CapHuman: Capture Your Moments in Parallel Universes

Supplementary Material

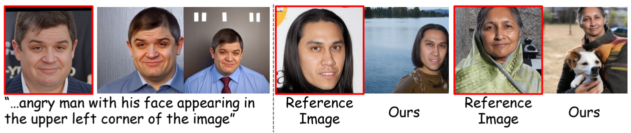
“...angry man with his face appearing in the upper left corner of the image”
Reference Image
Ours
Reference Image
Ours

Figure I. **Left:** The detailed prompt struggles to control the head position and facial expression. **Right:** Maintain the hairstyle.

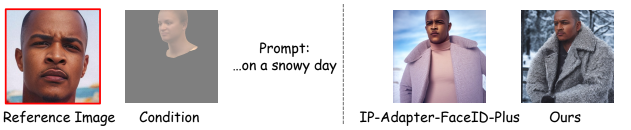

Figure II. **Visual comparison with IP-Adapter.** Our method shows better ID preservation and head control while following the given prompt.

A. More Qualitative Results

Can the detailed prompt achieve the head control as well? In Figure I (Left), we show the detailed prompt still struggles to control the human head, *e.g*. position, and facial expression.

Maintain the hairstyle. In Figure I (Right), we show that our model can keep the hairstyle via minor modifications, that is, keep the hair area in the ID features and masks.

Visual comparison with IP-Adapter. As shown in Figure II, we compare our method with IP-Adapter [15]. Our method shows better ID preservation and head control while following the given prompt.

Visual comparisons. We show more visual comparisons with the established baselines [9–12, 14] in Figure IV. Our CapHuman can generate well-identity-preserved, photo-realistic, and high-fidelity portraits with various head positions and poses in different contexts.

Facial expression control. In Figure V, we provide more examples, demonstrating the facial expression control ability of our CapHuman.

B. More Quantitative Results

More reference images. We compare our method with fine-tuning methods that take more reference images as input. The results are presented in Table I. Our method still outperforms LoRA [10] and DreamBooth [12] with better identity preservation and shorter personalization time.

Method	#ref.	↑ ID sim.	↓ Personalization time (s)
LoRA [10]	5	0.6298	1223
DreamBooth [12]	5	0.7457	1321
Ours	1	**0.8429**	**7**

Table I. **Comparisons with fine-tuning methods with more reference images.** Ours still outperforms other baselines with higher identity similarity and faster speed.

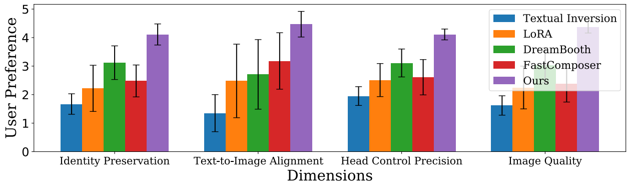

Figure III. **User Study.** Users prefer our method in all four dimensions: identity preservation, text-to-image alignment, head control precision, and image quality.

Method	↑ ID sim.	↑ CLIP score	↑ Prompt acc.	↓ Shape	↓ Pose	↓ Exp.	↓ Light.
IP-Adapter-FaceID-Plus [15]	0.8125	0.2056	61.01%	0.1293	0.0641	0.1519	0.1447
Ours	**0.8363**	**0.2256**	**74.17%**	**0.1020**	**0.0436**	**0.1241**	**0.0965**

Table II. **Comparison with IP-Adapter.** Our method outperforms IP-Adapter in all aspects.

User Study. We invite 50 users to score 20 groups of results from each method in terms of the following four dimensions: identity preservation, text-to-image alignment, head control precision, and image quality. Figure III shows our method is much more preferred by the users.

Comparison with IP-Adapter. We compare our method with IP-Adapter [15]. The results are presented in Table II. Our method outperforms IP-Adapter [15] in all aspects.

Ablation on the global ID feature. For the choice of the global ID feature, we compare FaceNet [13] and ArcFace [7]. FaceNet outperforms ArcFace. The ID similarity is 0.8367 (0.8091) measured by FaceNet and 0.4819 (0.4737) measured by ArcFace.

C. More Applications

Stylization by adaptation to other pre-trained models. Benefitting from the nature of open-source in the community, we can inherit the rich pre-trained models. Our CapHuman can be adapted to other pre-trained models [1–4] in the community flexibly, which can generate identity-preserved portraits with various head positions, poses, and facial expressions in different styles. More results are presented in Figure VI, VII, VIII, and IX.

Stylization by style prompts. We also showcase portraits with different styles driven by style prompts in Figure X.

Multi-Human image generation. Our CapHuman supports multi-human image generation. The generated results are presented in Figure XI.

Simultaneous head and body control. Combined with the pose-guided ControlNet [16], our CapHuman can control the head and the body simultaneously with identity preservation. More results are presented in Figure XII.

Photo ID generation. Photo ID is widely used in passports, ID cards, etc. There are typically some requirements for these photos, *e.g.* plain background, formal wearing, and standard head pose. As shown in Figure XIII, our CapHuman can generate standard ID photos by adjusting the head conditions and providing the proper prompts conveniently.

D. HumanIPHC Benchmark Details

We introduce more details about our HumanIPHC benchmark in this section.

ID split. 100 IDs used in our benchmark are listed in Table III.

Prompts. We list the prompts used in the benchmark:

- a photo of a person.
- a photo of a person with red hair.
- a photo of a person standing in front of a lake.
- a photo of a person holding a dog.
- a photo of a person running on a rainy day.
- a closeup of a person playing the guitar.
- a photo of a person wearing a suit on a snowy day.
- a photo of a person playing basketball.
- a photo of a person wearing a scarf.
- a photo of a person on a cobblestone street.
- a photo of a person with a sheep in the background.
- a photo of a person sitting on a purple rug in a forest.
- a photo of a person with a tree and autumn leaves in the background.
- a photo of a person with the Eiffel Tower in the background.
- a photo of a person wearing a red sweater.
- a photo of a person wearing a spacesuit.
- a photo of a person wearing a green coat.
- a photo of a person wearing a blue hoodie.
- a photo of a person wearing a santa hat.
- a photo of a person wearing a yellow shirt.
- a photo of a person with a city in the background.
- a photo of a person with a mountain in the background.
- a photo of a person on the beach.
- a photo of a person in the jungle.
- a photo of a person riding a horse.
- a photo of a person holding a bottle of a red wine.
- a photo of a person swimming in the pool.
- a photo of a person holding flowers.
- a photo of a person with a cat.
- a photo of a person reading a book.
- a photo of a person in a chief outfit.
- a photo of a person in a police outfit.
- a photo of a person in a firefighter outfit.
- a photo of a person in a purple wizard outfit.
- a photo of a person wearing a necklace.

