



Human Image Personalization with High-fidelity Identity Preservation

Shilong Zhang¹, Lianghua Huang², Xi Chen¹, Yifei Zhang²
Zhi-Fan Wu², Yutong Feng², Wei Wang², Yujun Shen³, Yu Liu², and Ping Luo¹

¹The University of Hong Kong, ²Alibaba Group, ³Ant Group

Abstract. This work presents **FlashFace**, a practical tool with which users can easily personalize their own photos on the fly by providing one or a few reference face images and a text prompt. Our approach is distinguishable from existing human photo customization methods by *higher-fidelity identity preservation* and *better instruction following*, benefiting from two subtle designs. First, we encode the face identity into a series of feature maps instead of one image token as in prior arts, allowing the model to retain more details of the reference faces (*e.g.*, scars, tattoos, and face shape). Second, we introduce a disentangled integration strategy to balance the text and image guidance during the text-to-image generation process, alleviating the conflict between the reference faces and the text prompts (*e.g.*, personalizing an adult into a “child” or an “elder”). Extensive experimental results demonstrate the effectiveness of our method on various applications, including human image personalization, face swapping under language prompts, making virtual characters into real people, *etc*. The page of this project is <https://jshilong.github.io/flashface-page>.

Keywords: Human photo customization · Identity preserving

1 Introduction

Image generation has advanced significantly in recent years. The overall framework has evolved from GANs [10, 29] to Diffusion models [7, 12], with improved training stability and generation quality. At the same time, large-scale multimodal datasets [28, 38] containing image-caption pairs connect the tasks of visual and linguistic processing. In this way, text starts to act as a control signal for image generation, bringing algorithms and products such as Stable Diffusion [27, 32], DALL-E series to the forefront and sparking a massive productivity revolution with diversified applications.

Among the applications of image generation, human image customization is a popular and challenging topic. It requires generating images with consistent face identity. Rather than relying solely on language conditions, this task integrates the human identity from multiple reference face images to guide the generation. Early explorations [9, 14, 35] require test-time fine-tuning, which takes about half



(a) High fidelity for facial details, and strong prompt following ability



(b) High fidelity for facial shape

Fig. 1: Diverse human image personalization results produced by our proposed FlashFace, which enjoys the features of (1) preserving the identity of reference faces *in great details* (e.g., tattoos, scars, or even the rare face shape of virtual characters) and (2) *accurately following the instructions* especially when the text prompts contradict the reference images (e.g., customizing an adult to a “child” or an “elder”). Previous SoTA refers to PhotoMaker [21].

an hour to achieve satisfactory results for each user. Recent methods [21, 45–47] investigate the zero-shot setting to make the model work on the fly (usually seconds). As shown in Fig. 2, these methods extract face ID information using a face encoder and convert it into one or several tokens. These tokens are then injected into the generation process along with text tokens. To gather sufficient training data, they also simplify data collection by cropping the face region from the target image as a reference.

These zero-shot methods make human image customization practical with significantly faster speed. However, they still have two limitations. First, they struggle to preserve the face shape and details. This is due to the loss of spatial representations when encoding the reference faces into one or a few tokens. Secondly, it is challenging for these methods to achieve precise language control. For instance, they usually fail to follow the prompt when the prompt requests the generation of an elderly person but provides a young person as the reference

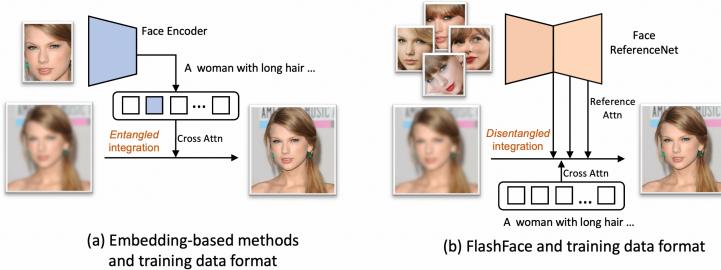


Fig. 2: Concept comparison between FlashFace and previous embedding-based methods. We encode the face to a series of feature maps instead of several tokens to preserve finer details. We do the disentangled integration using separate layers for the reference and text control, which can help to achieve better instruction following ability. We also propose a novel data construction pipeline that ensures facial variation between the reference face and the generated face.

image. We analyze that this issue is caused by treating face tokens and textual tokens equally and integrating them into U-Net at the same position. This design makes the two types of control signals entangled. Additionally, their data construction pipeline, which involves cropping the target image to construct a reference face image, leads the model to prioritize copying reference images rather than following the language prompt.

This study addresses these two major limitations at once. First, as shown in Fig. 2, to enhance facial identity fidelity, we leverage a reference network [49] to encode the reference image into a series feature map. Compared with the “token representation” applied by previous methods [21, 45–47], these feature maps retain the spatial shape and encompass richer facial shape and detail information(as shown in Fig. 1). Second, to balance the control signal of the reference image and the text prompt, we inject them in a disentangled manner. Specifically, we insert additional reference attention layers in the U-Net decoder to integrate the reference features (as shown in Fig. 2). These layers are separate from the cross-attention layer used for text conditions, ensuring that the two signals are disentangled. We also allow users to adjust the weight of the features produced by the reference attention layers. When incorporating this re-weighing operation with the classifier-free guidance [13], we achieve a smooth control of the face reference intensity. Besides the improvements of the model structure, we also propose a new data construct pipeline that gathers multiple images of a single person. This kind of data ensures the variation between the faces of the reference image and the target image during training, thereby mitigating the network’s tendency to directly copy the reference face. In this way, the model is forced to acquire guidance from the text prompt, and the prompt-following ability could be further improved.

In summary, our work makes the following contributions:

- We propose to encode the reference face image into a series of feature maps instead of one or several tokens to maintain more details, which assist the model in generating high-fidelity results.

- We incorporate reference and text control signals in a disentangled manner with separate layers. Together with our new data construction pipeline, we achieve precise language control, even in cases where there is a conflict between the prompt and the reference face images.
- We demonstrate the remarkable effectiveness of FlashFace across a wide range of downstream tasks, including human image customization, face swapping under different language prompts, bringing virtual characters as real people, etc. These capabilities open up infinite possibilities for the community.

