

To overcome the limitations of the tuning-based method, another line of research focuses on one generalized model without identity-specific finetuning. This direction of work extracts vision embedding from the reference image and injects it to the diffusion process. ELITE (Wei et al., 2023) extracts vision features from reference image and converts it to the text-embedding space through a local and a global mapping. PhotoMaker (Li et al., 2023) merges the vision and text tokens and replaces the original text tokens for cross-attention. PhotoVerse (Chen et al., 2023) incorporates an image adapter and a text adapter to merge the vision and language tokens, respectively. IP-Adapter-FaceID-Plus (Ye et al., 2023) leverages face embedding and clip vision encoder for identity preservation. InstantID is a control-based method that (Wang et al., 2024) adds ControlNet (Zhang et al., 2023) to further control the pose and facial expression. MoA (Ostashev et al., 2024) proposes a mixture of attention architecture to better fuse the vision reference and the text prompts.

### 3 Method

Our proposed Imagine yourself takes a single face image of a specific subject, and generates visually appealing personalized images guided by text prompts. Our method can follow complex prompt guidance and generate images with diverse head and body poses, expressions, style, and layout.

To push the boundaries of personalized image generation, our approach begins by identifying three key facets crucial to eliciting a satisfying human visual experience: identity preservation, prompt alignment, and visual appeal (Section 3.2). We then introduce novel techniques tailored to enhance each of these aspects. Specifically, we propose a novel synthetic paired data generation mechanism (Section 3.3), new fully parallel architecture that incorporates three text encoders and a trainable vision encoder for optimizing identity preservation and text-alignment (Section 3.4), a novel coarse-to-fine multi-stage finetuning methodology designed to progressively enhance visual appeal, thereby pushing the visual appeal boundary of generated images. (Section 3.5). Finally, we demonstrate Imagine yourself is generalizable to multi-subjects personalization in Section 3.6.

#### 3.1 Preliminary

Text-to-Image diffusion models gradually turn a noise $\epsilon$ to a clear image $x_0$. While the diffusion process can happen in the pixel space (Ramesh et al., 2022; Saharia et al., 2022), a common practice is to have latent diffusion models (LDM) perform diffusion process in a latent space $z = E(x_0)$. During training, the LDM models optimize the reconstruction loss in the latent space:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(\mu, \sigma^2)} \left\| \epsilon - \epsilon_\theta(z_t), t \right\|^2_2$$  \hspace{1cm} (1)

where $\mathcal{L}_{\text{diffusion}}$ is the diffusion loss. $\epsilon_\theta$ represents the diffusion model. $z_t$ is the noised input to the model at timestep $t$.

It is a common practice to use text or other condition signals $C$ to guide the diffusion process. Thus, the conditioned diffusion process generates images following the condition signals. Usually, the text condition is incorporated with the diffusion model through cross-attention mechanism:

$$\text{Attn}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V$$  \hspace{1cm} (2)

where $K = W_K C$, $V = W_V C$ represents transforms that map the condition $C$ to the cross-attention key and values. $Q = W_Q \phi(x_t)$ represents the hidden state of the diffusion model.
3.2 Overview

Figure 2 provides an illustration of the proposed model architecture. The key of using diffusion models for personalized image generation is incorporating the reference identity as an additional control signal to the diffusion model. We propose to extract the identity information from the reference image through a trainable clip patch encoder. The identity vision signal is then added to the text signals through a parallel cross attention module. To better preserve the high visual quality of the foundation model, we leveraged low-rank adapters (LoRA) to freeze the self-attention and text cross-attention modules while only fine-tuning the adapters.

3.3 Synthetic Paired Data (SynPairs)

We observed that one critical issue during training is the use of unpaired data, i.e., the cropped image as input and the original image as target. It can introduce a severe copy-paste effect, making the model hard to learn the true identity relationship between input and output more than duplicating the reference image. Thus, the model is not able to generate images that follow hard prompts, e.g., change expression or head orientation.

