GUAVA: Generalizable Upper Body 3D Gaussian Avatar

Dongbin Zhang[1,2*] Yunfei Liu[2†] Lijian Lin[2] Ye Zhu[2] Yang Li[1]
Minghan Qin[1] Yu Li[2‡] Haoqian Wang[1†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]International Digital Economy Academy (IDEA)

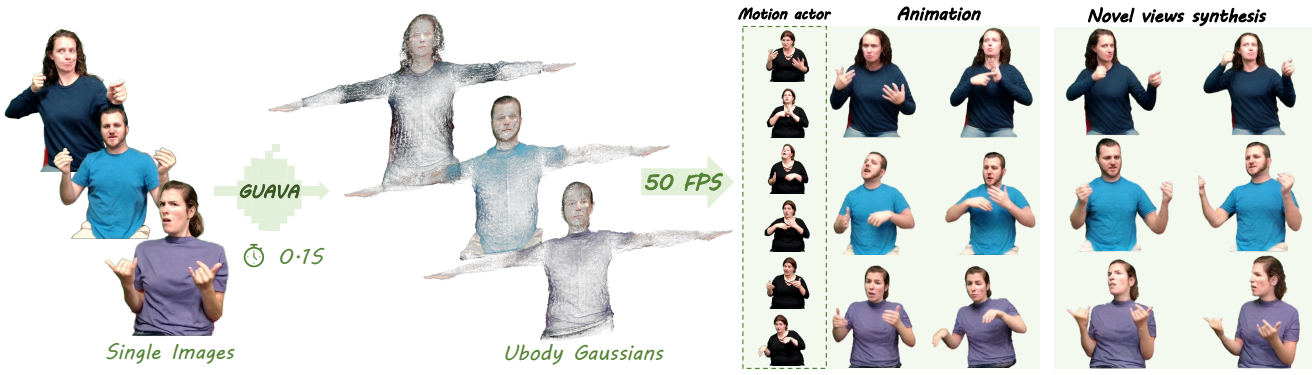

Figure 1. From a single image with a tracked pose, GUAVA can reconstruct a 3D upper-body Gaussian avatar via feed-forward inference within sub-second time, enabling real-time expressive animation and novel view synthesis at 512×512 resolution.

Abstract

Reconstructing a high-quality, animatable 3D human avatar with expressive facial and hand motions from a single image has gained significant attention due to its broad application potential. 3D human avatar reconstruction typically requires multi-view or monocular videos and training on individual IDs, which is both complex and time-consuming. Furthermore, limited by SMPLX's expressiveness, these methods often focus on body motion but struggle with facial expressions. To address these challenges, we first introduce an expressive human model (EHM) to enhance facial expression capabilities and develop an accurate tracking method. Based on this template model, we propose GUAVA, the first framework for fast animatable upper-body 3D Gaussian avatar reconstruction. We leverage inverse texture mapping and projection sampling techniques to infer Ubody (upper-body) Gaussians from a single image. The rendered images are refined through a neural refiner. Experimental results demonstrate that GUAVA significantly outperforms previous methods in rendering quality and offers significant speed improvements, with reconstruction times in the sub-second range (∼ 0.1s), and supports real-time animation and rendering. Code and video demos are available at the Project page.

1. Introduction

Creating realistic and expressive upper-body human avatars is essential for applications like films, games, virtual meetings, and digital media. These avatars are expected to exhibit high realism with expressiveness, such as detailed facial expressions and rich hand gestures. Efficiency, ease of creation, and real-time rendering are also critical. However, achieving these goals, especially from a single image, remains a significant challenge in computer vision.

Recently, this area has gained growing attention, with progress in both 2D and 3D-based methods. Several works [13, 25, 50, 51, 60] directly predict the coefficients of the human template model SMPLX [65] from images, enabling rapid human mesh reconstruction. However, due to the template model's coarse textures, these methods struggle to render photorealistic images. With the breakthrough of NeRF [59] in novel view synthesis, methods like [27, 39, 54, 68, 113] combine template models like FLAME [47] or SMPLX to reconstruct high-quality head or whole body avatars from multi-view or monocular videos. While achieving re-

* Intern at IDEA. ‡ Project lead. † Corresponding authors.

alistic effects, these methods suffer from slow training and rendering speeds. The rise of 3D Gaussian splatting (3DGS) [41] has led to methods like [34, 48, 70, 81, 97], which leverage 3DGS for real-time, high-quality avatars reconstruction. However, these methods still exist several limitations. *e.g.*, **ID-specific training:** Each person requires individual training; **Training complexity:** The process is time-consuming and requires calibrated multi-view or monocular videos; **Limited expressiveness:** Head reconstruction methods lack body motion representation, while full-body methods neglect detailed facial expressions.

