PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding
– *Supplementary Materials* –

Zhen Li[*1,2], Mingdeng Cao[*2,3], Xintao Wang[†2], Zhongang Qi[2], Ming-Ming Cheng[†4,1], Ying Shan[2]
[1]VCIP, CS, Nankai University [2]ARC Lab, Tencent PCG [3]The University of Tokyo
[4]NKIARl, Shenzhen Futian
https://photo-maker.github.io/

Contents

1. Dataset Details

Training dataset. Based on Sec. 3.3 in the main paper, following a sequence of filtering steps, the number of images in our constructed dataset is about 112K. They are classified by about 13,000 ID names. Each image is accompanied by a mask for the corresponding ID and an annotated caption.

Evaluation dataset. The *image dataset* used for evaluation comprises manually selected additional IDs and a portion of MyStyle [13] data. For each ID name, we have four images that serve as input data for comparative methods and for the final metric evaluation (*i.e.*, DINO [3], CLIP-I [5], and Face Sim. [4]). For single-embedding methods (*i.e.*, FastComposer [17] and IPAdapter [18]), we randomly select one image from each ID group as input. Note that the ID names exist in the training image set, utilized for the training of our method, and the test image set, do not exhibit any overlap. We list ID names for evaluation in Tab. 1. For *text prompts* used for evaluation, we consider six factors: clothing, accessories, actions, expressions, views, and background, which make up 40 prompts that are listed in the Tab. 2.

Evaluation IDs	
① Alan Turing	⑭ Kamala Harris
② Albert Einstein	⑮ Marilyn Monroe
③ Anne Hathaway	⑯ Mark Zuckerberg
④ Audrey Hepburn	⑰ Michelle Obama
⑤ Barack Obama	⑱ Oprah Winfrey
⑥ Bill Gates	⑲ Renée Zellweger
⑦ Donald Trump	⑳ Scarlett Johansson
⑧ Dwayne Johnson	㉑ Taylor Swift
⑨ Elon Musk	㉒ Thomas Edison
⑩ Fei-Fei Li	㉓ Vladimir Putin
⑪ Geoffrey Hinton	㉔ Woody Allen
⑫ Jeff Bezos	㉕ Yann LeCun
⑬ Joe Biden	

Table 1. **ID names used for evaluation.** For each name, we collect four images totally.

Figure 1. **Visual comparisons on non-celebrities.** We used the face image [2] generated by GAN as the reference image.

2. Non-Celebrities Results

Our method also performs well in non-celebrities inputs. We gathered 12 sets of IDs, including images of our colleagues and faces generated by generative models (*i.e.*, GAN and diffusion model), with approximately 1-3 images per ID set. We selected 10 prompts for evaluation. As demonstrated in Fig. 1 and Tab. 3, our method has the second best text consistency (CLIP-T) and ID similarity (DINO and Face Sim.), while generating the most diverse face regions (Face Div.). This demonstrates the comprehensiveness of our method.

Category	Prompt
General	a photo of a <class word>
Clothing	a <class word> wearing a Superman outfit
	a <class word> wearing a spacesuit
	a <class word> wearing a red sweater
	a <class word> wearing a purple wizard outfit
	a <class word> wearing a blue hoodie
Accessory	a <class word> wearing headphones
	a <class word> with red hair
	a <class word> wearing headphones with red hair
	a <class word> wearing a Christmas hat
	a <class word> wearing sunglasses
	a <class word> wearing sunglasses and necklace
	a <class word> wearing a blue cap
	a <class word> wearing a doctoral cap
	a <class word> with white hair, wearing glasses
Action	a <class word> in a helmet and vest riding a motorcycle
	a <class word> holding a bottle of red wine
	a <class word> driving a bus in the desert
	a <class word> playing basketball
	a <class word> playing the violin
	a <class word> piloting a spaceship
	a <class word> riding a horse
	a <class word> coding in front of a computer
	a <class word> playing the guitar

(a)

Category	Prompt
Expression	a <class word> laughing on the lawn
	a <class word> frowning at the camera
	a <class word> happily smiling, looking at the camera
	a <class word> crying disappointedly, with tears flowing
	a <class word> wearing sunglasses
View	a <class word> playing the guitar in the view of left side
	a <class word> holding a bottle of red wine, upper body
	a <class word> wearing sunglasses and necklace, close-up, in the view of right side
	a <class word> riding a horse, in the view of the top
	a <class word> wearing a doctoral cap, upper body, with the left side of the face facing the camera
	a <class word> crying disappointedly, with tears flowing, with left side of the face facing the camera
Background	a <class word> sitting in front of the camera, with a beautiful purple sunset at the beach in the background
	a <class word> swimming in the pool
	a <class word> climbing a mountain
	a <class word> skiing on the snowy mountain
	a <class word> in the snow
	a <class word> in space wearing a spacesuit

(b)

Table 2. **Evaluation text prompts categorized by (a) general setting, clothing, accessory, action, (b) expression, view, and background.** The `class word` will be replaced with `man`, `woman`, `boy`, etc. For each ID and each prompt, we generated four images for evaluation.

Methods	CLIP-T↑	DINO↑	Face Sim.↑	Face Div.↑
DreamBooth	**30.1**	44.4	37.7	47.5
FastComposer	25.9	**54.2**	**69.2**	39.1
IP-Adapter	23.3	47.5	61.4	37.7
PhotoMaker (Ours)	29.5	50.5	66.7	**52.5**

Table 3. Non-celebrities comparisons.