2 Related Work

Text-to-Image Generation. Text-to-image generation [3, 10, 16, 29, 30, 32, 36, 37] has significantly progressed in recent years, which can be primarily attributed to the development of diffusion models [7, 12, 27, 39, 41] and auto-regressive models [2, 31]. The state-of-the-art text-to-image generation models are capable of generating images precisely following text prompts provided by the user [27, 36]. Meanwhile, the generated images are remarkably realistic while maintaining a high level of aesthetic appeal [30]. However, these existing methods are limited to generating images solely from text prompts, and they do not meet the demand for efficiently producing customized images with the preservation of identities.

Subject-Driven Image Generation. Subject-driven image generation focuses on creating new images of a particular subject based on user-provided images. Methods like DreamBooth [35] and textual-inversion [9] along with their subsequent works [20, 23, 24, 44] utilize optimization-based approaches to embed subjects into diffusion models. However, these methods usually necessitate approximately 30 minutes of fine-tuning per subject, posing a challenge for widespread utilization. Zero-shot subject-driven image generation [4, 25, 42, 47] aims to generate customized images without extra model training or optimization process. Nevertheless, an obvious issue is that compared to optimization-based methods, these zero-shot approaches often sacrifice the ability to preserve identity, i.e., the generated images could lose the fine details of the original subject.

Human Image Personalization. In this paper, we mainly focus on the identity preservation of human faces, as it is one of the most widely demanded and challenging directions in subject-driven image generation. Recently, methods like IP-Adapter-FaceID [47], FastComposer [45], and PhotoMaker [21] show promising results on zero-shot human image personalization. They encode the reference face to one or several tokens as conditions to generate customized images. Yet, these methods usually fail to maintain the details of human identities. In contrast, our proposed method precisely retains details in human faces and accurately generates images according to text prompts in a zero-shot manner.

3 Method

In this section, we first introduce a universal pipeline for the construction of person ID datasets. Distinct from previous dataset [17, 22, 50], we incorporate

individual ID annotation for each image, allowing the sampling of multiple images of the same individual throughout the training process. The tools used in this pipeline are publicly accessible on the Internet, simplifying replication for fellow researchers. Next, we present **FlashFace** framework, which efficiently preserves facial details, and enables high-fidelity human image generation.

3.1 ID Data Collection Pipeline

We introduce a generic methodology for assembling a person ID dataset, which includes numerous images of each individual. Our approach involves data collection, data cleaning, and annotation. On average, the dataset presents approximately one hundred images per person, ensuring a diverse range of expressions, accessories, hairstyles, and poses. We believe that this pipeline will greatly contribute to the advancement of human image personalization.

Data collection. Our initial step is to collect public figures who are well-represented online, including actors, musicians, and other celebrities. To construct our ID list, we extract names from ranking platforms like IMDb [15], Douban [8], YouGov [48], among others. Utilizing these names, we construct a rich repository of publicly accessible images. To capture a wide set of visual data, we collect at least 500 images for each individual.

Data cleaning. To remove noise images that feature individuals other than the intended person within each ID group, we employ a face detection model (Retinaface-R50 [5]) to detect all faces. We then utilize a face recognition model (ArcFace-R101 [6]) to obtain face embeddings. By applying the k-means clustering algorithm, we select the center embedding of the largest cluster as the indicator embedding for the intended person. Subsequently, we filter out similarity scores below a specified threshold (0.6) within the largest cluster and all images in other clusters, effectively removing the noise images.

Image annotation. To ensure our image annotations are diverse and rich, we leverage Qwen-VL [1], a powerful vision-language model that is capable of accurately describing individuals in the images. We prompt it with a specific query: “Can you provide a detailed description of the person in the photo, including their physical appearance, hairstyle, facial expression, pose or action, accessories, clothing style, and background environment?” This allows us to generate captions that offer detailed descriptions of images, thus enabling precise control over the appearance of individuals.

Dataset statistics. Finally, leveraging our scalable pipeline, we managed to compile a dataset of 1.8 million images featuring 23,042 individuals. Each individual is represented by an average of 80 images, providing a rich individual context that enables our model to learn identification features from the reference images instead of simply copying the original image onto target images. We construct a validation dataset consisting of 200 individuals, while the remaining data is used for training. For additional statistical details about the dataset, please refer to our *Supplementary Material*.

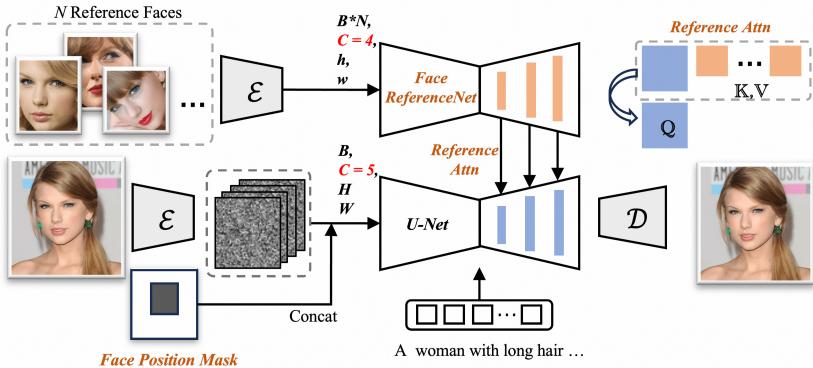


Fig. 3: The overall pipeline of FlashFace. During training, we randomly select B ID clusters and choose $N + 1$ images from each cluster. We crop the face region from N images as references and leave one as the target image. This target image is used to calculate the loss. The input latent of Face ReferenceNet has shape $(B * N) \times 4 \times h \times w$. We store the reference face features after the self-attention layer within the middle blocks and decoder blocks. A face position mask is concatenated to the target latent to indicate the position of the generated face. During the forwarding of the target latent through the corresponding position in the U-Net, we incorporate the reference feature using an additional reference attention layer. During inference, users can obtain the desired image by providing a face position(optional), reference images of the person, and a description of the desired image.

3.2 FlashFace

As shown in Fig. 3, the overall framework is based on the widely used SD-V1.5 project [33]. All images are projected into the latent space with a stride 8 auto-encoder [19]. A U-Net [34] is employed for denoise. The language branch of CLIP [28] is introduced to encode the language prompt and then use the cross-attention operation to do the integration. A notable feature of our framework is the inclusion of a Face ReferenceNet [49], which extracts detailed facial features that retain the spatial shape and incorporate them into the network using additional reference attention layers.