Head conditions. In Figure XIV, we show the head conditions of a specific individual in our benchmark, including Surface Normals, Albedos, and Lambertian renderings.

E. User Study Details

We asked the participants to fill out the questionnaires. Every participant is required to score for each question. The score ranges from 1 to 5. The questions are listed as follows:

- Given the reference image and generated image, score for the identity similarity. (1: pretty dissimilar, 5: pretty similar).
- Given the text prompt and generated image, score for the text-to-image alignment. (1: the image is pretty inconsistent with the text prompt, 5: the image is pretty consistent with the text prompt).
- Given the reference image, head condition, and generated image, score for the head control precision from the view of the shape, pose, position, lighting, and facial expression. (1: pretty bad, 5: pretty good).
- Given the generated image, score for the image quality. (1: pretty far away from the real image, 5: pretty close to the real image).

F. Limitations and Social Impact

Limitations. Although our proposed method can achieve promising generative results, it still has several limitations. Our basic generative capabilities come from the pre-trained model, suggesting that our model might fail to generate the scenario out of the pre-training distribution. On the other hand, our 3D facial representation reconstruction relies on the estimation accuracy of DECA [8]. We find it struggles for some extreme poses and facial expressions. This can cause the misalignment of our generated images and the expected head conditions in some cases. Besides, the text richness is limited in our training data. It might be the reason that the text-to-image alignment performance degrades after training. Utilizing permissioned internet data might help alleviate this issue. We leave it for future research.

Social Impact. Generative AI has drawn exceptional attention in recent years. Our research aims to provide an effective tool for human-centric image synthesis, especially for portrait personalization with head control in a flexible, fine-grained, and 3D-consistent manner. We believe it will play an important role in many potential entertainment applications. Like other existing generative methods, our method is susceptible to the bias from the large pre-trained dataset as well. Some malicious parties might have the potential to exploit this vulnerability for bad purposes. We encourage future research to address this concern. Besides, our model is at risk of abuse, *e.g*. synthesizing politically relevant images. This risk can be mitigated by some deepfake detection methods [5, 6] or by controlling the release of the model strictly.

References

[1] Realistic vision v3.0. https://huggingface.co/SG161222/Realistic_Vision_V3.0_VAE, 2023. 1, 4

[2] comic-babes. https://civitai.com/models/20294/comic-babes, 2023.

[3] disney-pixar-cartoon. https://civitai.com/models/65203/disney-pixar-cartoon-type-a, 2023.

[4] toonyou. https://civitai.com/models/30240/toonyou, 2023. 1

[5] Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. Interpretable-through-prototypes deepfake detection for diffusion models. In *ICCV*, pages 467–474, 2023. 3

[6] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, pages 1–5, 2023. 3

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 1

[8] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 40(8), 2021. 2

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim 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

[10] 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. *arXiv preprint arXiv:2106.09685*, 2021. 1

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[12] 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*, pages 22500–22510, 2023. 1

[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 1

[14] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv*, 2023. 1

[15] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 1

[16] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 12

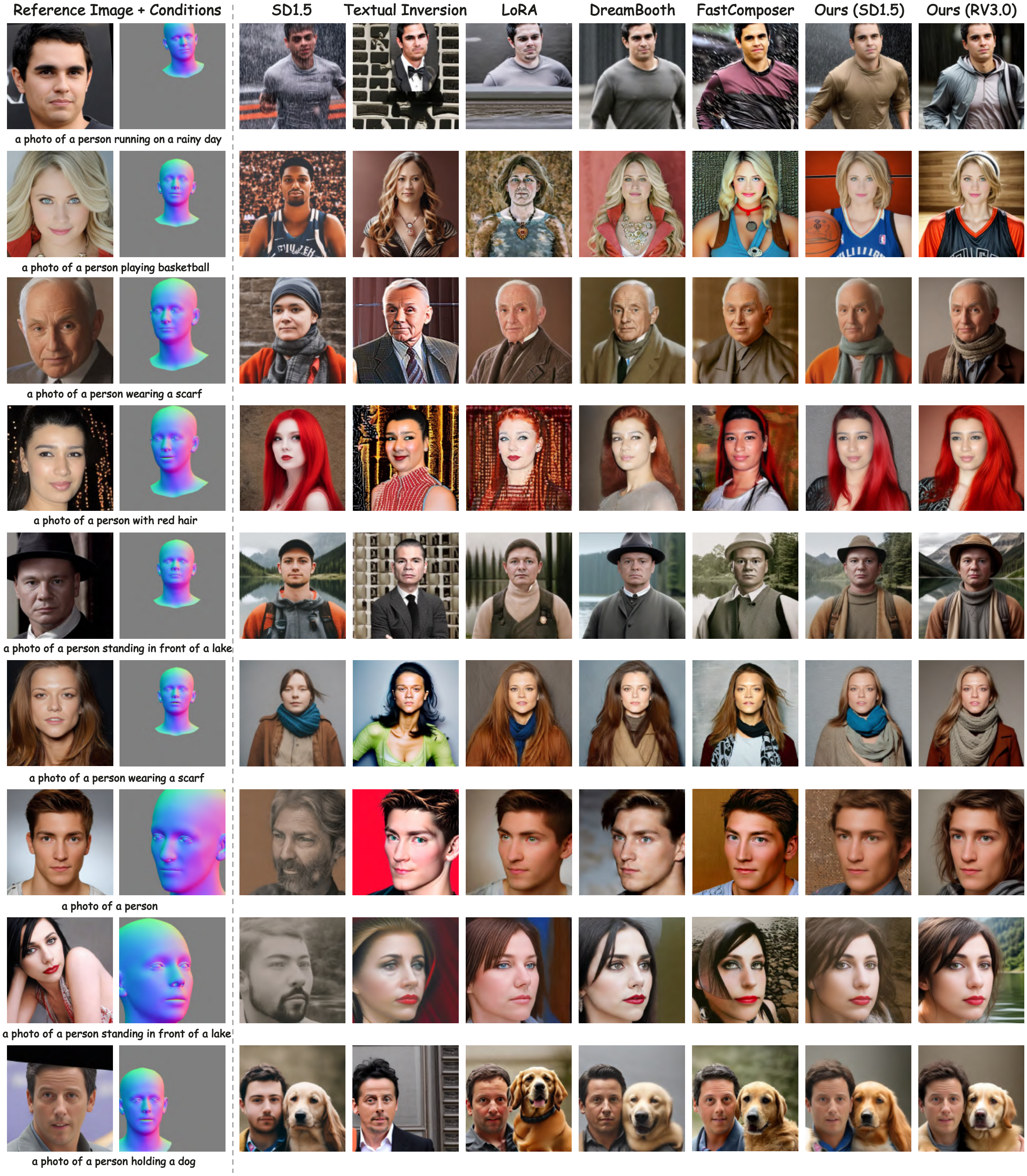
Reference Image + Conditions
SD1.5
Textual Inversion
LoRA
DreamBooth
FastComposer
Ours (SD1.5)
Ours (RV3.0)
a photo of a person running on a rainy day
a photo of a person playing basketball
a photo of a person wearing a scarf
a photo of a person with red hair
a photo of a person standing in front of a lake
a photo of a person wearing a scarf
a photo of a person
a photo of a person standing in front of a lake
a photo of a person holding a dog

Figure IV. **More qualitative results.** Our CapHuman can produce well-identity-preserved, photo-realistic, and high-fidelity portraits with various head positions and poses in different contexts, compared with the baselines. Note that our model can be combined with other pre-trained models, *e.g*. RealisticVision [1] in the community flexibly. For the head condition, we only display the Surface Normal here.

