

# PersonaMagic: Stage-Regulated High-Fidelity Face Customization with Tandem Equilibrium

Xinzhe Li<sup>1</sup>, Jiahui Zhan<sup>1,2\*</sup>, Shengfeng He<sup>3</sup>, Yangyang Xu<sup>4</sup>, Junyu Dong<sup>1</sup>, Huaidong Zhang<sup>5</sup>, Yong Du<sup>1†</sup>

<sup>1</sup>Ocean University of China, China

<sup>2</sup>Shanghai Jiao Tong University, China

<sup>3</sup>Singapore Management University, Singapore

<sup>4</sup>Harbin Institute of Technology (Shenzhen), China

<sup>5</sup>South China University of Technology, China

{lixinzhe, zhanjiahui}@stu.ouc.edu.cn, shengfenghe@smu.edu.sg, cnnlstm@gmail.com, huaidongz@scut.edu.cn, {dongjunyu, csyongdu}@ouc.edu.cn

## Abstract

Personalized image generation has made significant strides in adapting content to novel concepts. However, a persistent challenge remains: balancing the accurate reconstruction of unseen concepts with the need for editability according to the prompt, especially when dealing with the complex nuances of facial features. In this study, we delve into the temporal dynamics of the text-to-image conditioning process, emphasizing the crucial role of stage partitioning in introducing new concepts. We present PersonaMagic, a stage-regulated generative technique designed for high-fidelity face customization. Using a simple MLP network, our method learns a series of embeddings within a specific timestep interval to capture face concepts. Additionally, we develop a Tandem Equilibrium mechanism that adjusts self-attention responses in the text encoder, balancing text description and identity preservation, improving both areas. Extensive experiments confirm the superiority of PersonaMagic over state-of-the-art methods in both qualitative and quantitative evaluations. Moreover, its robustness and flexibility are validated in non-facial domains, and it can also serve as a valuable plug-in for enhancing the performance of pretrained personalization models.

## Introduction

As a natural extension of research on controllability in diffusion models (Ramesh et al. 2022; Rombach et al. 2022), personalized text-to-image generation (Gal et al. 2023a; Ruiz et al. 2023) has emerged as a prominent task. By providing multiple images of a specific subject, users can introduce new concepts to a pre-trained text-to-image diffusion model, enabling it to synthesize the same subject in various contexts. However, despite recent advancements, current methods struggle to align generated outputs with users’ envisioned images, particularly in face customization. The primary challenge lies in preserving the identity of a given face. While it is relatively straightforward to assess the consistency of generated results for common subjects (e.g.,

man-made objects, animals) based on contours and textures, achieving similar consistency in personalizing human faces, with their intricate features, is a more complex task.

New concepts are typically represented as word embeddings (Gal et al. 2023a) and integrated into prompts for customization. When training on a single image, the learned embedding often represents the entire image including background details rather than focusing on the target concept. As a result, personalized face generation usually requires 3 to 5 images of an individual in various scenes and poses to guide the cross-attention maps derived from the learned embedding toward the facial region. Several approaches (Kumari et al. 2023; Tewel et al. 2023) attempt to fine-tune cross-attention layer parameters to adjust the attention of new concepts, but this requires substantial memory costs during training and introduces language drift (Lee, Cho, and Kiela 2019; Lu et al. 2020). Leveraging human-centric datasets, recent methods (Ye et al. 2023; Li et al. 2024) have trained personalization models to generate customized results with accurate identities and more natural human poses. However, we observe a decline in identity preservation when generating images of individuals not included in the training set.

In this paper, we address the fidelity-editability challenge using only a single image. Our main idea is first rooted in the understanding that consistent textual conditioning manifests varying control capabilities across different temporal phases, as reflected in the cross-attention maps. During an intermediate stage of the denoising process, the learned embedding tends to focus on the facial region rather than the entire image, even when training on a single image. Additionally, as noted in previous research (Alaluf et al. 2023), word embeddings capture different concept details across timesteps. This insight inspired us to develop dynamic embeddings within a specific timestep interval to achieve personalized face generation. During training, we divide the reverse process of diffusion models into dynamic and static stages, based on changes in cross-attention maps over time. In the dynamic stage, we introduce a lightweight network to acquire dynamic embeddings at different timesteps, effectively capturing user-provided face information. Conversely,

\*Equal contribution.

†Corresponding author.



Figure 1: *PersonaMagic* seamlessly generates images of new roles, styles, or scenes based on a user-provided portrait. By learning stage-regulated embeddings through a Tandem Equilibrium strategy, our method accurately captures and represents unseen concepts, faithfully creating personas aligned with the provided prompts while minimizing identity distortion.

the static stage employs fixed word embeddings corresponding to supercategory to stabilize training.