To this end, we proposed a new synthetic data generation recipe to create high-quality paired data (same identity with varying expression, pose, and lighting conditions, etc.) for training. Compared to directly sourcing real paired data, which is not readily available, our study shows that curating the paired data synthetically allows us to retain higher quality data to further enhancing several aspects of the Imagine yourself model.

To generate SynPairs data, we first obtain a dense image caption of the real reference image via a multi-modal LLM. The caption then flows through a caption rewrite stage based on Llama3 (Meta AI) to inject more gaze and pose diversity in the caption. The rewritten caption is then fed to a text-to-image generation tool such as Emu (Dai et al., 2023) as the prompt to produce a high quality synthetic images. Next, we refine the
generated image identity based on the reference image identity. After a large number of curated synthetic pairs is generated they go through an automatic filter based on similarity.

3.4 Model Architecture

3.4.1 Vision Encoder

We propose to use a trainable CLIP ViT-H patch vision encoder to extract the identity control signal from the reference image. Unlike previous work that heavily relied on face embedding, we observed that a general trainable vision encoder can provide adequate information to preserve the identity.

To further improve the identity preservation capability, we crop the face area and mask the corresponding background of the reference image to avoid the model attending to the non-critical areas, e.g., image background and non-face area in the cropped image. Figure 2 illustrates the vision embedding workflow. We also proposed to use zero
_conv as initialization to avoid adding noisy control signals at the beginning of training.

3.4.2 Text Encoders

We employ three distinct text-encoders: CLIP ViT-L (Radford et al., 2021) text encoder, T5-XXL (Raffel et al., 2020), and ByT5 (Xue et al., 2022), as the text conditioning mechanisms. The selection of these encoders is driven by their respective strengths and suitability for specific tasks. The CLIP text encoder, for instance, shares a common space with the CLIP vision encoder, facilitating enhanced identity preservation. To capitalize on this alignment, we initialize the cross-attention module of the vision encoder with the pre-trained CLIP text encoder. Meanwhile, T5-XXL is specifically chosen for its proficiency in comprehending long and intricate text prompts, making it instrumental in handling complex input data. Furthermore, the ByT5 model is integrated for its supreme capability in encoding characters. We leverage ByT5 to improve visual text generating in the image, e.g., text on a signage.
3.4.3 Fully Parallel Image-Text Fusion

We investigated a parallel attention architecture to incorporate the vision and text conditions. Specifically, the newly added vision condition from the reference image and the spatial features fuse through a new vision cross-attention module. The output of the new vision cross-attention module is then added to the text cross-attention output. In our experiments, this design better balances the vision and text control than concatenating the text and vision controls.

3.4.4 LoRA

To preserve the visual quality from the foundation model, we leveraged low-rank adapters (LoRA) on top of the cross-attention module. The self-attention and text cross-attention modules in the foundation Unet are frozen. We observed that this design not only better preserves the foundation model’s image generation capability, but also accelerates the convergence speed by up to 5x.

3.5 Multi-Stage Finetune

We propose a multi-stage finetuning with interleaved real and synthetic data that help us achieve the best trade-off between editability and identity preservation. In the first two stages, we leverage large-scale data (nine millions) to pretrain the model to be able to condition on a reference identity. For the later stages, we finetune our pretrained checkpoint with high-quality, aesthetic images collected through Human-In-The-Loop (HITL). Empirically, we found training with real images gives the best identity preservation, while training with synthetic images gives better prompt alignment (editability). Synthetic images are generated from its respective prompt, therefore the image-text alignment is high and there is less noisy information during training, but the identity information is not as rich as in real data. This is why we adopt the interleaved training recipe, as shown in Figure 5. After the 1st real data pretraining, the model is able to condition on
image, after 2nd synthetic data pretraining, the prompt alignment is high but identity is not perfect, after the 3rd high-quality real-data finetune, the identity is good but prompt alignment drops, the 4th high-quality synthetic data finetune achieves the best trade-off between identity and editability.

