Infinite-ID: Identity-preserved Personalization via ID-semantics Decoupling Paradigm

Yi Wu[1⋆], Ziqiang Li[1,2⋆], Heliang Zheng[1], Chaoyue Wang[3⋆⋆], and Bin Li[1⋆⋆]

[1] University of Science and Technology of China, China
[2] Nanjing University of Information Science and Technology, China
[3] The University of Sydney, Australia
wuyi2021@mail.ustc.edu.cn, iceli@mail.ustc.edu.cn,
zhenghl@mail.ustc.edu.cn,
chaoyue.wang@outlook.com, binli@ustc.edu.cn

A Supplementary Materials

A.1 Training and evaluation datasets.

Our training dataset comprises 200K images, categorized into 20K distinct identities via face recognition. For evaluation, similar to PhotoMaker, which constructs an evaluation dataset containing 25 celebrity identities, we curated a dataset of 24 diverse identities (as shown in Fig. 1): 12 celebrities and 12 ordinary individuals randomly sampled from FFHQ dataset. Evaluation scenarios feature individuals in different costumes and settings like Times Square, living rooms, and beaches, engaging in activities such as holding flowers or running, totally including 30 prompts.

Fig. 1: Evaluation dataset. The dataset contains 24 diverse identities: 12 identities collected independently and 12 identities randomly sampled from FFHQ dataset. These evaluation identities encompass various races, skin tones, appearances and genders, including celebrities and ordinary people.

A.2 Evaluation metrics

Evaluation of face similarity. To assess face similarity, we utilize the face alignment module $\mathrm{FA}(\cdot)$, the face recognition backbone $\mathrm{E}_{\mathrm{face}}(\cdot)$, and the CLIP

⋆ First two authors contributed equally to this work.
⋆⋆ Corresponding Authors.

image encoder $\mathrm{E}_{\mathrm{clip}}(\cdot)$ to compute the metrics M_{FaceNet} and CLIP-I. Specifically, for each generated image $\mathrm{I}_{\mathrm{gen}}$ and its corresponding identity image I_{id}, we first employ the $\mathrm{FA}(\cdot)$ module to detect the face. Subsequently, we calculate the pairwise identity similarity using $\mathrm{E}_{\mathrm{face}}(\cdot)$ and $\mathrm{E}_{\mathrm{clip}}(\cdot)$, respectively:

$$\begin{aligned} M_{\mathrm{FaceNet}} &= cos(\mathrm{E}_{\mathrm{face}}(\mathrm{FA}(\mathrm{I}_{\mathrm{gen}})), \mathrm{E}_{\mathrm{face}}(\mathrm{FA}(\mathrm{I}_{\mathrm{id}}))), \\ \text{CLIP-I} &= cos(\mathrm{E}_{\mathrm{clip}}(\mathrm{FA}(\mathrm{I}_{\mathrm{gen}})), \mathrm{E}_{\mathrm{clip}}(\mathrm{FA}(\mathrm{I}_{\mathrm{id}}))), \end{aligned} \tag{1}$$

where $cos(\cdot, \cdot)$ is the cosine similarity function. Furthermore, in order to visualize (as shown in Fig. 2) the quantitative comparison in Table 1 of the main paper, we integrate both M_{FaceNet} and CLIP-I by utilizing z-score normalization:

$$mean(\text{z-score}(M_{\mathrm{FaceNet}}), \text{z-score}(\text{CLIP-I})), \tag{2}$$

where $\text{z-score}(x) = (x - \mu)/\sigma$, μ and σ are the average and standard deviation of the x, respectively.

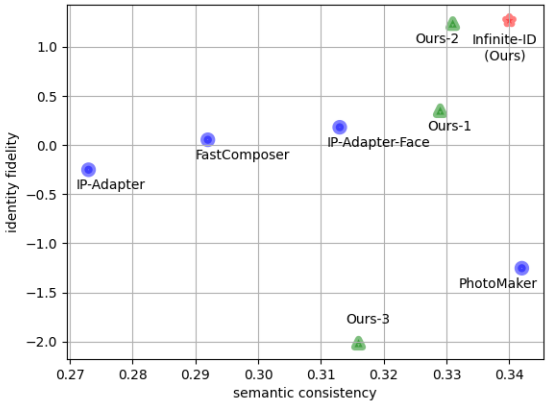

Fig. 2: Visualization of the quantitative comparison in Table 1 of the main paper. The compared methods including IP-Adapter, IP-Adapter-Face, FastComposer, PhotoMaker, and ablation versions of our method including w/o identity-enhanced training (Ours-1), w/o mixed attention (Ours-2) and mixed attention ⇒ mutual attention (Ours-3).