With the diffusion model frozen, the learned embeddings become the sole variable for minimizing denoising loss during training. This focus on learned embeddings may cause the diffusion model to overlook the semantics other than new concept. We observe that these neglected word embeddings have lower attention weights in the cross-attention layers of U-Net, resulting from the disproportionately high attention given to the learned embeddings in the text encoder. To address this, we introduce a Tandem Equilibrium (TE) strategy. During training, we input diverse text prompts into the text encoder and balance the attention weights of the new concept with other tokens, ensuring complete semantic representation. Unlike prior methods that required multiple images to generalize the expressiveness of the learned embedding across different scenarios, our TE strategy achieves this by directly operating within the text encoder, eliminating the need to pass latent image features to U-Net. This approach allows us to generate identity-preserving and semantically complete results with a single image, as shown in Fig. 1.

In summary, the contributions of this work are threefold:

- We propose learning dynamic embeddings within a specific range to enable high-fidelity personalized face generation. Even from a single image, the learned embeddings generate the desired cross-attention maps, effectively preserving identity while improving efficiency.
- We introduce a TE strategy to regulate self-attention maps in the text encoder, ensuring that personalized results align closely with textual descriptions without the need for multiple images.
- Extensive quantitative and qualitative experiments validate our effectiveness, demonstrating a strong balance between textual alignment and identity preservation.

## Related Work

**Text-to-Image Diffusion Models.** Diffusion models (Song and Ermon 2019; Song et al. 2021; Tang et al. 2022) have

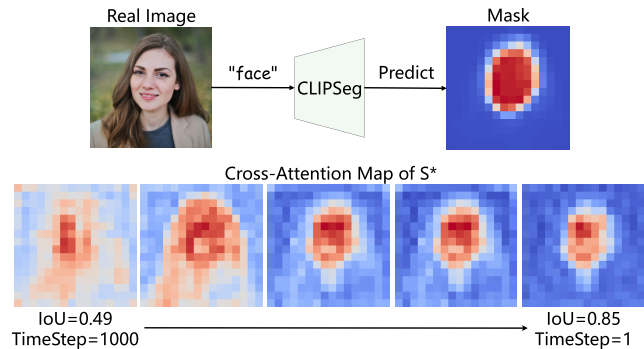


Figure 2: Cross-attention maps of  $S^*$  at each timestep. We calculate IoU with facial mask for stage partition.

recently excelled in generating images from text (Ramesh et al. 2021; Gu et al. 2022; Yu et al. 2022; Saharia et al. 2022). Noteworthy examples include GLIDE (Nichol et al. 2022), which crafts high-resolution images using diverse diffusion models, and DALL-E 2 (Ramesh et al. 2022), generating CLIP (Radford et al. 2021) image embeddings from text through a diffusion model. Imagen (Saharia et al. 2022) enriches semantic information with a pre-trained text encoder (Raffel et al. 2020). Stable Diffusion (Rombach et al. 2022) proposes denoising in a low-dimensional latent space through an autoencoder (Esser, Rombach, and Ommer 2021). While these methods can generate images aligned with text prompts, customizing a specific subject remains challenging. Our aim is to train an efficient network to introduce concept information about an unseen face into a pre-trained text-to-image diffusion model, enabling face customization across various scenes or styles.

**Personalized Image Generation.** Personalized generation methods find widespread applications across computer vision and graphics. Previous methods (Yang et al. 2021; Xu et al. 2021; Nitzan et al. 2022; Roich et al. 2022; Alaluf et al. 2022; Song et al. 2022) typically rely on GANs (Kar-

