HiFi-Portrait: Zero-shot Identity-preserved Portrait Generation with High-fidelity Multi-face Fusion

Supplementary Material

Appendix

In this appendix, we provide the following details:

(1) A comprehensive statistics of our proposed dataset, presented in Sec. A.

(2) A detailed description of the HiFi-Net architecture, featured in Sec. B.

(3) Supplementary experiments, presented in Sec. C.

(4) Further discussions, found in Sec. D.

A. Detailed Dataset Statistics

To train our proposed HiFi-Portrait, we construct a high-quality ID-based dataset. In this section, we count identities, the number of images, identity cluster size, face area, age and gender distribution, to highlight the diversity of our dataset.

Statistics for identities and images. As illustrated in Tab. 6, the dataset consists of three parts: VGGFace2-HQ [13], IMDb-Face [50], Web. Among them, VGGFace2-HQ is the high-resolution version of VGGFace2 [10]. Its images utilize face alignment; therefore, the face area is relatively large. To improve diversity, we download the source images of IMDb-Face and collect additional data from the Internet (abbreviated as "Web"). After extensive data cleansing and processing, our final dataset consists of 34k IDs and 960k images, averaging 28.2 images per ID, underscoring the diversity of identities.

Table 6. **Statistics of identities and images of our dataset**. It consists of three parts: VGGFace2-HD [13], IMDb-Face [50], Web (data we collected from the Internet).

Dataset	Identities	Images	Images/ID
VGGFace2-HD	8k	440k	55.0
IMDb-Face	21k	380k	18.1
Web	5k	140k	28.0
All	34k	960k	28.2

Identity cluster size. Fig. 11 presents the distribution of each ID cluster size. Ince the model training requires at least 1 target image and 4 reference images, ID clusters with fewer than five images are excluded.

Face area. Fig. 12 displays the distribution of face area proportions within the dataset, where the face area is defined

