MagiCapture: High-Resolution Multi-Concept Portrait Customization

Junha Hyung\textsuperscript{1,}, Jaeyo Shin\textsuperscript{2,}, and Jaegul Choo\textsuperscript{1}

\textsuperscript{1}KAIST AI  
\textsuperscript{2}Sogang University  
\{sharpeeee, jchoo\}@kaist.ac.kr, tlswody123@sogang.ac.kr

Abstract

Large-scale text-to-image models including Stable Diffusion are capable of generating high-fidelity photorealistic portrait images. There is an active research area dedicated to personalizing these models, aiming to synthesize specific subjects or styles using provided sets of reference images. However, despite the plausible results from these personalization methods, they tend to produce images that often fall short of realism and are not yet on a commercially viable level. This is particularly noticeable in portrait image generation, where any unnatural artifact in human faces is easily discernible due to our inherent human bias. To address this, we introduce MagiCapture, a personalization method for integrating subject and style concepts to generate high-resolution portrait images using just a few subject and style references. For instance, given a handful of random selfies, our fine-tuned model can generate high-quality portrait images in specific styles, such as passport or profile photos. The main challenge with this task is the absence of ground truth for the composed concepts, leading to a reduction in the quality of the final output and an identity shift of the source subject. To address these issues, we present a novel Attention Refocusing loss coupled with auxiliary priors, both of which facilitate robust learning within this weakly supervised learning setting. Our pipeline also includes additional post-processing steps to ensure the creation of highly realistic outputs. MagiCapture outperforms other baselines in both quantitative and qualitative evaluations and can also be generalized to other non-human objects.

Introduction

To obtain high-quality portrait images suitable for resumes or wedding events, individuals typically have to visit a photo studio, followed by a costly and time-consuming process of photo retouching. Imagine a scenario where all that’s required is a few selfie images and reference photos, and you could receive high-quality portrait images in specific styles, such as passport or profile photos. This paper aims to automate this process.

Recent advancements in large-scale text-to-image models, such as Stable Diffusion (Rombach et al. 2022) and Image (Saharia et al. 2022), have made it possible to generate high-fidelity, photorealistic portrait images. The active area of research dedicated to personalizing these models seeks to synthesize specific subjects or styles using provided sets of train images. In this work, we formulate our task as a multi-concept customization problem. Here, the source content and reference style are learned respectively, and the composed output is generated. Unlike text-driven editing, using reference images allows users to provide fine-grained guidance, making it more suitable for this task.

However, despite the promising results achieved by previous personalization methods, they often produce images that lack realism and fall short of commercial viability. This problem primarily arises from attempting to update the parameters of large models using only a small number of images. This decline in quality becomes even more evident in a multi-concept generation, where the absence of ground truth images for the composed concepts frequently leads to the unnatural blending of disparate concepts or deviation from the original concepts. This issue is particularly conspicuous in portrait image generation, as any unnatural artifacts or shifts in identity are easily noticeable due to our inherent human bias.

To address these issues, we present MagiCapture, a multi-concept personalization method for the fusion of subject and style concepts to generate high-resolution portrait images with only a few subject and style references. Our method employs composed prompt learning, incorporating the composed prompt as part of the training process, which enhances the robust integration of source content and reference style. This is achieved through the use of pseudo labels and auxiliary loss. Moreover, we propose the Attention Refocusing loss in conjunction with a masked reconstruction objective, a crucial strategy for achieving information disentanglement and preventing information leakage during inference. MagiCapture outperforms other baselines in both quantitative and qualitative evaluations and can be generalized to other non-human objects with just a few modifications.

The main contributions of our paper are as follows:
• We introduce a multi-concept personalization method capable of generating high-resolution portrait images that faithfully capture the characteristics of both source and reference images.

• We present a novel Attention Refocusing loss combined with masked reconstruction objective, effectively disen-tangling the desired information from input images and preventing information leakage during the generation process.

• We put forth a composed prompt learning approach that leverages pseudo-labels and auxiliary loss, facilitating the robust integration of source content and reference style.

• In both quantitative and qualitative assessments, our method surpasses other baseline approaches and, with minor adjustments, can be adapted to generate images of non-human objects.

### Related Work

**Text-to-image diffusion models**

Diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2019; Song et al. 2020; Song, Meng, and Ermon 2020) have recently achieved remarkable success in image generation, driving advancements in various applications and fields. Their powerful performance has significantly propelled the field of text-guided image synthesis (Nichol et al. 2021; Kim, Kwon, and Ye 2022; Saharia et al. 2022; Ramesh et al. 2022) forward. In particular, large-scale text-to-image diffusion models, trained on extensive text-image pair datasets, have set new benchmarks. Notable examples include Stable diffusion (von Platen et al. 2022) and Imagen (Saharia et al. 2022). Our work is built upon the pre-trained stable diffusion model.