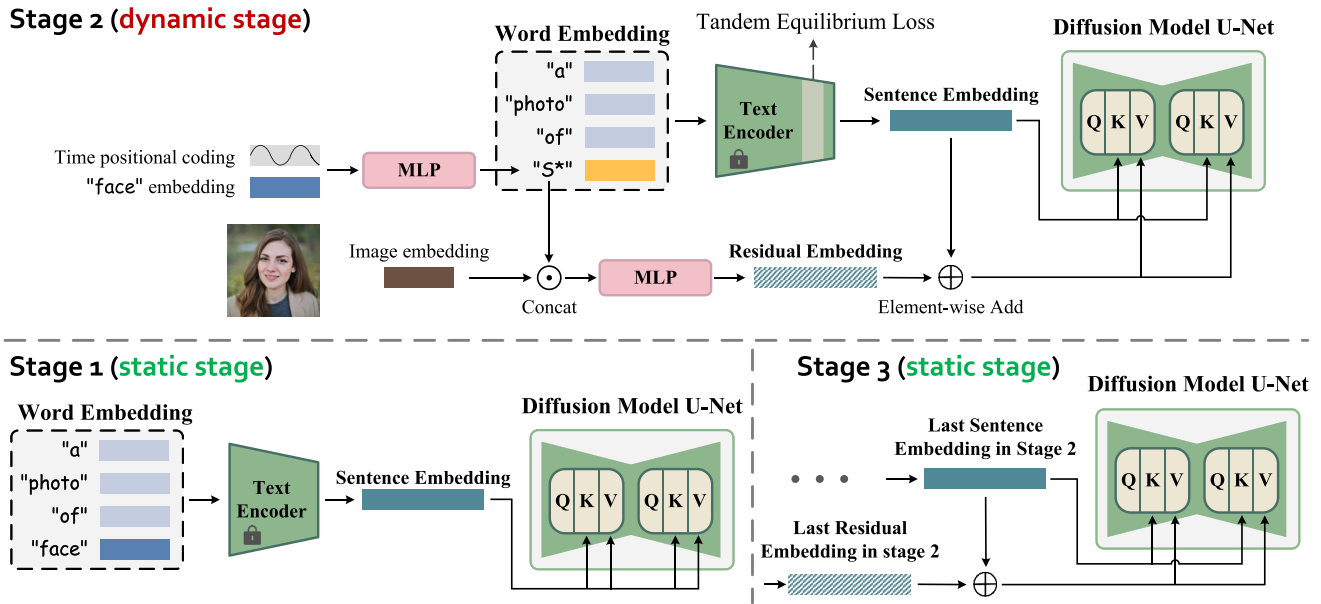


Figure 3: Overview of our pipeline. Given a single image, we learn a series of embeddings during dynamic stage to capture identity information effectively, while employing fixed embeddings in the static stages. The proposed TE strategy is applied in the text encoder, ensuring further alignment of personalized results with textual descriptions.

ras, Laine, and Aila 2019; Karras et al. 2020; Goodfellow et al. 2020), encountering difficulties in handling out-of-domain images. Based on text-to-image diffusion models, Textual Inversion (Gal et al. 2023a) proposed to optimize word embeddings for personalization. DreamBooth (Ruiz et al. 2023) recommends fine-tuning diffusion model parameters to introduce new concepts. However, these methods struggle to excel in both identity preservation and text similarity simultaneously. In contrast, we propose TE strategy, leading to trade-off in both aspects. Recently, several studies (Ye et al. 2023; Gal et al. 2023b; Li et al. 2024) have trained on human-centric datasets, acquiring more comprehensive prior knowledge for personalization. For instance, IP-Adapter (Ye et al. 2023) introduces a decoupled cross-attention strategy for semantic guidance. Photomaker (Li et al. 2024) extracts ID embeddings from facial images to provide identity information to diffusion model. However, the performance of these approaches may be influenced by biased datasets, resulting in poor identity preservation for unseen faces. Instead, our method, guided by a self-contained stage-regulated text conditioning mechanism, efficiently customizes unseen persons with precise identity preservation.

## Method

**Stage-regulated Textual Conditioning.** We conduct an experiment to understand the temporal dynamics of the text-to-image conditioning process. Given a facial image, we use a rare token  $S^*$  to represent this new concept. Through a simple fully connected network and the textual condition “A photo of  $S^*$ ”, we learn a set of word embeddings that vary across timesteps. We then visualize cross-attention maps of

$S^*$  at each timestep, as shown in Fig. 2. We observe that when timestep is large, cross-attention map of  $S^*$  is inaccurate, spreading its focus across the entire image. This indicates that the word embedding might be capturing background details, which is detrimental to personalized face generation. As the timestep decreases, cross-attention map of  $S^*$  progressively narrows down to a more precise facial region, suggesting that  $S^*$  can more effectively acquire accurate identity information at these timesteps.

We believe that the reason for these phenomena is that in the early stage of the denoising process, noised image  $X_t$  contains more noise, preventing  $S^*$  from accurately focusing on the facial region, which results in undesired cross-attention maps. Consequently, the learned embeddings struggle to capture useful concept information during this phase. As the noise in  $X_t$  diminishes, the interference from redundant information decreases as well. However, prior research (Balaji et al. 2022) indicates that the control ability of text prompts over  $X_t$  becomes weaker in the late stage of the denoising process. This suggests that learning word embeddings in the middle of the time schedule is a better choice than learning over the entire time schedule.

