



CapHuman: Capture Your Moments in Parallel Universes

Chao Liang¹ Fan Ma¹ Linchao Zhu¹ Yingying Deng² Yi Yang^{1†}
¹ReLER, CCAI, Zhejiang University, Zhejiang, China ²Huawei Technologies Ltd., China

† Corresponding author

{cs.chaoliang, zhulinchao, yangyics}@zju.edu.cn, flower.fan@foxmail.com, dyy15@outlook.com

<https://caphuman.github.io>

Reference Image

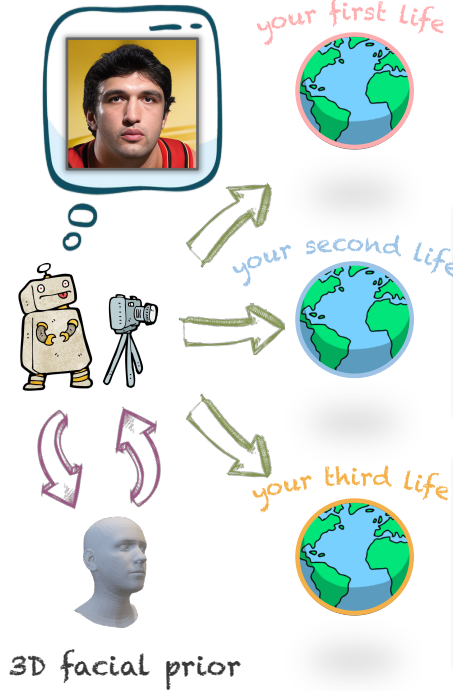


Figure 1. Given only one reference facial photograph, our CapHuman can generate photo-realistic specific individual portraits with content-rich representations and diverse head positions, poses, facial expressions, and illuminations in different contexts.

Abstract

We concentrate on a novel human-centric image synthesis task, that is, given only one reference facial photograph, it is expected to generate specific individual images with diverse head positions, poses, and facial expressions in different contexts. To accomplish this goal, we argue that our generative model should be capable of the following favorable characteristics: (1) a strong visual and semantic understanding of our world and human society for basic object and human image generation. (2) generalizable identity preservation ability. (3) flexible and fine-grained head control. Recently, large pre-trained text-to-image diffusion models have shown remarkable results, serving as a powerful generative foundation. As a basis, we aim to unleash

the above two capabilities of the pre-trained model. In this work, we present a new framework named CapHuman. We embrace the “encode then learn to align” paradigm, which enables generalizable identity preservation for new individuals without cumbersome tuning at inference. CapHuman encodes identity features and then learns to align them into the latent space. Moreover, we introduce the 3D facial prior to equip our model with control over the human head in a flexible and 3D-consistent manner. Extensive qualitative and quantitative analyses demonstrate our CapHuman can produce well-identity-preserved, photo-realistic, and high-fidelity portraits with content-rich representations and various head renditions, superior to established baselines. Code and checkpoint will be released at <https://caphuman.github.io>

1. Introduction

John Oliver: "... Does that mean there is a universe out there where I am smarter than you?"

Stephen Hawking: "Yes. And also a universe where you're funny."

– Last Week Tonight

There are infinite possibilities in parallel universes. The parallel universe, *i.e.* multiverse, is a many-worlds interpretation of quantum mechanics. When mapping into the realism framework, it means there might be thousands of different versions of our lives out here, living simultaneously. Our human beings are naturally imaginative. We are strongly eager for our second life to play different roles that have never been explored yet. Have you ever dreamed that you are a pop singer in the spotlight? Have you ever dreamed that you become a scientist, working with Stephen Hawking and Geoffrey Hinton? Or, have you ever dreamed that you act as an astronaut and have a chance to travel around the vast universe fearlessly? It will be quite satisfactory to capture our different moments in parallel universes if possible. To make our dreams come true, we raise an open question: can we resort to the help of the current machine intelligence and is it ready?

Thanks to the rapid development of advanced image synthesis technology in generative models [27, 30–32, 34, 43], the recent large text-to-image diffusion models bring the dawn of possibilities. They show promising results in generating photo-realistic, diverse, and high-quality images. To achieve our goal, we first analyze and decompose the fundamental functionalities of our model. In our scenario (see Figure 1), an ideal generative model should have the following favorable properties: (1) *a strong visual and semantic understanding of our world and human society*, which can provide the basic capabilities of object and human image generation. (2) *generalizable identity preservation ability*. Identity information is often described as a kind of visual content. It is represented as even only one reference photograph in some extreme situations, in order to meet the user’s preference. This requires our generative model to learn to extract key identity features, well-generalizable to new individuals. (3) *flexible to put the head everywhere with any poses and expressions in fine-grained control*. Human-centric image generation demands our model to support the geometric control of facial details. Then, we dive deep into the existing methods and investigate their availability. Poorly, all of them cannot meet all the aforementioned requirements. On the one hand, a number of works [14, 17, 33] attempt to personalize the pre-trained text-to-image model by fine-tuning at test-time, suffering from the overfitting problem in the one-shot setting. They

are insufficient to supply the head control as well. On the other hand, some works [12, 26, 42] focus on the head control. However, these approaches cannot preserve the individual identity or are trained from scratch without a good vision foundation and lack of text control, so as to constrain their generative ability.

In this work, we propose a novel framework CapHuman to accomplish our target. Our CapHuman is built upon the recent pre-trained text-to-image diffusion model, Stable Diffusion [32], which serves as a general representative vision generator. As a basis, we aim to unlock its potential for generalizable identity preservation and fine-grained head control. Instead of test-time fine-tuning the pre-trained model, we embrace the “encode then learn to align” paradigm, which guarantees generalizable identity preservation for new individuals without cumbersome tuning at inference. Specifically, our CapHuman encodes the global and local identity features and then aligns them into the latent feature space. Additionally, our generative model is equipped with fine-grained head control by leveraging the 3D Morphable Face Model [22]. It provides a flexible and 3D-consistent way to control the head via the parameter tuning, once we build the 3D facial representation to the reference image correspondence. With the 3D-aware facial prior, the local geometric details are better preserved.

We introduce HumanIPHC, a new challenging and comprehensive benchmark for identity preservation, text-to-image alignment, and head control precision evaluation. Our CapHuman achieves impressive qualitative and quantitative results compared with other established baselines, demonstrating the effectiveness of our proposed method.

Overall, our contributions can be summarized as follows:

- We outline a blueprint for a novel human-centric image synthesis task that generates specific individual portraits with various head positions, poses, and facial expressions in different contexts given one reference facial image.
- We propose a new framework CapHuman. We embrace the “encode then learn to align” paradigm for generalizable identity preservation without tuning at inference, and introduce 3D facial representation to provide fine-grained head control in a flexible and 3D-consistent manner.
- To the best of our knowledge, our CapHuman is the first framework to preserve individual identity while enabling text and head control in human-centric image synthesis.
- We introduce a new benchmark HumanIPHC to evaluate identity preservation, text-to-image alignment, and head control ability. Our method outperforms other baselines.

2. Related Work

2.1. Text-to-Image Synthesis

There has been significant advancement in the field of text-to-image synthesis. With the emergence of large-scale data

collections such as LAION-5B [36] and the support of powerful computation resources, large generative models bloom in abundance. One of the pathways is driven by diffusion models. Diffusion models [16] are easily scalable without instability and mode collapse of adversarial training [15]. They have achieved amazing results in generating photo-realistic and content-rich images with high fidelity. Imagen [34], GLIDE [27], and DALL-E 2 [31] directly operate the denoising process in the pixel space. Instead, Stable Diffusion [32] performs it in the latent space to enable training under the limited resources scenarios while retaining the capability of high-quality image generation. Besides, some works research on auto-regressive modeling [41] or masked generative modeling [10]. Recently, GigaGAN [19] has explored the potential of the traditional GAN framework [20] for large-scale training on the same large datasets and can synthesize high-resolution images as well.

2.2. Personalized Image Generation

Given a small subset of reference images, the personalization for text-to-image diffusion models aims to endow the pre-trained models with the capability of preserving the identity of a specific subject. Although large text-to-image diffusion models have learned strong semantic priors, they are still lacking the ability of identity preservation. A series of approaches are proposed to compensate for this missing ability by fine-tuning the pre-trained models. Textual Inversion [14] introduces a new word embedding for the user-provided concept. However, too few parameters limit the expressiveness of the output space. DreamBooth [33] fine-tunes the entire UNet backbone with a unique identifier. A class-specific prior preservation loss is further used to overcome the overfitting problem, due to the limited number of reference images. Considering the efficiency of fine-tuning, LoRA [17] only learns the residual of the model with low-rank matrices. These methods follow the “test-time fine-tuning” paradigm and need to personalize the pre-trained model for each subject. As a result, all of them fall short of fast and generalizable personalization. To address the aforementioned problem, a few works [18, 37, 39] pursue a tuning-free method. The main idea is to learn a generalizable encoder for the novel subject and preserve the text control, free from additional fine-tuning at test time.

2.3. Controllable Human Image Generation

The text condition is awesome, but still unsatisfactory for real-world applications like human image generation, which requires more fine-grained control. The challenge is how to structurally control the existing pre-trained text-to-image models. ControlNet [42] and T2I Adapter [26] design an adapter to align the new and external control signal with the original internal representation of the pre-trained text-to-image models. They both provide pose-guided condi-

tional generation but fail to preserve the identity. In addition, DiffusionRig [12] supports personalized facial editing with head control. The proposed framework cannot provide the text editing ability, limiting its generative capability.

3. Method

3.1. Preliminary

Stable Diffusion [32] is a popular open-source text-to-image generation framework, that achieves great progress in high-resolution and content-rich image generation. It has attracted considerable interest and is applied in several tasks [7, 23, 40]. Stable Diffusion belongs to the family of the latent diffusion models. By compressing the data into the latent space, it enables more efficient scalable model training and image generation. This framework is composed of two stages. First, it trains an autoencoder \mathcal{E} to map the original image x into the lower-dimensional latent representation $z = \mathcal{E}(x)$. Then, in the latent space, a time-conditional UNet denoiser predicts the added noise at different timesteps. For the text condition, this model employs the cross-attention mechanism [38] to understand the semantics of text prompts. Put it together, the denoising objective can be formulated as follows:

$$\mathcal{L}_{LDM} = E_{z,c,\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}), t \sim \mathcal{U}(1,T)} [\|\epsilon_\theta(z_t, t, c) - \epsilon\|_2], \quad (1)$$

where z_t is the noisy latent code, c is the text embedding, ϵ is sampled from the standard Gaussian distribution, and t is the timestep. Pre-trained on large-scale internet data, Stable Diffusion has learned strong semantic and relation priors for natural and high-quality image generation.