**Personalization of Text-to-image Models.**

Personalizing generative models for specific concepts is a key goal in the vision field. With the rise of GANs, there have been efforts to fine-tune GANs, like Pivotal Tuning (Roich et al. 2022), based on GAN inversion (Zhu et al. 2020).

More recently, studies have sought to personalize diffusion models using small image datasets, typically 3 ∼ 5 images, associated with a particular object or style and incorporating specialized text tokens to embed such concepts. For instance, when customizing models for a specific dog, the prompt “a [V1] dog” is used so that the special token can learn information specific to the dog. DreamBooth (Ruiz et al. 2023) fine-tunes entire weights, Textual Inversion (Gal et al. 2022) adjusts text embeddings, and Custom Diffusion (Kumari et al. 2023) adapts the mapping matrix for the cross-attention layer. While effective in learning concepts, these models sometimes generate less realistic or identity-losing images. Methods like ELITE (Wei et al. 2023) and InstantBooth (Shi et al. 2023) employ a data-driven approach for encoder-based domain tuning, which is not directly comparable to our approach.

Our method differs from concurrent works like SVDiff (Han et al. 2023), FastComposer (Xiao et al. 2023), and Break-A-Scene (Avrahami et al. 2023), which use similar techniques like attention loss or composed prompts. Unlike SVDiff’s collage approach (Cut-Mix-Unmix), our method is tailored for style-mixed outputs, enhancing the quality
Figure 2: The overall pipeline of MagiCapture, where the training process is formulated as multi-task learning of three different tasks: source, reference, and composed prompt learning. In the composed prompt learning, reference style images serve as pseudo-labels, along with auxiliary identity loss between the source and predicted images. Attention Refocusing loss is applied to all three tasks. After training, users can generate high-fidelity images with integrated concepts and can further manipulate them using varying text conditions.

of multi-concept portraits. Distinct from FastComposer and Break-A-Scene, our attention loss only targets regions in the attention map not present in the ground-truth mask ($A_k[i, j]$ for all $(i, j) \in \{(i, j) \mid M_v[i, j] = 0\}$), allowing for the varying optimal values for other areas.

**Preliminaries**

**Diffusion Models.** Diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2019; Song et al. 2020; Song, Meng, and Ermon 2020) are a class of generative models that create images through an iterative denoising process. These models comprise a forward and backward pass. During the forward pass, an input image $x(0)$ is progressively noised using the equation $x(t) = \sqrt{\alpha_t}x(0) + \sqrt{1 - \alpha_t}\epsilon$, where $\epsilon$ represents standard Gaussian noise and $\{\alpha_t\}$ is a pre-defined noise schedule with timestep $t$, $1 < t < T$. During backward pass, the generated image is obtained by denoising the starting noise $x_T$ using a UNet $\epsilon_\theta(x(t), t)$, which is trained to predict noise at the input timestep $t$. Latent diffusion models (LDM) (Rombach et al. 2022) are a variant of diffusion models where the denoising process occurs in the latent space. An image encoder $\mathcal{E}$ transforms the input image $x$ into a latent representation $z = \mathcal{E}(x)$. The denoised latent representation is decoded into the final image $x^{(0)'} = \mathcal{D}(z^{(0)})$, with a decoder $\mathcal{D}$. Stable diffusion (von Platen et al. 2022) is a text-guided latent diffusion model (LDM) trained on large-scale text-image pairs. It has the following objective:

$$L_{LDM} = \mathbb{E}_{z,c,t} \left[ ||\epsilon_\theta(z^{(t)}, t, c) - \epsilon||^2 \right],$$

where $c$ refers to the text condition.

**Attention maps** Large-scale text-to-image diffusion models utilize cross-attention layers for text-conditioning. In Stable Diffusion (Rombach et al. 2022), CLIP text encoder (Radford et al. 2021) is used to produce text embedding features. These text embeddings are then transformed to obtain the key $K$ and value $V$ for the cross-attention layer through linear mapping, and spatial feature of image is projected to query $Q$. The attention map of the cross-attention layer is computed as:

$$A = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right).$$

The attention map corresponding to a specific token with index $k$ can be obtained as $A_k = A[k]$. Such attention maps are useful for visualizing the influence of individual tokens in the text prompt. Moreover, they can be altered or manipulated for the purpose of image editing, as demonstrated in Prompt-to-Prompt (Hertz et al. 2022).