Face ReferenceNet. We utilize an identical U-Net architecture with the same pre-trained parameters (SD-V1.5 [33]) as our Face ReferenceNet. During training, we randomly sample B ID clusters, and then random sample N reference images from each ID cluster. We crop the face regions and resize them to 224×224 . After VAE encoding, these faces are converted to the input latents with the shape of $(B * N) \times 4 \times 28 \times 28$. As the latent feature passes through the middle and decoder blocks, we cache all features after the self-attention layer, which have shape $(B * N) \times (h_r * w_r) \times C$. The number of reference images N is randomly set between 1 and 4 during training.

U-Net with reference attention layer. During the training phase, we encode the selected B target images as initial latent representations with 4 channels. For our face inpainting model, we fill the latent representation inside the face region with 0 and concatenate it with the noised latent representation to form

the input with 8 channels. To facilitate the composition for the user, we initialize a face position mask, which is a one-channel zero mask only the region within the face bounding box filled 1. The face mask is then connected to the noised latent representation as the input latent. When the latent feature forwards to the same position as the cached reference features, we use an extra reference attention layer to incorporate reference face features into the U-Net. As shown in Eq. (1), the query(Q) remains as the U-Net latent feature(y_g), which shapes $B \times (h_g * w_g) \times C$, while the key(K) and value(V) are concatenated features consisting of the U-Net and all reference face features(y_{ri} , i refer to the i_{th} reference face image). All the reference features are contacted with the U-Net feature along the spatial dimension, and get the K, V with length $h_g * w_g + N * h_r * w_r$. To maintain the original image generation capability, we randomly drop the face position mask and reference feature with a probability of 0.1 during training.

$$y_g = y_g + \begin{cases} Attn(Q = y_g, K, V = y_g), & \text{if } random < 0.1 \\ Attn(Q = y_g, K, V = cat[y_g, y_{r1}, y_{r2}...]), & \text{if } random \geq 0.1 \end{cases} \quad (1)$$

Face reference strength. We also provide one way to control the reference strength in the inference phase. This allows users to adjust the balance of the language prompt and references in case of conflicts. For example, to make a person a bronze statue as seen in Fig. 9.

Firstly, in Eq. (2), we introduce λ_{feat} to re-weight the two reference attention layer results: One result incorporates the reference features in its K and V , while the other does not. A higher value of λ_{feat} assigns greater weight to the result with reference feature, resulting in a higher ID fidelity face in the generated images. This operation provides smooth control over the reference strength. It also facilitates the **identity mixing** by easily changing to re-weight attention results associated with two different individuals, we show the case in our *Supplementary Material*.

$$y_g = y_g + (1 - \lambda_{feat}) * Attn(y_g, y_g) + \lambda_{feat} * Attn(y_g, cat[y_g, y_{r1}, y_{r2}...]) \quad (2)$$

Secondly, we also involve classifier-free guidance [13] by combining three inference results under different conditions: no reference image or text prompt (Y_{None}), only text prompt (Y_{Text}), and both text prompt and reference image ($Y_{Text\&Ref}$). The final results are obtained by combining these results using the Eq. (3), and λ_{text} , λ_{ref} are weighting factors for text guidance and reference guidance. Reference guidance is not that smooth like λ_{feat} , but it helps to preserve more details(refer to *Supplementary Material*). In this study, $\lambda_{feat} = 0.85$, $\lambda_{ref} = 2$, and $\lambda_{text} = 7.5$ by default.

$$Y = Y_{None} + \lambda_{text} * (Y_{Text} - Y_{None}) + \lambda_{ref} * (Y_{Text\&Ref} - Y_{Text}) \quad (3)$$

4 Experiments

4.1 Implementation details

In this study, we employ the SD-V1.5 model [32]. For both human customization and face inpainting models, we use Adam [18] to do the parameter optimization

on 8 NVIDIA A100 GPUs, with a batch size of 32, for a total of 50000 iterations. The learning rate is set to 2e-5 for both the Face ReferenceNet and U-Net modules. During training, we randomly select N reference images and one target image from the same ID cluster for training purposes (N is randomly set between 1 and 4). We crop the face region of the reference images and resize them to 224×224 . The target images are also resized to 768×768 to represent the face details better. To improve generation quality through classifier-free guidance [13], there is a 10% chance of using a null-text embedding. For sampling, we utilize 50 steps of the DDIM sampler [40]. The feature strength λ_{feat} is set to 0.85 by default. The scale of text classifier-free guidance [13] is set to 7.5 by default, and the reference guidance is set to 2 by default.

4.2 Applications

In this section, we demonstrate four applications that highlight the strong customization ability of our method. With our novel human customization model, we can generate realistic human images of different attributes, ages, and genders while keeping the identity of individuals. Additionally, we can do the transformation between the artwork and real people with high face ID fidelity. Finally, we showcase the impressive performance of our face inpainting model, which seamlessly places your face to anybody even with an additional language prompt. **FlashFace** and PhotoMaker [21] utilize four reference images as input, while the remaining methods use only one reference image as input. The reference images for each figure can be found in our *Supplementary Material*.

General human image customization. As shown in Fig. 4, we present image generation results of “Dwayne Johnson” using different language prompts. We note a tendency of copy-pasting and same appearance among the generated faces of IP-Adapter-faceid [47], FastComposer [45], and InstantID [43]. They fail to achieve natural face variations based on the language prompt. For instance, “singing” and “clown makeup”. **FlashFace** maintains high fidelity and accurately follows the language prompt. With SD-V1.5 [32], and achieve comparable performance to the concurrent work, PhotoMaker [21], which uses a much larger model (SDXL [27]).

Change the age and gender. As shown in Fig. 5, **FlashFace** demonstrates a remarkable capability to alter age and gender, which conflicts with the reference images. By utilizing four reference images of “Taylor Swift” and reference strength $\lambda_{feat} = 0.85$ and guidance scale $\lambda_{ref} = 2$, **FlashFace** can generate images that seamlessly combine the youthful facial features of a “girl”, the aging wrinkles of “old”, and the masculine characteristics of a “man” into the generated face while maintaining ID information. In comparison, IP-Adapter-faceid [47] shares the same face part for all three age-related prompts and even fails to maintain facial fidelity. For our concurrent work, Photomaker [21] and InstantID [43] successfully preserve identity information. However, when it comes to different ages and genders, the facial details remain unaltered and continue to resemble a young woman. (Please zoom in to examine the facial details closely)



Fig. 4: Image personalization results of “Dwayne Johnson”. FlashFace maintains high fidelity and accurately follows the language prompt. The methods marked with * are our concurrent works.