3. User Study

In this section, we conduct a user study to make a more comprehensive comparison. The comparative methods we have selected include DreamBooth [15], FastComposer [17], and IPAdapter [18]. We use SDXL [14] as the base model for both DreamBooth and IPAdapter because of their open-sourced implementations. For each pair of inputs, we have four randomly generated images of each method. Each user is requested to answer four questions for these 20 sets of results: 1) Which method is *most similar* to the input person's *identity*? 2) Which method produces the *highest quality* generated images? 3) Which method generates *the most diverse* facial area in the images? 4) Which method generates images that *best match* the input *text prompt*? We have anonymized the names of all methods and randomized the order of methods in each set of responses. We had a total of 40 candidates participating in our user study, and we received 3,200 valid votes. The results are shown in Fig. 2.

We find that our PhotoMaker has advantages in terms of ID fidelity, generation quality, diversity, and text fidelity, especially the latter three. In addition, we found that DreamBooth is the second-best algorithm in balancing these four evaluation dimensions, which may explain why it was more prevalent than the embedding-based methods in the past. At the same time, IPAdapter shows a significant disadvantage in terms of generated image quality and text consistency, as it focuses more on image embedding during the training phase. FastComposer has a clear shortcoming in the diversity of the facial region for their single-embedding training pipeline. The above results are generally consistent with Tab. 1 in the main paper, except for the discrepancy in the

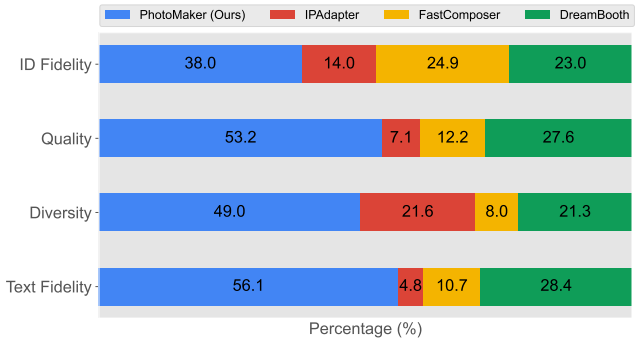

Figure 2. **User preferences on ID fidelity, generation quality, face diversity, and text fidelity for different methods.** For ease of illustration, we visualize the proportion of total votes that each method has received. Our PhotoMaker occupies the most significant proportion in these four dimensions.

Figure 3. **More comparisons on identity mixing.**

CLIP-T metric. This could be due to a preference for selecting images that harmonize with the objects appearing in the text when manually choosing the most text-compatible images. In contrast, the CLIP-T tends to focus on whether the object appears. This may demonstrate the limitations of the CLIP-T. We also provide more visual samples in Fig. 10-13 for reference.

4. More Ablations

The influence about the number of input ID images. We explore the impact that forming the proposed stacked ID embedding through feeding different numbers of ID images. In Tab. 4, we visualize this impact across different metrics. We conclude that using more images to form a stacked ID embedding can improve the metrics related to ID fidelity. This improvement is particularly noticeable when the number of input images is increased from one to two. Upon the input of an increasing number of ID images, the growth rate of the values in the ID-related metrics significantly decelerates. Additionally, we observe a linear decline on the CLIP-T metric. This indicates there may exist a trade-off between text controllability and ID fidelity. From Fig. 5, we see that increasing the number of input images enhances the similarity of the ID. Therefore, the more ID images to form the stacked ID embedding can help the model perceive more comprehensive ID information, and then more accurately represent the ID to generate images.

Variants	CLIP-T↑	DINO↑	Face Sim.↑	Face Div.↑
50% IDs	28.4	45.1	63.4	54.2
50% size	28.5	43.9	61.6	55.1
Full data & IDs	**28.6**	**45.3**	**63.9**	**55.6**

Table 4. **Comparisons on different variants**.

Besides, as shown by the Dwayne Johnson example, the gender editing capability decreases, and the model is more prone to generate images of the original ID's gender.

Adjusting the ratio during identity mixing. For identity mixing, our method can adjust the merge ratio by either controlling the percentage of identity images within the input image pool or through the method of prompt weighting [7, 9]. In this way, we can control that the person generated with a new ID is either more closely with or far away from a specific input ID. Fig. 6 shows how our method customizes a new ID by controlling the proportion of different IDs in the input image pool. For a better description, we use a total of 10 images as input in this experiment. We can observe a smooth transition of images with the two IDs. This smooth transition encompasses changes in skin color and age. Next, we use four images per generated ID to conduct prompt weighting. The results are shown in Fig. 7. We multiply the embedding corresponding to the images related to a specific ID by a coefficient to control its proportion of integration into the new ID. Compared to the way to control the number of input images, prompt weighting requires fewer photos to adjust the merge ratio of different IDs, demonstrating its superior usability. Besides, the two ways of adjusting the mixing ratio of different IDs both demonstrate the flexibility of our method.

More identity mixing comparisons. In addition to the results of DreamBooth (using a single model trained on the two identities) and SDXL itself shown in the main paper, we have displayed the results of more methods on identity mixing. These include mixing two identity LoRAs [8, 16], linear interpolation of two Textual Inversion [5] embeddings, and Custom Diffusion [10] results. Since these methods are *less stable* in identity mixing compared to ours, we have chosen the best results generated by these methods to compare with the *non-cherry-picked* results of our method. Fig. 3 shows that our method can generate a new ID that better retains the features of the two IDs. Most multi-concept methods [6, 11, 12] only facilitate the multi-object existence or style mixing and struggle to merge fine-grained IDs. Besides, by simply controlling the input image pool, our method can manage the proportion of different IDs involved, demonstrating superior *flexibility*.