In the 2D domain, diffusion-based models [4, 8, 9, 43] have demonstrated remarkable results in video generation. Meanwhile, ControlNet [107] has expanded the controllability of generative models by adding extra conditions to guide the generation process of stable diffusion model (SD) [77]. Consequently, some researchers [33, 40, 91] adapt it for human motion transfer. They use conditions like keypoints or SMPLX to transfer the target image's human pose to the source image's ID, enabling the creation of human animation videos. Despite these methods achieving good visual effects, they still face several limitations. *e.g.*, **ID consistency:** They struggle to maintain a consistent ID without 3D representations, especially with large pose changes; **Efficiency:** High computational costs and multiple denoising steps result in slow inference, hindering real-time applications; **Viewpoint control:** 2D methods cannot easily adjust the camera pose, limiting viewpoint control.

To address the above challenges, we propose GUAVA, a framework for creating upper-body 3D Gaussian avatars from single images, as shown in Fig. 1. GUAVA enables fast reconstruction of animatable, expressive avatars, supporting real-time pose, expression control, and rendering. Unlike traditional 3D avatar reconstruction methods, GUAVA completes the reconstruction in a single forward pass. Compared to 2D-based methods, we use 3D Gaussians for consistent avatar representation in canonical space, overcoming issues of ID consistency and enabling real-time rendering.

Like other 3D-based body reconstruction methods, we rely on the human template model SMPLX to construct the upper-body avatar. This requires aligning each image with the template model via tracking [50, 60]. However, current SMPLX parameter estimation methods struggle with accurately tracking hand movements and fine-grained facial expressions. Besides, SMPLX has limited facial expression capability. Therefore, we introduce EHM (Expressive Human Model), combining SMPLX and FLAME, and develop an optimization-based tracking method for accurate parameter estimation. Based on these tracked results, we design a reconstruction model with two branches: one uses the EHM's vertices and their projection features to create template Gaussians, while the other applies inverse texture mapping to transfer screen-space features into UV space for decoding UV Gaussians. This approach reduces the sparsity of template Gaussians and captures finer texture details. To further improve rendering quality, each Gaussian is equipped with a latent feature to generate a coarse map, which is then refined using a learning-based refiner.

We train GUAVA on monocular upper-body human videos with diverse IDs to ensure good generalization for unseen IDs during inference. Extensive experiments show that GUAVA outperforms previous 2D- and 3D-based methods in visual effects. Additionally, our method reconstructs the upper body in $\sim \mathbf{0.1}s$, and supports real-time animation and rendering. Our main contributions are as follows:

- We propose GUAVA, the first framework for generalizable upper-body 3D Gaussian avatar reconstruction from a single image. Using projection sampling and inverse texture mapping, GUAVA enables fast feed-forward inference to reconstruct Ubody Gaussians from the image.
- We introduce an expressive human template model with a corresponding upper-body tracking framework, providing an accurate prior for reconstruction.
- Extensive experiments show that GUAVA outperforms existing methods in rendering quality, and significantly outperforms 2D diffusion-based methods in speed, offering fast reconstruction and real-time animation.

2. Related work

2.1. 3D based Avatar Reconstruction

Traditional human or head reconstruction primarily focuses on mesh reconstruction. Researchers have constructed human template models like SMPL [56], the facial model FLAME [47], and the hand model MANO [78] based on thousands of 3D scans. These models are mesh-based and represent shape and expression variations in a linear space, with rotational movement modeled using joints and linear blend skinning. SMPLX [65] extends SMPL by integrating FLAME and MANO, enhancing facial and hand expressiveness. Since SMPLX is trained on full-body scans and may overlook facial details, its expressive capability for face is still limited compared to FLAME. By combining deep learning, some methods [11, 16, 22, 26, 29, 42, 75, 80, 103, 109] predict the parameters of human, hand, and face models to achieve fast image-to-mesh reconstruction.