Building on our analysis, we propose partitioning the time schedule in the reverse process of diffusion models into three intervals based on cross-attention maps. However, variations in facial regions across images affect the identity information captured by the learned embeddings, resulting in differences in the selected intervals. To reasonably define these stages, we initially train across the entire time schedule. After several iterations, the learned embedding, while not yet capturing accurate identity information, begins to show significant differences in cross-attention maps at different



Figure 4: The overlooked semantic yield suboptimal attention map. Attention weights are annotated in the lower left corner of cross-attention maps.

timesteps, without exhibiting overfitting. We utilize an existing semantic segmentation model, CLIPSeg (Lüddecke and Ecker 2022), to extract facial mask and calculated IoU with cross-attention maps. As shown in Fig. 2, we observe that IoU gradually increases as timestep progresses. Based on this, we set two thresholds,  $\lambda_1$  and  $\lambda_2$  ( $\lambda_1 < \lambda_2$ ), to update the training intervals. We designate the phase where the IoU falls below  $\lambda_1$  as the first static stage. During this stage, we cease training and instead use the supercategory embedding (e.g., face) directly for inference. Once the IoU surpasses the threshold  $\lambda_2$ , the learned embedding accurately identifies the facial region, but this also indicates that the noised image contains minimal noise, reducing controllability over the diffusion model. Consequently, we define this as the second static stage, where training ceases, and a fixed word embedding is used during testing.

We designate the remaining intermediate interval as the dynamic stage. In this phase, we utilize a lightweight network that takes as input a time embedding and a supercategory word embedding to generate dynamic embeddings, as illustrated in Fig. 3. Similar to NeTI (Alaluf et al. 2023), we incorporate a residual embedding into the sentence embedding output from the text encoder to serve as  $V$  for the cross-attention layer, providing additional information that the text encoder alone cannot capture. Notably, we introduce the CLIP image embedding of the training image into the network to generate the residual embedding. This is because features extracted from images are beneficial for learning conceptual information, a finding supported by existing research (Ye et al. 2023; Li et al. 2024).

**Face Customization With Tandem Equilibrium.** The embeddings learned during the dynamic stage can more accurately focus on the facial region. This not only helps avoid overfitting but also improves training efficiency by narrowing the range of timesteps. However, we observe that even when  $S^*$  focus on the desired region, the diffusion model may still produce results inconsistent with the given text description under certain textual conditions. As shown in the first row of Fig. 4, although the result accurately preserves the identity of the individual, it deviates significantly from the semantics of “garland”. We use L2 norm as a global metric to measure the weight of each text token in the cross-attention map. The weight of  $S^*$  (0.9596) is notably higher than that of “garland” (0.5407), indicating that the diffusion model has neglected the semantic information of “garland”.

In U-Net, the cross-attention map is computed from  $Q$

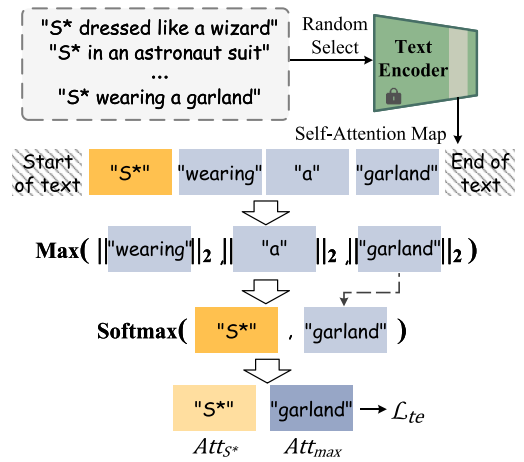


Figure 5: Illustration of the proposed Tandem Equilibrium.

and  $K$ , where  $Q$  is derived from the latent image features and  $K$  from the sentence embedding. During training, we freeze diffusion model parameters to prevent language drift, making sentence embedding the sole variable affecting the cross-attention map. We further analyze the self-attention maps in the text encoder to understand this phenomenon. We find that in the shallow layers of text encoder, the attention weights for each text token are relatively uniform. As the layers deepen, nouns begin to dominate the attention, while prepositions have lower weights, due to the concrete semantic content typically associated with nouns. We believe that during training, in an effort to reduce denoising loss, text encoder overly emphasizes the semantics contained in the learned embedding while neglecting those in the original embedding. This is evident in the excessively high attention weights of  $S^*$ , which may lead diffusion model to generate results that do not align with the given text description.

One straightforward solution is to minimize the L2 norm of the self-attention map corresponding to  $S^*$  within the text encoder. However, directly constraining the self-attention map can easily cause the diffusion model to overlook the semantics of  $S^*$ , leading to results with inaccurate identities. Another solution is to manually set a threshold to constrain the attention weight of  $S^*$  within a specific range. However, since it varies across different text prompts, using a fixed threshold is not ideal. Therefore, the constraint should be adjusted dynamically based on text prompt.