Definition of semantic consistency. We adopt the CLIP-T metric to assess semantic consistency. Specifically, for a generated image I_{gen} paired with its corresponding prompt P, we compute the CLIP-T metric utilizing both the CLIP image encoder E_{clip} and the CLIP text encoder E_{text}:

$$\text{CLIP-T} = cos(E_{\mathrm{clip}}(I_{gen}), E_{\mathrm{text}}(P)), \tag{3}$$

where the $cos(\cdot, \cdot)$ is the cosine similarity function.

Table 1: Quantitative ablation of cross-attention merge. The metrics includes CLIP-T (higher is better) measuring the semantic consistency, CLIP-I (higher is better) and M_{FaceNet} (higher is better) which are both reflect the identity fidelity. The best result is shown in **bold**.

	CLIP-T↑	CLIP-I↑	M_{FaceNet}↑
Ours w/o Cross-attention merge	0.335	0.910	0.681
Ours	**0.340**	**0.913**	**0.689**

Fig. 3: Quantitative ablation of cross-attention merge. It is obvious that the cross-attention merge helps to improve the semantic consistency.

A.3 More Results on Ablation Study

Ablation Study of Cross-attention Merge. We conduct ablation experiments on the cross-attention merge to evaluate its effectiveness. As depicted in Fig. 3 and Table 1, the incorporation of cross-attention merge demonstrates improvement in semantic consistency.

Ablation Study of Input ID Images' Resolution. We perform an ablation study on the resolution of input ID images to assess the robustness of our method. Specifically, we utilize images with varying resolutions while maintaining the same text prompt for personalization. As illustrated in Fig. 4, the identity fidelity exhibits only a marginal decrease with decreasing image resolution, while semantic consistency remains stable across all resolutions. In conclusion, our method demonstrates robustness to changes in input image resolution.

A.4 Comparison with DreamBooth and InstantID.

We compare our method with DreamBooth, representing optimization-based methods and tuning-free methods representing InstantID, as shown in Fig. 5 and Table 2. In our comparisons for raw and style photo generations, DreamBooth excels in semantic consistency but struggles with identity fidelity and stylization. It often requires multiple training images to prevent training collapse, which can occur with a single ID image input. To address this, we use early-stop and train the LoRA weights for 300 iterations in DreamBooth. InstantID excels in identity fidelity but lacks semantic consistency due to overfitting to reference images

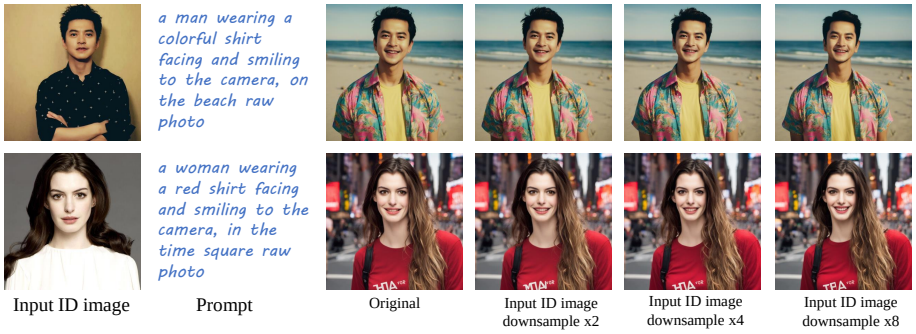

Fig. 4: Ablation study of input ID images' resolution. The identity fidelity slightly drops along with the lower image resolution and the semantic consistency is stable for all the resolution. Our method is robust to the resolution of input ID image.