With the development of Neural Radiance Fields (NeRF) [59], studies have combined NeRF with template models for more realistic 3D reconstructions of heads [1, 2, 73, 111] or humans [46, 90, 95]. NeRFcae [27] extends NeRF to a dynamic form by introducing expression and pose parameters as conditioning inputs for driveable avatar reconstruction. Neural Body [69] proposes a new representation for dynamic humans by encoding posed human meshes into latent code volumes. Recently, 3DGS [41] has made breakthroughs in real-time rendering, leading to its rapid applica-

tion in various fields [28, 37, 72, 88, 105], leveraging its efficient rendering capabilities, including avatar reconstruction [35, 53, 70, 79, 104, 106]. GART [45] uses 3D Gaussians to represent deformable subjects, employing learnable blend skinning to model non-rigid deformations and generalizing to more complex deformations with novel latent bones. GaussianAvatar [34] learns pose-dependent Gaussians using a 2D pose encoder to represent human avatar. ExAvatar [61] combines mesh and Gaussian representations, treating each Gaussian as a vertex in the mesh with pre-defined connectivity, enhancing facial expression expressiveness.

However, these methods require training for each individual ID and lack generalization. To address this, some works have explored generalizable networks for single image to avatar reconstruction [15, 19, 20]. GAGAvatar [14] combines the FLAME and introduces a dual-lifting method for inferring 3D Gaussians from a single image, enabling feed-forward reconstruction of head avatars. Yet, single-image 3D human reconstruction remains underexplored. Methods like [10, 38, 63, 98] achieve avatar reconstruction from a one-shot image but rely on generating a series of images through generative models, followed by optimization for reconstruction, which is time-consuming. Other methods [36, 49, 112] focus on generalizable models for quick novel view synthesis from sparse images. In this context, we propose GUAVA, to the best of our knowledge, the first method enables fast and generalizable upper-body avatar reconstruction from a single image, while also addressing the lack of body expressiveness in generalizable head avatars.

2.2. 2D based Human Animation

Early motion transfer frameworks are primarily based on Generative Adversarial Networks (GANs) [30], such as those for facial expression transfer [32, 71, 83, 92] and human motion transfer [21, 52, 82, 87, 94]. These methods typically achieve motion transfer by warping the source image's feature volume according to the target motion. However, they often generate artifacts in unseen regions and struggle with generalizing to out-of-distribution images. Recently, diffusion models [76, 77, 85, 86, 89] have shown remarkable diversity, high generalization, and superior quality in image generation. By learning from vast image datasets, these models have developed strong priors, enabling them to generalize well to downstream tasks. Some works have integrated pre-trained image generation with temporal layers [31, 84, 93, 96] or transformer structures [3, 58, 100, 102] to achieve video generation. For human motion transfer tasks, these video or image generation models can provide powerful priors. Inspired by these advancements, several works have leveraged generative models to create controllable human animation videos based on pose information extracted from OpenPose [6] or DWPose [101]. For example, DreamPose [40], based on Stable Diffusion [77], integrates CLIP [74] and VAE to encode images, using an adapter to generate animation videos from pose sequences. MagicPose [7], also an SD-based model, introduces a two-stage training strategy to better decouple human motions and appearance. For more accurate pose control, Champ [115] integrates SMPL as a 3D guide, using depth, normal, and semantic maps as motion control conditions. MimicMotion [110] employs an image-to-video diffusion model as a prior, proposing a confidence-aware pose guidance to enhance generation quality and inter-frame smoothness, while utilizing a progressive fusion strategy for longer videos generation. Although these 2D-based methods ensure high-quality image synthesis, they all struggle with maintaining ID consistency. In contrast, our method uses 3D Gaussians as the 3D representation for avatars, which ensures better identity consistency.

3. Method

In Sec. 3.1, we introduce the EHM template model, which enhance SMPLX's [65] facial expressiveness, along with an upper-body tracking method. Sec. 3.2 explains how Ubody Gaussians are predicted from the source image. In Sec. 3.3, we deform the Ubody Gaussians into pose space using the tracked parameters and render them. Lastly, Sec. 3.4 outlines the training strategy and loss functions. The overall pipeline is illustrated in Fig. 2.

3.1. EHM and Tracking

3D body reconstruction often relies on models like SMPLX as priors. Accurately estimating shape and pose parameters to align each frame with SMPLX is essential. However, previous full-body pose estimation methods struggle with robustness and accuracy for wild images, making their results unreliable. Additionally, while SMPLX struggles to capture fine facial expressions [61], FLAME [47] excels at this but isn't compatible with SMPLX. This means parameters estimation results based on FLAME cannot be applied to SMPLX. To solve this, we propose EHM (Expressive Human Model), which replaces the SMPLX head part with FLAME for more accurate facial expression representation. Furthermore, we design the tracking in two stages: first, pre-trained models provide a coarse estimate, followed by refinement primarily using 2D keypoint loss.