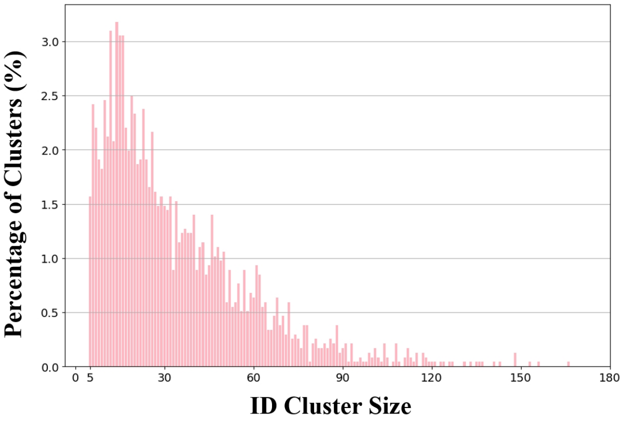


Figure 11. **Distribution of ID cluster size.**

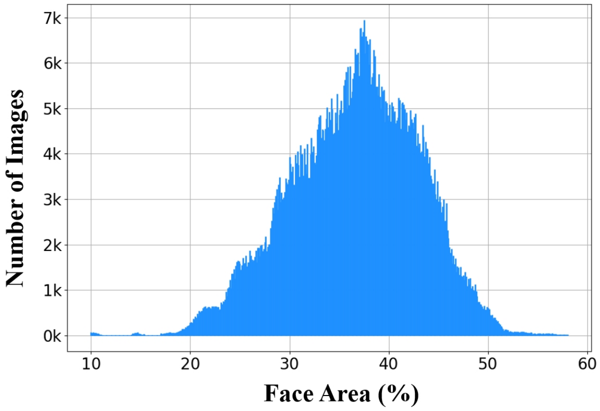


Figure 12. **Distribution of face area.**

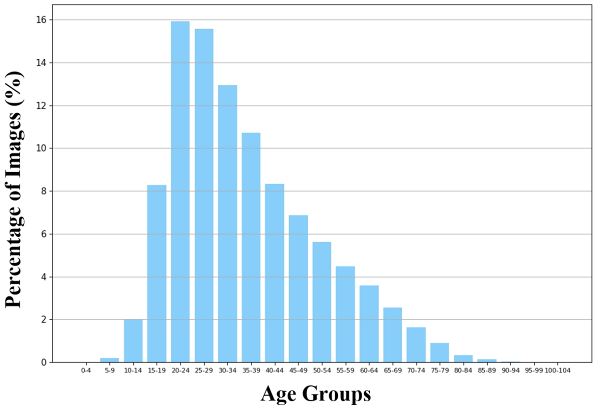


Figure 13. **Age distribution.**

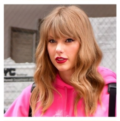
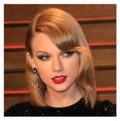
Images	Prompts
	A young woman with blonde, wavy hair, wearing black-rimmed glasses and a patterned shirt, looking calmly at the camera with a blurred indoor background.
	A woman with styled blonde hair and red lipstick, wearing an elegant, strapless black gown with intricate lace-up detailing, poses in front of a dark background with white text.
	A woman with straight blonde hair and bangs, wearing red lipstick and a white blazer over a light-colored top, stands outdoors in a brightly lit urban setting with a blurred background.
	A woman with long, wavy blonde hair and bangs, wearing a yellow and black plaid corset-style dress with a matching choker, stands in front of a dark, patterned background with a bold design.
	A woman with wavy blonde hair and bangs, wearing red lipstick and a bright pink hoodie with printed lettering, poses outdoors with a blurred urban background.
	A woman with styled blonde hair swept to the side, wearing red lipstick and statement earrings, dressed in a black glittery outfit, looking off-camera against a warm-toned, striped background.

Figure 14. **Some samples from the same ID group.**

by the bounding box area predicted by RetinaFace [16]. The spacing on the x-axis is set at 0.1. We observe that images with small faces often feature blurry and distorted faces; therefore, we filter out images where the face area is less than 10%.

Age and gender distribution. Fig. 13 and 16 outline the distribution of age groups and genders, respectively. Here, age is estimated by Facelib [15]. For gender, both IMDb-Face and Web contain names, and we can retrieve gender by name. For images where gender could not be directly retrieved, we use Facelib to obtain their gender.

Dataset sample. In Fig. 14, we present several samples from the same ID group in the dataset.

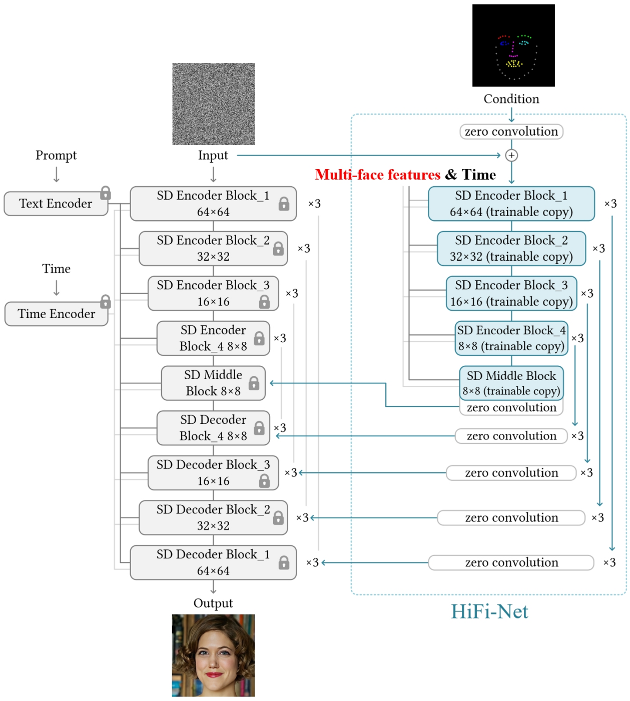


Figure 15. **The architecture of HiFi-Net.** The input condition is the face landmarks, and the input cross-attention condition is multi-face features. The landmarks guide the multi-face features to be effectively aligned and fused in HiFi-Net.