FLAME [22] is one of the expressive 3D Morphable Face Models [6, 8, 9, 22, 28]. It is a statistical parametric face model that captures variations in shape, pose, and facial expression. Given the coefficients of shape β , pose θ , and expression ψ , the model can be described as:

$$M(\beta, \theta, \psi) = W(T_P(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}), \quad (2)$$

where T_P is rotated around joints J linearly smoothed by blendweight \mathcal{W} . Here, T_P denotes the template with added shape, pose, and expression offsets. In other words, it is flexible for us to control the facial geometry by adjusting or tuning the parameters of β , θ , and ψ within a range.

3.2. Overview

In this work, we consider a novel human-centric image synthesis task. Given only one reference face image I indicating the individual identity, our goal is to generate photo-realistic and diverse images for the specific identity with different head positions, poses, and facial expressions in different contexts, driven by the text prompt \mathcal{P} and the head condition \mathcal{H} . Input as a triplet data pair $(I, \mathcal{P}, \mathcal{H})$, we learn a model \mathcal{G} as our generative model to produce a new image

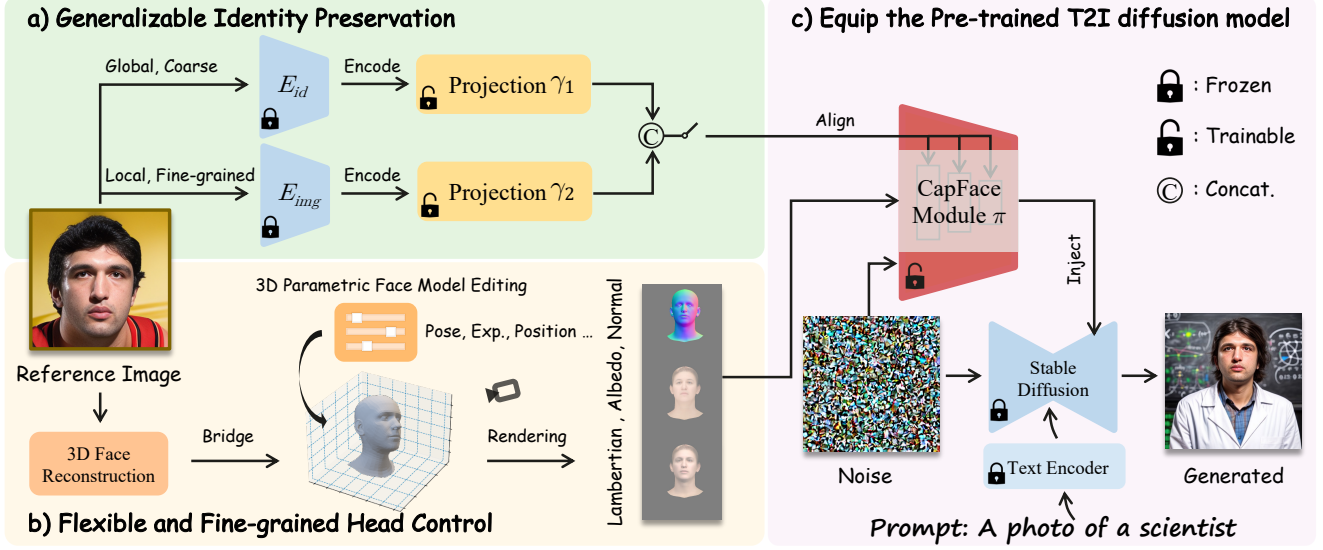


Figure 2. **Overview of CapHuman.** Our CapHuman stands upon the pre-trained T2I diffusion model. a) We embrace the “encode then learn to align” paradigm for generalizable identity preservation. b) The introduction of the 3D parametric face model enables flexible and fine-grained head control. c) We learn a CapFace module π to equip the pre-trained T2I diffusion model with the above capabilities.

\hat{I} . The pipeline can be defined as:

$$\hat{I} = \mathcal{G}(I, \mathcal{P}, \mathcal{H}). \quad (3)$$

To accomplish this task, ideally, the model \mathcal{G} should be equipped with the following functionalities: (1) basic object and human image generation capability. (2) generalizable identity preservation ability. (3) flexible and fine-grained head control. Recently, large pre-trained text-to-image diffusion models [31, 32, 34] have shown incredible and impressive generative ability. They are born with the implicit knowledge of our world and human society, which serves as a good starting point for our consolidation. We propose a new framework CapHuman, which is built upon the pre-trained text-to-image diffusion model, Stable Diffusion [32]. Although Stable Diffusion has the in-born generation capability, it still lacks the ability of identity preservation and head control, limiting its application in our scenario. We aim to endow the pre-trained model with the above two abilities by introducing a CapFace module π . § 3.3 introduces the generalizable identity preservation module. § 3.4 concentrates on the flexible and fine-grained head control capability. § 3.5 presents the training and inference process. The overall framework is shown in Figure 2.

3.3. Generalizable Identity Preservation

The most straightforward solution [14, 17, 33] is to fine-tune the pre-trained model with the given reference image. Though the model can preserve the identity in this case, it sacrifices the generality. The fine-tuning process forces the model to memorize the specific individual. When a new individual comes, it needs to re-train the model, which is cumbersome. Instead, we advocate the “encode then learn

to align” paradigm, that is, we treat identity preservation as one of the generalizable capabilities that our model should have. We formulate it as a learning task. The task requires our model to learn to extract the identity information from the reference image and preserve the individual identity in the image generation. We break it down into two steps.

Encode global and local identity features. In the first step, the reference face image I is encoded into identity features at different granularities. Here, we consider two types of identity features: (1) **global coarse feature** represents the key and typical characteristics of the human face. We use the feature extractor E_{id} pre-trained on the face recognition task [35] to obtain the global face embedding $\mathbf{f}_{global} = E_{id}(I) \in \mathbb{R}^{1 \times d_1}$. The global feature captures the key information to help distinguish it from other identities, but some appearance details might be overlooked. (2) **local fine-grained feature** depicts more facial details, which can further enhance the fidelity of face image generation. We leverage the CLIP [29] image encoder E_{img} to extract local patch image feature $\mathbf{f}_{local} = E_{img}(I) \in \mathbb{R}^{N \times d_2}$. Note that we only keep the face area by masking and the irrelevant background is removed.

Learn to align into the latent space. In the second step, our model π learns to align the identity features into its feature space. We embed the global and local features into the latent identity feature \mathbf{f}_{id} :

$$\mathbf{f}_{id} = [\gamma_1(\mathbf{f}_{global}); \gamma_2(\mathbf{f}_{local})] \in \mathbb{R}^{(1+N) \times d}, \quad (4)$$

where γ_1, γ_2 are projection layers and $[\cdot]$ denotes the concatenation operation. Then, the latent identity feature is processed by the cross-attention mechanism [38], attending to

the latent feature \mathbf{f}_l in π , as formulated in the following way:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (5)$$

where the query, key and value are defined as $Q = \phi_Q(\mathbf{f}_l)$, $K = \phi_K(\mathbf{f}_{id})$, $V = \phi_V(\mathbf{f}_{id})$. And ϕ_Q, ϕ_K, ϕ_V are linear projections. By inserting the identity features into the latent feature space in the denoising process, our model can preserve the individual identity in the image synthesis. The combination of global and local features not only strengthens the recognition of individual identity but also complements the facial details in the human image generation. The “encode then learn to align” paradigm guarantees our model is generalizable for new individuals without the need for extra tuning in the inference time.

3.4. Flexible and Fine-grained Head Control

Human-centric image generation favors flexible, fine-grained, and precise control over the human head. It is desirable to have the ability to put the head everywhere in any pose and expression in the human image synthesis. However, the powerful pre-trained text-to-image diffusion model lacks this control. It is believed that the pre-trained model has learned internal structural priors about generating different human images with different head positions, poses, and facial expressions. We tend to unlock its capability by introducing a proper control signal as a trigger. The first question is what is a good representation for this signal?

Bridge 3D facial representation. We pay attention to the popular 3D parametric face model FLAME [22]. It constructs a compact latent space to represent the shape, pose, and facial expression separately. As a by-product, it provides a friendly and flexible interface to edit the facial geometry easily, *e.g.* changing the head pose, and expression with varied parameters. In our setting, we need to bridge the input reference image I and the 3D facial representation. We use DECA [13] to reconstruct the specific 3D head model with detailed facial geometry from a single image. Then, we transform it into a set of pixel-aligned condition images including Surface Normal, Albedo, and Lambertian rendering. They contain the position, local geometry, albedo, and illumination information [12].

Equip with 3D-consistent head control. We attempt to equip the pre-trained generative model with the ability to respond to the control signal. Given the head condition $\mathcal{H} = \{I_{Normal}, I_{Albedo}, I_{Lambertian}\}$, we obtain the feature map \mathcal{F}_t . The process is defined as:

$$\mathcal{F}_t = \pi(z_t, t, \mathcal{H}, \mathbf{f}_{id}). \quad (6)$$

Because the head condition images are coarse facial appearance representations, we incorporate the identity features to strengthen the local details. In order to force the CapFace module π to focus on the facial area, we predict the facial

mask \mathcal{M} from the head condition \mathcal{H} . Finally, the masked feature map $\mathcal{F}_t \odot \mathcal{M}$ is injected into the original feature space of the pre-trained model. By embedding the new control signal, the pre-trained model is endowed with the ability of head control. The introduction of the 3D parametric face model enables 3D-consistent control of the human head.

3.5. Training and Inference

Training objective. We calculate the denoising loss between the predicted and groundtruth noise, with the mask prediction loss. The training objective for the model optimization is formulated as:

$$\mathcal{L} = \|\epsilon_\theta(z_t, t, c, \pi(z_t, t, \mathcal{H}, \mathbf{f}_{id})) - \epsilon\|_2 + \lambda \|\mathcal{M} - \mathcal{M}_{gt}\|_2, \quad (7)$$

where \mathcal{M}_{gt} is the groundtruth facial mask, and we set $\lambda = 1$. We keep ϵ_θ frozen and train the CapFace module π .

Time-dependent ID dropout. In the learning process, our model might focus more on the identity features due to the entanglement of the head pose information in the reference image, which results in weak control of the head condition. Inspired by the fact that the denoising process in the diffusion model is progressive and the appearance is concentrated at the later stage [16], we propose a time-dependent ID dropout regularization strategy that discards the identity feature at the early stage to alleviate the issue. We formulate the training strategy in the following:

$$\mathcal{F}_t^\dagger = \begin{cases} \pi(z_t, t, \mathcal{H}, \mathbf{f}_{id}), & t < \tau, \\ \pi(z_t, t, \mathcal{H}, \emptyset), & \text{otherwise,} \end{cases} \quad (8)$$

where τ is the start timestep and \mathcal{F}_t^\dagger is the feature map.

