This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CapHuman: Capture Your Moments in Parallel Universes

Chao Liang[1] Fan Ma[1] Linchao Zhu[1†] Yingying Deng[2] Yi Yang[1]
[1]ReLER, CCAI, Zhejiang University [2]Huawei Technologies Ltd.
[†] Corresponding author
{cs.chaoliang, zhulinchao, yangyics}@zju.edu.cn, flower.fan@foxmail.com, dyy15@outlook.com
https://caphuman.github.io

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, facial expressions, and illuminations 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://github.com/VamosC/CapHuman.

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 [29, 33–35, 37, 51], 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 [12, 16, 36] 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 [10, 28, 49] 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 [35], 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, 46, 53]. 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 propose a novel human-centric image synthesis task that generates specific individual portraits with various head positions, poses, facial expressions, and illuminations in different contexts given one reference 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 [39] and the support of powerful computation resources, large generative models bloom in abundance. One of the pathways is driven by diffusion

models. Diffusion models [15] are easily scalable without instability and mode collapse of adversarial training [14]. They have achieved amazing results in generating photo-realistic and content-rich images with high fidelity. Imagen [37], GLIDE [29], and DALL-E 2 [34] directly operate the denoising process in the pixel space. Instead, Stable Diffusion [35] 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 [48] or masked generative modeling [9]. 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 [12] introduces a new word embedding for the user-provided concept. However, too few parameters limit the expressiveness of the output space. DreamBooth [36] 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 [16] 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, 41, 43] 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

Text-conditioned methods [13, 17, 40, 45] have shown remarkable capability in human/avatar 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 [49] and T2I Adapter [28] 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 conditional generation but fail to preserve the identity. In addition, DiffusionRig [10] 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 [35] 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 [6, 25, 31, 44, 47, 50, 52]. 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 [42] to understand the semantics of text prompts. Put it together, the denoising objective can be formulated as follows:

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

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 Models (3DMM) [5, 7, 8, 22, 30]. 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, facial expressions, and illuminations 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

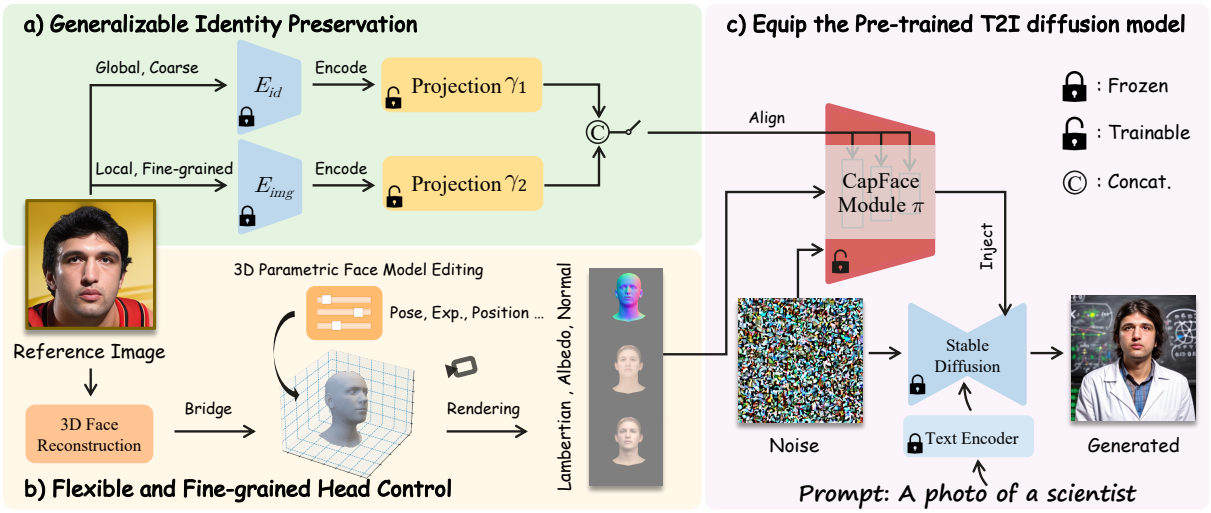

Figure 2. **Overview of CapHuman.** Our CapHuman stands upon the pre-trained T2I diffusion model. a) We embrace the "encode then learn to align" paradigm for generalizable identity preservation. b) The introduction of the 3D parametric face model enables flexible and fine-grained head control. c) We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

produce a new image $\hat{I}$. The pipeline can be defined as:

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

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 [34, 35, 37] 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 [35]. 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. Our pipeline exhibits several advantages: *well-generalizable* identity preservation that needs no time-consuming fine-tuning for each new individual, *3D-consistent* head control that incorporates 3DMM to support fine-grained control, and *plug-and-play* property that is compatible with rich off-the-shelf base models. § 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 [12, 16, 36] 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 one 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 [38] 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 [32] 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 segmentation [23, 24] 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. As identity features contain high-level semantic information, we inject them like Stable Diffusion [35] treats the text. We embed the global and local features into the latent identity feature $\mathbf{f}_{id}$:

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

where γ_1, γ_2 are projection layers and $[;]$ denotes the concatenation operation. Then, the latent identity feature is processed by the cross-attention mechanism [42], attending to the latent feature $\mathbf{f}_l$ in π, as formulated in the following way:

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

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 regarding the generation of diverse human images with varying head positions, poses, facial expressions, and illuminations. We aim to unlock its capability by introducing an appropriate control signal as a trigger. The first question is: what constitutes a good representation for this signal?

Bridge 3D facial representation. We pay attention to the popular 3DMM FLAME [22]. It constructs a compact latent space to represent the shape, pose, and facial expression separately. It provides a friendly and flexible interface to edit the facial geometry, *e.g.* changing the head pose, and facial expression with varied parameters. In our setting, we bridge the input reference image I and the 3D facial representation. We use DECA [11] 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 [10].

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:

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

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. Considering the low-level characteristics of head control and plug-and-play property, we adopt the side network design like ControlNet [49]. CapFace module π shares a similar structure with the Stable Diffusion encoder. The feature map is element-wise aligned with that in the decoder part of Stable Diffusion for each layer. 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:

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

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. 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 [15], 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 strategy in the following:

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

where t is the timestep in the diffusion process, 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 $\pi^\star$ at inference:

