IDAdapter: Learning Mixed Features for Tuning-Free Personalization of Text-to-Image Models

Supplementary Material

1. Implementation Details

Adapter Layer In our proposed approach, the adapter layer involves linear mapping of the MFF vision embedding, a gated self-attention mechanism, a feedforward neural network, and normalization before the attention mechanism and the feedforward network.

Model Details The model in this work refers to the trainable structures, including the MFF module, a multi-layer perceptron, key and value projection matrices in each cross-attention block. The total model size is 262M, which is smaller than Subject Diffusion [2] (700M) and Dreambooth [3] (983M). We set the sampling step as 50 for inference. Our method is tuning-free during testing, enabling the synthesis of 5 images within half a minute.

2. Subject Personalization Results

Our method achieve very effective editability, with semantic transformations of face identities into high different domains, and we conserve the strong style prior of the base model which allows for a wide variety of style generations. We show results in the following domains. The images for visualization is from SFHQ dataset [1] and we use the unique facial image for each identity in the dataset as a reference to generate multiple images.

Age Altering We are able to generate novel faces of a person with different appearance of different age as Figure 1 shows, by including an age noun in the prompt sentence:"*[class noun] is a [age noun]*". We can see in the example that the characteristics of the man is well preserved.

Recontextualization We can generate novel images for a specific person in different contexts (Figure 2) with descriptive prompts ("*a [class noun] [context description]*"). Importantly, we are able to generate the person in new expressions and poses, with previously unseen scene structure and realistic integration of the person in the scene.

Expression Manipulation Our method allows for new image generation of a person with modified expressions that are not seen in the original input image by prompts "*[class noun] is [expression adjective]*". We show examples in Figure 3.

Art Renditions Given a prompt "*[art form] of [class noun]*", we are able to generate artistic renditions of the person. We show examples in Figure 4. We select similar viewpoints for effect, but we can generate different angles of the woman with different expressions actually.

Accessorization We utilize the capability of the base model to accessorize subject persons. In Figure 5, we show examples of accessorization of a man. We prompt the model with a sentence:"*[class noun] in [accessory]*" to fit different accessories onto the man with aesthetically pleasing results.

View Synthesis We show several viewpoints for facial view synthesis in Figure 6, using prompts as "*[class noun] [viewpoint]*" in the figure.

Property Modification We are able to modify facial properties. For example, we show a different body type, hair color and complexion in Figure 7. We prompt the model with the sentences "*[class noun] is/has [property description]*". In particular, we can see that the identity of the face is well preserved.

Lighting control Our personalization results exhibit natural variation in lighting and we can also control the lighting condition by prompts like "*[class noun] in [lighting condition]*", which may not appear in the reference images. We show examples in Figure 8.

Body Generation Our model has the ability to infer the body of the subject person from facial features and can generate specific poses and articulations in different contexts based on the prompts combined with "*full/upper body shot*" as Figure 9 shows. In essence, we seek to leverage the model's prior of the human class and entangle it with the embedding of the unique identifier.

References

[1] David Beniaguev. Synthetic faces high quality (sfhq) dataset. https://github.com/SelfishGene/SFHQ-dataset, 2022. 1

[2] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion:open domain personalized text-to-image generation without test-time fine-tuning, 2023. 1

[3] 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. 2022. 1

Figure 1. **Age altering.** We present photos of the same person at different age stages by prompting our generative model.

Figure 2. **Recontextualizaion.** We generate images of the subject person in different environments, with high preservation of facial details and realistic scene-subject interactions.