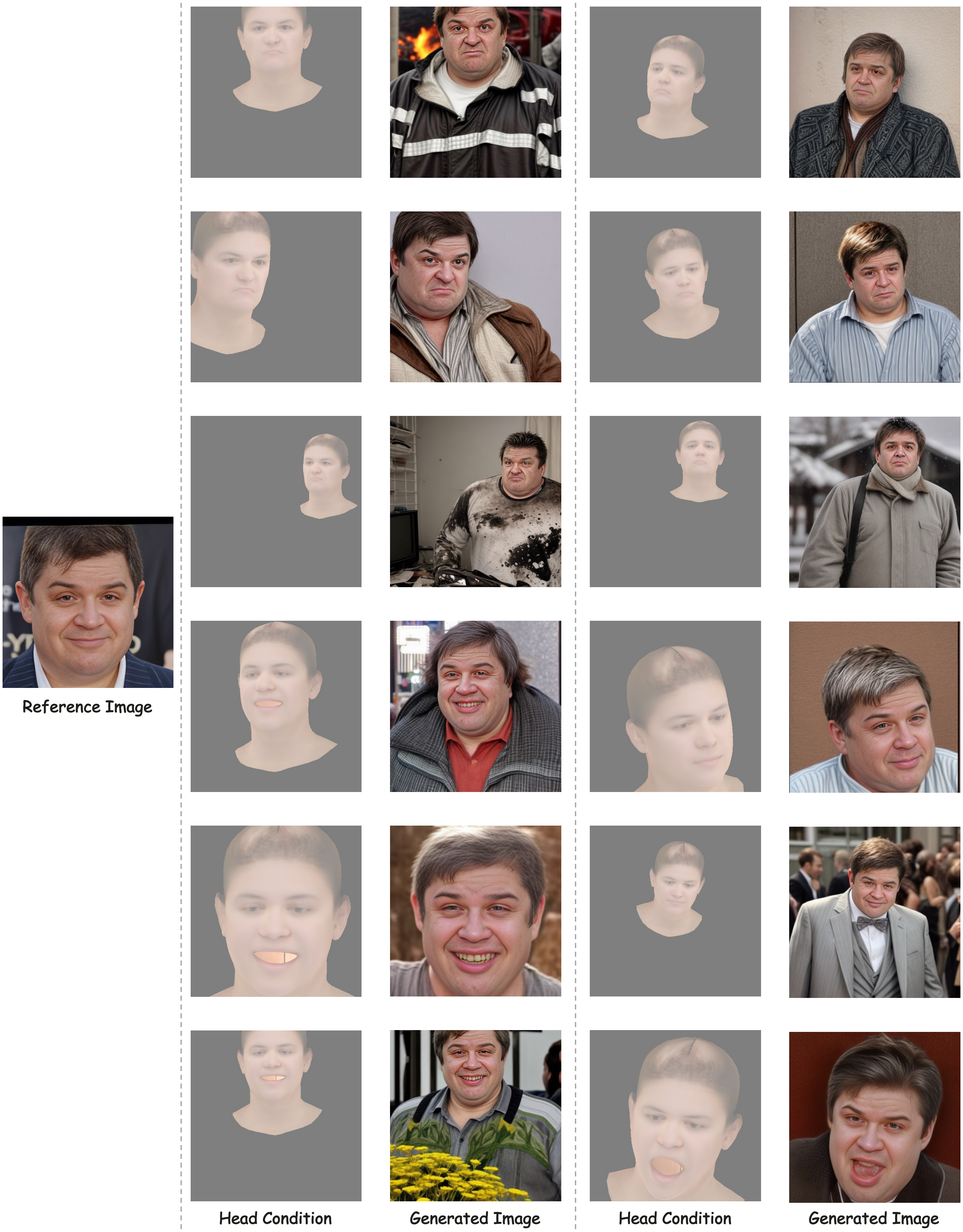

Figure V. **More results with different and rich facial expressions.** Our CapHuman can provide facial expression control in a flexible and fine-grained manner.