As shown in Fig. 5, we randomly input text prompt into the text encoder and extract self-attention maps in the final layer. Ignoring the “start of text” and “end of text” tokens, we identify the original embedding with the highest attention weight. We then applying a softmax function to its self-attention map with that of  $S^*$ , obtaining  $Att_{max}$  and  $Att_{S^*}$ . We calculate tandem equilibrium loss  $\mathcal{L}_{te}$  as follows:

$$\mathcal{L}_{te} = -\phi(Att_{S^*}) \times \phi(Att_{max}), \quad (1)$$

where  $\phi(\cdot)$  denotes the summation function. Since the total sum remains constant, maximizing the product requires  $\phi(Att_{S^*})$  and  $\phi(Att_{max})$  to be as close as possible. This ensures that the diffusion model balances its attention between



Figure 6: Qualitative comparison with state-of-the-art methods on celebrities.

$S^*$  and other text tokens effectively.

**Loss Functions.** We introduce a mask  $M$  indicating the face region to calculate  $\mathcal{L}_{mse}$ , enforcing diffusion model to focus on denoising the masked region. It is formulated as:

$$\mathcal{L}_{mse} = \|(\epsilon_{\theta}(X_t, t, y_t) - \epsilon) \cdot M\|^2. \quad (2)$$

To ensure the preservation of identity information from the given image  $X_0$ , we assess the similarity between the identity features of the noised image  $X_t$  estimated to the clean image  $X_{0|t}$  at time  $t$  and  $X_0$ . Identity features are extracted using Arcface (Deng et al. 2019), and the loss  $\mathcal{L}_{id}$  is defined as follows:

$$\mathcal{L}_{id} = 1 - \cos(\text{Arcface}(X_{0|t}), \text{Arcface}(X_0)). \quad (3)$$

As the diffusion model faces challenges in recovering an accurate clean image when  $t$  is large, affecting the effectiveness of  $\mathcal{L}_{id}$ , we introduce a hyperparameter schedule,  $\lambda_{id}(t)$ :

$$\lambda_{id}(t) = \cos\left(\frac{t}{2T}\pi\right). \quad (4)$$

Finally, our total objective  $\mathcal{L}(t)$  is formulated as follows:

$$\mathcal{L}(t) = \mathcal{L}_{te} + \mathcal{L}_{mse} + \lambda_{id}(t)\mathcal{L}_{id}. \quad (5)$$

## Experiments

**Competitors.** We compare our method with several state-of-the-art personalization methods, including Textual Inversion (Gal et al. 2023a), DreamBooth (Ruiz et al. 2023), Custom Diffusion (Kumari et al. 2023), NeTI (Alaluf et al. 2023), and Perfusion (Tewel et al. 2023). Since official implementations were unavailable for Perfusion and DreamBooth, we employed unofficial versions (von Platen et al.

2022; ChenDarYen 2023) for our comparisons. All comparative methods were conducted using their default settings.

**Metrics.** We evaluate performance from two perspective to assess the effectiveness of proposed method. For text similarity, we leverage the pre-trained CLIP (Radford et al. 2021) model and calculate CLIPScore (Hessel et al. 2021) between prompts and customized results, where the placeholder “ $S^*$ ” in prompts is substituted with “face”. For identity preservation, we use MTCNN (Zhang et al. 2016) to detect faces in unaligned generated images and then apply CosFace (Wang et al. 2018) to evaluate identity similarity to the given faces. It is noteworthy that if no face is generated, the score is set to the minimum value of -1.

**Qualitative Evaluation.** In order to visually demonstrate our generated effects, we collected some public images from the Internet, comprising some celebrity close-ups and portraits of non-celebrities, all of which are unaligned and encompass approximately 30 individuals. We present generated outcomes under varies text prompts in Figs. 6 and 7. For each individual, we collected 3-5 images and randomly selected one as the input for one-shot setting. To ensure fairness, we also present the results of default few-shot setting for competitors. Considering that all inputs in Fig. 6 are celebrities, we also showcase the results of directly inputting prompts into pretrained diffusion model by replacing the placeholders “ $S^*$ ” with the names of celebrities. While generating directly from prompts leverages the extensive prior knowledge of Stable Diffusion to create celebrities with distinctive features, it may introduce biases from the given images, such as inconsistencies in Steve Jobs’ hair or Leonardo DiCaprio’s age. It is evident that DreamBooth shows strong identity preservation, but its results often misalign with text prompts. This issue arises from training across the entire