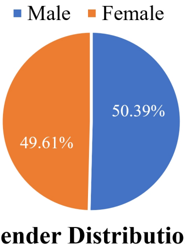


Figure 16. **Gender distribution.**

B. HiFi-Net

The architecture of HiFi-Net is depicted in Fig. 15. We made two critical modifications compared to ControlNet [63]:

1. Face landmarks are used as input conditions to guide multi-face fusion, which enables more precise control over facial expressions and poses, thereby enhancing ID fidelity as illustrated in Fig. 6 and 9.
2. The cross-attention condition is multi-face features derived from the face refiner. Multi-face features capture more facial details than local and global features, as shown in Fig. 8 and 9. In addition, compared with single-face features, multi-face features contain richer ID information, as shown in Fig. 10.

In summary, we employ HiFi-Net for effective multi-face fusion.

C. Supplementary experiments

C.1. Comparison with LoRA-based methods

Fig. 17 shows the qualitative comparison between HiFi-Portrait and LoRA-based methods [26, 34]. Our method demonstrates higher ID fidelity and precisely controls the face expression and pose.

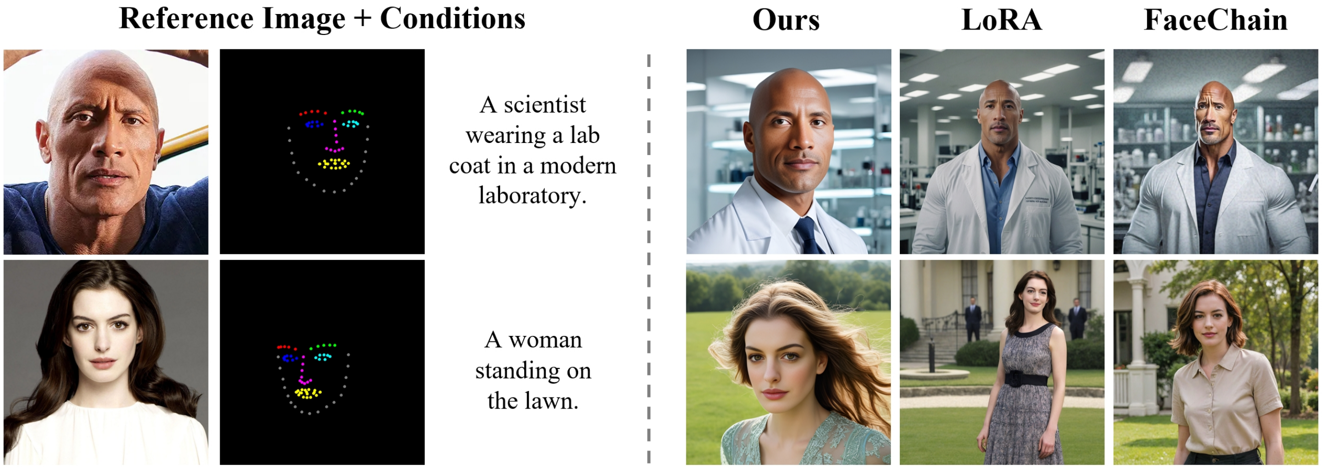

Reference Image + Conditions
Ours
LoRA
FaceChain
A scientist wearing a lab coat in a modern laboratory.
A woman standing on the lawn.

Figure 17. **Comparison with LoRA-based methods.** Our method exhibits higher ID fidelity.