Post-hoc Head Control Enhancement. To enhance the head control of our generative model, we optionally fuse the feature map with others from the head control model π^* at inference:

$$\mathcal{F}_t^\ddagger = \pi(z_t, t, \mathcal{H}, \mathbf{f}_{id}) + \alpha \cdot \pi^*(z_t, t, \mathcal{H}, \emptyset), \quad (9)$$

where α is the control scale and \mathcal{F}_t^\ddagger is the feature map.

4. Experiments

4.1. Training setup

We train our model on CelebA [24], which is a large-scale face dataset with more than 200K celebrity images, covering diverse pose variations. For data preprocessing, we crop and resize the image to the size of 512×512 resolution. Following [20], we crop and align the face region for the reference image. We use BLIP [21] for image captioning. We choose ViT-L/14 as the CLIP [29] image encoder. Our model is based on Stable Diffusion V1.5 [32]. The learning rate is 0.0001 and the batch size is 128. We use AdamW [25] for the optimization.

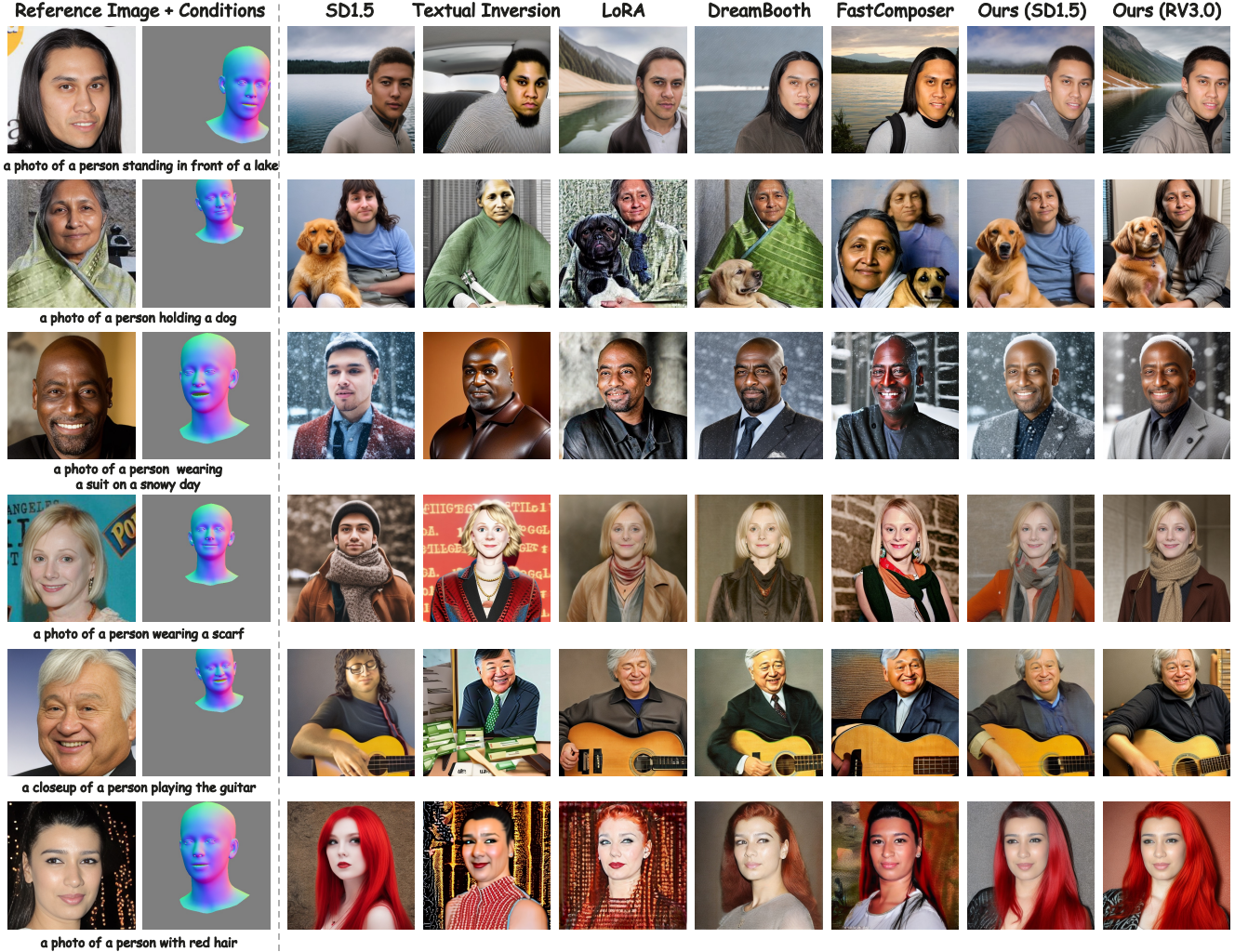


Figure 3. **Qualitative results.** Our CapHuman can produce identity-preserved, photo-realistic portraits with various head positions and poses in different contexts. Also, our model can be combined with the pre-trained model, *e.g.* RealisticVision [1] in the community flexibly.

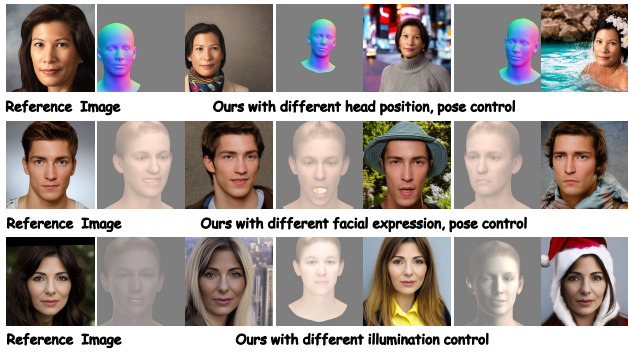


Figure 4. **Head position, pose, facial expression, and illumination control.** Our method offers the 3D-consistent head control.

4.2. Qualitative Analysis

Visual comparisons. We focus on the one-shot setting where only one reference image is given. We compare our method with the established techniques including Textual

Inversion [14], DreamBooth [33], LoRA [17] and FastComposer [39]. These methods are designed for personalization and lack of head control. Also, the original Stable Diffusion model [32] is one of our baselines. For fair comparisons, we combine them with ControlNet [42], since ControlNet can provide facial landmark-driven control. The visual qualitative results are presented in Figure 3. Obviously, Stable Diffusion cannot preserve the individual identity. The fine-tuning baselines can preserve the individual identity to a certain extent. However, they suffer from the overfitting issue. The input prompt might not take effect in some cases. It suggests that these methods sacrifice the diversity for the identity memorization. Compared with the state-of-the-art approaches, our method shows competitive and impressive generative results with good identity preservation. Given only one reference photo, our CapHuman can produce photo-realistic and well identity preserved images with various head positions and poses in different contexts.



Figure 5. **Adapt our model to other pre-trained models.** Our model can be adapted to generate portraits in different styles.

Head control capability. Figure 4 shows the head control capability of our CapHuman. The visual results demonstrate our CapHuman can offer 3D-consistent control over the human head in position, pose, and facial expressions. More results can be found in Figure 12 in the supplementary material.

Adapt to other pre-trained models. Our model can be adapted to other pre-trained models [2–4] in the community seamlessly. The results are presented in Figure 5. More results can be found in Figure 13, 14, 15, 16 in the supplementary material.

4.3. Quantitative Analysis

Benchmark. We introduce a new challenging and comprehensive benchmark HumanIPHC for identity preservation, text-to-image alignment, and head control precision evaluation. We select 100 identities from the CelebA [24] test split. They consist of different ages, genders, and races. We collect 10 diverse prompts and 10 different head conditions with various positions and poses. For each (identity, prompt, head condition) combination, we generate three different images, totaling 30,000 images.

Evaluation metrics. We evaluate the effectiveness of our proposed method in the following three dimensions: (1) Identity Preservation. We apply a face recognition network [35] to extract the facial identity feature from the face region. The cosine similarity between the reference image and the generated image is used to measure the facial identity similarity. (2) Text-to-Image Alignment. We use the CLIP score as the metric. The CLIP [29] score is calculated as the pairwise cosine similarity between the image and text features. In addition, we report the prompt accuracy. It is the classification accuracy between the generated image and a set of candidate prompts. (3) Head Control Precision. We compute the root mean squared error (RMSE) between the DECA [13] code estimated from the generated image and the given condition. We divide the DECA code into four groups: Shape, Pose, Expression, and Lighting.

Quantitative results. Table 1 shows the evaluation results on our benchmark. For identity preservation, Textual In-

Method	Identity Preservation		Text-to-Image Alignment		Head Control Precision			
	Generalizable	↑ ID sim.	↑ CLIP score	↑ Prompt acc.	↓ Shape	↓ Pose	↓ Exp.	↓ Light.
Stable Diffusion [32]	✗	0.0462	0.2398	88.88%	0.2725	0.0494	0.3643	0.2732
Textual Inversion [14]	✗	0.5003	0.1566	19.42%	0.2041	0.0513	0.2507	0.2563
LoRA [17]	✗	0.5759	0.1841	40.49%	0.1669	0.0435	0.2068	0.1688
DreamBooth [33]	✗	0.6661	0.1806	45.93%	0.1595	0.0435	0.1996	0.1792
FastComposer [39]	✓	0.6183	0.1977	61.22%	0.1851	0.0608	0.2111	0.1875
Ours	✓	0.8281	0.2148	73.26%	0.1030	0.0422	0.1250	0.0982

Table 1. **Comparisons with the established state-of-the-art methods.** Our CapHuman outperforms other baselines for better identity preservation and better head control. Compared with other personalization methods, our method can still keep a high level of prompt control. **Bold** denotes the best result.

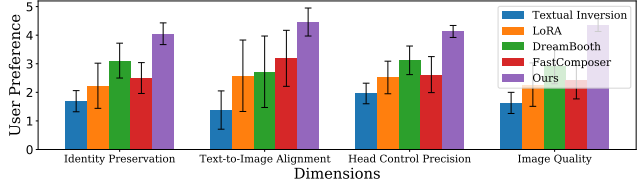


Figure 6. **User Study.** Users prefer our CapHuman in all four dimensions: identity preservation, text-to-image alignment, head control precision, and image quality.

sion [14], LoRA [17], and DreamBooth [33] can improve the performance on identity similarity. Their abilities depend on the scale of the trainable parameters. DreamBooth fine-tunes the entire backbone while Textual Inversion only trains the word embedding. As a result, DreamBooth shows better results. By learning to encode the identity information, our model achieves generalizable identity preservation capability, surpassing DreamBooth [33] and FastComposer [39] by 16% and 21%, respectively. For text-to-image alignment, the fine-tuning methods fall into the overfitting problem under the one-shot setting. They sacrifice prompt diversity for better identity preservation. In contrast, our method can still maintain a high level of prompt control. For head control precision, our method shows remarkable improvement in Shape, Expression, and Lighting metrics, *i.e.*, 6%, 7%, 7% compared with the second best results. We attribute this to the introduction of the 3D facial prior.