Transformation between the artwork and real people. FlashFace also supports the transformation between virtual persons and real individuals with high face-shape fidelity. For virtual persons into real, we only require a single reference image, and a simple description(such as "a man in a black suit"). When converting a real person into an artwork, we utilize four reference images. For simplicity, only one image is displayed in the Fig. 6. Additionally, the prompt should specify the type of artwork, for example, "a marble statue of a man" or "a bronze statue of a man".

ID face inpainting. FlashFace_{inpainting} model can seamlessly paint one's face onto another person under the control of a language prompt. As shown in Fig. 7, by utilizing four reference images of “Leonardo DiCaprio” along with language prompts, the model ensures the preservation of facial identity while



Fig. 5: Personalization results of “Taylor Swift” with different ages and genders. For different prompts, the facial details of Photomaker [21] and InstantID [43] remain unaltered and continue to resemble a young woman. Our FlashFace seamlessly combines the characteristics represented by age and gender words with the original person ID. The methods marked with * are our concurrent works.



Fig. 6: Transformation between artwork and real people. FlashFace even can do transformation between artwork and real people with high face fidelity. We can see the face shape is well preserved in both transformations.

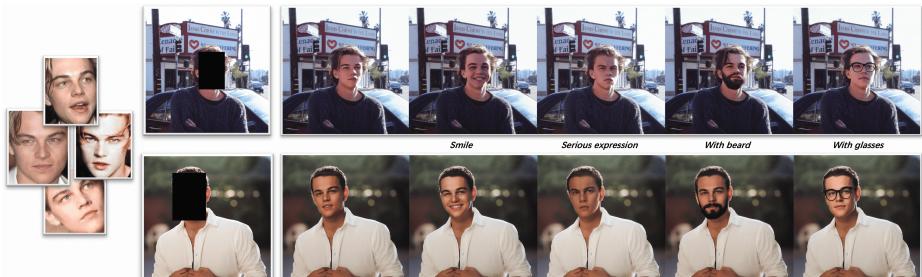


Fig. 7: ID face inpainting with different prompts. FlashFace_{inpainting} is capable of inpainting a person’s face onto another individual with high face ID fidelity. Additionally, during the inpainting process, it allows using language to specify facial details like facial expressions, glasses, beards, etc.

Table 1: Performance comparison of FlashFace and previous methods. We report the face similarity score between the generated image and the target image (Sim_{Target}), as well as the maximum similarity between the generated face and all reference faces (Sim_{Ref}). We also provide the CLIP similarity score of the image. A higher value of $Paste = Sim_{Ref} - Sim_{Target}$ indicates a greater likelihood of pasting one reference face. Our FlashFace achieves the best results in terms of Sim_{Target} , demonstrating a high fidelity of ID and better language following ability.

Method	Sim_{Ref}	$CLIP \uparrow$	$Paste \downarrow$	$Sim_{Target} \uparrow$	$FID \downarrow$
IP-Adaptor(SD1.5) [47]	33.4	57.5	7.7	25.7	121.1
FastComposer(SD1.5) [45]	33.3	65.2	7.5	25.8	104.9
PhotoMaker(SDXL) [21]	45.8	72.1	5.8	40.0	84.2
InstantID(SDXL) [43]	68.6	62.7	16.0	52.6	106.8
FlashFace(SD1.5)	63.7	74.2	7.6	56.1	85.5
FlashFace(SD1.5)_{inpaiting}	65.6	89.2	3.8	61.8	26.6

introducing various facial details. Prompts may encompass modifications in facial expressions, as well as the addition of glasses or a beard.

4.3 Benchmark Results

We create a benchmark with the 200 ID clusters in our validation split. We select 4 reference images and 1 target image from each cluster. To evaluate the generated images, we measure the face similarity using a face recognition model (ArcFace-R101 [6]). The most important metric is the face similarity between the generated image and the target image, denoted as Sim_{Target} . Additionally, we calculate the maximum face similarity between the generated images and all reference faces, denoted as Sim_{Ref} . We also report the CLIP similarity score. To demonstrate that the model can extract ID information and vary following language prompts instead of simply pasting a reference face onto the target image, we calculate $Paste = Sim_{Ref} - Sim_{Target}$ as a metric. A higher value of $Paste$ indicates a greater likelihood of pasting one of the reference images. For the single reference image method, we simply sample the first reference image as input. We also report the FID score [11, 26] as a metric of image quality.

As shown in Tab. 1, IP-Adapter [47], FastComposer [45] and PhotoMaker [21] demonstrate a low score for both Sim_{Ref} and Sim_{Target} , indicating challenges in preserving facial details when using their embedding-based pipeline. For our concurrent work, InstantID [43], a very high value $Paste$ suggests a strong tendency of reference face pasting instead of following the language prompt. Our method outperforms others in terms of Sim_{Target} with a reasonable $Paste$. This suggests that the FlashFace is capable of doing facial variations based on the language prompt while maintaining a high level of identity fidelity.

4.4 Ablation study

Details preserving ability of ReferenceNet. We assess the unbounded ability of ReferenceNet to maintain face details. We train our model using

Table 2: Identity-preserving comparison under the same training paradigm.
When employing the same training paradigm as previous methods, which involve cropping the human face from the target as a reference, **FlashFace** achieves a high Sim_{Ref} . This indicates **FlashFace** has a strong ability to preserve details.

	IP-Adaptor [47]	FastComposer [45]	Instant-ID [43]	FlashFace
Sim_{Ref}	33.4	33.3	68.6	82.0

the cropped target face as the input reference, similar to IP-Adapter [47] and FastComposer [45]. As shown in Tab. 2, **FlashFace** can achieve much higher Sim_{Ref} than other methods, indicating the strong ability of ReferenceNet to preserve face details.

Face reference strength. Fig. 8.(A) reveals a fidelity(Sim_{Target}) improvement when enhancing the reference strength with λ_{feat} and $\lambda_{guidance}$. However, the *Paste* also increases as shown in Fig. 8.(B). This suggests that there is also an increasing tendency to copy the reference image rather than incorporating language prompts. To achieve a balance between fidelity and language prompting ability, we suggest setting λ_{feat} between 0.75 and 1, and $\lambda_{guidance}$ between 2 and 2.5.