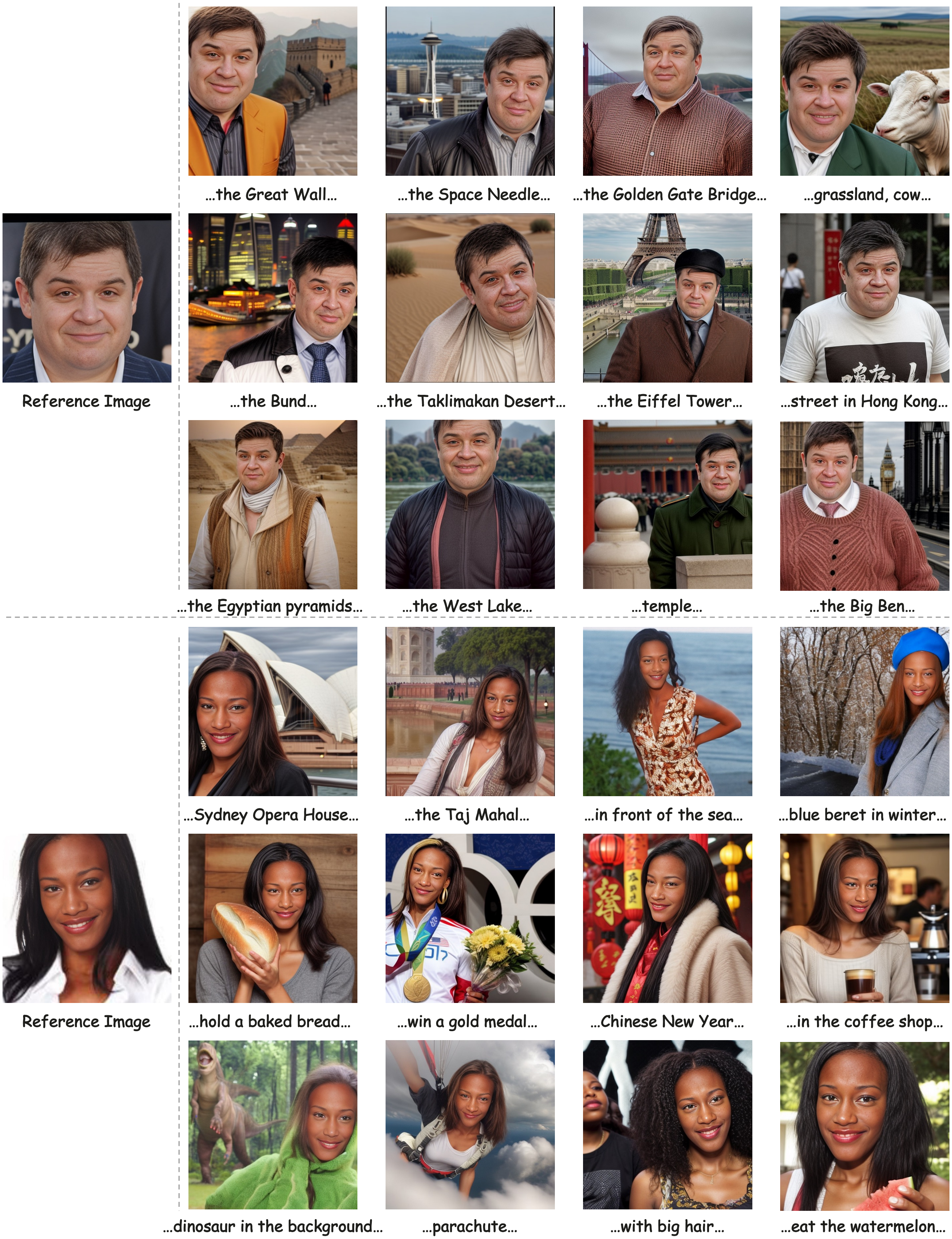

Figure VI. **More results in the realistic style.** Our CapHuman can be adapted to produce various identity-preserved and photo-realistic portraits with diverse head positions, poses, and facial expressions.

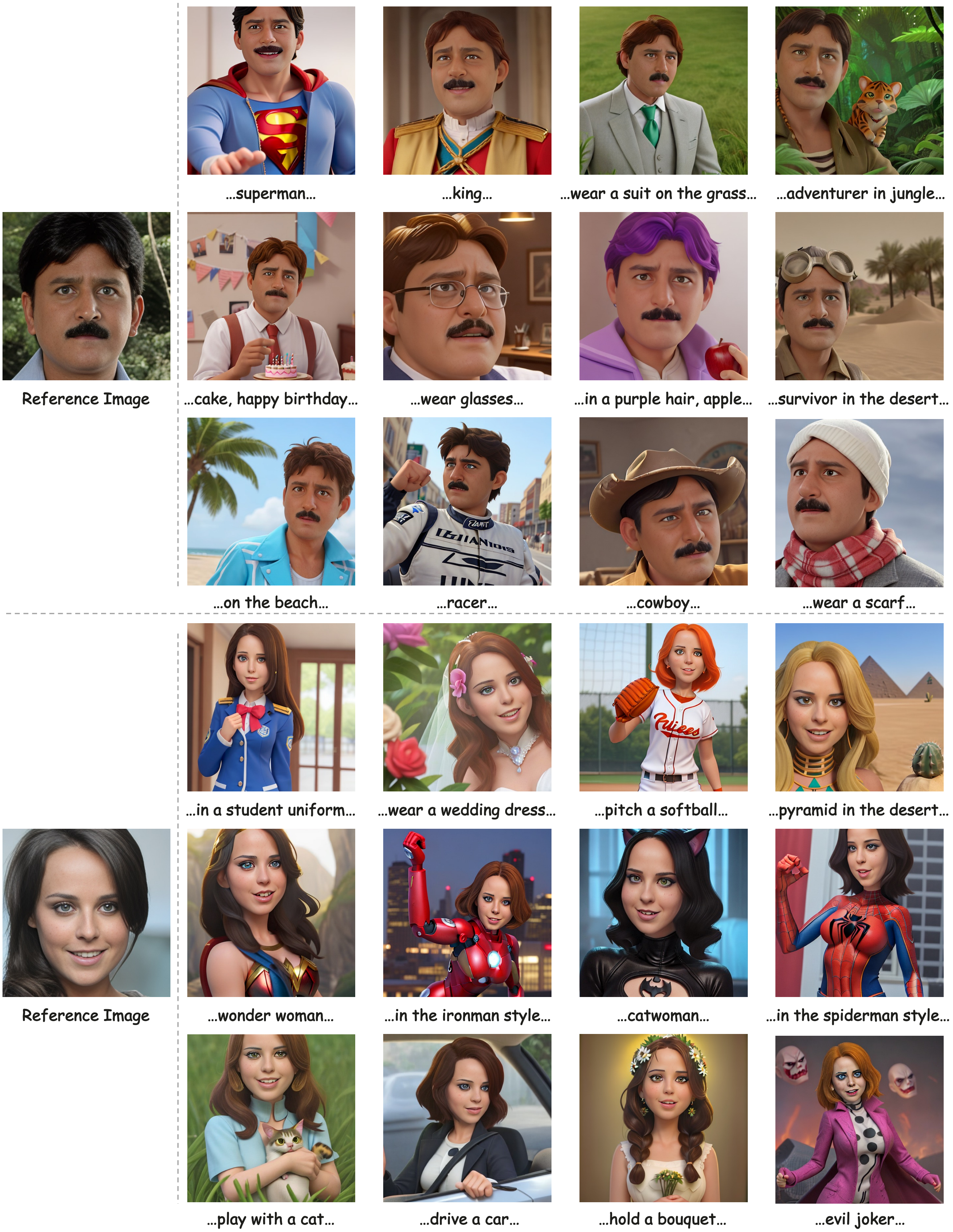
...superman...
...king...
...wear a suit on the grass...
...adventurer in jungle...
Reference Image
...cake, happy birthday...
...wear glasses...
...in a purple hair, apple...
...survivor in the desert...
...on the beach...
...racer...
...cowboy...
...wear a scarf...
...in a student uniform...
...wear a wedding dress...
...pitch a softball...
...pyramid in the desert...
Reference Image
...wonder woman...
...in the ironman style...
...catwoman...
...in the spiderman style...
...play with a cat...
...drive a car...
...hold a bouquet...
...evil joker...

Figure VII. **More results in the Disney cartoon style.** Our CapHuman can be adapted to produce various identity-preserved portraits with diverse head positions, poses, and facial expressions.