User Study. We invite 12 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 6 shows our CapHuman is much more preferred by the users.

4.4. Ablation Studies

We perform the ablation studies on a small subset with 10 identities to study the effectiveness of our design.

Effect of global and local identity features. We investigate the importance of global and local features for identity preservation. In Table 2, we present the identity similarity comparison. As expected, both global and local identity features contribute to identity preservation. The performance drops when removing the global or local feature individually. Furthermore, we illustrate the effectiveness of the identity features in Figure 7. We can observe that our model cannot preserve the individual identity if no identity

Method	\uparrow ID sim.	Num. N	\uparrow ID sim.
w/o global & local feat.	0.3915	32	0.8370
w/o local feat.	0.7725	64	0.8376
w/o global feat.	0.8095	128	0.8182
w global & local feat.	0.8429	257	0.8429

Table 2. Ablation on ID features.

Table 3. Effect of N .

Figure 7. Visual results of global and local identity features. Both global and local features contribute to identity preservation.

Method	\downarrow Shape	\downarrow Pose	\downarrow Exp.	\downarrow Light.
ControlNet [42]	0.2909	0.0501	0.3967	0.2899
Our Head Control	0.1381	0.0262	0.1639	0.1196

Table 4. Comparison with ControlNet. Our head control model achieves significant performance on the head control precision.

features are involved during the image generation. With the global identity feature, we can recognize the identity basically. Additionally, the local feature complements the details and enhances the facial fidelity.

Effect of the number N in the local identity feature.

We study the effect of the number N in the local identity feature. As reported in Table 3, we find the compression of the local identity feature can hurt the performance of identity preservation. This indicates the local feature captures rich local facial details. It is better to make full use of the local identity features in human face image generation.

Comparison with ControlNet on the head control. We compare our method with ControlNet [42] to validate the effectiveness of the head control ability. We remove the identity preservation module. Table 4 shows the head control results. Our method outperforms ControlNet by a large margin for all the metrics. The gap is 15%, 2%, 23%, and 17% in Shape, Pose, Expression, and Lighting, respectively. This is because ControlNet only relies on the facial landmark for head control. Instead, we leverage the 3D parametric face model to provide a strong facial prior. The introduction of the 3D facial representation brings more information like local geometry and illumination. Figure 8 further confirms more precise head control of our method.

Influence of the ID dropout start timestep τ . We study the influence of the ID dropout start timestep τ . As shown in Table 5, with more time identity features participate in the denoising process, our model shows stronger identity preservation capability. However, the pose metric gets

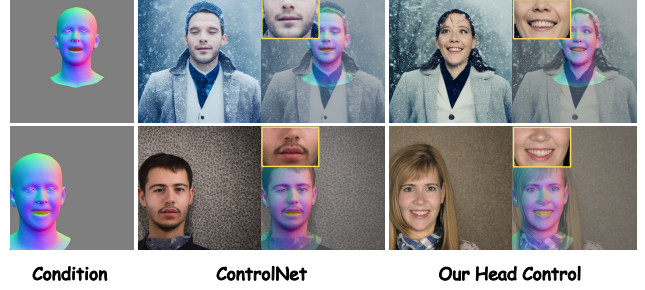


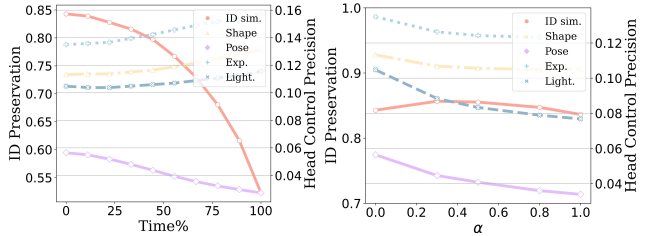
Figure 8. Visual comparison with ControlNet. Our head control model shows more fine-grained control results with local details.

Method	\uparrow ID sim.	\downarrow Shape	\downarrow Pose	\downarrow Exp.	\downarrow Light.
$\tau = 0$	0.3915	0.1381	0.0262	0.1639	0.1196
$\tau = 300$	0.6600	0.1257	0.0292	0.1493	0.1124
$\tau = 500$	0.7589	0.1185	0.0343	0.1450	0.1074
$\tau = 700$	0.7986	0.1165	0.0467	0.1409	0.1033
$\tau = 1000$	0.8429	0.1132	0.0564	0.1349	0.1047

Table 5. Ablation on the ID dropout start timestep τ . The time-dependent ID dropout training strategy plays a role in the tradeoff between identity preservation and pose control.

Method	\uparrow ID sim.	\downarrow Shape	\downarrow Pose	\downarrow Exp.	\downarrow Light.
w/o Post-hoc Enhance.	0.8429	0.1132	0.0564	0.1349	0.1047
+ ControlNet [42]	0.8386	0.1118	0.0427	0.1377	0.1032
+ Our Head Control	0.8338	0.1060	0.0358	0.1263	0.0795

Table 6. Post-hoc Head Control Enhancement at inference. Head control metrics are boosted with the head control model.

Figure 9. Left: The utilization time (%) of the head control model at inference. Using the head control model at the early stage can improve the pose control but sacrifice the identity similarity. Right: Ablation on the control scale α . With the control scale α increasing, head control metrics are improved at a negligible cost of identity preservation.

worse. In the learning process, our model might concentrate more on the identity feature and overlook the pose condition. The experimental results prove that the time-dependent ID dropout strategy plays a role in the tradeoff between identity preservation and head pose control.

Post-hoc Head Control Enhancement. We further explore the possibilities of enhancing the head pose control in the inference time. We train a head control model without the identity preservation module. First, we use the head control model for the early denoising stage, and then our model with the identity preservation module. We vary the start timestep. The evaluation results are shown in Figure 9. It improves the pose metric by sacrificing the ID preserva-

tion capability. Second, we enhance our model with ControlNet or the head control model by fusing the feature map. Table 6 presents the results. As we can see, the pose metric further boosts when we combine our model with the head control model. Last, we perform the ablation studies on the control scale α . Figure 9 shows the head control model can strengthen the pose control at a negligible loss of identity.

5. Conclusion

In this paper, we propose CapHuman, a generalizable identity preservation and fine-grained head control human-centric image synthesis framework, based on the powerful pre-trained text-to-image diffusion model. We embrace the “encode then learn to align” paradigm to give birth to its generalizable identity preservation capability without further fine-tuning. By incorporating the 3D facial representation, it endows the pre-trained model with flexible and fine-grained head control. Given one reference face image, our CapHuman can generate identity-preserved, high-fidelity, and photo-realistic human portraits with diverse head positions, poses, and facial expressions in different contexts.

References

- [1] Realistic vision v3.0. https://huggingface.co/SG161222/Realistic_Vision_V3.0_VAE, 2023. 6, 1, 3, 4
- [2] comic-babes. <https://civitai.com/models/20294/comic-babes>, 2023. 7
- [3] disney-pixar-cartoon. <https://civitai.com/models/65203/disney-pixar-cartoon-type-a>, 2023.
- [4] toonyou. <https://civitai.com/models/30240/toonyou>, 2023. 7, 1
- [5] Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 467–474, 2023. 2
- [6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 3
- [7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3
- [8] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 3
- [9] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2): 233–254, 2018. 3
- [10] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3
- [11] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2
- [12] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023. 2, 3, 5
- [13] 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. 5, 7, 1
- [14] 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 *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 4, 6, 7, 1
- [15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5
- [17] 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. 2, 3, 4, 6, 7, 1
- [18] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 3
- [19] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 3
- [20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3, 5
- [21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 5

- [22] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 3, 5
- [23] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023. 3
- [24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5, 7
- [25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5
- [26] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhong-gang Qi, Ying Shan, and Xiaoohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3
- [27] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3
- [28] Pascal Paysan, Reinhard Kothke, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 3
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5, 7
- [30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2
- [31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3, 4
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4, 5, 6, 7, 1
- [33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3, 4, 6, 7, 1
- [34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3, 4
- [35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4, 7
- [36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3
- [37] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instant-booth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 3
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4
- [39] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv*, 2023. 3, 6, 7, 1
- [40] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaoohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 3
- [41] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 3
- [42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 6, 8, 1, 13
- [43] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 1795–1803. ijcai.org, 2023. 2

CapHuman: Capture Your Moments in Parallel Universes

Supplementary Material

6. More Qualitative Results

Visual comparisons. We show more visual comparisons with the established baselines [14, 17, 32, 33, 39] in Figure 10 and 11. 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 12, we provide more examples, demonstrating the facial expression control ability of our CapHuman.

7. 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 13, 14, 15, and 16.

Stylization by style prompts. We also showcase portraits with different styles driven by style prompts in Figure 17.

Multi-Human image generation. Our CapHuman supports multi-human image generation. The generated results are presented in Figure 18.

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 19, our CapHuman can generate standard ID photos by adjusting the head conditions and providing the proper prompts conveniently.

Simultaneous head and body control. Combined with the pose-guided ControlNet [42], our CapHuman can control the head and the body simultaneously with identity preservation. More results are presented in Figure 20.

8. 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 7.

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 wearing a scarf.
- a photo of a person playing basketball.
- a photo of a person running on a rainy day.
- a photo of a person wearing a suit on a snowy day.
- a photo of a person holding a dog.
- a closeup of a person playing the guitar.
- a portrait of a person with a comic book.

Head conditions. In Figure 21, we show the head conditions of a specific individual in our benchmark, including Surface Normals, Albedos, and Lambertian renderings.

9. 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).

10. 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 [13]. 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, 11] or by controlling the release of the model strictly.