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

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

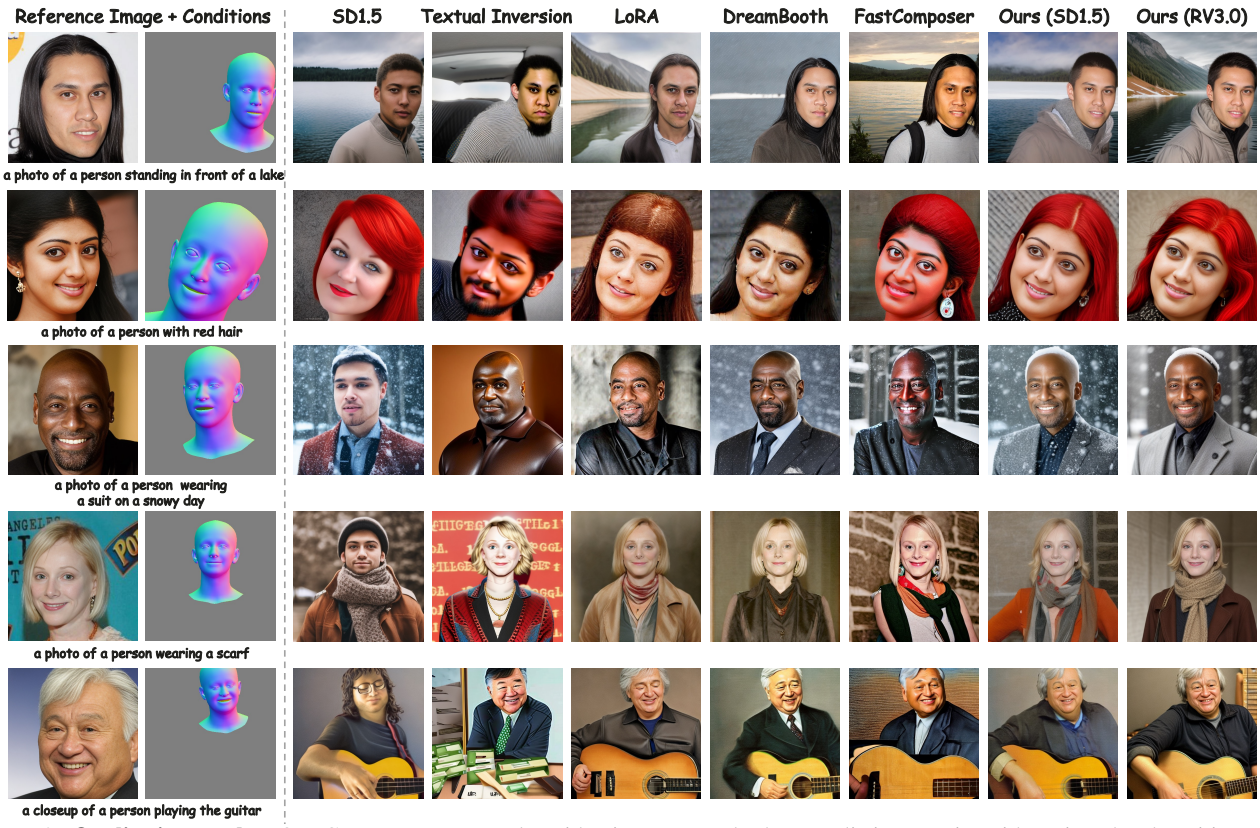

Figure 3. **Qualitative results.** Our CapHuman can produce identity-preserved, photo-realistic portraits with various head positions and poses in different contexts. Also, our model can be combined with the pre-trained model, *e.g.* RealisticVision [1] in the community flexibly.

4. Experiments

4.1. Training setup

We train our model on CelebA [26], 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 $\times$ 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 [32] image encoder. Our model is based on Stable Diffusion V1.5 [35]. The learning rate is 0.0001 and the batch size is 128. We use AdamW [27] for the optimization.

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 [12], DreamBooth [36], LoRA [16] and FastComposer [43]. These methods are designed for personalization and lack of head control. For fair comparisons, we combine them with ControlNet [49], since ControlNet can provide facial landmark-driven control. Also, landmark-guided ControlNet [49] is one of our baselines. The visual qualitative results are presented in Figure 3. Obviously, landmark-guided ControlNet 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.

Head control capability. Figure 4 shows the head control capability of our CapHuman. The results demonstrate our CapHuman can offer 3D-consistent control over the human head in diverse positions, poses, facial expressions, and illuminations. More results can be found in the appendix.

Adapt to other pre-trained models. The plug-and-play property enables our model can be adapted to other pre-trained models [2–4] in the community seamlessly. The results are presented in Figure 5. More visual results with more styles can be found in the appendix.

4.3. Quantitative Analysis

Benchmark. We introduce a new challenging and comprehensive benchmark HumanIPHC for identity preserva-

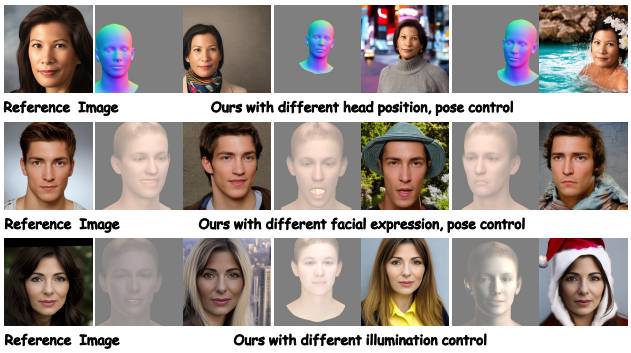

Figure 4. **Head position, pose, facial expression, and illumination control.** Our method offers the 3D-consistent head control.