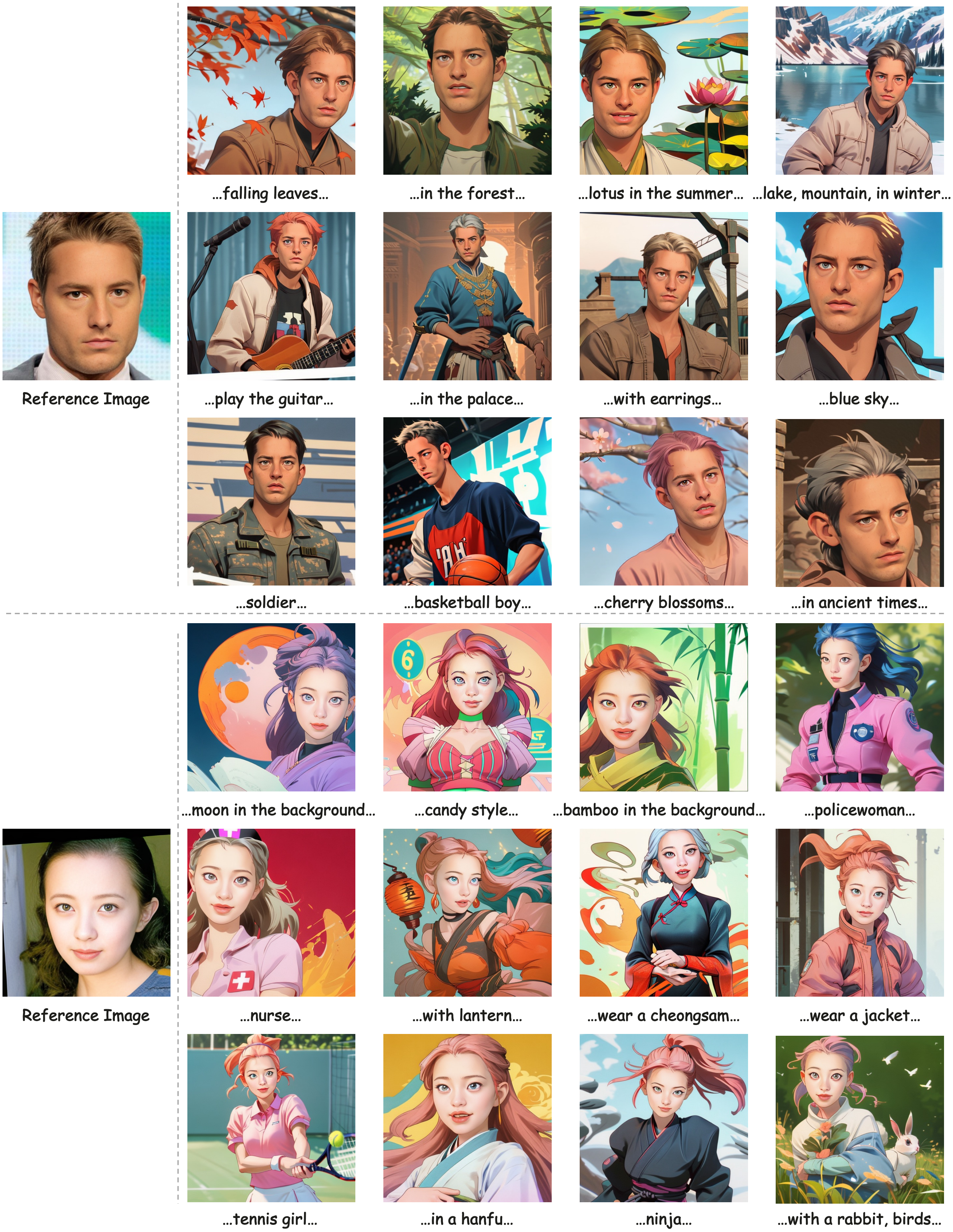

Figure VIII. **More results in the animation style.** Our CapHuman can be adapted to produce various identity-preserved portraits with diverse head positions, poses, and facial expressions.

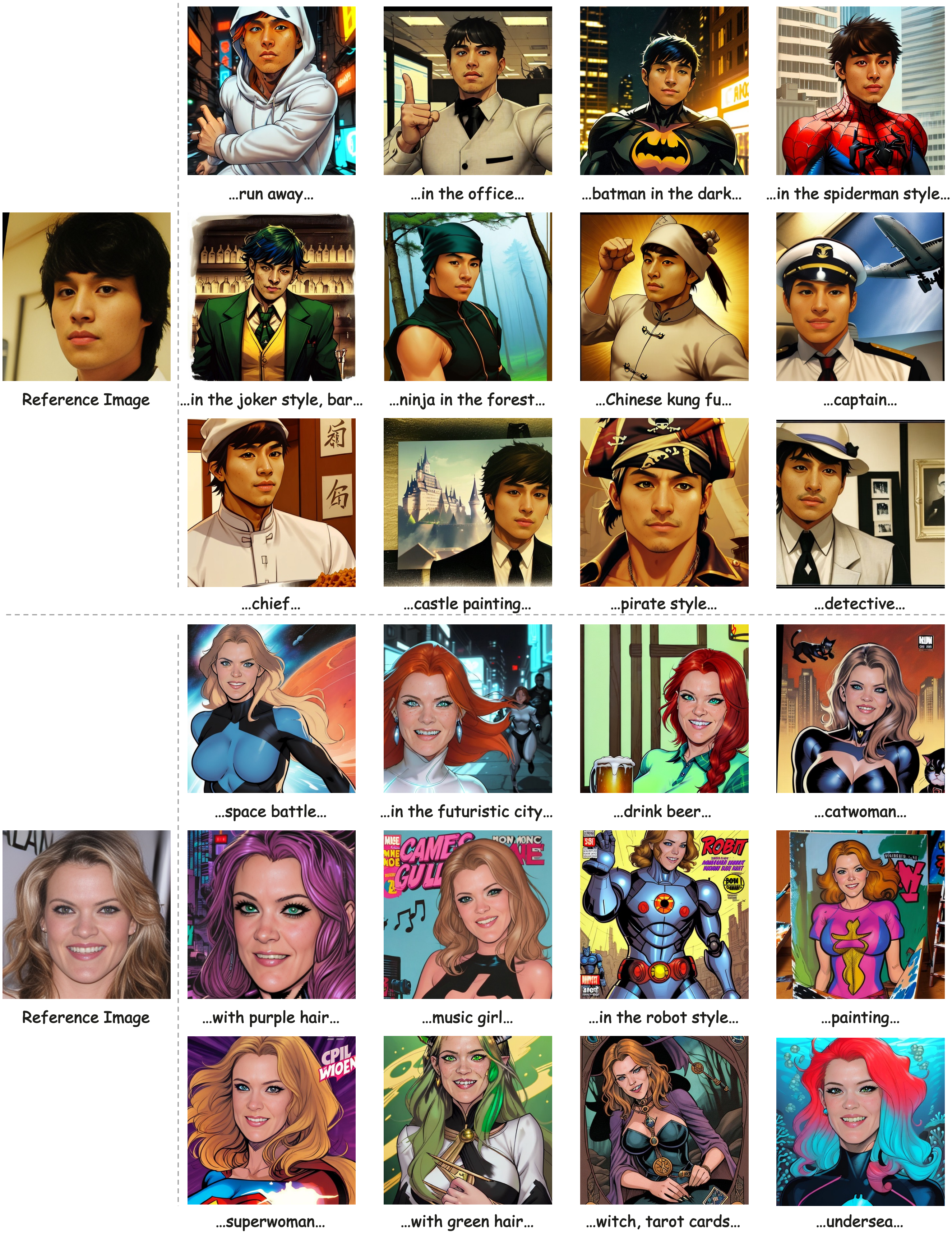

Figure IX. **More results in the comic style.** Our CapHuman can be adapted to produce various identity-preserved portraits with diverse head positions, poses, and facial expressions.