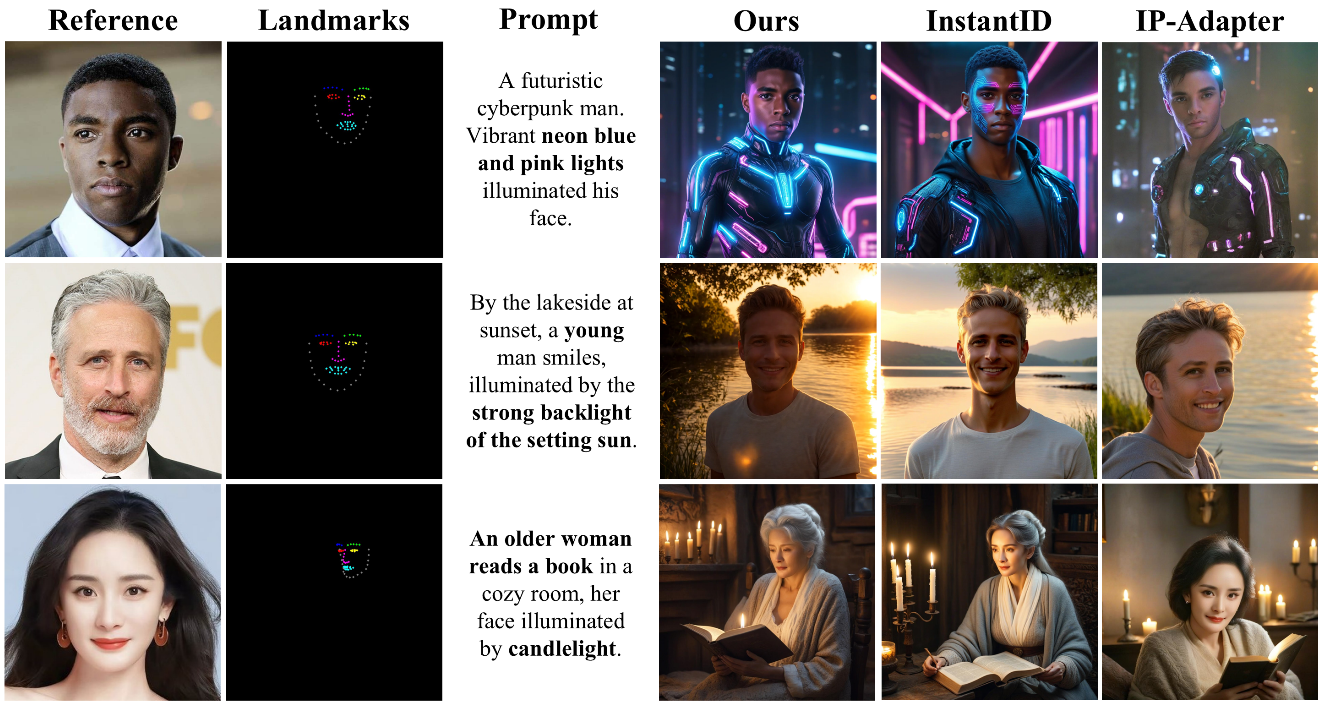

Reference
Landmarks
Prompt
Ours
InstantID
IP-Adapter
A futuristic cyberpunk man. Vibrant **neon blue and pink lights** illuminated his face.
By the lakeside at sunset, a **young** man smiles, illuminated by the **strong backlight of the setting sun**.
An older woman reads a book in a cozy room, her face illuminated by **candlelight**.

Figure 18. HiFi-Portrait generates **lighting** that is more consistent with prompt. Zoom in the image for a better visual experience.

C.2. More challenging cases

Figs. 18–20 illustrates varying lighting, facial occlusion, and extreme expressions. Fig. 18 shows that HiFi-Portrait generates text-consistent lighting from multiple angles. Fig. 19 demonstrates more natural facial occlusions. Fig. 20 confirms that HiFi-Portrait effectively remains extreme expressions.

C.3. Input condition analysis

Fig. 21 reports results under different conditions. Col. 5 without text prompts or landmarks. Col. 6 uses text only. Comparing Col. 7 (uses landmarks only) and Col. 8 (uses text and landmarks) demonstrates that using text prompt improves the alignment between target and generated expressions.