Study of the dataset. In Tab. 4, we conducted two sets of experiments with the same amount of training data, one with the ID halved and the other with the dataset size halved

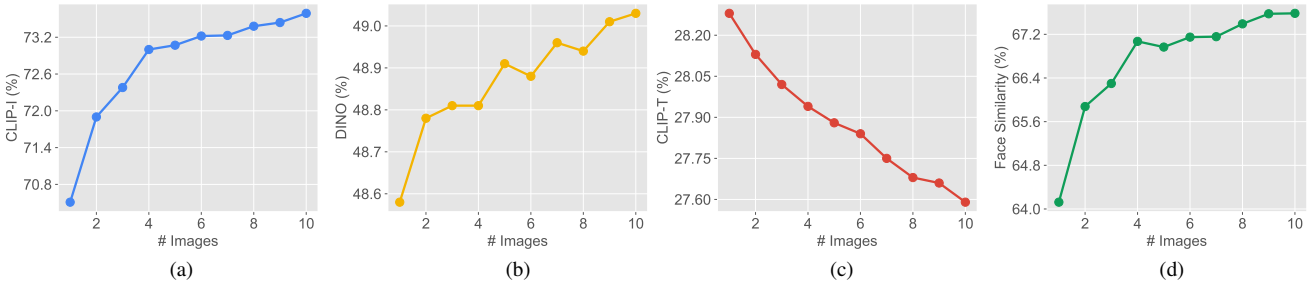

Figure 4. **The impact of the number of input ID images on (a) CLIP-I, (b) DINO, (c) CLIP-T, and (d) Face Similarity, respectively.**

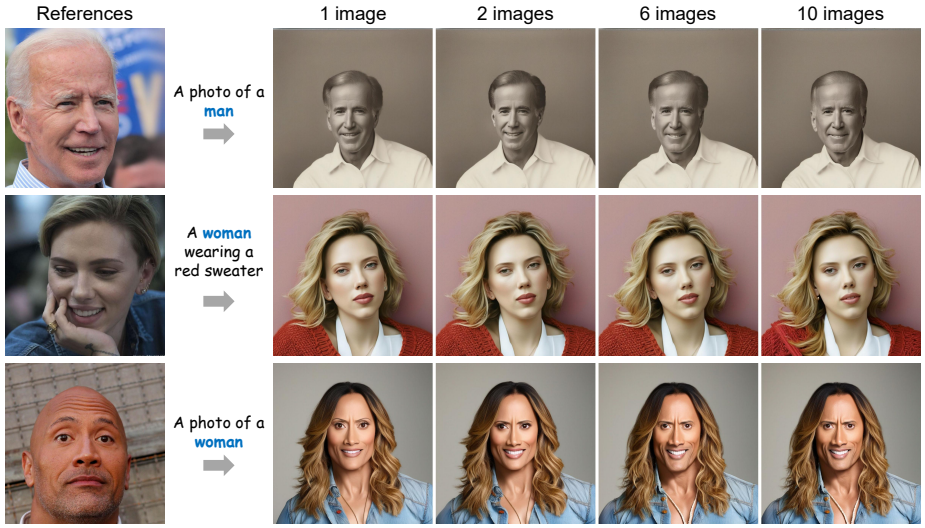

Figure 5. **The impact of varying the quantity of input images on the generation results.** It can be observed that the fidelity of the ID increases with the quantity of input images.

while keeping the ID quantity unchanged. We found that the number of IDs mainly affects diversity, while the number of images per ID influences similarity more.

5. Stylization Results

Our method not only possesses the capability to generate realistic human photos, but it also allows for stylization while preserving ID attributes. This demonstrates the robust generalizability of the proposed method. We provide the stylization results in Fig. 8.

6. More Visual Results

Recontextualization. We first provide a more intuitive comparison in Fig. 10. We compare our PhotoMaker with DreamBooth [15], FastComposer [17], and IPAdapater [18], for universal recontextualization cases. Compared to other methods, the results generated by our method can simultaneously satisfy high-quality, strong text controllability, and high ID fidelity. We then focus on the IDs that SDXL can not generate itself. We refer to this scenario as the "non-celebrity" case. Compared Fig. 11 with Fig. 9, our method can successfully generate the corresponding input IDs for this setting.

Bringing person in artwork/old photo into reality. Fig. 12-13 demonstrate the ability of our method to bring past celebrities back to reality. It is worth noticing that our method can generate photo-realistic images from IDs in statues and oil paintings. Achieving this is quite challenging for the other methods we have compared.

Changing age or gender. We provide more visual results for changing age or gender in Fig. 14. As mentioned in the main paper, we only need to change the class word when we conduct such an application. In the generated ID images changed in terms of age or gender, our method can well preserve the characteristics in the original ID.