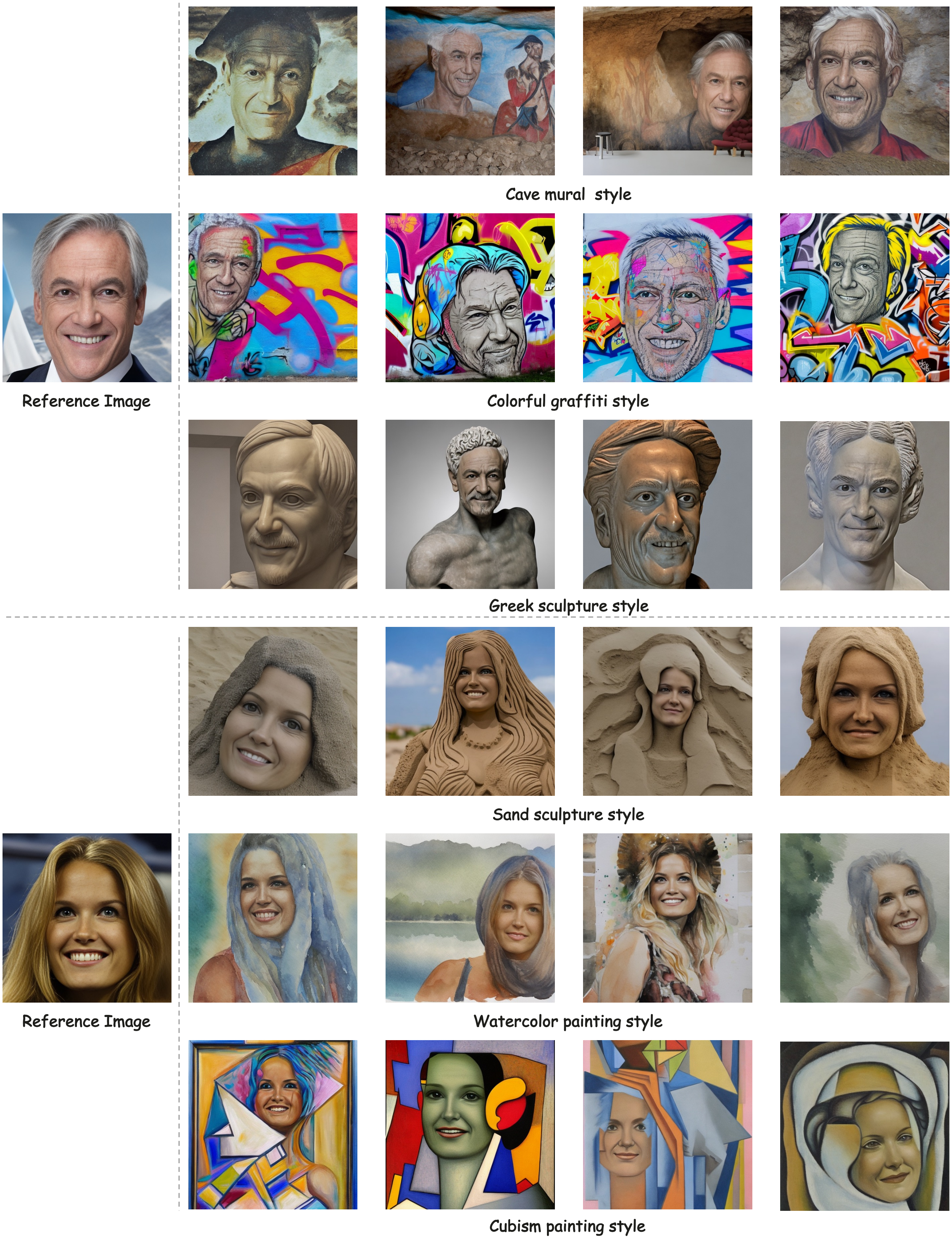

Figure X. **Stylization by style prompts.** Our CapHuman can generate identity-preserved portraits with different styles by style prompts.