3.6 Extension to Multi-Subject Personalization

The previously introduced fully parallel image-text fusion pipeline (Section 3.4.3) can be flexibly extended to accommodate multi-subject personalization. In the two-person scenario for example, instead of passing the global embedding and patch embedding of the single reference image into the $K$ and $V$ components as shown in the top left branch of Figure 4, we can concatenate the vision embedding from both reference images and passing it into the $K$ and $V$ components. Given this setup, through training, the network learns how to map from reference$_i$ to subject$_j$ in the group photo while generating prompt-induced image context accordingly. Some examples of the two-person personalization results are shown in Figure 11.

4 Experiments

In this section, we perform both qualitative and quantitative evaluations of our model. We also compare our model to the SOTA personalization models. Results show that our model outperforms the existing models on all axes setting the new state-of-the-art.

4.1 Qualitative Evaluation

We show examples of our model generated image in Figures 6-10. Our model generates visually appealing images that both preserve the identity and follow the prompt faithfully.

4.2 Quantitative Evaluation

4.2.1 Evaluation Dataset

To quantitatively evaluate Imagine yourself, we created an evaluation set consisting of two parts: (i) reference images, and (ii) eval prompts. To have a comprehensive comparison in all representative cases, we collected a total of 51 reference identities covering different gender, race, and skin tone. We created a list of 65 prompts to evaluate the model. It widely covers a wide range of usage scenarios, and also including hard prompts that require face expression or pose changes, camera motions, and stylization. These prompts help to assess the model’s ability to engage in more complex and nuanced interactions, diverse pose generation, and harmonization. Each identity is paired with all 65 prompts, so a total of $51 \times 65 = 3315$ generations for one round of human evaluation. The distributions of the prompts is shown in Figure 12.
Figure 12 Distribution of the evaluated prompts. It widely covers a wide range of usage scenarios, and also including hard prompts that require face expression or pose changes, camera motions, etc.

4.2.2 Benchmarked Methods

We benchmarked the SOTA adapter-based personalization model and the SOTA control-based model. For the adapter-based method, we select the best one that strikes the best balance among visual appeal, identity preservation, and prompt alignment, the three axes that we evaluate our models on. For control-based method, we noticed that the choice of the pose image plays an important role in how the final generated image is composed, i.e., for some prompts, a carefully chosen pose image can make the generated images look better or worse. For a fair comparison, we use the reference image itself as the pose condition.

4.2.3 Human Evaluation

To evaluate the quality of the generated images, we conducted a large-scale annotation process that assessed various aspects of the images. We used human annotation as the gold standard to assess the model’s performance (standalone evaluation) and compare it with other models (head-to-head evaluation).

In the standalone evaluation, we presented annotators with the input image, prompt, and a generated image and asked them three questions to rate on a scale of Strong Pass / Weak Pass / Fail. (1) Identity Similarity: Does the subject in the output image appear to have the same identity as the subject in the original image? (2) Prompt Alignment: Does the output image follow the personalization prompt faithfully? (3) Visual Appeal: Is the output image visually appealing? In the head-to-head model evaluation, we compared one model against another on the same three axes.

As shown in Table 1, Imagine yourself outperforms the two state-of-the-art methods adapter-based model and control-based model by a significant margin in most axes. Specifically, Imagine yourself is significantly better in prompt alignment, with a +45.1% and +30.8% improvement over the SOTA adapter-based model and the SOTA control-based model, respectively. However, we observed that the control-based model is better in identity preservation than Imagine yourself, due to its hard copy-pasting of the reference image at the center of the image, resulting in unnatural images despite the high identity metric.

4.3 Ablation Study

In our ablation study, we examined the effectiveness of various components within our proposed Imagine yourself. Main ablation results are shown in Table 2.

4.3.1 Impact of Multi-stage Finetune

The ablation results highlight the impact of multi-stage fine-tuning. Reducing the multi-stage fine-tuning to a single stage significantly degrade all metrics, especially 25.5% in prompt alignment and 42.0% in visual appeal. Moreover, we observe that, the synthetic fine-tune stages provides better prompt alignment and the real data fine-tune stage improves the identity preservation capability.
Table 2  Ablation study of different component in Imagine yourself.