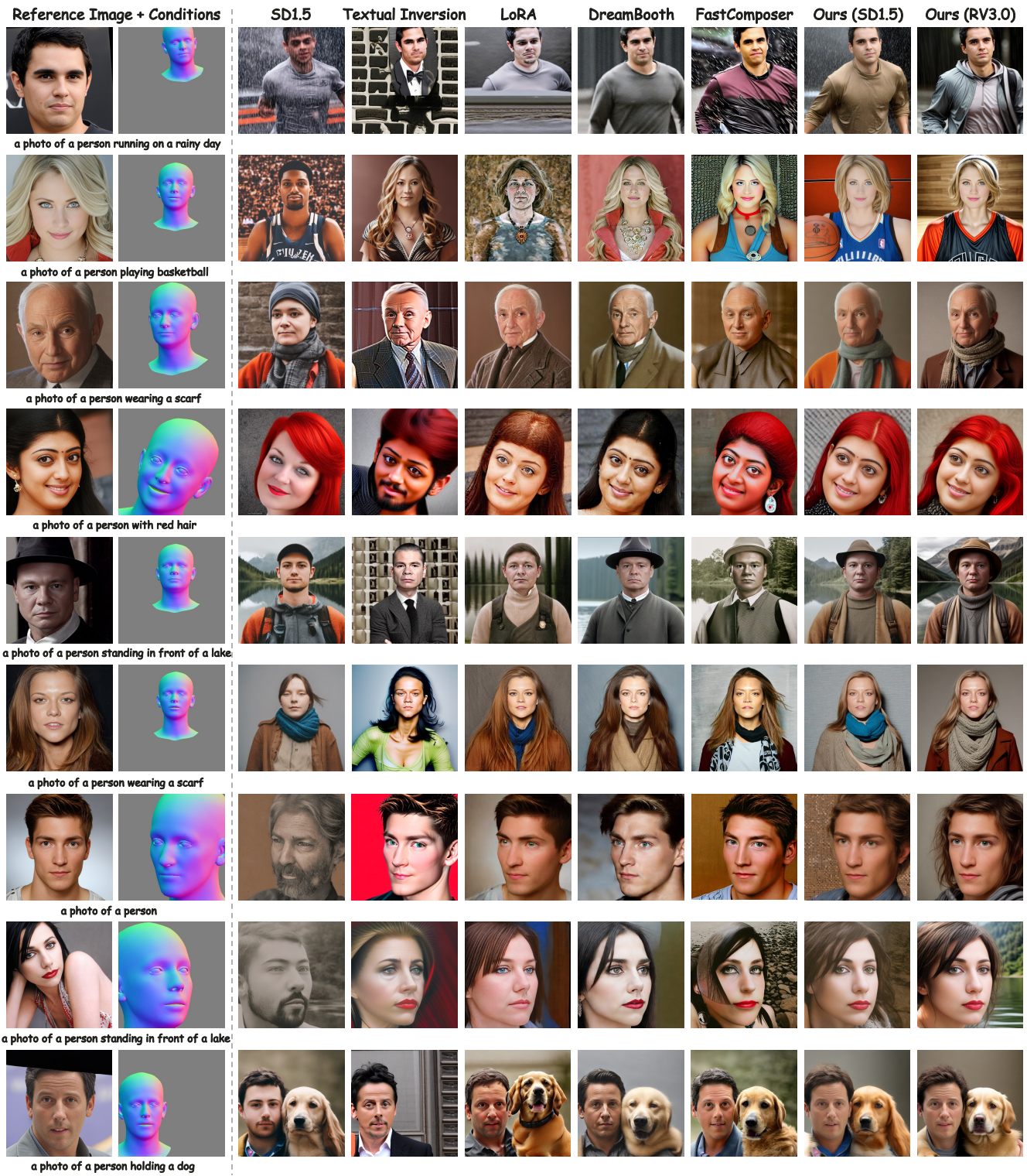


Figure 10. **More qualitative results.** Our CapHuman can produce well-identity-preserved, photo-realistic, and high-fidelity portraits with various head positions and poses in different contexts, compared with the baselines. Note that our model can be combined with other pre-trained models, *e.g.* RealisticVision [1] in the community flexibly. For the head condition, we only display the Surface Normal here.



Figure 11. **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.

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

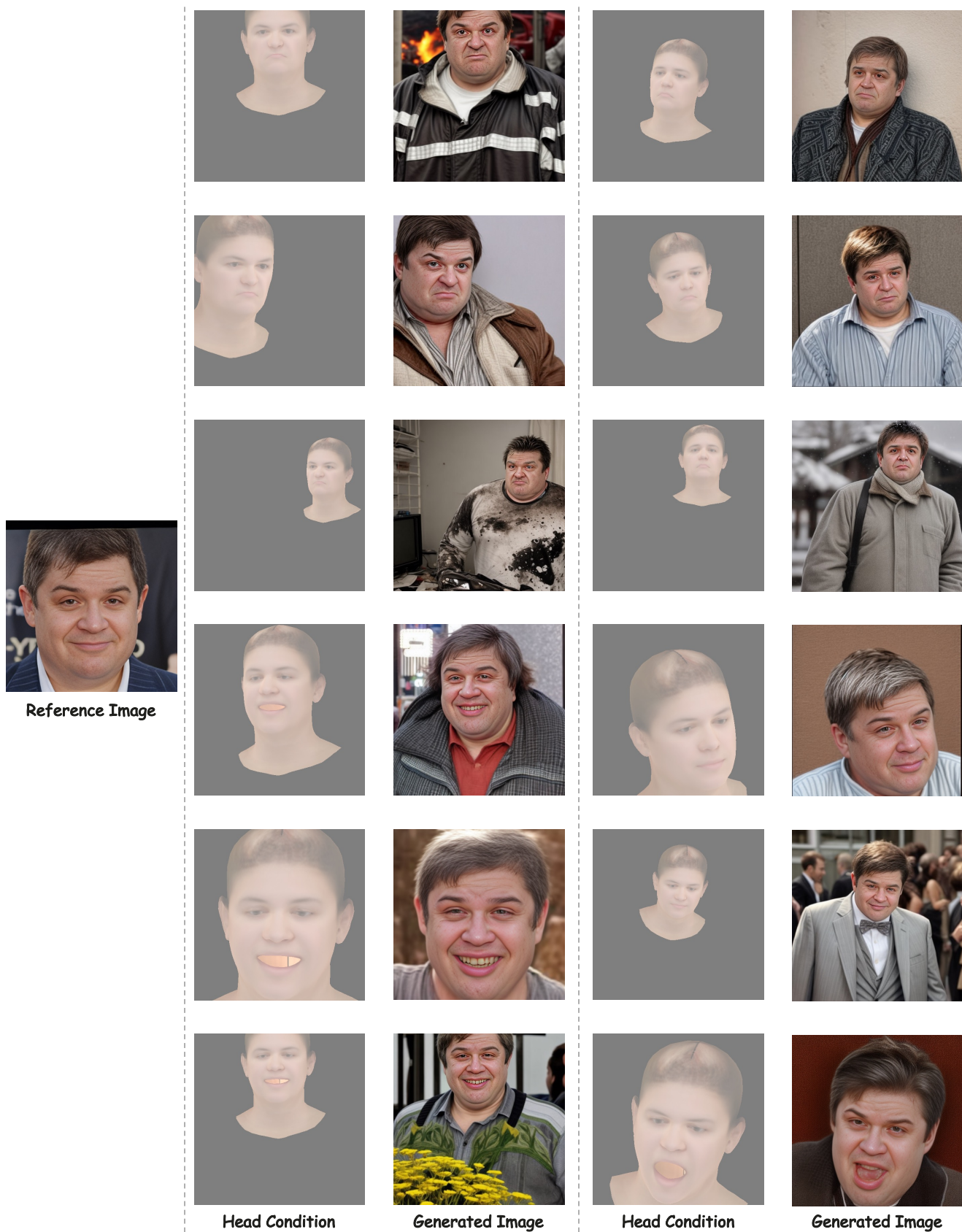


Figure 12. More results with different and rich facial expressions. Our CapHuman can provide facial expression control in a flexible and fine-grained manner.

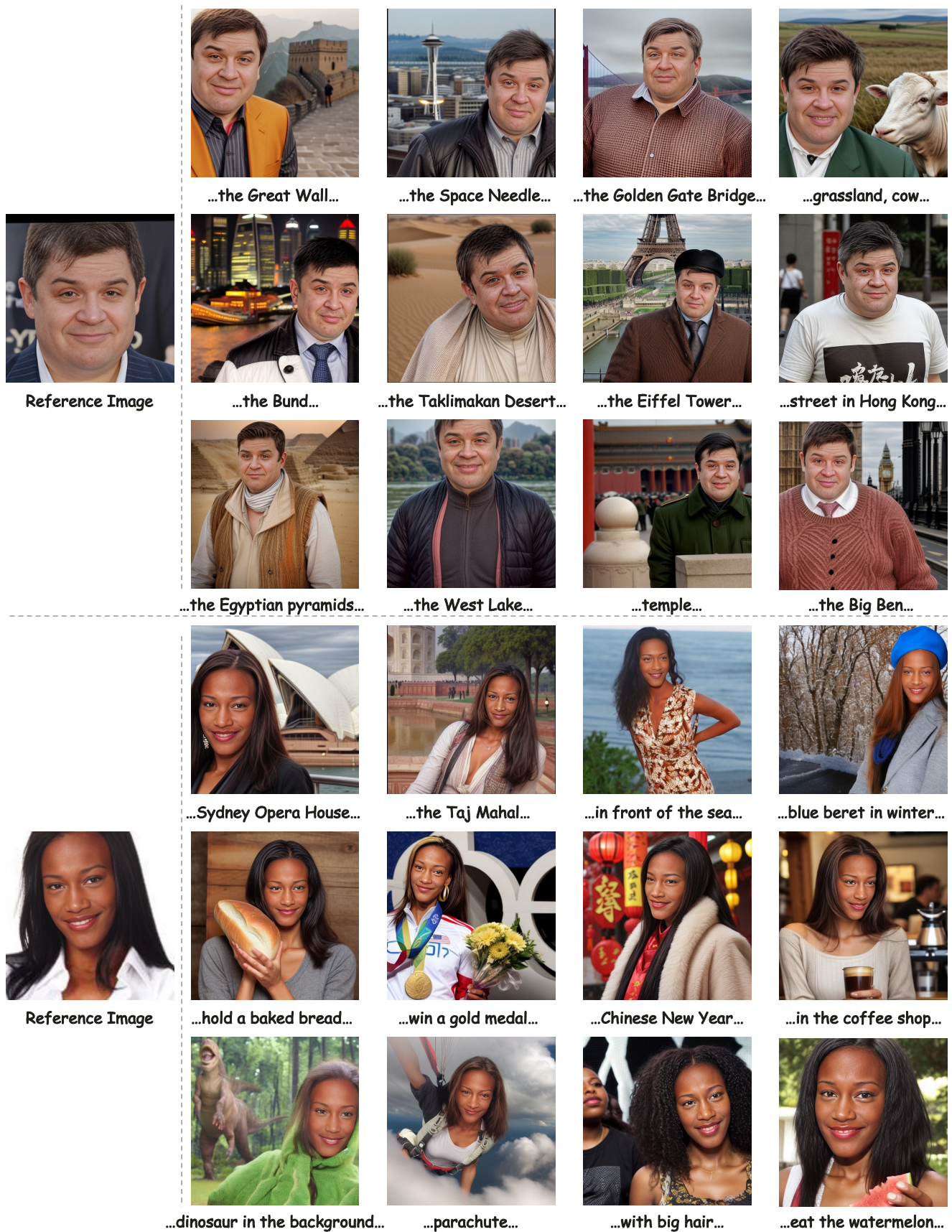


Figure 13. **More results in the realistic style.** Our CapHuman can be adapted to produce various identity-preserved and photo-realistic portraits with diverse head positions, poses, and facial expressions.