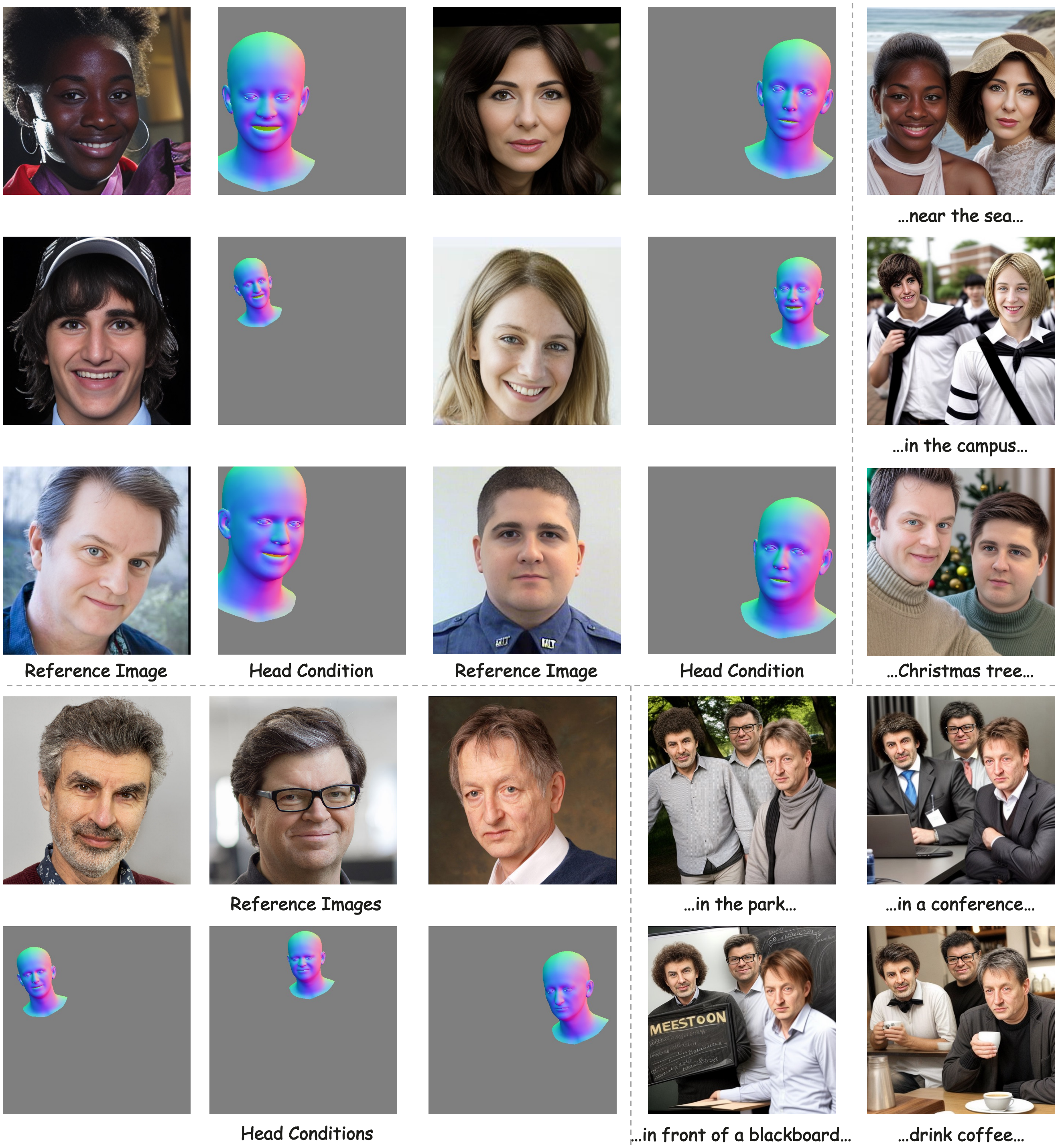

Figure XI. **Multi-Human image generation.** Given reference images, our CapHuman can generate various identity-preserved multi-human images, consistent with the corresponding head conditions.

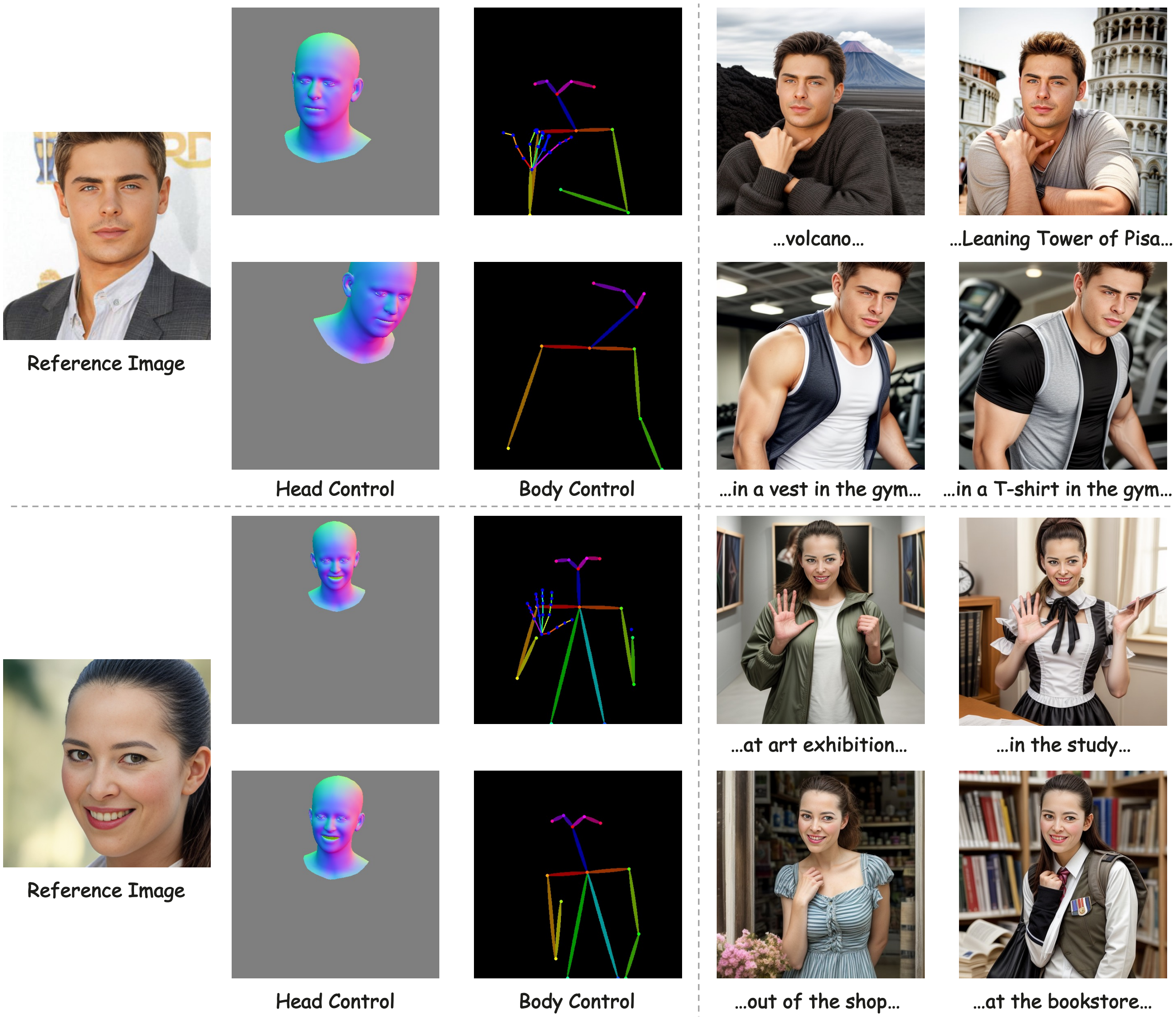

Figure XII. **Simultaneous head and body control with identity preservation.** Our CapHuman can control the head and body simultaneously with the pose-guided ControlNet [16] with identity preservation.

182723	182765	182828	182879	183243	183262	183344	183401	184642	184712
184713	184848	184858	184998	185120	185758	185827	186101	186436	186479
186538	186862	186981	187031	187083	187958	187990	188016	188082	188346
188646	189420	189454	189597	189635	189888	189913	189930	190093	190146
190971	190986	191153	191611	191663	191847	192006	192254	192279	192541
192816	192904	193230	193793	194155	194303	194309	194330	194629	194656
195350	195514	196047	196099	196205	196251	196475	196824	197119	197129
197168	197210	197464	197630	197829	198143	198223	198234	198413	198614
198869	198909	199377	199538	199621	199732	200305	200504	200505	201191
201546	201703	201731	201737	201915	201962	202244	202338	202459	202515

Table III. **ID list.** We list all the IDs used in our HumanIPHC benchmark.