<table>
<thead>
<tr>
<th>Model</th>
<th>Prompt Alignment</th>
<th>Identity Preservation</th>
<th>Visual Appeal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imagine yourself w/o multi-stage finetune</td>
<td>55.3%</td>
<td>90.7%</td>
<td>45.7%</td>
</tr>
<tr>
<td>Imagine yourself w/o fully parallel attention</td>
<td>75.6%</td>
<td>81.9%</td>
<td>65.7%</td>
</tr>
<tr>
<td>Imagine yourself w/o SynPairs</td>
<td>57.0%</td>
<td>95.5%</td>
<td>83.4%</td>
</tr>
<tr>
<td>Imagine yourself</td>
<td>80.8%</td>
<td>83.3%</td>
<td>87.7%</td>
</tr>
</tbody>
</table>

4.3.2 Impact of Fully Parallel Attention

We ablate removing the full parallel attention to a standard token concatenate design to show the impact of the fully parallel attention architecture. We observed that all metrics, specifically 5.2% in prompt alignment, 1.4% in identity preservation, and 22.0% in visual appeal, respectively. This shows the importance incorporating all three text encoders and the vision encoder through fully parallel attention.

4.3.3 Impact of Synthetic Pairs

SynPairs increase the diversity of the generated images by eliminating the copy-paste effect. Our ablation verifies this assumption and demonstrate better prompt-alignment compared to the model without synthetic paired training. We observed that it is especially effective for the complex prompts that require strong changes to the original images, e.g., expression change, covering the face, or turning head, etc. However, we observed a regression in identity preservation with SynPair training because the faces in the corresponding reference and target pair are not exactly the same. Future work will focus on improving the face similarity of SynPair training data.

5 Future Work

We would like to continue research and explore the following directions: 1) extend personalized image to video generation. The key is to consistently preserve the identity and scene in video generation. 2) While Imagine yourself has improved the prompt-alignment against existing models, we observed that it still has limitation in following prompts describing very complex poses, e.g., jumping from a mountain. Future work will focus on improving the generated images’ quality on these prompts.

6 Societal Risk

Actions have been taken to mitigate these risks during the Imagine yourself model development, including harmful content filtering, unbiased data fine-tuning, and private data filtering.

7 Conclusion

In this study, we introduce Imagine yourself, a pioneering model tailored for personalized image generation. Unlike traditional tuning-based approaches, Imagine yourself operates as a tuning-free solution, offering a shared framework accessible to all users without the need for individual adjustments. Imagine yourself overcomes the prior research limit in handling the intricate balance between preserving identity, following complex prompts, and maintaining visual quality by introducing 1) a novel synthetic paired data generation mechanism to foster image diversity, 2) a fully parallel attention architecture featuring three text encoders and a fully trainable vision encoder to enhance text faithfulness, and 3) a novel coarse-to-fine multi-stage fine-tuning methodology to progressively enhance visual quality. We perform large-scale human evaluation on thousands of examples and showcase that Imagine yourself outperforms state-of-the-art personalization models, demonstrating superior capabilities in identity preservation, visual quality, and text alignment.
Acknowledgment

We extend our gratitude to the following people for their contributions: Xiaoliang Dai, Ji Hou, Kunpeng Li, Sam Tsai, Jialiang Wang, Zijian He for text-to-image foundation model; Eric Alamillo, Xiao Chu, Yangwen Liu, Yan Yan, Jonas Kohler, Artsiom Sanakoyeu, Ali Thabet, Arantxa Casanova Paga, Zhipeng Fan for model evaluation and discussion; Aaron Nissenbaum, Becky McMahon, Mo Metanat, Danny Trinh, Jack Hanlon, Tali Zvi, Manohar Paluri, Prashant Ratanchandani, and Ahmad Al-Dahle for support and leadership.

References


Figure 6  More visualizations (1/5) of generated personalized images using Imagine yourself.
Figure 7  More visualizations (2/5) of generated personalized images using Imagine yourself.
Figure 8  More visualizations (3/5) of generated personalized images using Imagine yourself.
Figure 9  More visualizations (4/5) of generated personalized images using Imagine yourself.
Figure 10 More visualizations (5/5) of generated personalized images using Imagine yourself.
Figure 11 Multi-player personalization using Imagine yourself.