Identity mixing. We provide more visual results for identity mixing application in Fig. 15. Benefiting from our stacked ID embedding, our method can effectively blend the characteristics of different IDs to form a new ID. Subsequently, we can generate text controlled based on this new

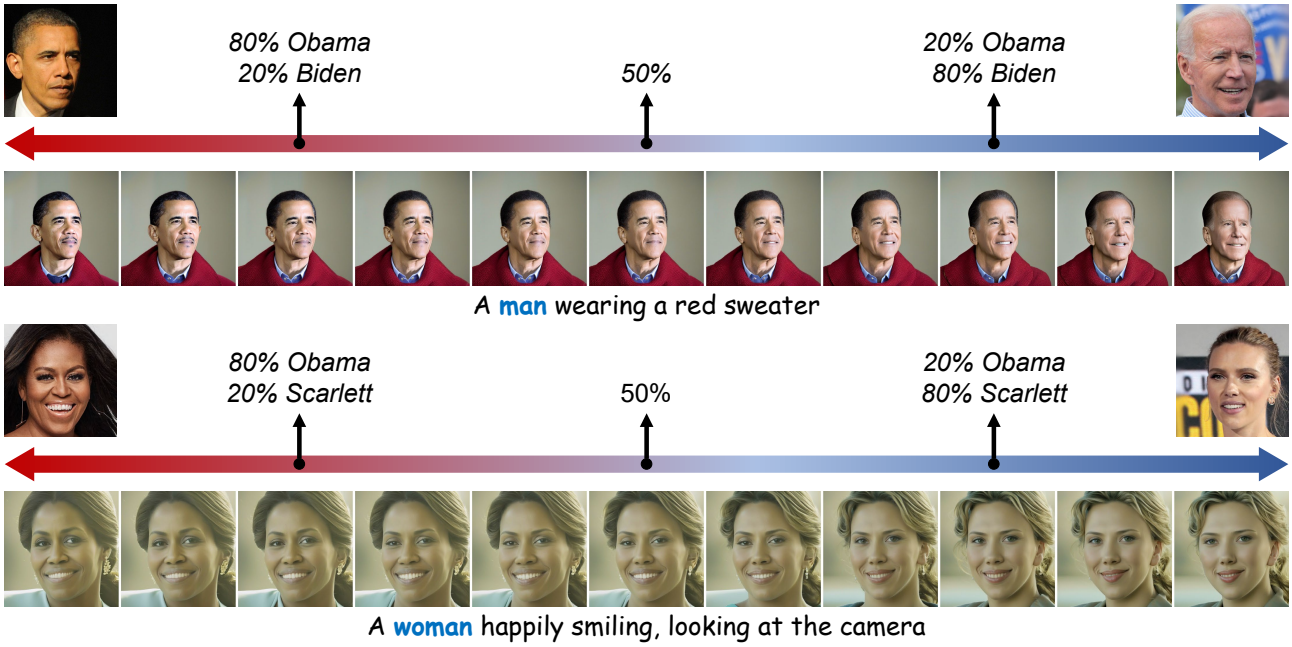

Figure 6. **The impact of the proportion of images with different IDs in the input sample pool on the generation of new IDs.** The first row illustrates the transition from Barack Obama to Joe Biden. The second row depicts the shift from Michelle Obama to Scarlett Johansson. To provide a clearer illustration, percentages are used in the figure to denote the proportion of each ID in the input image pool. The total number of images contained in the input pool is 10. *(Zoom-in for the best view).*

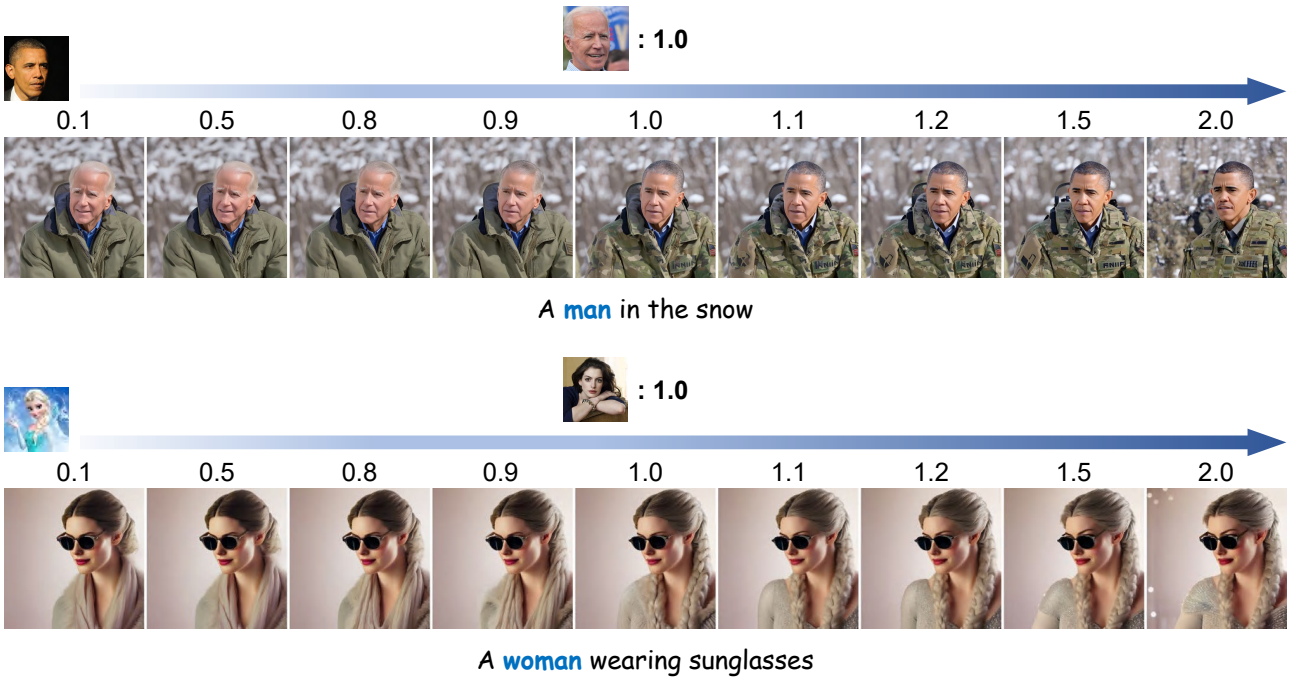

Figure 7. **The impact of prompt weighting on the generation of new IDs.** The first row illustrates a blend of Barack Obama and Joe Biden. The first row from left to right represents the progressive increase in the weight of the ID image embedding corresponding to Barack Obama in the image. The second row illustrates a blend of Elsa (Disney) and Anne Hathaway. The weight for Elsa is gradually increased. *(Zoom-in for the best view).*