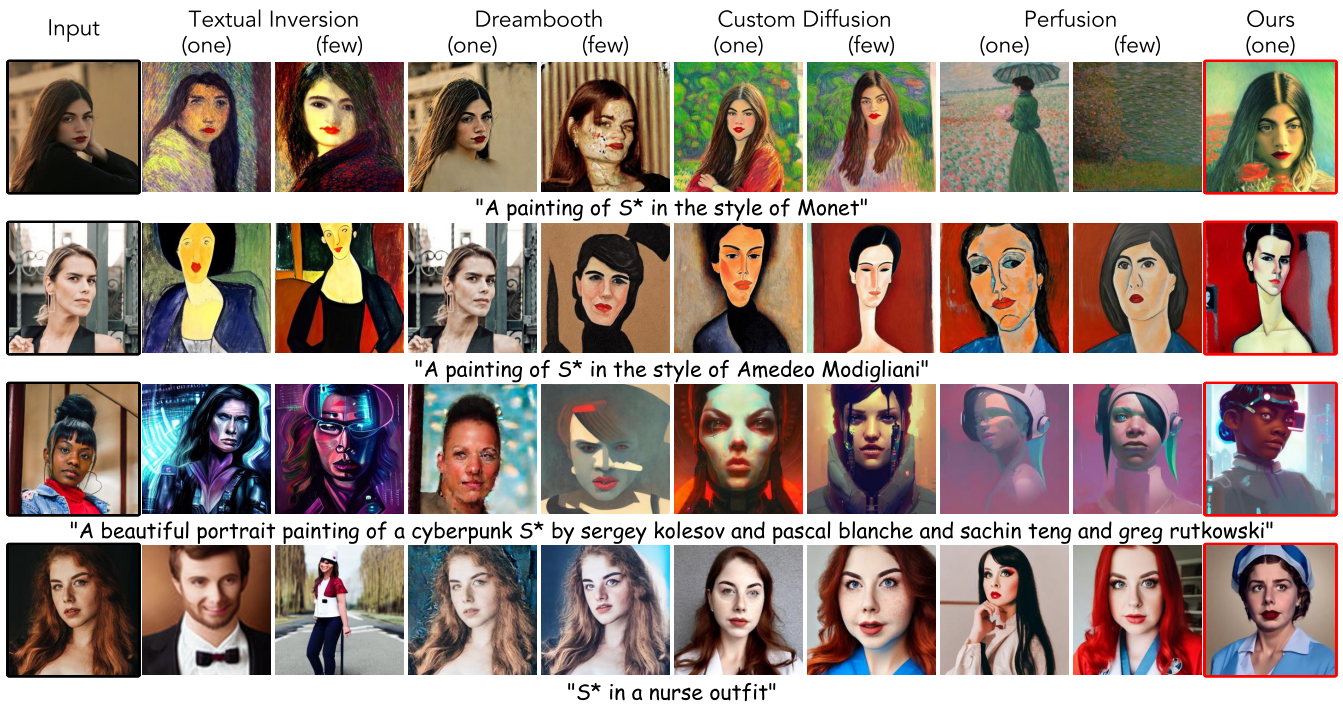


Figure 7: Qualitative comparison with state-of-the-art methods on non-celebrities.

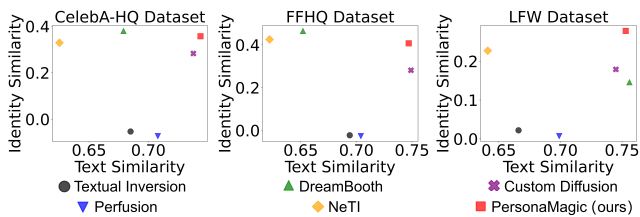


Figure 8: Quantitative evaluation on CelebA-HQ, FFHQ, and LFW datasets shows that our method sits on the Pareto front, highlighting its superiority over competitors.

time schedule, leading to overfitting where cross-attention maps focus on the entire image. Contrastingly, our method confines training to the dynamic stage, directing embeddings to focus exclusively on facial regions. Textual Inversion, Custom Diffusion, and Perfusion generate results that align with text prompts but lack accurate identity. In contrast, our approach not only utilizes dynamic embeddings across multiple timesteps to convey intricate character information but also employs a lightweight network for learning, which better captures the correspondence between the embedding space and facial attributes.

**Quantitative Evaluation.** For quantitative evaluation of one-shot setting, we randomly selected 100 distinct images from CelebA-HQ (Karras et al. 2018) and FFHQ (Karras, Laine, and Aila 2019) datasets respectively and repeated five times for fair comparisons, gotten 500 images of each dataset. For few-shot setting, we use the LFW (Huang et al. 2008) dataset, focusing on individuals that contain multiple

Method	CelebA-HQ		LFW	
	T-Sim.	I-Pre.	T-Sim.	I-Pre.
Custom Diffusion (CD)	0.737	0.283	0.744	0.179
Vanilla	0.610	0.334	0.631	0.226
+Stage Regulation	0.672	<b>0.363</b>	0.665	<b>0.289</b>
+Tandem Equilibrium	0.691	0.284	0.668	0.206
PersonaMagic	<b>0.744</b>	<u>0.358</u>	<b>0.753</b>	<u>0.278</u>

Table 1: Ablation study on CelebA-HQ and LFW datasets.

images. We collected 17 prompts involving diverse personalized modifications, with each individual and prompt inferred four times, resulting in 68 outcomes per person. We then calculated text similarity and identity preservation for each set, with the average score presented in Fig. 8. Our method outperforms competitors, sitting on the Pareto front. Textual Inversion and Perfusion, struggling to capture intricate identity details, demonstrate low identity preservation but show lower text similarity. In contrast, we introduce the Tandem Equilibrium strategy to balance the semantic representation of text tokens, leading to better text similarity. Custom Diffusion balances concept fidelity and text alignment, but its performance remains below ours.