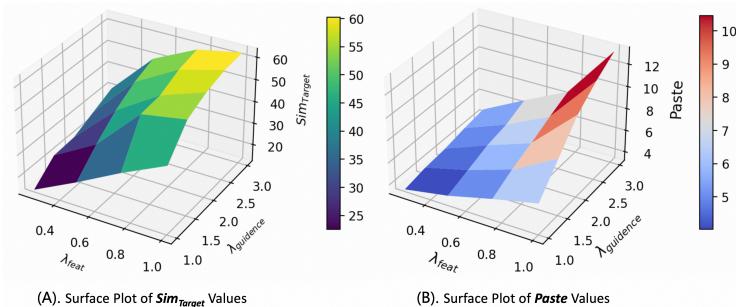


Fig. 8: The surface plot of Sim_{Target} and $Paste$ with respect to λ_{feat} and $\lambda_{guidance}$. Both values, $Paste$ and Sim_{Target} , increase as λ_{feat} and $\lambda_{guidance}$ increase. To balance fidelity and language prompting ability, we recommend setting λ_{feat} between 0.75 and 1, and $\lambda_{guidance}$ between 2 and 2.5.

We also note that λ_{feat} exhibits smoother Sim_{Target} changes as it operates in the feature space. We visualize the results for λ_{feat} ranging from 0 to 1 under $\lambda_{guidance} = 2$ to show the smooth visual similarity improvement (as shown in Fig. 9). Notably, our model consistently produces natural results across the entire λ_{feat} range. Another noteworthy observation is that in cases of conflicting conditions, such as when attempting to create a bronze-like face of “Dwayne Johnson”, the optimal result is not achieved at $\lambda_{feat} = 1$. At that point, the facial features lose their metallic luster. This highlights an advantage of our framework: the ability to balance conflicts by adjusting the control strength between the text and reference. In our *Supplementary Material*, we demonstrate



Fig. 9: Smooth ID fidelity improvement as λ_{feat} ranges from 0 to 1. Using four reference face images of “Dwayne Johnson”, we increase the λ_{feat} from 0 to 1, moving from left to right, while $\lambda_{guidance}$ is set to 2. Throughout this process, we notice a smooth improvement in ID fidelity.

Table 3: Performance of FlashFace under different numbers of reference images. Increasing the number of reference images led to a significant improvement in face fidelity.

# Ref	Sim_{Ref}	$CLIP \uparrow$	$Paste \downarrow$	$Sim_{Target} \uparrow$	$FID \downarrow$
1	49.9	69.6	9.8	40.1	85.3
2	58.7	72.2	9.2	49.4	85.1
3	62.3	73.4	8.0	54.3	85.4
4	63.7	74.2	7.6	56.1	85.5

the smooth change characteristics of λ_{feat} for identity mixing and highlight the role of $\lambda_{guidance}$ in preserving details.

Number of input reference faces. Tab. 3 presents the evaluation results of FlashFace using different numbers of reference images. Increasing the number of reference faces leads to a noticeable improvement in identity enhancement. Specifically, the improvement is significant when changing from 1 reference face to 2, resulting in a 9.3% increase. Another crucial aspect is the reduction of *Paste*. This reduction implies that the model can learn more precise identity information and generate high-quality images based on the prompt, rather than simply relying on reference images.

Tab. 3 presents the evaluation results of FlashFace using different numbers of reference images. Increasing the number of reference faces leads to a noticeable improvement in identity enhancement. Specifically, the improvement is significant when changing from 1 reference face to 2, resulting in a 9.3% increase. Another crucial aspect is the reduction of *Paste*. This reduction implies that the model can learn more precise identity information and generate high-quality images based on the prompt, rather than simply relying on reference images.

The consistent visualization results can be found in Fig. 10, which contains generated images with the prompt “A young woman in a white dress”. Increasing the number of reference images from 1 to 2 indeed significantly enhances the fidelity of identity. When only one reference image is used, the model can capture certain facial features such as the nose and eyes. However, it struggles to accurately depict the overall face shape (The generated face is not round enough).



Fig. 10: Generated image visualization under different numbers of reference faces. The number of reference images increases from 1 to 4, progressing from left to right. We observe a significant improvement in fidelity, particularly when changing the number from 1 to 2.

Table 4: Performance of incorporating reference attention layers to various positions in the U-Net. We achieve the best performance by adding it to the decoder blocks.

Model Part	Sim_{Ref}	$CLIP \uparrow$	$Paste \downarrow$	$Sim_{Target} \uparrow$	$FID \downarrow$
Encoder	8.4	61.9	1.6	6.8	90.3
Encoder & Decoder	60.3	73.0	8.1	52.2	88.1
Decoder	63.7	74.2	7.6	56.1	85.5

Nevertheless, as the number of reference images increases, the model gradually improves in capturing all these details.

Position of the reference attention layer. As shown in Tab. 4, we conduct ablation experiments by adding reference attention layers to different positions of the U-Net, including the encoder blocks, decoder blocks, or both(Adding reference attention layer to middle blocks by default). We observe that when reference attention layers are only added to the encoder blocks, the model is hard to capture the face identity. This can be attributed to the encoder being further away from the output, making it difficult to align with the convergence direction indicated by the loss. Corresponding to this, when reference attention layers are added solely to the decoder blocks, we achieve the best performance even compared to adding them in both the encoder and decoder.

5 Conclusion

In this paper, we present **FlashFace**, a zero-shot human image personalization method. It excels in *higher-fidelity identity preservation* and *better instruction following*. We encode the reference face into a series of feature maps of different spatial shapes with a Face ReferenceNet, which effectively preserves both face shape and details. Additionally, we utilize a disentangled condition integration strategy that balances the text and image guidance during the text-to-image generation process, thereby achieving powerful language control even if there are conflicts between the reference image and language prompt. Experiments prove we achieve exceptional personalization capabilities, offering endless possibilities for related applications.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 5
2. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M., Murphy, K.P., Freeman, W.T., Rubinstein, M., Li, Y., Krishnan, D.: Muse: Text-to-image generation via masked generative transformers. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 4055–4075 (2023) 4
3. Chen, S., Xu, M., Ren, J., Cong, Y., He, S., Xie, Y., Sinha, A., Luo, P., Xiang, T., Perez-Rua, J.M.: Gentron: Delving deep into diffusion transformers for image and video generation (2023) 4
4. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023) 4
5. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 5
6. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5962–5979 (Oct 2022). <https://doi.org/10.1109/tpami.2021.3087709>, <http://dx.doi.org/10.1109/TPAMI.2021.3087709> 5, 11
7. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (NeurIPS). pp. 8780–8794 (2021) 1, 4
8. douban: douban. <https://m.douban.com/movie/> 5
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 1, 4
10. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014) 1, 4
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018) 11
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020) 1, 4
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 3, 7, 8, 20
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 1, 18
15. IMDb: Imdb. <https://www.imdb.com/> 5
16. Kang, M., Zhu, J., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10124–10134 (2023) 4
17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019) 4
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 7