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

tion, text-to-image alignment, and head control precision evaluation. We select 100 identities from the CelebA [26] test split. They consist of different ages, genders, and races. We collect 35 diverse prompts and 10 different head conditions with various positions and poses. Three different images are generated for each combination.

Evaluation metrics. We evaluate the effectiveness of our proposed method in the following three dimensions: (1) Identity Preservation. We apply a face recognition network [38] 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 [32] 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. We check whether the prompt with the largest CLIP score is the prompt used to generate or not. (3) Head Control Precision. We compute the root mean squared error (RMSE) between the DECA [11] 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 Inversion [12], LoRA [16], and DreamBooth [36] 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 [36] and FastComposer [43] by 15% 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.*, 5%, 7%, 7% compared with the second best results. We attribute this to the introduction of the 3D facial prior.

Method	Identity Preservation Generalizable	↑ ID sim.	Text-to-Image Alignment ↑ CLIP score	↑ Prompt acc.	Head Control Precision ↓ Shape	↓ Pose	↓ Exp.	↓ Light.
ControlNet [49]	✗	0.0534	**0.2479**	**90.32%**	0.2722	0.0494	0.3584	0.2718
Textual Inversion [12]	✗	0.4857	0.1561	13.70%	0.2075	0.0516	0.2530	0.2579
LoRA [16]	✗	0.5860	0.1897	35.96%	0.1648	0.0446	0.2039	0.1634
DreamBooth [36]	✗	0.6860	0.1873	39.21%	0.1542	0.0441	0.1922	0.1729
FastComposer [43]	✓	0.6191	0.2150	68.52%	0.1851	0.0611	0.2119	0.1861
Ours	✓	**0.8363**	0.2256	74.17%	**0.1020**	**0.0436**	**0.1241**	**0.0965**

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.

Method	↑ ID sim.
w/o global & local feat.	0.3915
w/o local feat.	0.7725
w/o global feat.	0.8095
w/ global & local feat.	**0.8429**

Table 2. **Ablation on ID features.**

Num. N	↑ ID sim.
32	0.8370
64	0.8376
128	0.8182
257	**0.8429**

Table 3. **Effect of N.**

Figure 6. **Visual results of global and local identity features.** Both global and local features contribute to identity preservation.

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

Method	↓ Shape	↓ Pose	↓ Exp.	↓ Light.
w/o 3DMM	0.2909	0.0501	0.3967	0.2899
w/ 3DMM (Ours)	**0.1381**	**0.0262**	**0.1639**	**0.1196**

Table 4. **Ablation on 3DMM.** Ours with 3DMM achieves significant improvement in head control precision.

Figure 7. **Visual comparison on 3DMM.** Ours with 3DMM shows more fine-grained control results with local details.

the identity features in Figure 6. We can observe that our model cannot preserve the individual identity if no identity 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. It is better to make full use of the local identity features in human face image generation.

Ablation on 3DMM. We validate the effectiveness of 3DMM. We remove the identity preservation module. Table 4 shows the results. With 3DMM, our method shows significant improvement in head control precision. The introduction of the 3D facial representation brings more information such as local geometry and illumination. Figure 7 confirms the 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 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 8.

Method	↑ ID sim.	↓ Shape	↓ Pose	↓ Exp.	↓ 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	↑ ID sim.	↓ Shape	↓ Pose	↓ Exp.	↓ Light.
w/o Post-hoc Enhan.	**0.8429**	0.1132	0.0564	0.1349	0.1047
+ w/o 3DMM model	0.8386	0.1118	0.0427	0.1377	0.1032
+ w/ 3DMM model	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.