**Method**

Given a small set of source images and reference style images, the goal of this paper is to synthesize images that integrate the source content with the reference style. While our method is primarily designed for generating portrait images, it can be easily adapted to handle other types of content with minor modifications. We utilize the customization of each concepts during the optimization phase and employ a composed prompt during inference to generate multi-concept images. A comprehensive overview of our approach is depicted in Fig. 2, and the details of our method will be elaborated upon in the subsequent sections.
Two-phase Optimization. Similar to Pivotal Tuning (Roich et al. 2022) in GAN inversion, our method consists of two-phase optimization. In the first phase, we optimize the text embeddings for the special tokens $[V^*]$ using the reconstruction objective as in (Gal et al. 2022). While optimizing the text embeddings is not sufficient for achieving high-fidelity customization, it serves as a useful initialization for the subsequent phase. In the second phase, we jointly optimize the text embeddings and model parameters with the same objective. Rather than optimizing the entire model, we apply the LoRA (Hu et al. 2021), where only the residuals $\Delta W$ of the projection layers in the cross-attention module are trained using low-rank decomposition. Specifically, the updated parameters are expressed as:

$$W' = W + \Delta W, \quad \Delta W = UV^T,$$

where $U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}$, and $r \ll n, m$. Empirically, we find that this two-phase optimization coupled with LoRA strikes a favorable balance between reconstruction and generalization. It preserves the model’s generalization capabilities for unseen prompts while effectively capturing the finer details of the source images.

Masked Reconstruction. In our approach, a source prompt $c_s$ (e.g., A photo of a $[V1]$ person.) and a reference prompt $c_r$ (e.g., A photo of a person in the $[V2]$ style.) are used to reconstruct the source image $I_s$ and a target style image $I_r$, respectively. It is crucial to disentangle the identity of the source subject from non-facial regions, such as the background and clothing, to prevent this unwanted information from being encoded into the special token $[V1]$. Similarly, we need to disentangle the reference image to ensure that the facial details of the person in the reference image are not embedded into the special token $[V2]$. To achieve this, we propose to use a masked reconstruction loss. Specifically, we employ a mask that indicates the relevant region and apply it element-wise to both the ground truth latent code and the predicted latent code. In the context of portrait generation, a source mask $M_s$ indicates the facial region of the image $I_s$, and a target mask $M_r$ denotes the non-facial areas of the reference image $I_r$. Formally, the masked reconstruction loss for the source and the reference prompts are given by:

$$L_{mask}^s = \mathbb{E}_{z_s, c_s, t} \left[ ||e \odot M_s - \theta_{e}(z^{(t)}_s, t, c_s) \odot M_s||^2_2 \right], \quad (4)$$

$$L_{mask}^r = \mathbb{E}_{z_r, c_r, t} \left[ ||e \odot M_r - \theta_{e}(z^{(t)}_r, t, c_r) \odot M_r||^2_2 \right], \quad (5)$$

where $z^{(t)}_s$ and $z^{(t)}_r$ are the source and reference noised latent at timestep $t \sim \text{Uniform}(1, T)$ and $e \sim \mathcal{N}(0, I)$.

Composed Prompt Learning. Generating images with a composed prompt $c_c$ such as “A photo of a $[V1]$ person in the $[V2]$ style,” leads to undefined behavior because the model had not been customized on such prompts. Typically, the resulting images generated using these unseen composed prompts suffer from a shift in the identity of the source subject and a decline in output quality. To address this issue, we include training on the composed prompt. However, no ground truth image exists for such a prompt. We approach this challenge as a weakly-supervised learning problem, where there are no available ground truth labels. We craft pseudo-labels and develop an auxiliary objective function to suit our needs. In the context of the portrait generation task, we want to retain the overall composition, pose, and appearance from the reference style image, excluding the facial identity. To achieve this, we employ the masked reconstruction objective given by:

$$L_{mask}^c = \mathbb{E}_{z_r, c_c, t} \left[ ||e \odot M_r - \theta_{e}(z^{(t)}_c, t, c_c) \odot M_r||^2_2 \right]. \quad (6)$$

For the facial regions, we use an auxiliary identity loss that utilizes a pre-trained face recognition model (Deng et al. 2019) $R$ and cropping function $B$ conditioned by the face detection model (Deng et al. 2020):

$$L_{id} = \mathbb{E}_{z^{(0)}_r, I_s} \left[ 1 - \cos(R(B(\hat{z}^{(0)})), R(B(I_s))) \right], \quad (7)$$

where $\cos$ denotes the cosine similarity and $\hat{z}^{(0)} = D(z^{(0)})$ refers to the estimated clean image from $z_r^{(t_i)}$ using Tweedie’s formula (Kim and Ye 2021). Timestep $t_{id}$ is sampled as $t_{id} \sim \text{Uniform}(1, T')$, where $T' < T$, to avoid blurry and inaccurate $\hat{z}^{(0)}$ estimated from noisy latent with large timesteps, which can impair cropping or yield odd facial embeddings.