19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 6
20. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion (2023) 4
21. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. arXiv preprint arXiv:2312.04461 (2023) 2, 3, 4, 8, 10, 11, 18, 22
22. Liu, X., Ren, J., Siarohin, A., Skorokhodov, I., Li, Y., Lin, D., Liu, X., Liu, Z., Tulyakov, S.: Hyperhuman: Hyper-realistic human generation with latent structural diffusion. arXiv preprint arXiv:2310.08579 (2023) 4
23. Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones: Concept neurons in diffusion models for customized generation. arXiv preprint arXiv:2303.05125 (2023) 4
24. Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones 2: Customizable image synthesis with multiple subjects. arXiv preprint arXiv:2305.19327 (2023) 4
25. Pan, X., Dong, L., Huang, S., Peng, Z., Chen, W., Wei, F.: Kosmos-g: Generating images in context with multimodal large language models. arXiv preprint arXiv:2310.02992 (2023) 4
26. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation (2022) 11
27. Podell, D., English, Z., Lacev, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 1, 4, 8, 23
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021) 1, 6
29. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016) 1, 4
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022) 4
31. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 8821–8831. PMLR (2021) 4
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022) 1, 4, 7, 8, 23
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022) 6
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 6
35. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 1, 4

36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022) [4](#)
37. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 30105–30118 (2023) [4](#)
38. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022) [1](#)
39. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the International Conference on Machine Learning (ICML). vol. 37, pp. 2256–2265 (2015) [4](#)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) [8](#)
41. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 11895–11907 (2019) [4](#)
42. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. arXiv preprint arXiv:2312.13286 (2023) [4](#)
43. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024) [8](#), [10](#), [11](#), [12](#), [22](#)
44. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023) [4](#)
45. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023) [2](#), [3](#), [4](#), [8](#), [11](#), [12](#), [22](#)
46. Yan, Y., Zhang, C., Wang, R., Zhou, Y., Zhang, G., Cheng, P., Yu, G., Fu, B.: Facestudio: Put your face everywhere in seconds. arXiv preprint arXiv:2312.02663 (2023) [2](#), [3](#)
47. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arxiv:2308.06721 (2023) [2](#), [3](#), [4](#), [8](#), [11](#), [12](#), [22](#)
48. YouGov: Yougov. <https://today.yougov.com/ratings/international/fame/all-time-people/all> [5](#)
49. Zhang, L.: Reference-only controlnet. <https://github.com/Mikubill/sd-webui-controlnet/discussions/1236> (2023) [3](#), [6](#)
50. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. arXiv preprint arXiv:2112.03109 (2021) [4](#)

FlashFace: Human Image Personalization with High-fidelity Identity Preservation

(*Supplementary Material*)

Shilong Zhang¹, Lianghua Huang², Xi Chen¹, Yifei Zhang²
Zhi-Fan Wu², Yutong Feng², Wei Wang², Yujun Shen³, Yu Liu², and Ping Luo¹

¹The University of Hong Kong, ²Alibaba Group, ³Ant Group

- Sec. 1: All the reference images for each figure in the main script.
- Sec. 2: More statistics about the dataset.
- Sec. 3: Identity mixing visualization.
- Sec. 4: How the λ_{ref} helps to preserve more details.
- Sec. 5: Comparison of artwork-real transformation with PhotoMaker.
- Sec. 6: LoRA version of FlashFace.
- Sec. 7: Does face position mask affect performance of FlashFace?
- Sec. 8: More visualization results.
- Sec. 9: Limitations and Social Impact.

In this *Supplementary Material*, we begin by showcasing all the reference images for each figure in the main script. Then, we provide more statistics about the dataset we collected. Additionally, we present identity mixing visualization results to demonstrate the smooth control of λ_{feat} in the feature space. Then, we show how the λ_{ref} helps to preserve more details. Furthermore, we compare the transformation between artwork and real people with the previous state-of-the-art PhotoMaker [21]. We also report the performance without the face position mask. Then we provide a LoRA [14] version of FlashFace and analyze its performance. We provide more visualization results that demonstrate our strong ability to maintain identity fidelity and follow language prompts across various applications under diverse prompts. Finally, we discuss our limitations, social impact, and responsibility to human subjects.

1 Reference Images for Each Figure

As shown in Fig. 1, We list all reference images for each individual who is presented in the figures of the main script. Real-person customization involves four reference images, while virtual person to real person transformation only requires one reference face image, cropped from the ID Images. For brevity, we omit the reference images for the virtual person in Figure 1.

2 Statistics about the Dataset

We provide statistics on the dataset collected using our ID data collection pipeline, which consists of 1.8 million images featuring 23,042 individuals. In



Fig.1

Fig.1



Fig.1

Fig.5



Fig.4 & Fig.6 & Fig.9

Fig.6

Fig. 1: Reference faces images for each figure in the main script. Most generated images have four reference images, while the virtual person to real person transformation has only one reference face image, which is cropped from the ID Images. Therefore, we omit them in this figure.

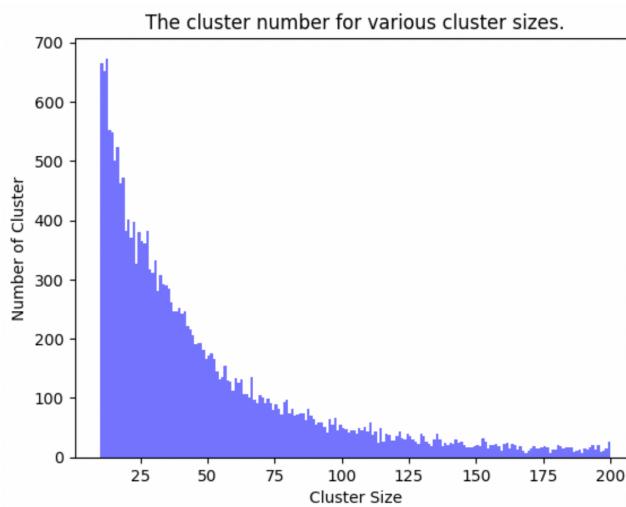


Fig. 2: Number of clusters for different cluster size

the Fig. 2, we display the number of individuals with specific cluster sizes. In Fig. 3, we also provide five images from the same cluster to demonstrate the variance within the cluster.