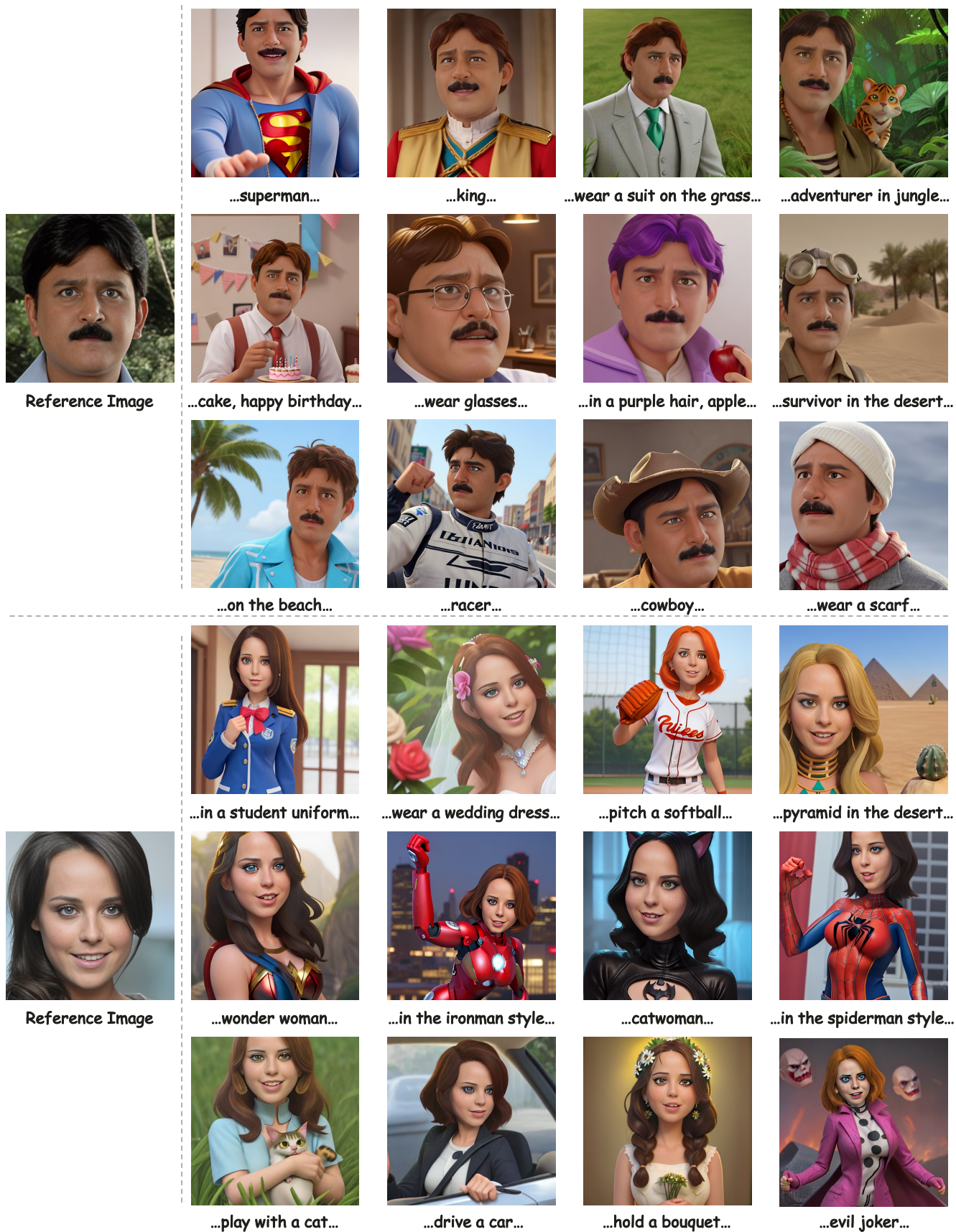


Figure 14. **More results in the Disney cartoon style.** Our CapHuman can be adapted to produce various identity-preserved portraits with diverse head positions, poses, and facial expressions.

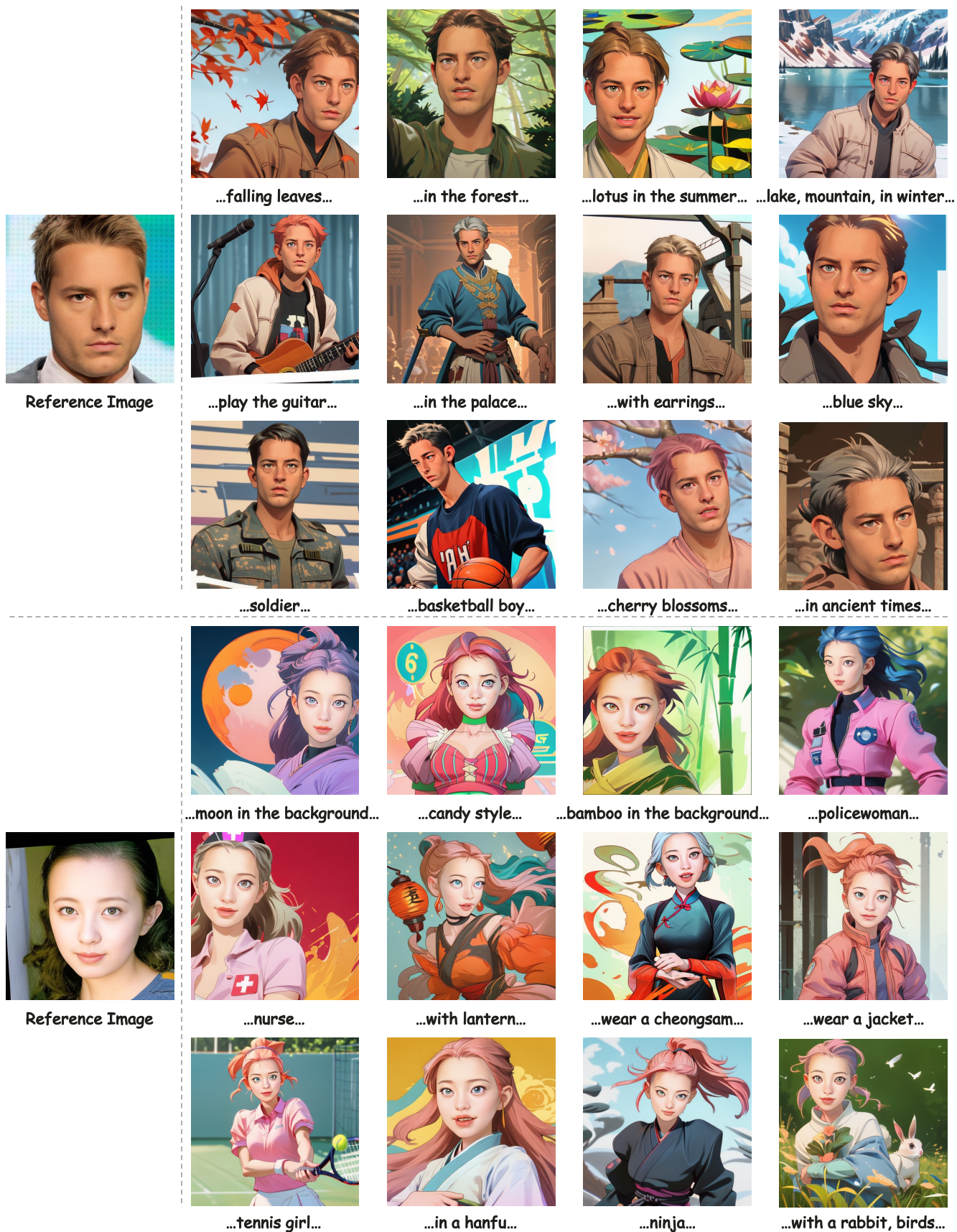


Figure 15. **More results in the animation style.** Our CapHuman can be adapted to produce various identity-preserved portraits with diverse head positions, poses, and facial expressions.

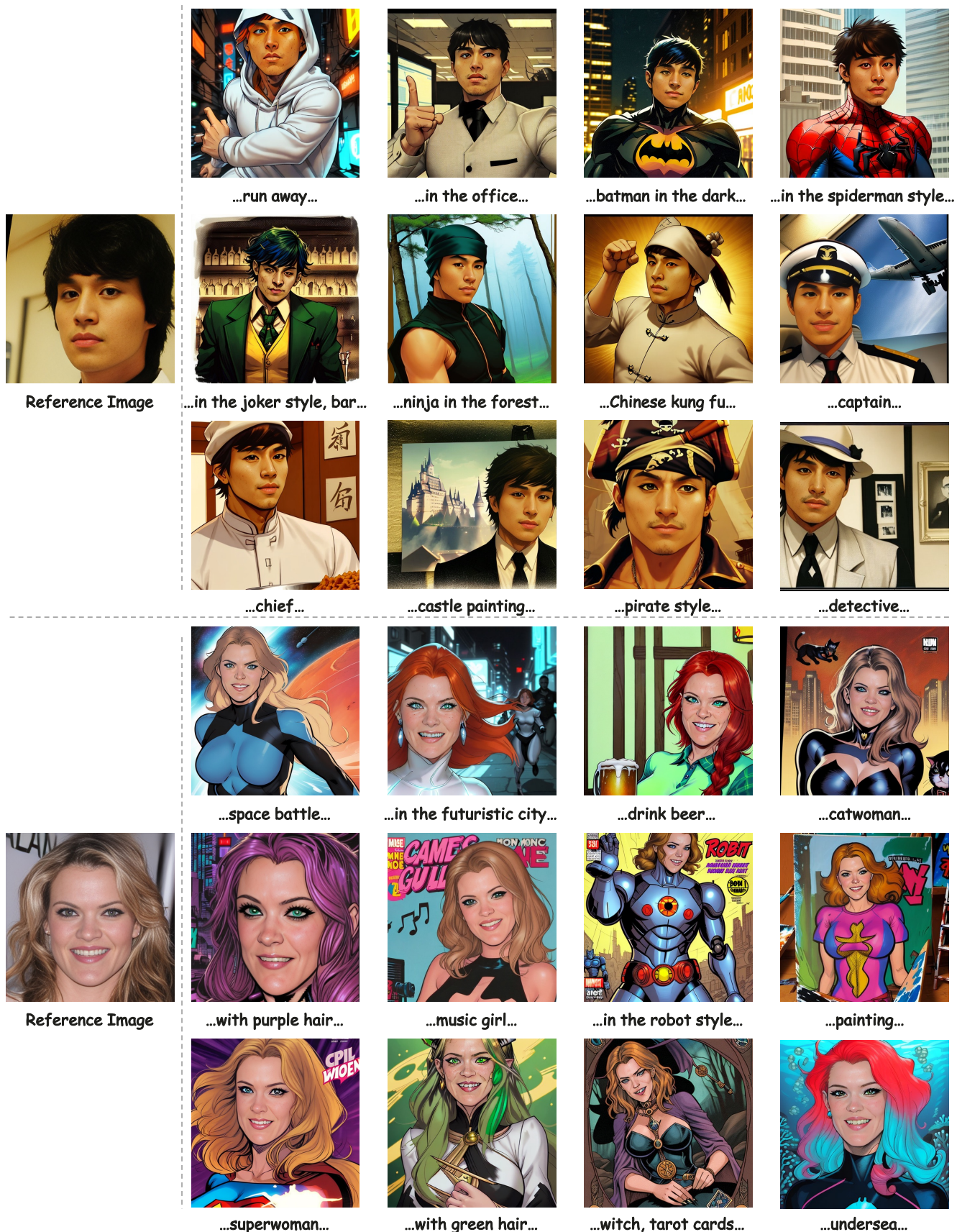


Figure 16. **More results in the comic style.** Our CapHuman can be adapted to produce various identity-preserved portraits with diverse head positions, poses, and facial expressions.