Specifically, we start by using existing models [25, 55, 67] to coarsely estimate the shape parameters β_b and body pose parameters θ_b of SMPLX, hand pose parameters θ_h of MANO, and facial shape β_f, expression ψ_f, and jaw pose θ_{jaw} of FLAME. Then, we extract human body keypoints K_b and facial keypoints K_f using keypoint detection models [5, 17, 57, 101] for further fine-tuning. We optimize facial parameters of FLAME primarily using keypoint loss: $\beta_f, \psi_f, \theta_{jaw} = \text{argmin}\ \mathcal{L}_1(\hat{K_f}, K_f)$, where $\hat{K_f}$ are the keypoints from the FLAME. This yields

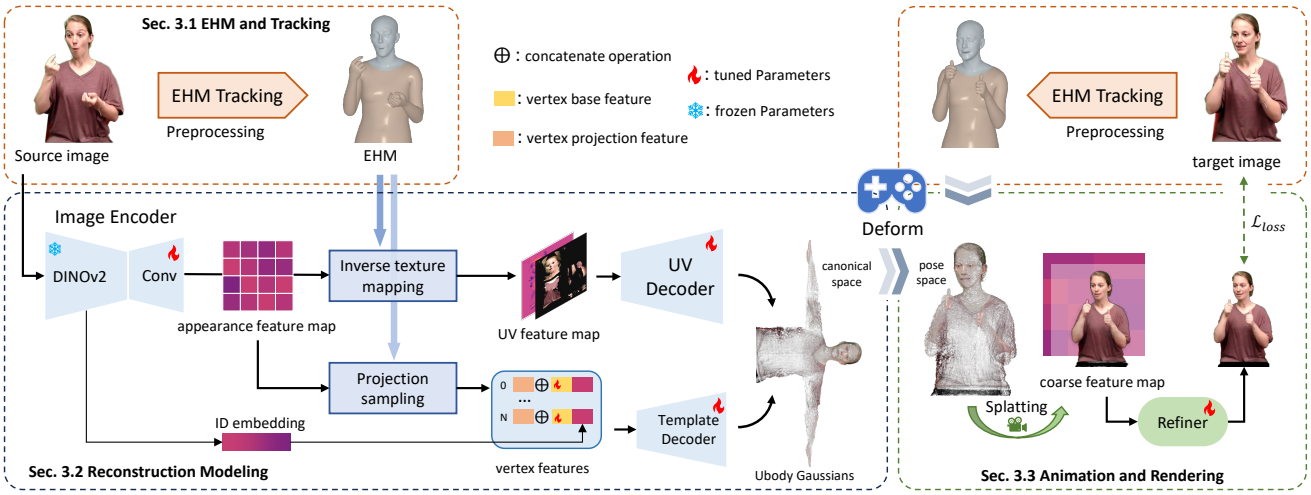

Figure 2. Given the source and target images, we first obtain the shape, expression, and pose parameters of the EHM template model through preprocessing tracking. The source image is then passed through an image encoder to extract an appearance feature map. Using these features and the tracked EHM, one branch predicts the template Gaussians, and the other predicts the UV Gaussians. These are concatenated to form the Ubody Gaussians in canonical space, which are then deformed into pose space using the tracked parameters from the target image. Finally, a coarse feature map is rendered and refined by a neural refiner to produce the final image.

a head model with expression in the neutral pose, denoted as $M_f(\beta_f, \psi_f, \theta_{jaw})$. Next, we replace the head of the SMPLX in the T-pose $M_b(\beta_f)$ with the head model $M_f(\beta_f, \psi_f, \theta_{jaw})$. The modified model then undergoes pose deformation through linear blend skinning ($\mathcal{LBS}$). Additionally, we introduce ID-specific joint offset parameters ΔJ for better joint alignment. Thus, the EHM model is formulated as:

$$M_{ehm} = \mathcal{LBS}\Big(\mathcal{U}\big(M_b(\beta_b), M_f(\beta_f, \psi_f, \theta_{jaw}) + \vec{e}\big), \theta_b, \theta_h, \mathcal{J}_b(\beta_b) + \Delta J\Big), \tag{1}$$