D. Further Discussion

D.1. Limitations and future works

The primary experiments in this study required approximately 1500 GPU hours on A800 GPUs. Therefore, we do not have enough computing power to explore some hyperparameter settings. Future research should focus on developing more

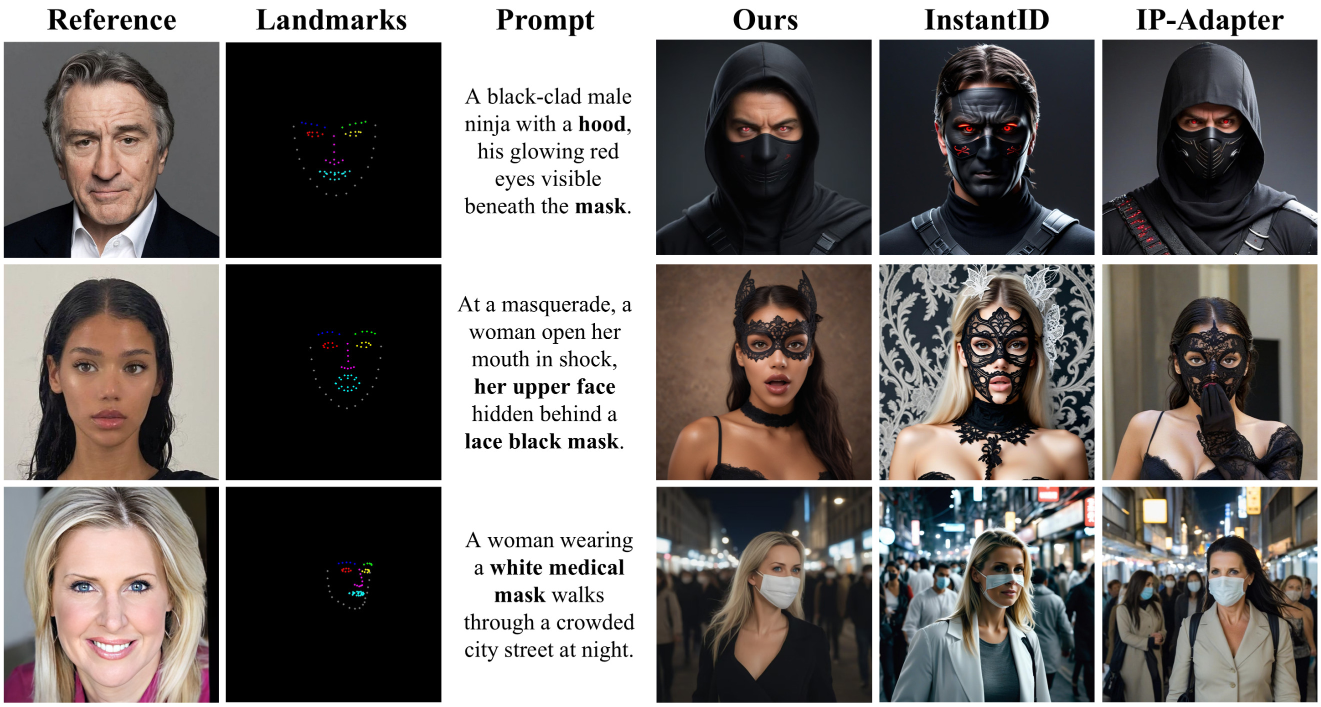

Reference
Landmarks
Prompt
Ours
InstantID
IP-Adapter
A black-clad male ninja with a hood, his glowing red eyes visible beneath the mask.
At a masquerade, a woman open her mouth in shock, her upper face hidden behind a lace black mask.
A woman wearing a white medical mask walks through a crowded city street at night.

Figure 19. Our method generates more natural **facial occlusions**.