Figure 8. **The stylization results of our PhotoMaker with different input IDs and different style prompts.** Our method can be seamlessly transferred to a variety of styles, concurrently preventing the generation of realistic results. The symbol **<class>** denotes it will be replaced by `man` or `woman` accordingly. *(Zoom-in for the best view).*

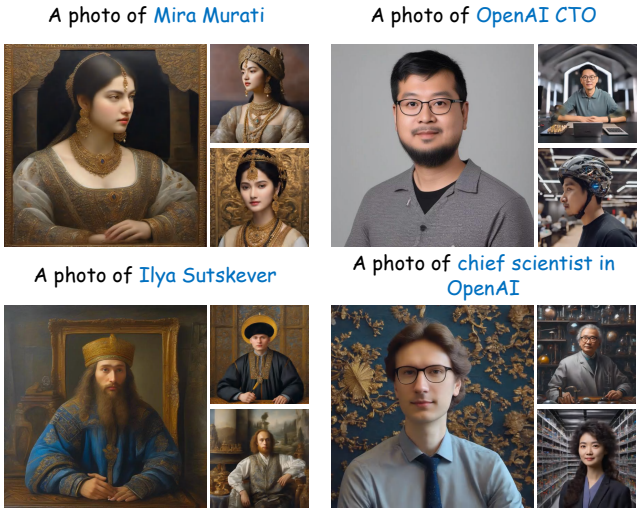

Figure 9. **Two examples that are unrecognizable by the SDXL.** We replaced two types of text prompts (*e.g.*, name and position) but were unable to prompt the SDXL to generate Mira Murati and Ilya Sutskever.

ID. Additionally, our method provides great flexibility during the identity mixing, as can be seen in Fig. 6-7. More importantly, we have explored in the main paper that existing methods struggle to achieve this application. Conversely, our PhotoMaker opens up a multitude of possibilities.

7. Limitations

First, our method only focuses on maintaining the ID information of a single generated person in the image, and cannot control multiple IDs of generated persons in one image simultaneously. Second, our method excels at generating half-length portraits, but is relatively not good at generating full-length portraits. Third, the age transformation ability of our method is not as precise as some GAN-based methods [1]. If the users need more precise control, modifications to the captions of the training dataset may be required. Finally, our method is based on the SDXL and the dataset we constructed, so it will also inherit their biases.

8. Broader Impact

In this paper, we introduce a novel method capable of generating high-quality human images while maintaining a high degree of similarity to the input identity. At the same time, our method can also satisfy high efficiency, decent facial generation diversity, and good controllability.

For the academic community, our method provides a strong baseline for personalized generation. Our data creation pipeline enables more diverse datasets with varied poses, actions, and backgrounds, which can be instrumental in developing more robust and generalizable computer vision models.

In the realm of practical applications, our technique has the potential to revolutionize industries such as entertainment, where it can be used to create realistic characters for movies or video games without the need for extensive CGI work. It can also be beneficial in virtual reality, providing more immersive and personalized experiences by allowing users to see themselves in different scenarios. It is worth noticing that everyone can rely on our PhotoMaker to quickly customize their own digital portraits.

However, we acknowledge the ethical considerations that arise with the ability to generate human images with high fidelity. The proliferation of such technology may lead to a surge in the inappropriate use of generated portraits, malicious image tampering, and the spreading of false information. Therefore, we stress the importance of developing and adhering to ethical guidelines and using this technology responsibly. We hope that our contribution will spur further discussion and research into the safe and ethical use of human generation in computer vision.

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *TOG*, 2021. 7

[2] David Beniaguev. Synthetic faces high quality (sfhq) dataset, 2022. 1

[3] 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. 1

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1

[5] 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. In *ICLR*, 2023. 1, 3

[6] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Chen Yunpeng, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Shan Ying, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2023. 3

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 3

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[9] Huggingface. Prompt weighting. https://huggingface.co/docs/diffusers/using-diffusers/weighted_prompts, 2023. 3

Figure 10. **More visual examples for recontextualization setting.** Our method not only provides high ID fidelity but also retains text editing capabilities. We randomly sample three images for each prompt. *(Zoom-in for the best view)*

[10] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[11] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 3