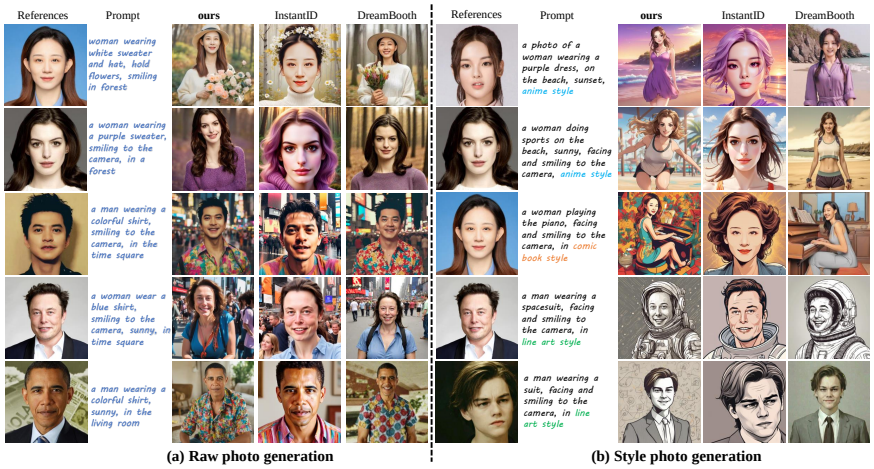

Fig. 5: Comparison with DreamBooth and InstantID on raw photo generation and stlye photo generation. Our method consistently maintains identity fidelity and achieves high-quality semantic consistency with just a single input image in raw photo generation. Our method also achieves precise stylization.

from IdentityNet, and exhibits unrealistic style in raw photo generation. While combining ControlNet and additional pose image inputs can improve semantic consistency for InstantID, this approach is not aligned with our study's use of only one reference image. In contrast, our method achieves high identity fidelity, semantic consistency, and precise stylization. Moreover, we also conduct the user study among different methods in the identity fidelity, image quality, text fidelity (semantic consistency) and stylization, as shown in Table 3.

Table 2: Quantitative comparison. We utilize CLIP-T metric to reflect the semantic consistency. The CLIP-I and M_{FaceNet} reflect the identity fidelity. FID score and LPIPS score are employed to reflect the quality and diversity of the generated faces. We also compare the speed of different methods.

	CLIP-T (↑)	CLIP-I (↑)	M_{FaceNet} (↑)	FID (↓)	LPIPS (↓)	Speed (↓)
PhotoMaker	<u>0.34</u>	0.81	0.50	<u>266.1</u>	<u>0.664</u>	**≈10s**
InstantID	0.30	<u>0.85</u>	**0.70**	269.1	0.725	<u>≈20s</u>
DreamBooth	**0.38**	0.80	0.48	268.5	0.677	≈360s
Ours	<u>0.34</u>	**0.91**	<u>0.68</u>	**216.6**	**0.637**	<u>≈20s</u>

Table 3: User study. Users preferences on Identity fidelity (identity), image quality (quality), text fidelity (semantics) and stylization (style). The best result is shown in **bold**, and the second best is <u>underlined</u>.

(a) Raw photo generation

	Identity	Quality	Semantics
PhotoMaker	12.5	<u>31.2</u>	<u>25.0</u>
InstantID	**41.2**	25.0	6.2
DreamBooth	7.5	0.0	22.5
Ours	<u>38.8</u>	**43.8**	**46.3**

(b) Style photo generation

Identity	Quality	Semantics	Style
27.5	25.0	<u>32.5</u>	<u>37.5</u>
36.2	<u>27.5</u>	7.5	11.2
3.8	2.5	12.5	2.5
<u>32.5</u>	**45.0**	**47.5**	**48.8**

A.5 Evaluation of the diversity and quality of the generated faces.

We utilize the FID score and LPIPS score to assess the quality and diversity of the generated faces, respectively. As shown in Table 2, our Infinite-ID surpasses other methods in both the diversity and quality of the generated faces.

A.6 Inference time and storage space.

Compared to the original text-to-image pipeline, our method requires more storage space and inference time. Notably, Infinite-ID does not necessitate two U-Net models. Instead, we integrate an additional cross-attention module, consisting of two single-layer fully connected layers, into the original U-Net. This allows Infinite-ID to toggle between handling text and identity information by activating or deactivating the image cross-attention. Moreover, our method is comparable to other personalization methods in terms of storage space and inference time (as shown in Table 2).

A.7 Identity mixing

Upon receiving multiple images from distinct individuals, we stack all the identity embeddings to merge corresponding identities, as depicted in Fig. 6. The

generated image can well retain the characteristics of different IDs, which releases possibilities for more applications. Additionally, by adjusting the interpolation of the identity embeddings, we can regulate the similarity between the generated identity and different input identities, as demonstrated in Fig. 7.

A.8 More Qualitative Results of Raw Photo Generation

Fig. 9 demonstrates the ability of our method to extract identity information from artworks while preserving identity for personalization purposes. Additionally, Fig. 8 illustrates the capability of our method to alter attributes of the extracted identities for raw photo generation. Additional visual samples for raw photo generation are provided in Fig. 10 and Fig. 11, showcasing identities of ordinary individuals sampled from the FFHQ dataset, spanning diverse races, skin tones, and genders.