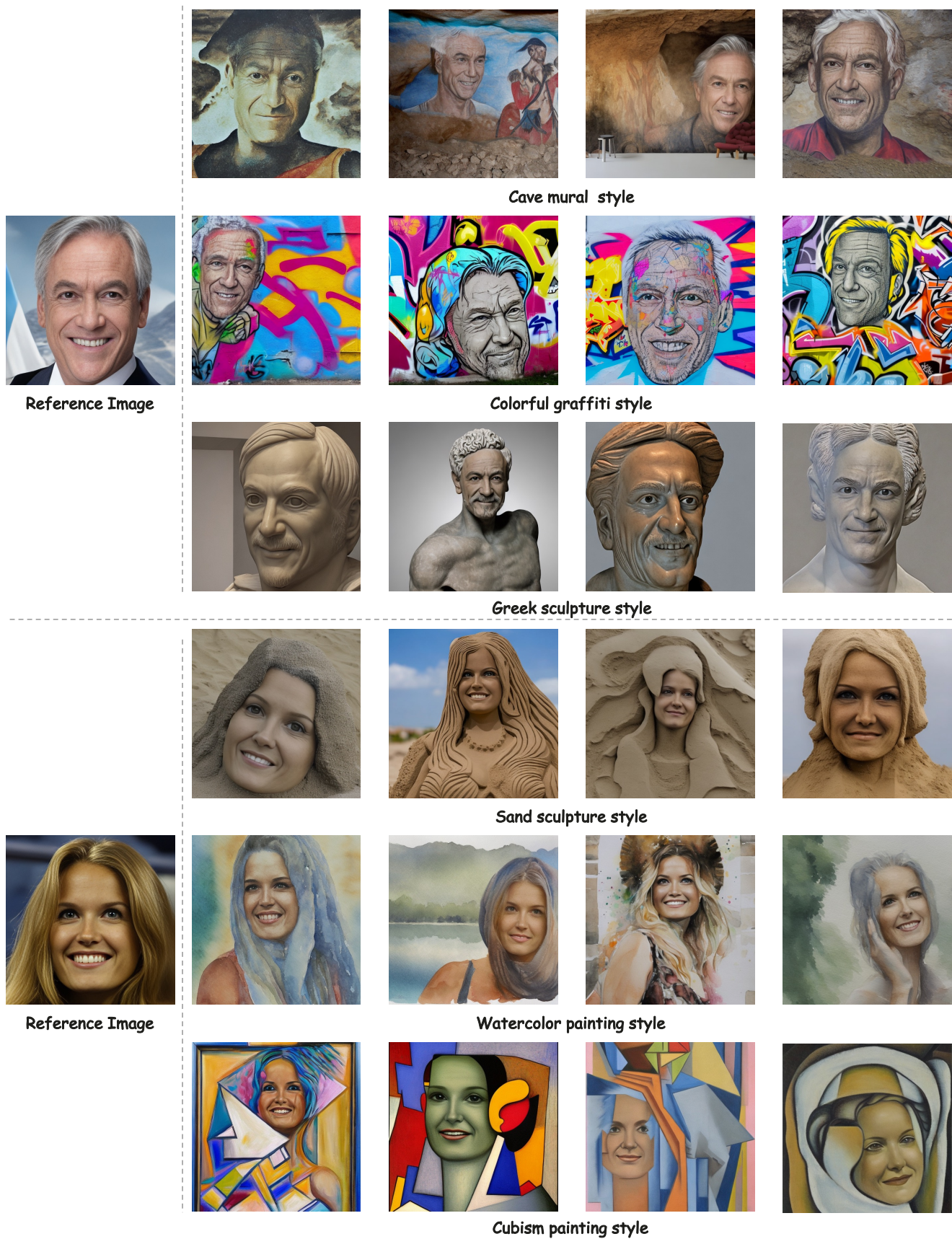


Figure 17. **Stylization by style prompts.** Our CapHuman can generate identity-preserved portraits with different styles by style prompts.

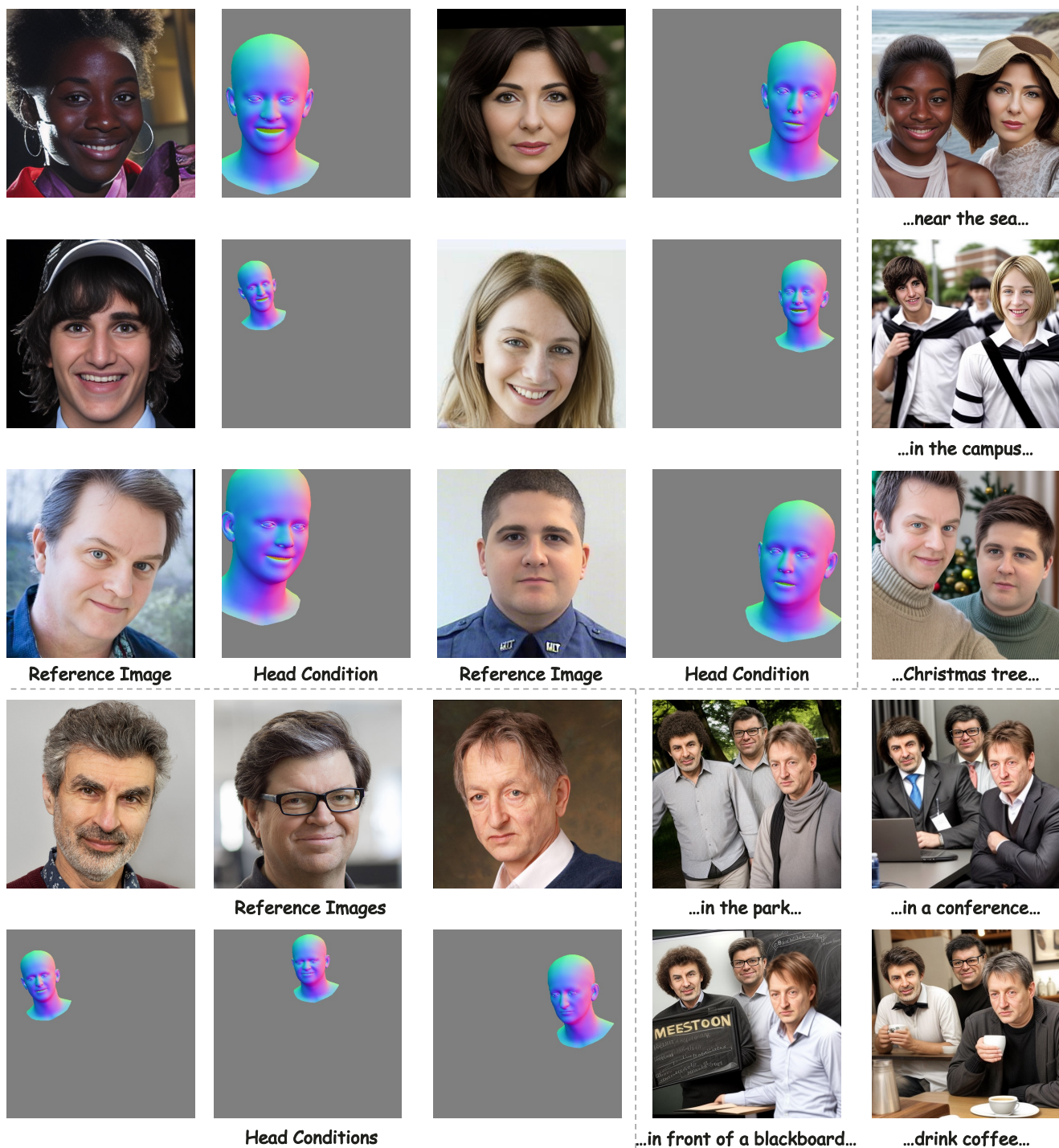


Figure 18. **Multi-Human image generation.** Given reference images, our CapHuman can generate various identity-preserved multi-human images, consistent with the corresponding head conditions.

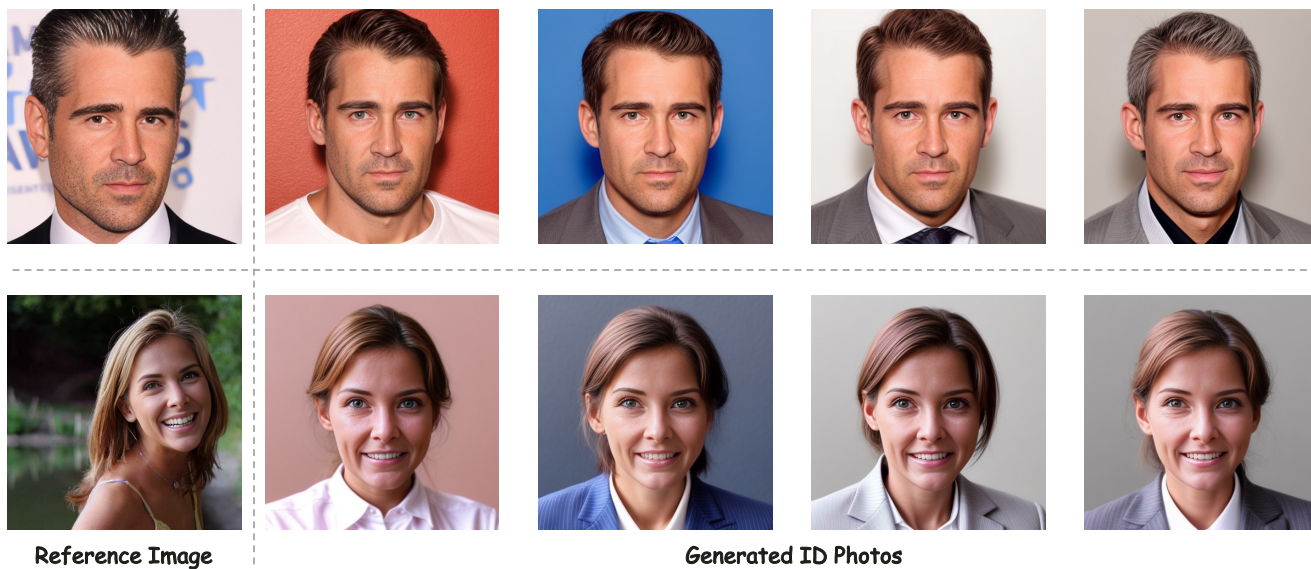


Figure 19. **Photo ID generation.** Our CapHuman can generate standard ID photos by adjusting the head conditions and providing the proper prompts.