[12] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023. 3

[13] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *TOG*, 2022. 1

[14] Dustin Podell, Zion English, Kyle Lacey, Andreas

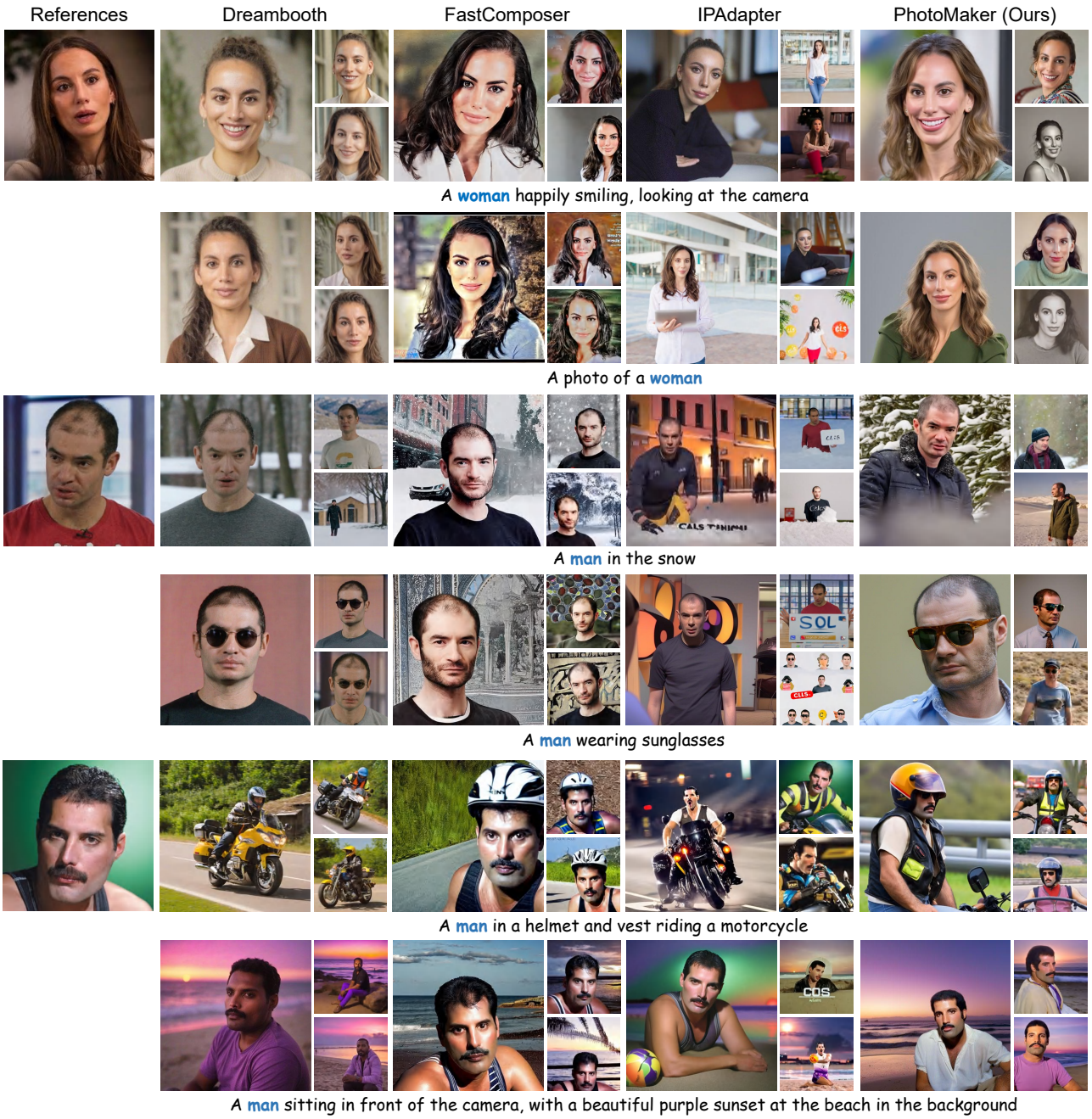

Figure 11. **More visual examples for recontextualization setting.** Our method not only provides high ID fidelity but also retains text editing capabilities. We randomly sample three images for each prompt. *(Zoom-in for the best view)*

Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[15] 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. 2, 4