A.9 More Qualitative Results of Style Photo Generation

Fig. 12 and Fig. 13 display the results of style photo generation. The identity samples consist of ordinary individuals randomly selected from the FFHQ dataset. A total of 12 stylization styles are employed, affirming the generalizability of our method.

Fig. 6: Identity mixing. When receiving multiple input ID images from different individuals, our method can mix these identities by stacking all the identity embeddings.

Fig. 7: Linear interpolation of different identities.

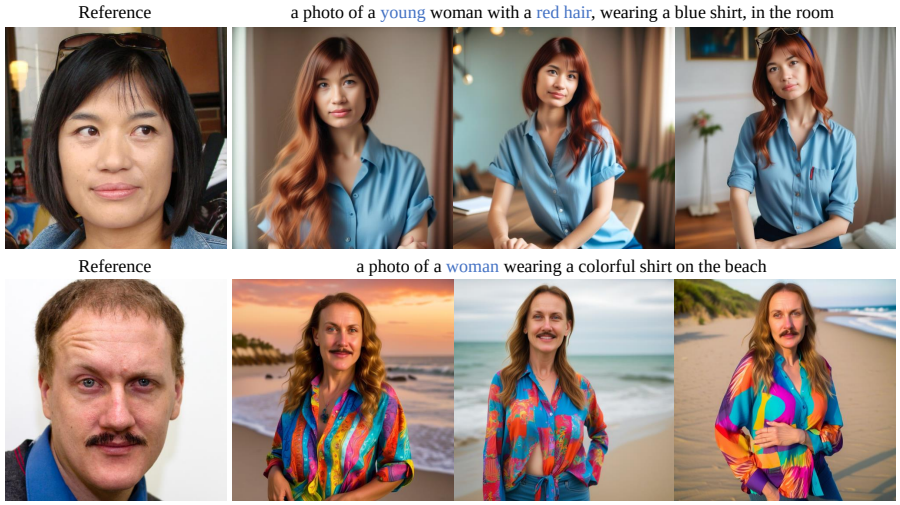

Fig. 8: Applications on attribute change.

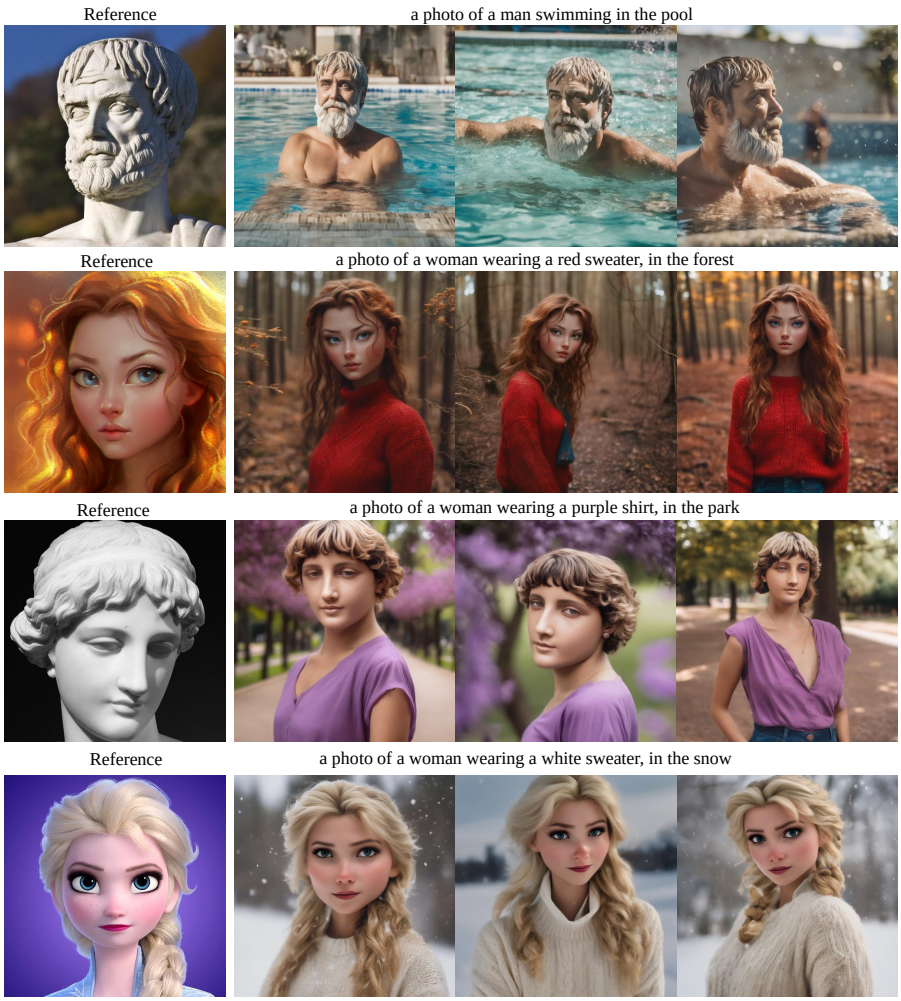

Fig. 9: Applications on artworks to raw photo.