**Ablation Study.** We validated different components of PersonaMagic through ablation studies (Tab. 1). First, we examined stage regulation’s impact. In one-shot experiments on CelebA-HQ, our method improved identity preservation

Setting	$\lambda_1=0.5$ $\lambda_2=1.0$	$\lambda_1=0.6$ $\lambda_2=1.0$	$\lambda_1=0.7$ $\lambda_2=1.0$	$\lambda_1=0.8$ $\lambda_2=1.0$	$\lambda_1=0.7$ $\lambda_2=0.9$	$\lambda_1=0.7$ $\lambda_2=0.8$
T-Sim.	0.691	0.727	0.744	0.749	0.741	0.744
I-Pre.	0.284	0.307	0.328	0.302	0.328	0.358

Table 2: Evaluation of stage partition variants on CelebA-HQ. We set  $\lambda_1 = 0.7$  and  $\lambda_2 = 0.8$ , as this configuration achieves the optimal balance between fidelity and editability.

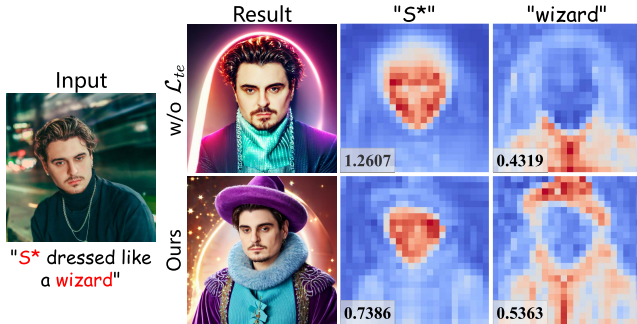


Figure 9: Customized results with and without  $\mathcal{L}_{te}$  during training. Attention weights are annotated in the lower left corner of cross-attention maps.

by 0.051 over the second best competitor CD due to learning a series of embeddings for new concepts, which outperforms optimizing a single embedding. However, text similarity decreased slightly due to the early use of learned embeddings, leading to misaligned spatial layouts with the prompt.

Stage regulation improved both text similarity and identity preservation, increasing text similarity from 0.610 to 0.672 by using supercategory embeddings in the initial static stage, reducing overfitting. The dynamic stage further enhanced ID preservation by focusing more on facial features.

To further explore stage regulation, we conducted a quantitative analysis (Tab. 2). Initially, with  $\lambda_1 = 0.5$  and  $\lambda_2 = 1.0$ , the denoising process was unpartitioned, using dynamic embeddings throughout, resulting in suboptimal performance. Setting  $\lambda_1$  to 0.6 or 0.7 introduced a two-stage process: a static stage below  $\lambda_1$  using supercategory embeddings and a dynamic stage above it with dynamic embeddings. Increasing  $\lambda_1$  improved text similarity but reduced identity preservation at  $\lambda_1 \geq 0.8$  due to limited dynamic stage timesteps. Based on the results, we chose  $\lambda_1 = 0.7$ . Further refinement of  $\lambda_2$  divided the process into three stages, with a third static stage emerging as  $\lambda_2$  decreased, using fixed embeddings. We observed that when  $\lambda_2 < 0.8$ , the dynamic stage became too constrained, impacting identity detail capture. Therefore, we set  $\lambda_1 = 0.7$  and  $\lambda_2 = 0.8$ . This strategy, based on IoU between cross-attention maps and real masks, generalizes well across datasets.

The TE strategy balanced text similarity and identity preservation by adjusting self-attention in the text encoder, aligning semantic strength between learned and original embeddings. Visualizations (Fig. 9) showed TE reduced mis-

Method	PhotoMaker	PhotoMaker w/ ours	IP-Adapter	IP-Adapter w/ ours
T-Sim.	0.775	<b>0.790</b>	0.762	<b>0.778</b>
I-Pre.	0.335	<b>0.352</b>	0.348	<b>0.359</b>

Table 3: Universality of our method as a plug-in for pre-trained personalization models.

alignment in cross-attention, improving personalized results.

Integrating both strategies, we surpassed CD in text similarity. Despite a slight reduction in identity preservation with TE, dynamic embeddings maintained accurate details, performing better CD. Few-shot experiments on LFW confirmed these strategies enhance performance, balancing text similarity and identity preservation.