[16] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2022. 3

[17] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv*

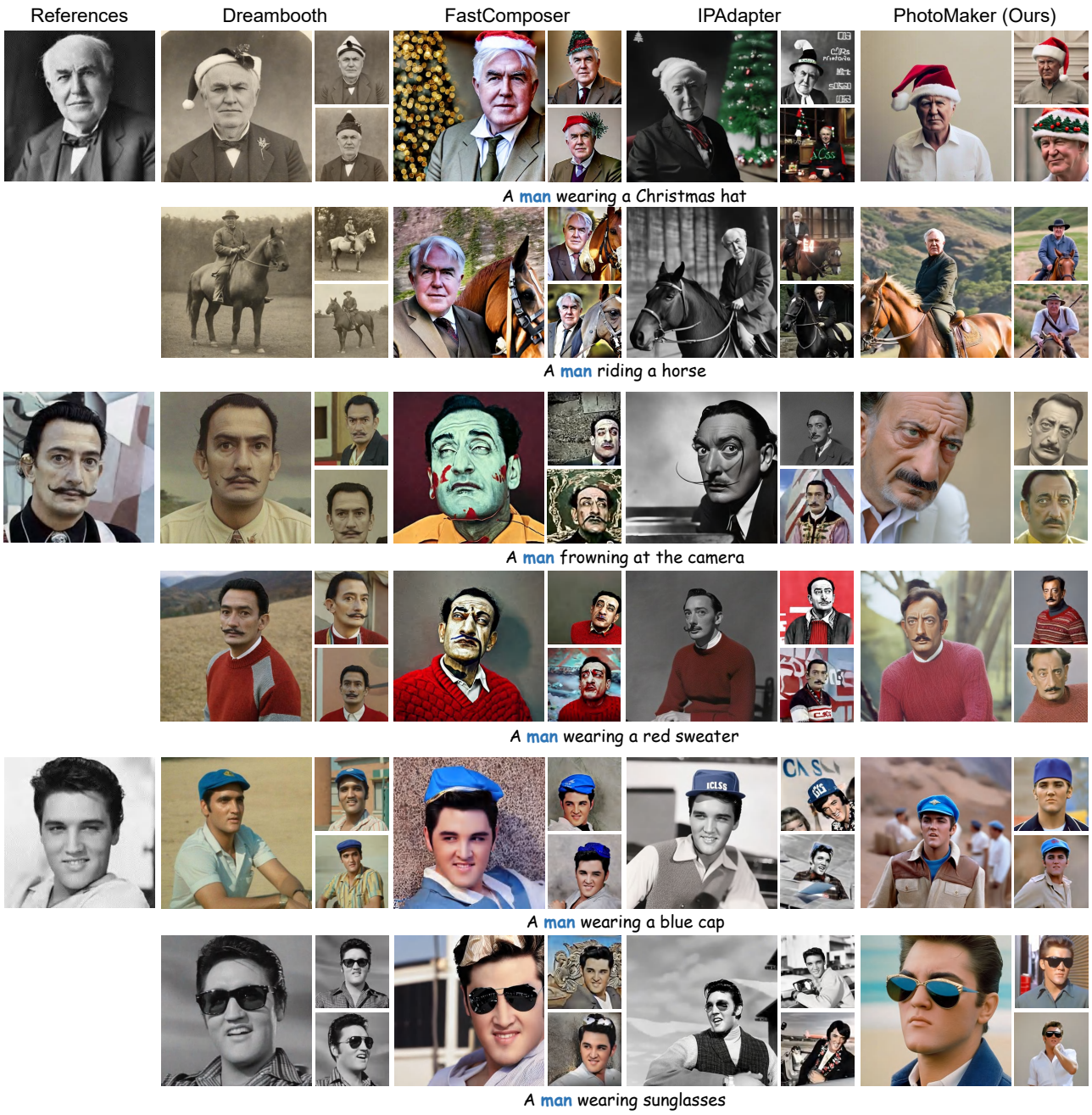

Figure 12. **More visual examples for bringing person in old-photo back to life.** Our method can generate high-quality images. We randomly sample three images for each prompt. *(Zoom-in for the best view)*

preprint arXiv:2305.10431, 2023. 1, 2, 4

[18] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 4

Figure 13. **More visual examples for bringing person in artworks back to life.** Our PhotoMaker can generate photo-realistic images while other methods are hard to achieve. We randomly sample three images for each prompt. *(Zoom-in for the best view)*