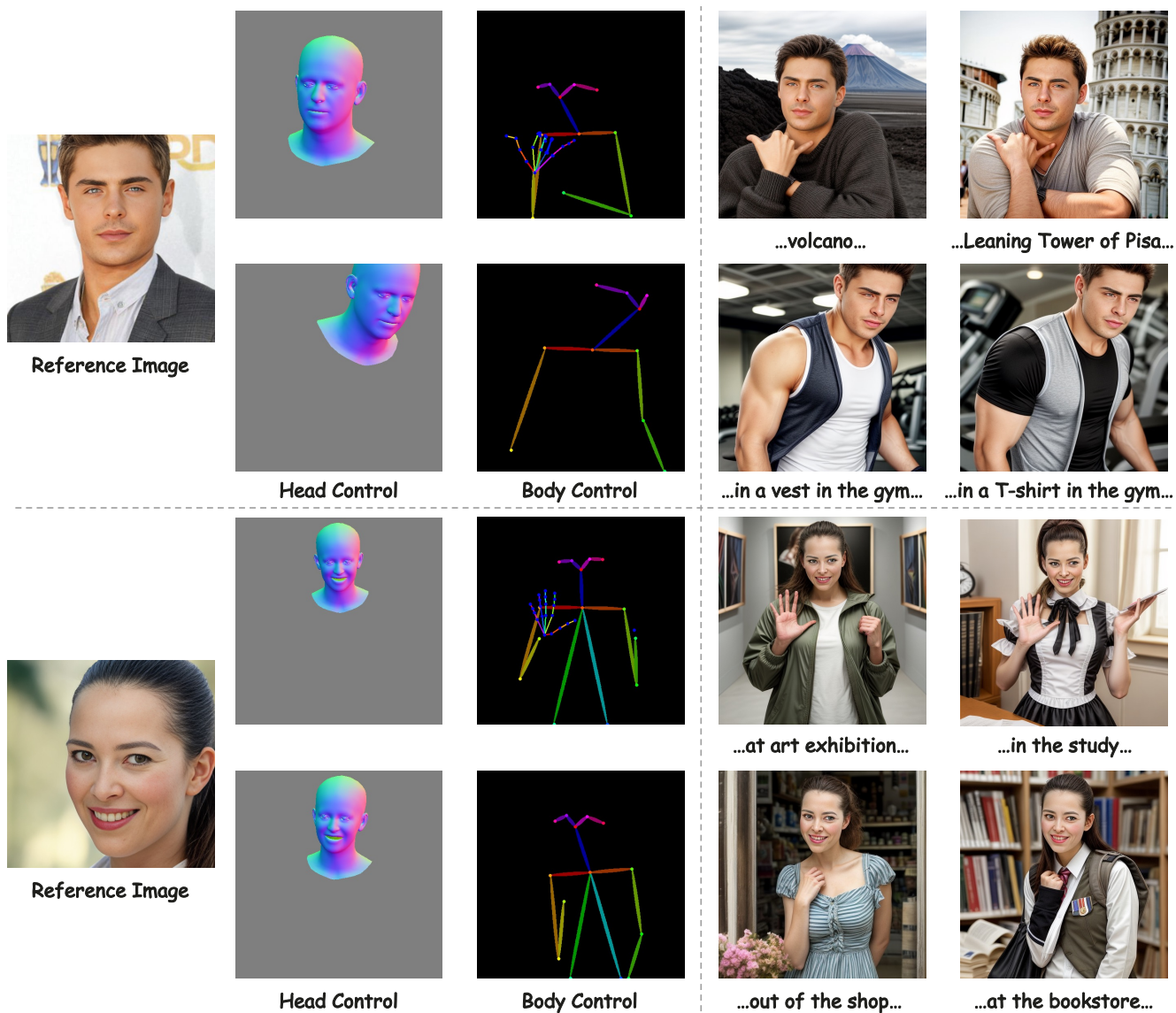
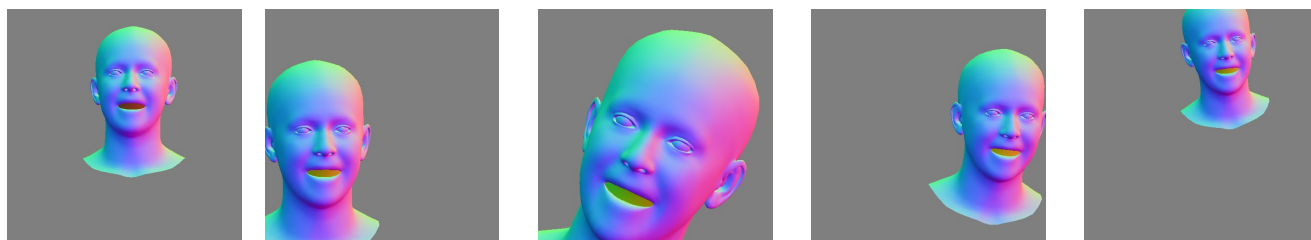
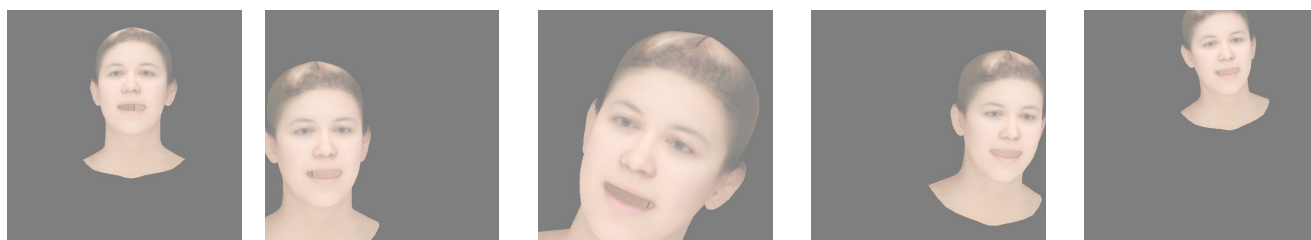
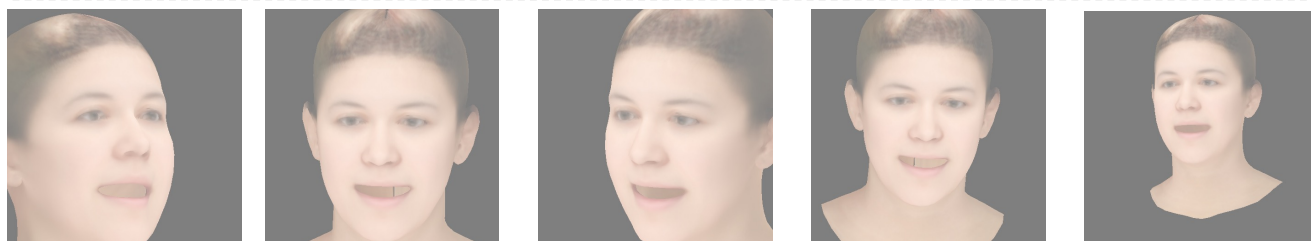


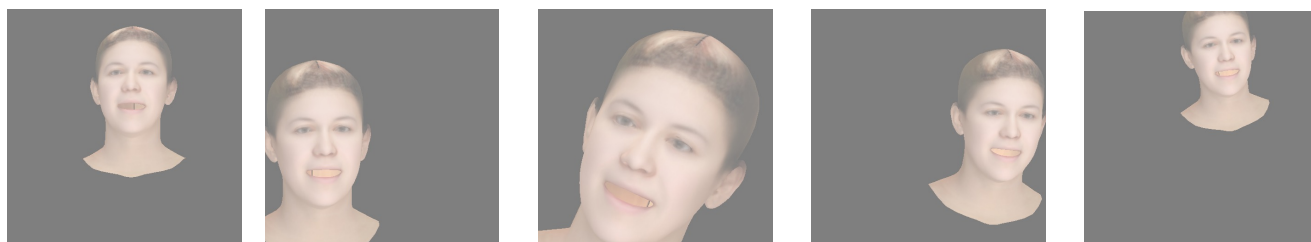
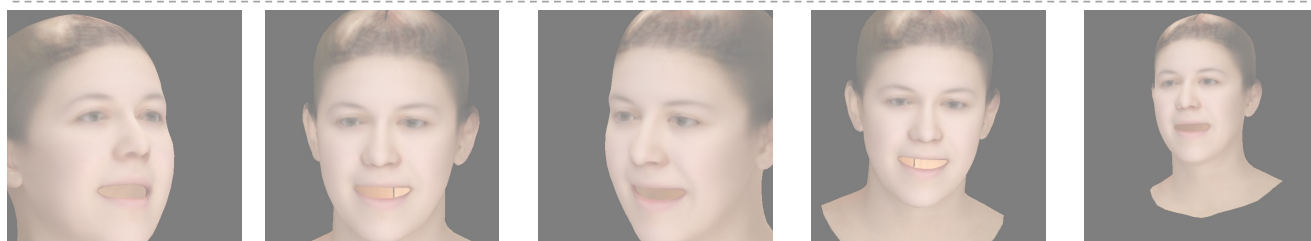
Figure 20. **Simultaneous head and body control with identity preservation.** Our CapHuman can control the head and body simultaneously with the pose-guided ControlNet [42] with identity preservation.



Surface Normals



Albedos



Lambertian renderings

Figure 21. **Head Conditions.** We list the head conditions of a specific individual in our HumanIPHC benchmark, including Surface Normals, Albedos, and Lambertian renderings.