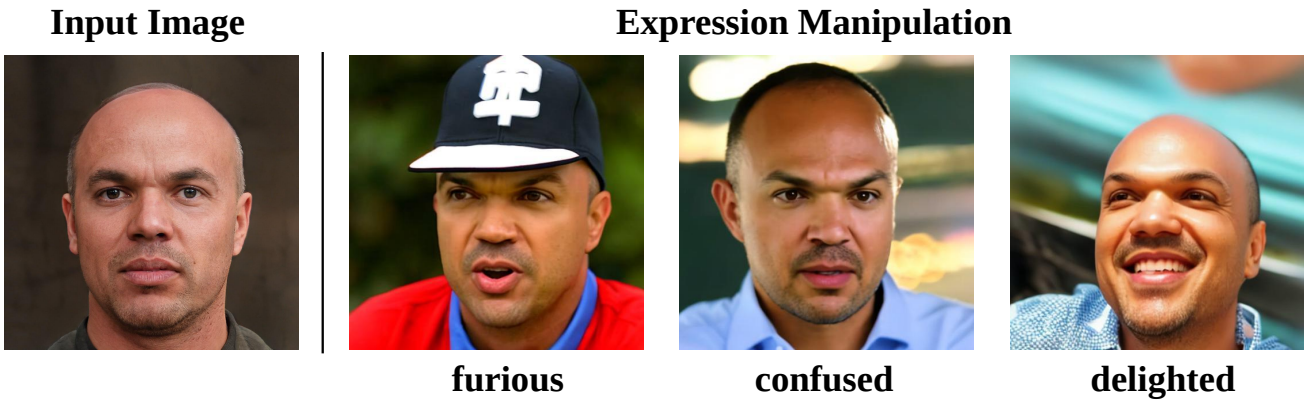

Figure 3. **Expression manipulation.** Our method can generate a range of expressions not present in the input images, showcasing the model's inference capabilities.

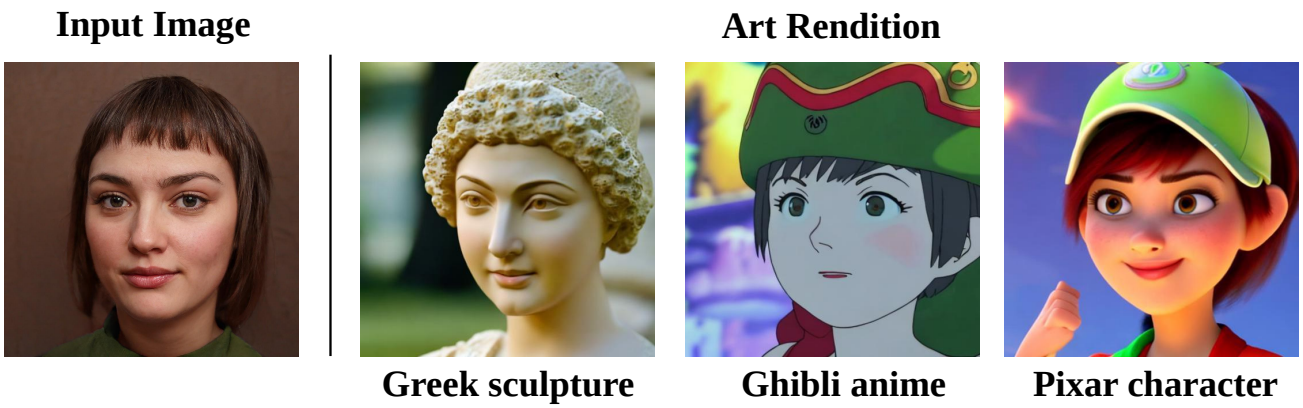

Figure 4. **Artistic renderings.** We can observe significant changes in the appearance of the character to blend facial features with the target artistic style.

Figure 5. **Outfitting a man with accessories.** The identity of the subject person is preserved and different outfits or accessories can be applied to the man given a prompt of type "*[class noun] in [accessory]*". We observe a realistic interaction between the subject man and the outfits or accessories.

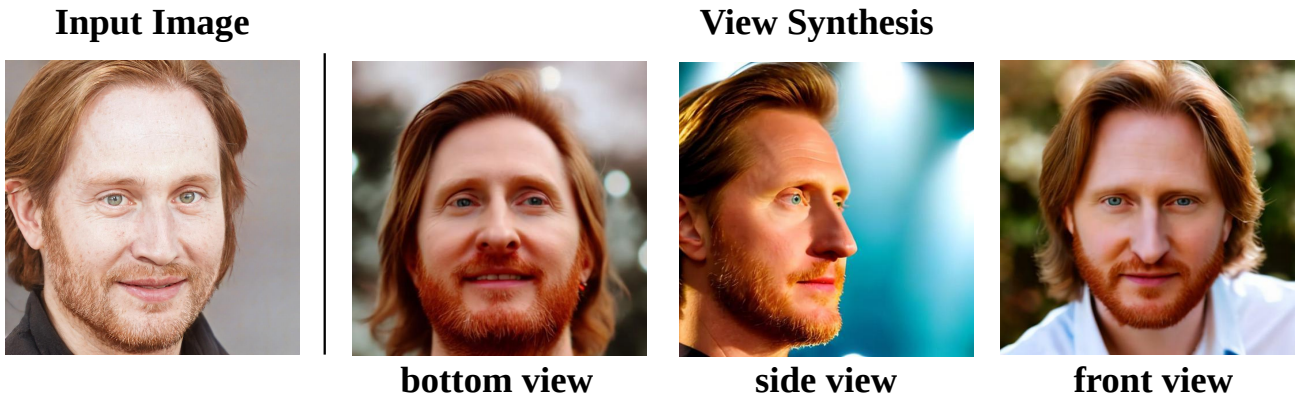

Figure 6. **View Synthesis.** Our technique can synthesize images with specified viewpoints for a subject person.