We augment the composed prompt $c_c$ by randomly selecting from predefined prompt templates to boost editing stability and generalization.

Attention Refocusing. When optimizing with training images, it is vital to achieve information disentanglement, ensuring that special tokens exclusively embed the information of the region of interest, denoted as $M_v$ for $v \in \{s, r\}$. However, the masked reconstruction objective falls short of this goal because the presence of transformer layers in the
Table 1: Quantitative comparison of our method against DreamBooth (Ruiz et al. 2023), Textual Inversion (Gal et al. 2022), and Custom Diffusion (Kumari et al. 2023). Our method outperforms other baselines in terms of identity similarity measured between the source images (CSIM), masked CLIP similarity measure (Style), and Aesthetic score (Schuhmann Aug 2022).

<table>
<thead>
<tr>
<th>Method</th>
<th>CSIM ↑</th>
<th>Style ↑</th>
<th>Aesthetic ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DreamBooth</td>
<td>0.102</td>
<td>0.720</td>
<td>5.770</td>
</tr>
<tr>
<td>Textual Inversion</td>
<td>0.224</td>
<td>0.623</td>
<td>5.670</td>
</tr>
<tr>
<td>Custom Diffusion</td>
<td>0.436</td>
<td>0.606</td>
<td>5.263</td>
</tr>
<tr>
<td>Ours w/o AR &amp; CP</td>
<td>0.429</td>
<td>0.726</td>
<td>6.178</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.566</strong></td>
<td><strong>0.730</strong></td>
<td><strong>6.218</strong></td>
</tr>
</tbody>
</table>

To solve this issue, we propose a novel Attention Refocusing (AR) loss, which steers the cross attention maps \(A_k\) of the special token \(V^+\) (where \(k = \text{index}(\{V^+\})\) using a binary target mask. Our AR loss incorporates two crucial details: First, it is applied only to regions where \(\neg M_e\), where the mask value is zero. For the attention map values \(A_k[i,j]\) where \((i,j) \in \{(i,j)|M_e[i,j] = 1\}\), the optimal values can vary across different UNet layers and denoising time steps, so they do not necessarily have to be close to 1. Conversely, for \(A_k[i,j]\) where \((i,j) \in \{(i,j)|M_e[i,j] = 0\}\), the values should be forced to 0 to achieve information disentanglement during training and minimize information spill in the inference stage. Second, it is essential to scale the attention maps to the \([0,1]\) range. Both of these techniques are re-
Figure 4: Curated results of MagiCapture.
Table 2: User study of our method against DreamBooth (Ruiz et al. 2023), Textual Inversion (Gal et al. 2022), and Custom Diffusion (Kumari et al. 2023). Our method outperforms other baselines in terms of identity similarity score (ID), style similarity measure (Style), and image fidelity score (Fidelity).

Ablation Study. As shown in Fig. 3, we find that Attention Refocusing loss effectively prevents attention maps from attending to unwanted regions, mitigating information spill and promoting information disentanglement. Empirically, we observe that the Attention Refocusing loss should only be applied during the second phase of training (LoRA training). We infer that text embeddings are not well-suited for learning geometric information related to attention maps. Moreover, without composed prompt learning, the generated images often exhibit undefined behaviors where only one of the source or reference sets is evident in the image, without blending. We present the evaluation metrics for both the presence and absence of composed prompt learning (CP) and Attention Refocusing (AR) in Table 1. For more results and detailed analysis, please refer to the supplement.

Applications. Since our method is robust to generalizations, users can further manipulate the composed results using prompts with more descriptions (e.g., $c' = \text{“A photo of [V1] person in the [V2] style, wearing sunglasses.”} \) We demonstrate such results in Fig. 6 and in the supplement. Furthermore, our method is adaptable for handling different types of content, including non-human images. For methodologies and results related to non-human content, please refer to the supplementary material.

Limitations and Conclusions

Our method occasionally produces abnormal body parts such as limbs, fingers, as shown in Fig. 7. Furthermore, the model tends to exhibit lower fidelity for non-white subjects and demonstrates a noticeable gender bias—for instance, it struggles to accurately generate images of men wearing wedding dresses. These issues are largely related to the inherent biases of the pre-trained text-to-image models, and addressing these problems within a few-shot setting represents a significant avenue for future research. We acknowledge the ethical implications of our work and are committed to taking them seriously. We are also proactive in leading and supporting efforts to prevent potential misuse of our contributions.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913), and Institute of Information & communications Technology Planning &
References