**The flexibility of PersonaMagic.** Pretrained personalization models (Li et al. 2024; Ye et al. 2023) tend to perform suboptimally when applied to individuals outside the training dataset. This issue is not well-addressed by fine-tune based methods (Ruiz et al. 2023; Kumari et al. 2023), as it risks disrupting the pre-learned semantic knowledge in diffusion model. In contrast, our approach avoids this limitation by freezing model parameters during training, making it a flexible plug-in that can be integrated into pretrained face personalization models to enhance their performance.

To validate this, We integrated PersonaMagic into Photomaker (Li et al. 2024) and IP-Adapter (Ye et al. 2023) and conducted quantitative experiments on our collected images of non-celebrities, as shown in Tab. 3. Introducing our method led to improvements in both text similarity and identity preservation for Photomaker and IP-Adapter. The increase in text similarity is largely attributed to our TE strategy, which guides the diffusion model to focus on overlooked semantics in the text prompt. Additionally, the Stage Regulation strategy enhances text similarity by helping Photomaker and IP-Adapter generate spatial layouts during early denoising that better align with the text description. This strategy also captures facial features that were previously missed, further improving identity preservation in personalized models. This conclusion is supported by visual comparisons, available in the supplementary materials.

## Conclusion

In this paper, we present PersonaMagic, a high-fidelity face customization technique that utilizes a stage-regulated textual conditioning strategy based on a comprehensive analysis. We introduce a lightweight network to implement this conditioning mechanism through dynamic word embeddings, effectively capturing identity information while avoiding overfitting. Furthermore, we propose a tandem equilibrium loss to address the trade-off between text alignment and identity preservation. Extensive experiments demonstrate the superior performance of our method compared to state-of-the-art approaches, excelling in both fidelity and editability, and showcasing its effectiveness across various downstream customization tasks.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62102381, 41927805); Shandong Natural Science Foundation (No. ZR2021QF035); the National Key R&D Program of China (No. 2022ZD0117201); the Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097); the AI Singapore Programme under the National Research Foundation Singapore (Grant AISG3-GV-2023-011); and the Lee Kong Chian Fellowships.

## References

- Alaluf, Y.; Richardson, E.; Metzger, G.; and Cohen-Or, D. 2023. A neural space-time representation for text-to-image personalization. *ACM TOG*, 42(6): 1–10.
- Alaluf, Y.; Tov, O.; Mokady, R.; Gal, R.; and Bermano, A. 2022. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *CVPR*, 18511–18521.
- Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.
- ChenDarYen. 2023. Key-Locked-Rank-One-Editing-for-Text-to-Image-Personalization. <https://github.com/ChenDarYen/Key-Locked-Rank-One-Editing-for-Text-to-Image-Personalization>.
- Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 4690–4699.
- Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *CVPR*, 12873–12883.
- Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023a. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*.
- Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023b. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG*, 42(4): 1–13.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.
- Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 10696–10706.
- Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.
- Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*.
- Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*.
- Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*, 4401–4410.
- Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*, 8110–8119.
- Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *CVPR*, 1931–1941.
- Lee, J.; Cho, K.; and Kiela, D. 2019. Countering language drift via visual grounding. In *EMNLP*.
- Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 8640–8650.
- Lu, Y.; Singhal, S.; Strub, F.; Courville, A.; and Pietquin, O. 2020. Countering language drift with seeded iterated learning. In *ICML*, 6437–6447. PMLR.
- Lüddecke, T.; and Ecker, A. 2022. Image Segmentation Using Text and Image Prompts. In *CVPR*, 7086–7096.
- Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 16784–16804. PMLR.
- Nitzan, Y.; Aberman, K.; He, Q.; Liba, O.; Yarom, M.; Gandselman, Y.; Mosseri, I.; Pritch, Y.; and Cohen-Or, D. 2022. Mystyle: A personalized generative prior. *ACM TOG*, 41(6): 1–10.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1): 5485–5551.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
- Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*, 8821–8831. PMLR.
- Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2022. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 42(1): 1–13.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 36479–36494.

Song, H.; Du, Y.; Xiang, T.; Dong, J.; Qin, J.; and He, S. 2022. Editing out-of-domain gan inversion via differential activations. In *ECCV*, 1–17. Springer.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 11895–11907.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-based generative modeling through stochastic differential equations. In *ICLR*.

Tang, Z.; Gu, S.; Bao, J.; Chen, D.; and Wen, F. 2022. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH*, 1–11.

von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 5265–5274.

Xu, Y.; Du, Y.; Xiao, W.; Xu, X.; and He, S. 2021. From continuity to editability: Inverting gans with consecutive images. In *ICCV*, 13910–13918.

Yang, H.; Chai, L.; Wen, Q.; Zhao, S.; Sun, Z.; and He, S. 2021. Discovering interpretable latent space directions of gans beyond binary attributes. In *CVPR*, 12177–12185.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *TMLR*.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.