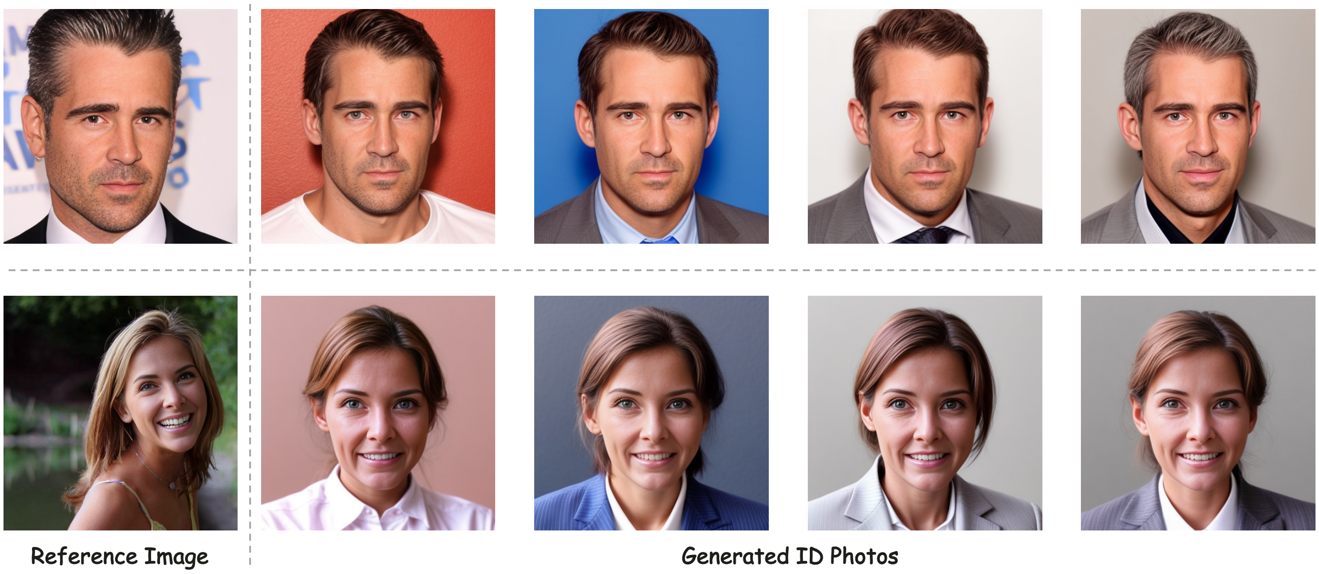

Figure XIII. **Photo ID generation.** Our CapHuman can generate standard ID photos by adjusting the head conditions and providing the proper prompts.

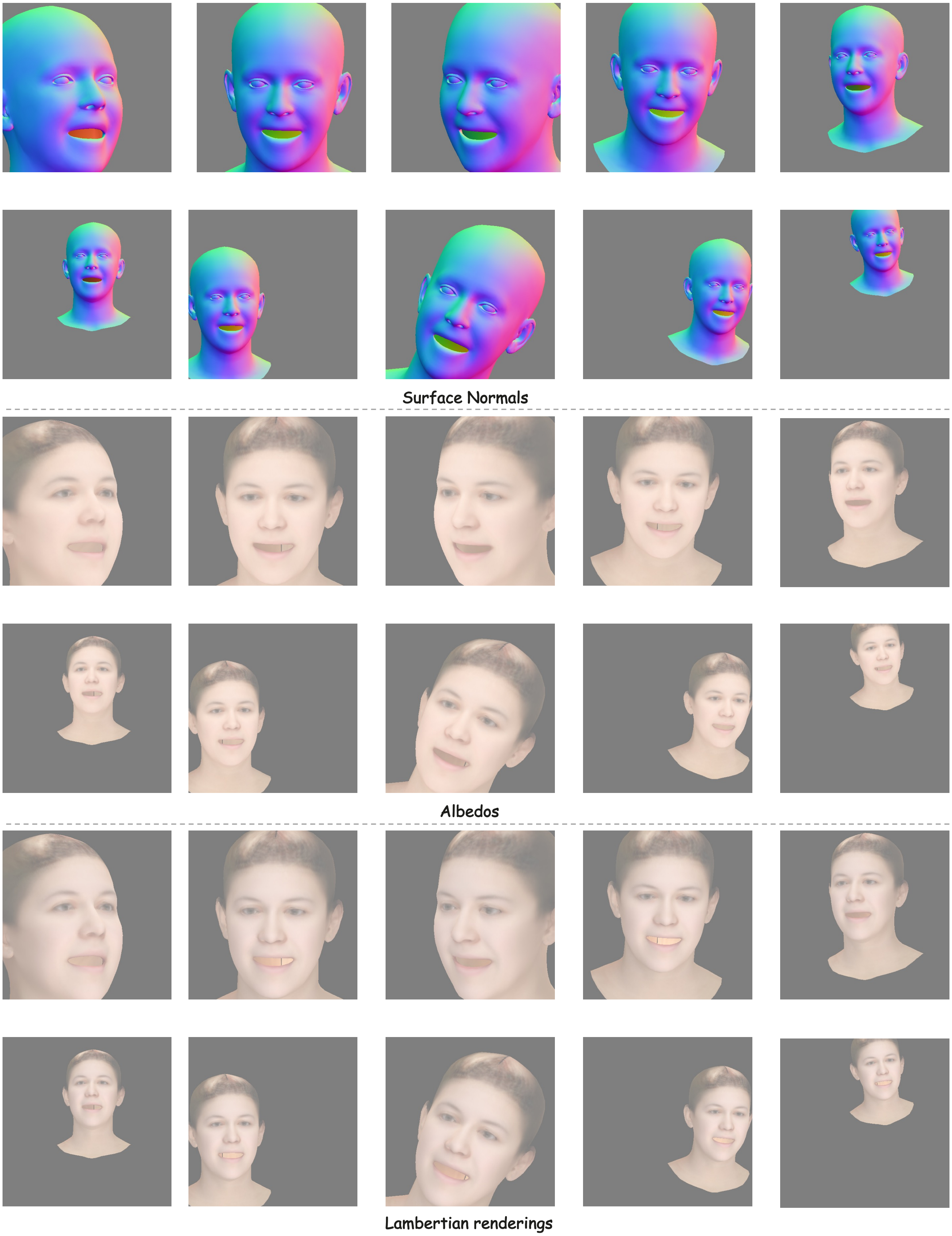

Figure XIV. **Head Conditions.** We list the head conditions of a specific individual in our HumanIPHC benchmark, including Surface Normals, Albedos, and Lambertian renderings.