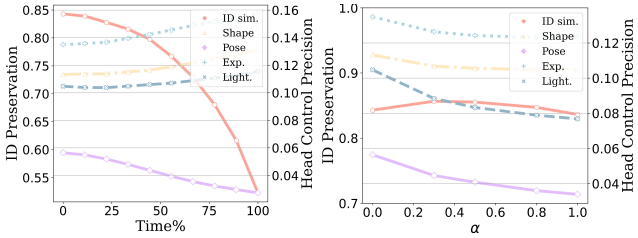

Figure 8. **Left: The utilization time (%) of the head control model at inference.** Using the head control model at the early stage can improve the pose control but sacrifice the identity similarity. **Right: Ablation on the control scale α.** With the control scale α increasing, head control metrics are improved at a negligible cost of identity preservation.

It improves the pose metric by sacrificing the ID preservation capability. Second, we study the effect of fusion with different head control models. Specifically, we set $\pi^\star = \varnothing$, or w/o 3DMM model, or w/ 3DMM model in Eq. 9. 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 8 shows the head control model can strengthen the pose control at a negligible loss of identity.

5. Conclusion

In this paper, we propose a novel framework CapHuman for the human-centric image synthesis with generalizable identity preservation and fine-grained head control. We embrace the "encode then learn to align" paradigm for generalizable identity preservation capability without further cumbersome fine-tuning. By incorporating the 3D facial representation, it enables flexible and 3D-consistent head control. Given one reference face image, our CapHuman can generate well-identity-preserved, high-fidelity, and photo-realistic human portraits with diverse head positions, poses, facial expressions, and illuminations in different contexts.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (T2293723, 62293554, U2336212).

References

[1] Realistic vision v3.0. https://huggingface.co/SG161222/Realistic_Vision_V3.0_VAE, 2023. 6

[2] comic-babes. https://civitai.com/models/20294/comic-babes, 2023. 6

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

[4] toonyou. https://civitai.com/models/30240/toonyou, 2023. 6

[5] 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

[6] 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 *CVPR*, pages 22563–22575, 2023. 3

[7] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, pages 5543–5552, 2016. 3

[8] 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

[9] 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

[10] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *CVPR*, pages 12736–12746, 2023. 2, 3, 5

[11] 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

[12] 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. 2, 3, 4, 6, 7

[13] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *CVPR*, pages 22634–22645, 2023. 3

[14] 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

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 3, 5

[16] 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

[17] Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. In *ACM MM*, pages 5734–5745, 2023. 3

[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 *CVPR*, pages 10124–10134, 2023. 3

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 3, 6

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 6

[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] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Guangliang Cheng, Pang Jiangmiao, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *arXiv pre-print*, 2023. 4

[24] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *CVPR*, 2024. 4

[25] 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 *ICCV (ICCV)*, pages 9298–9309, 2023. 3

[26] 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. 6, 7

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[28] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu 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

[29] 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

[30] Pascal Paysan, Reinhard Knothe, 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

[31] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *CVPR*, 2024. 3

[32] 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 *ICML*, pages 8748–8763. PMLR, 2021. 4, 6, 7

[33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 2

[34] 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

[35] 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. 2, 3, 4, 6

[36] 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. 2, 3, 4, 6, 7

[37] 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. *NeurIPS*, 35:36479–36494, 2022. 2, 3, 4

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 4, 7

[39] 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. *NeurIPS*, 35:25278–25294, 2022. 2

[40] Xiaolong Shen, Jianxin Ma, Chang Zhou, and Zongxin Yang. Controllable 3d face generation with conditional style code diffusion. In *AAAI*, 2024. 3

[41] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 3

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3, 5

[43] 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

[44] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *CVPR*, pages 20908–20918, 2023. 3

[45] Yuanyou Xu, Zongxin Yang, and Yi Yang. Seeavatar: Photorealistic text-to-3d avatar generation with constrained geometry and appearance. *arXiv preprint arXiv:2312.08889*, 2023. 3

[46] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021. 2

[47] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392*, 2024. 3

[48] 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

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 5, 6, 7

[50] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 2024. 3

[51] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *IJCAI*, 2023. 2

[52] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *CVPR*, 2024. 3

[53] Zhenglin Zhou, Fan Ma, Hehe Fan, and Yi Yang. Headstudio: Text to animatable head avatars with 3d gaussian splatting. *arXiv preprint arXiv:2402.06149*, 2024. 2