Fig. 3: Five images from a cluster

3 Identity Mixing

We do the identity-mixing by adjusting the weights of the reference layer attention results for two different individuals with λ_{ID1} and λ_{ID2} (as shown in Eq. (1),). These weights are constrained by the equation $\lambda_{ID1} + \lambda_{ID2} = 1$. Additionally, $y_r^{id_i}$ represents the j -th reference face feature of the i -th person .

$$\lambda_{ID1} * Attn(y_g, cat[y_g, y_{r1}^{id_1}, y_{r2}^{id_1}...]) + \lambda_{ID2} * Attn(y_g, cat[y_g, y_{r1}^{id_2}, y_{r2}^{id_2}...]) \quad (1)$$

We visualize the results of mixing two individuals, “Taylor Swift” and “Leonardo DiCaprio”. By increasing the weight of “Leonardo DiCaprio” from left to right, we observe a smooth transition from “Taylor Swift” to “Leonardo DiCaprio”.



Fig. 4: **Visualization of identity mixing results.** We present the smooth transition from "Taylor Swift" to "Leonardo DiCaprio" in the identity mixing process, displayed from left to right.

4 Visualization under Different λ_{ref} and λ_{feat}

We present the visualization results of "Mike Tyson" with varying values of λ_{ref} and λ_{feat} in the Fig. 5. Increasing λ_{ref} and λ_{feat} leads to a noticeable increase in fidelity.

An interesting observation is that λ_{ref} is crucial for preserving details. When we remove the reference classifier-free guidance [13] and use the inference method

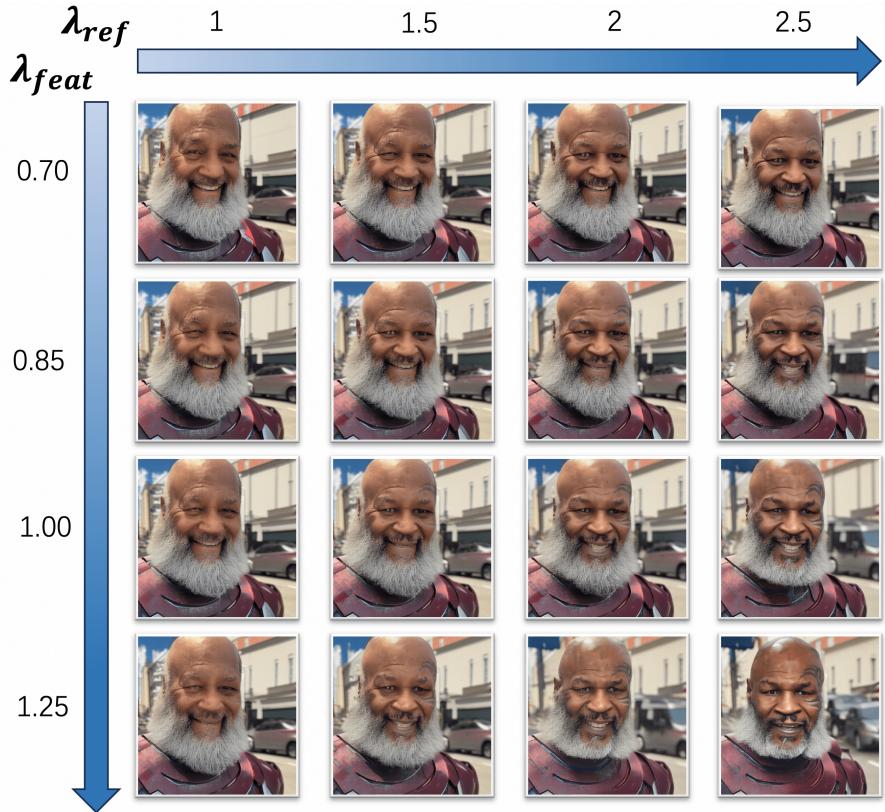


Fig. 5: Visualization under different values of λ_{ref} and λ_{feat} .

as Eq. (2), we notice that the face shape remains faithful to the identity, and get pretty good results for face similarity($\text{FlashFace}_{w/o-cfg}(SD1.5)$ in Tab. 1). However, certain details such as scars, tattoos, and other face details are challenging to preserve(as shown in Fig. 6).

$$Y = Y_{Ref} + \lambda_{text} * (Y_{Text\&Ref} - Y_{Ref}) \quad (2)$$



Fig. 6: Visualization of with and without reference classifier-free guidance.

Table 1: Performance of different versions of FlashFace. We report the performance of the LoRA version of FlashFace ($\text{FlashFace}_{LoRA}(SD1.5)$). Additionally, we evaluate the performance when removing the reference classifier free guidance($\text{FlashFace}_{w/o-cfg}(SD1.5)$) and when removing the face position mask($\text{FlashFace}_{w/o-pmask}(SD1.5)$). We report the face similarity score between the generated image and the target image(Sim_{Target}), as well as the maximum similarity between the generated face and all reference faces (Sim_{Ref}). We also provide the *CLIP* similarity score of the image. A higher value of $Paste = Sim_{Ref} - Sim_{Target}$ indicates a greater likelihood of pasting one reference face.