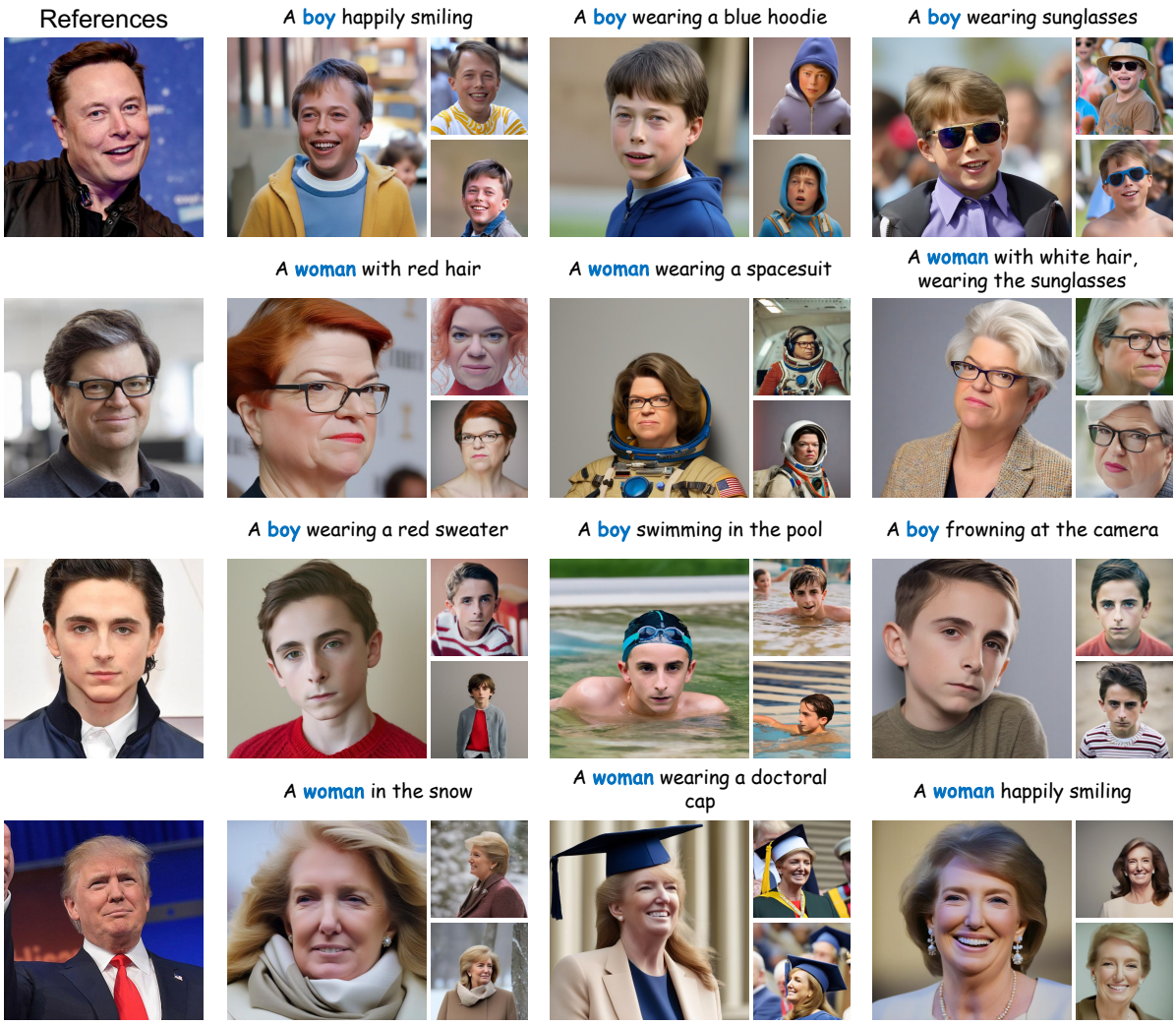

Figure 14. **More visual examples for changing age or gender for each ID.** Our PhotoMaker, when modifying the gender and age of the input ID, effectively retains the characteristics of the face ID and allows for textual manipulation. We randomly sample three images for each prompt. *(Zoom-in for the best view)*

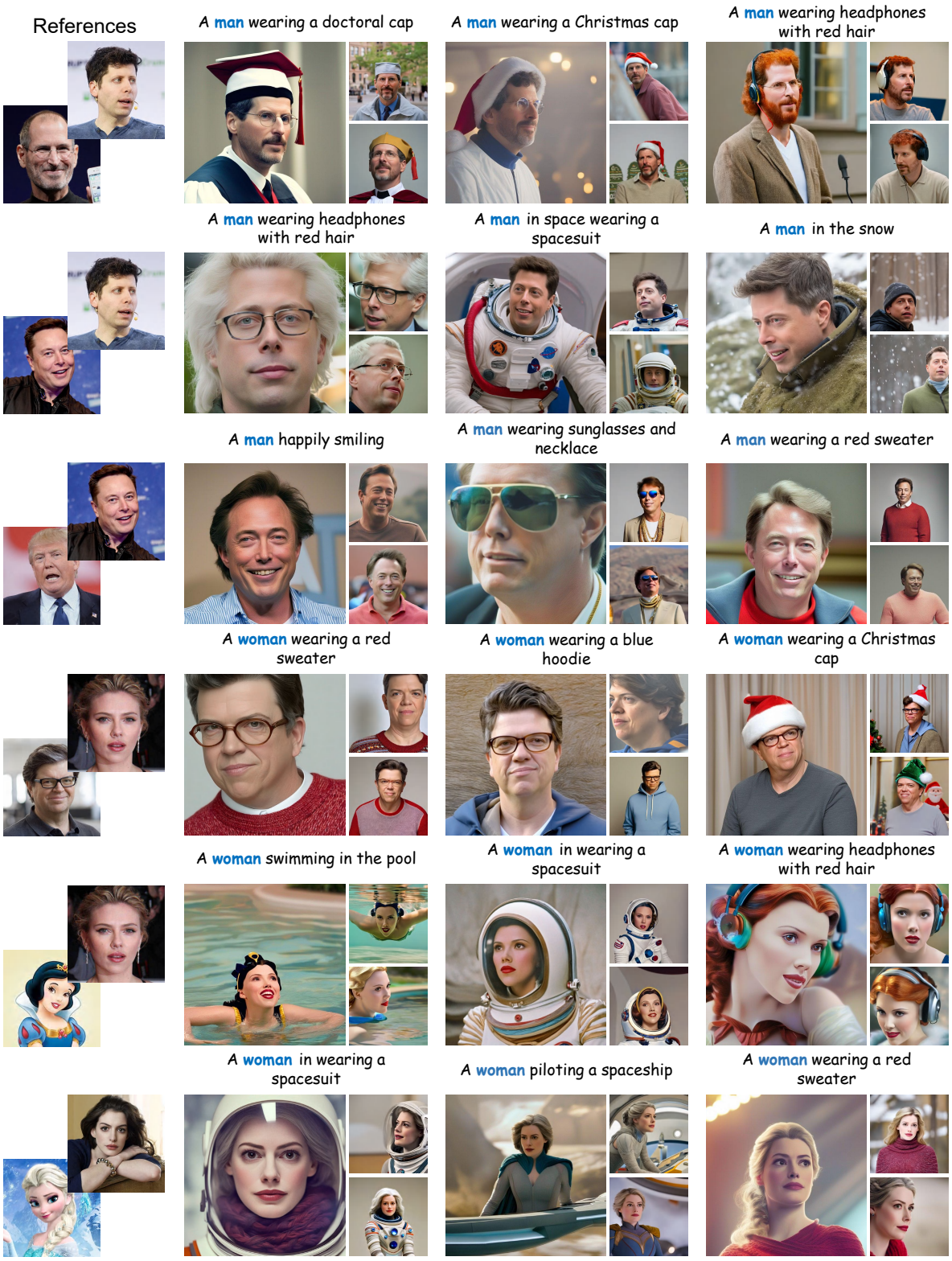

Figure 15. **More visual results for identity mixing applications.** Our PhotoMaker can maintain the characteristics of both input IDs in the new generated ID image, while providing high-quality and text-compatible generation results. We randomly sample three images for each prompt. *(Zoom-in for the best view)*