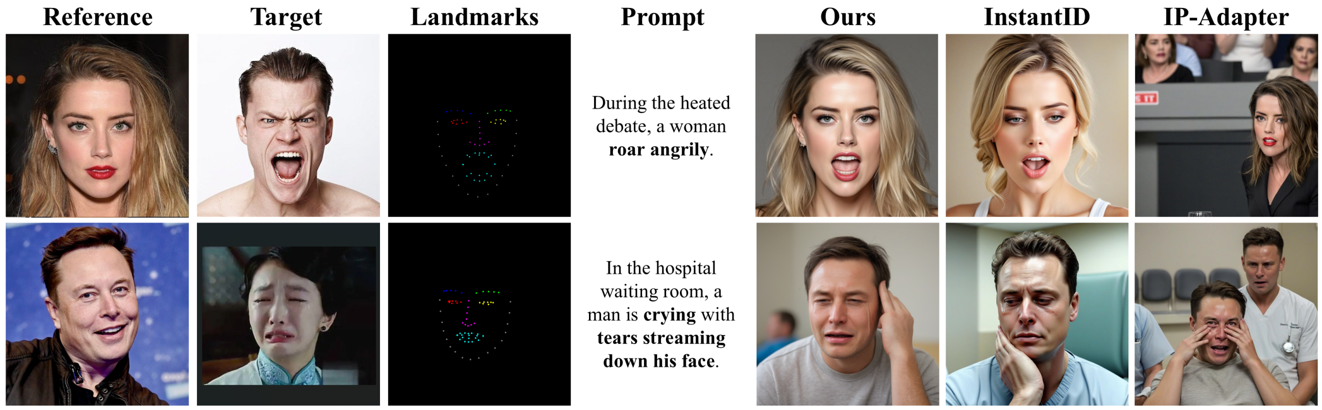

Reference
Target
Landmarks
Prompt
Ours
InstantID
IP-Adapter
During the heated debate, a woman roar angrily.
In the hospital waiting room, a man is crying with tears streaming down his face.

Figure 20. Our method effectively remains **extreme expressions**.

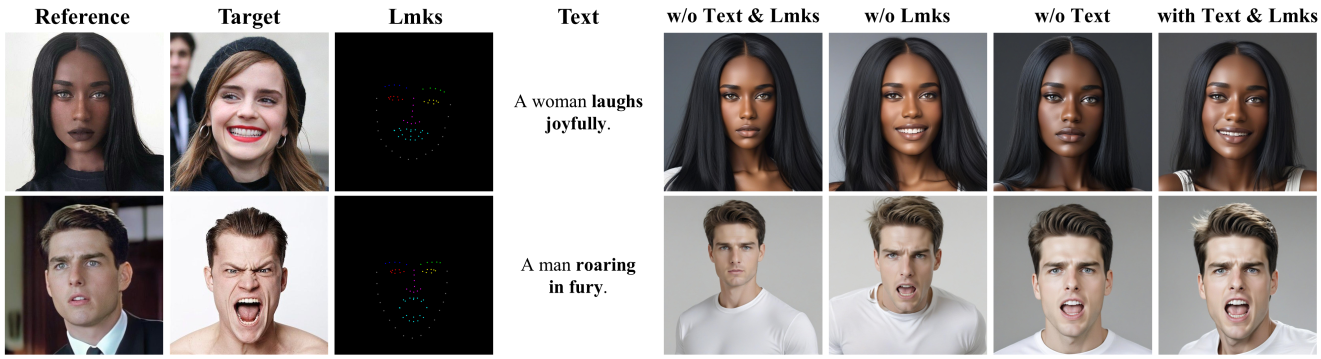


Figure 21. Using text prompt improves the alignment between the target and generated expressions.

straightforward and controlled models, rather than expending extensive efforts on hyperparameter tuning. Innovations such as a streamlined ControlNet-based method, ControlNext [37], or faster denoiser [33] could be beneficial. Additionally, experimenting with recent models like SD3 [17] or FLUX [2] may prove fruitful.

D.2. Broader impacts

This work contributes a high-fidelity framework for zero-shot ID-preserved generation to the open-source community, capable of customizing facial expressions and poses. Depending on the application context, this may have positive and negative impacts. On the one hand, our multi-face fusion strategy could advance the development of open-source ID-preservation models and use them in practical applications. On the other hand, the high fidelity of generated images could be misused for facial fraud.

D.3. Ethical considerations

To facilitate a more accurate understanding of human anatomy, SDXL utilizes a limited number of nude images for training. In our experiments, if clothing is not specified in the text prompts, there is a small probability that the generated portraits may be nude. Consequently, we recommend activating NSFW detection to minimize this issue. Additionally, the generative results have observed no other unethical or harmful behaviors. Finally, it is imperative to note that all data and models in this paper are intended strictly for research purposes and must not be used commercially.