Method	Sim_{Ref}	$CLIP \uparrow$	$Paste \downarrow$	$Sim_{Target} \uparrow$	$FID \downarrow$
IP-Adaptor(SD1.5) [47]	33.4	57.5	7.7	25.7	121.1
FastComposer(SD1.5) [45]	33.3	65.2	7.5	25.8	104.9
PhotoMaker(SDXL) [21]	45.8	72.1	5.8	40.0	84.2
InstantID(SDXL) [43]	68.6	62.7	16.0	52.6	106.8
FlashFace(SD1.5)	63.7	74.2	7.6	56.1	85.5
FlashFace_{w/o-pmask}(SD1.5)	62.3	73.2	7.1	55.2	85.8
FlashFace_{w/o-cfg}(SD1.5)	61.6	72.9	8.3	53.3	86.1
FlashFace_{LoRA}(SD1.5)	48.7	71.7	5.3	43.4	89.1

5 Comparison of Artwork and Real People Transformation.

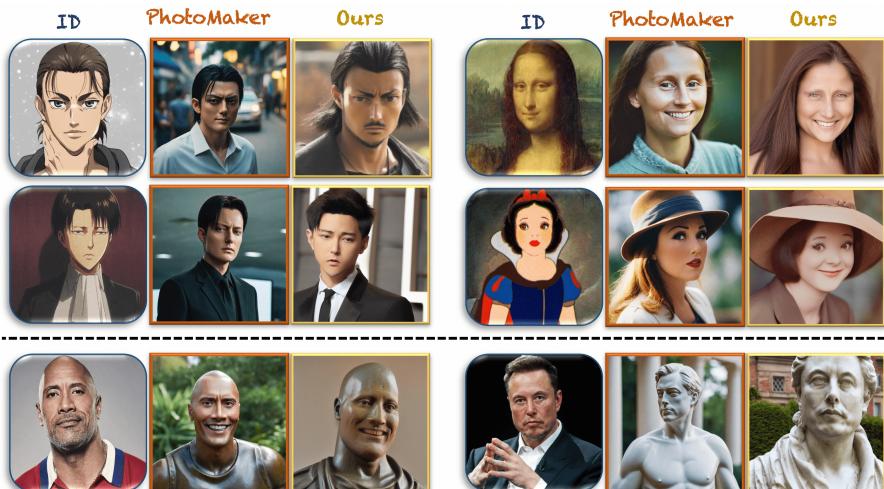


Fig. 7: Comparison of artwork and real People transformation.

We compare our results with the previous state-of-the-art PhotoMaker [21]. Using the same input reference image and prompt, we present the results in

Figure 7. Our method achieves comparable or even superior results compared to the previous state-of-the-art, utilizing a more compact model, SD-1.5 [32].

6 LoRA Version of FlashFace

We additionally provide a LoRA version of FlashFace by integrating LoRA into all attention layers of the U-Net. We maintain the same hyperparameters except for changing the resolution to 512*512. We solely train the LoRA, reference attention layers and face referencenet. As $\text{FlashFace}_{\text{LoRA}}(\text{SD1.5})$ in Tab. 1, this model also attains competitive results.



Fig. 8: Visualization of LoRA version of FlashFace

We provide a visualization of LoRA versions in Fig. 8. Due to its resolution limitation of 512*512, we can observe that its ability to preserve face shape is still satisfactory. But its image quality and ability to preserve details are inferior to the full fine-tuning version.

7 Face Position Mask

As $\text{FlashFace}_{w/o-pmask}(\text{SD1.5})$ in the Tab. 1, it can be observed that the face position mask has minimal impact on performance (we fill all zero values in the additional mask channel for this ablation). This aligns with our earlier mention that it is an optional user input primarily used for user-controlled composition.

8 Additional Visualization Results

We present more visualizations of FlashFace in Fig. 9. These include transforming “Elon Musk” into artwork, altering the age of “Leonardo DiCaprio”, and inpainting “Taylor Swift” face with different prompts.

9 Limitations and Social Impact

Limitations. Our method has limitations in terms of success rate, as artifacts may appear in some images. This could be attributed to the base model SD-1.5 [32], we may try to change to the bigger model SDXL [27]. Furthermore, we

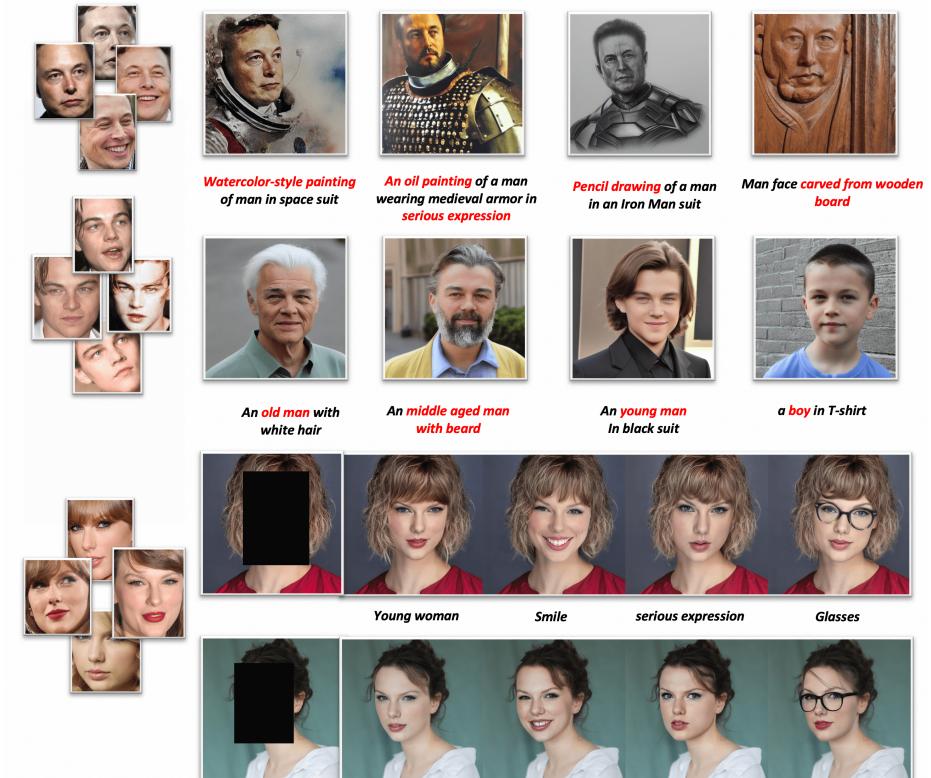


Fig. 9: More visualization of FlashFace

find that describing the pose of the head using language is challenging. In future versions, it may be beneficial to introduce improved controllability.

Social Impact. As a facial customization method for identity preservation, it can generate realistic character photos, providing an entertaining tool for the public. However, if misused, it has the potential to become a generator of false information. In the future, we should also implement checks on prompts and generated photos to minimize negative impacts.

Responsibility to Human Subjects. The collected human images in this study are sourced from publicly available images with open copyrights. Since most of them feature public figures, their portraits are relatively less sensitive. Additionally, our data algorithm is solely used for academic purposes and not for commercial use.