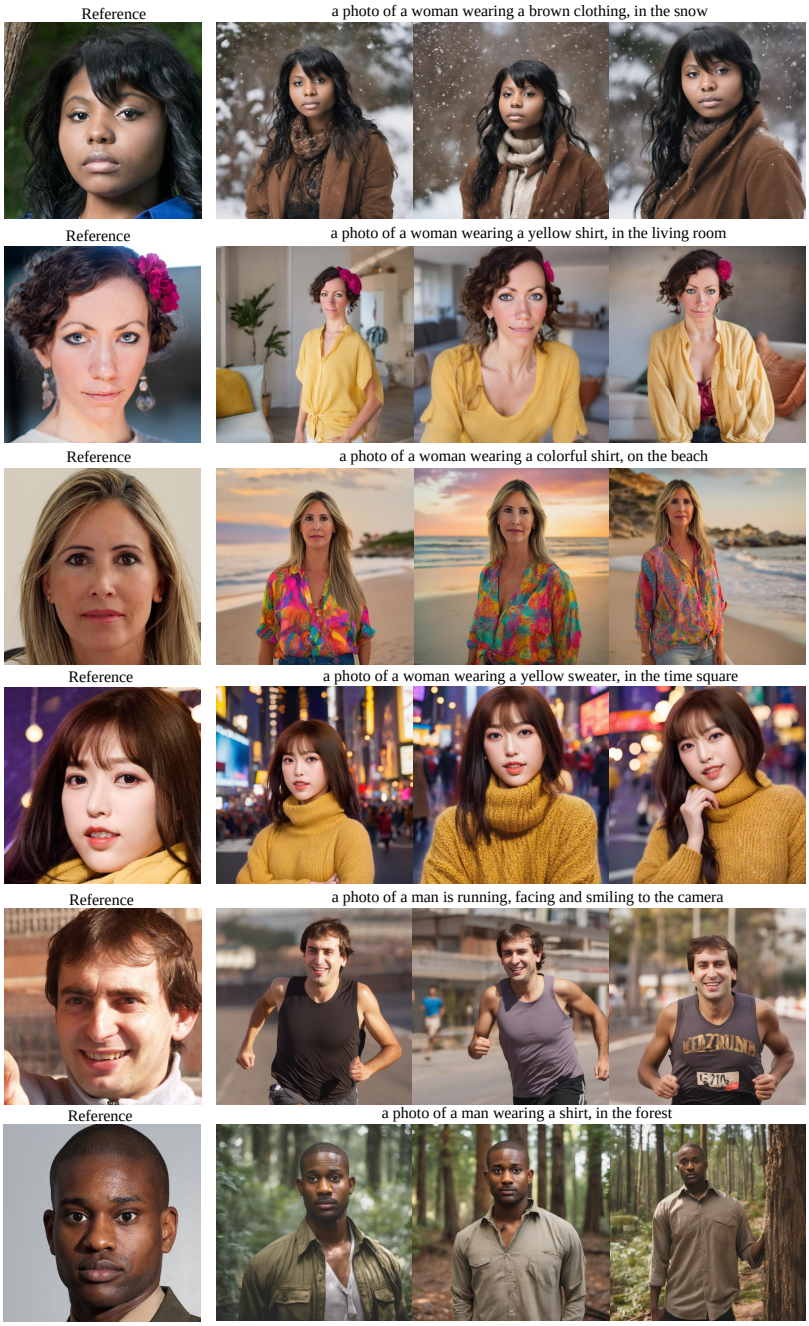

Fig. 10: Raw photo generation. These identities are ordinary people sampled from FFHQ dataset, including various races, skin colors, male and female.

Fig. 11: Raw photo generation. These identities are ordinary people sampled from FFHQ dataset, including various races, skin colors, male and female.

Fig. 12: More visual examples for stylization. These identities are ordinary people sampled from FFHQ dataset, including various races, skin colors, male and female.

Fig. 13: More visual examples for stylization. These identities are ordinary people sampled from FFHQ dataset, including various races, skin colors, male and female.