Figure 7. **Modification of subject properties.** We show modifications in the body type, hair color and complexion (using prompts "*[class noun] is/has [property description]*").

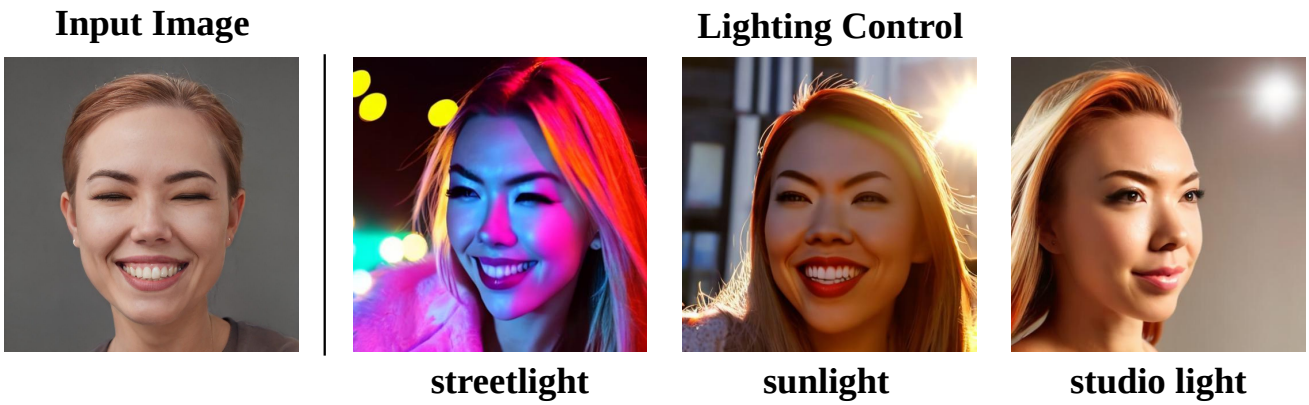

Figure 8. **Lighting control.** Our method can generate lifelike subject photos under different lighting conditions, while maintaining the integrity to the subject's key facial characteristics.

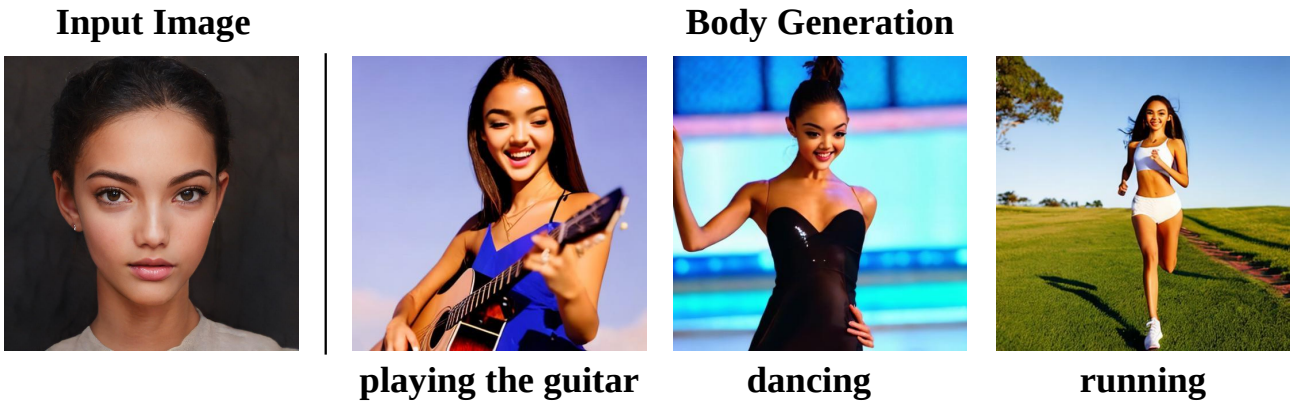

Figure 9. **Body generation.** We are able to generate the body of the subject person in novel poses and articulations with only a facial image.