where $\mathcal{U}$ represents replacing the SMPLX head with the FLAME, aligned by a displacement vector $\vec{e}$, which represents the displacement between the eye joints of both models, and $\mathcal{J}_b(\beta_b)$ denotes the joints' position. We then fine-tune the parameters mainly using keypoint loss: $\beta_h,\ \beta_f,\ \theta_b,\ \theta_h,\ \Delta J = \text{argmin}\mathcal{L}_1(\hat{K}_b, K_b)$, where $\hat{K}_b$ are the keypoints from the EHM. Notably, our tracking method is designed to optimize multiple frames in parallel to meet the speed requirement.

3.2. Reconstruction Modeling

Unlike person-specific reconstruction methods, our model is designed for single-image feedforward inference, enabling fast upper-body avatar reconstruction. We represent the avatar in a canonical space using Gaussians [41], where each Gaussian includes position, rotation, scale, opacity, and a latent appearance feature: $G = \{\mu, r, s, \alpha, c\}$. The avatar consists of two parts: template Gaussians based on EHM, which are fewer in number and handle coarse texture and geometry modeling, and UV Gaussians rigged on the triangulated mesh to capture finer details.

Template Gaussians. Specifically, we first extract features from the image using pre-trained DINOV2 [62] and obtain a global ID embedding f_{id}. Convolutional upsampling is then applied to generate an appearance feature map F_a matching the source image's resolution. Since the person in the source image is posed, we use the tracked EHM in the pose space, project each vertex onto the screen space, and employ linear interpolation $\mathcal{S}$ to sample the projection features:

$$f_p^i = \mathcal{S}\left(F_a, \mathcal{P}\left(v^i, RT_s\right)\right), \tag{2}$$

where v^i represents the i-th vertex and RT_s is the viewing matrix of the source image. Additionally, for each vertex, we adopt an optimizable base feature f_b to learn unique semantic information. Combining these three features, a Template decoder $\mathcal{D}_T$, consisting of MLPs, is used to predict the template Gaussian attributes: $\{r^i, s^i, \alpha^i, c^i\} = \mathcal{D}_T(f_p^i \oplus f_b^i \oplus f_{id})$, where $\oplus$ denotes concatenation. For μ^i, we directly take the vertex position v^i.

UV Gaussians. Using only the EHM's vertices to represent the avatar can not fully capture high-frequency details due to the limited number of vertices, and struggles to represent regions beyond the template model. To address these issues, we propose introducing additional Gaussians to capture fine details. Specifically, we use the UV texture map to construct UV Gaussians, where we predict a Gaussian for each valid pixel in the texture map. Unlike template Gaussians, which are directly represented in world coordinates, as in [70], we rig each UV Gaussian to the corre-

sponding mesh triangle. Each UV Gaussian is defined as $G_{uv} = \{\Delta\mu, r, s, \alpha, c, k, t\}$, where $\Delta\mu$ represents the local position of the Gaussian in the triangle, k is the triangle index, and t refers to interpolated position on the triangle using the pixel's barycentric coordinates. During rendering, these properties are transformed into world coordinates:

$$r' = R_t r,\ \mu' = \sigma R_t \Delta\mu + t,\ s' = \sigma s, \tag{3}$$

where σ and R_t represent the average edge length and orientation of the triangle in world coordinates. Rigging the Gaussians in the local coordinates of the triangle, they can flexibly model parts that the EHM template model cannot represent. Additionally, the increased number of Gaussians enhances the model's ability to capture finer details.

Inverse texture mapping. To effectively predict Gaussian attributes in the UV space based on the source image, we design an inverse texture mapping process. This explicitly maps the appearance feature map from screen space to the UV texture space, resulting in a UV feature map F_{uv}. Utilizing the tracked EHM mesh, we first map each pixel of the UV map to the mesh by interpolating the triangle's vertex positions with pixel's barycentric coordinates. Next, each mapped position t is projected into screen space using camera parameters, followed by linear sampling on the appearance feature map. These sampled features are then placed back into the UV space. Finally, we use a mesh rasterizer to retrieve the visible pixel region and filter out the invisible ones from the UV features. After obtaining the UV feature map, we input it into the UV decoder to predict the corresponding Gaussian attributes: $\{\Delta\mu, r, s, \alpha, c\} = \mathcal{D}_{UV}(F_{uv})$. The UV decoder consists of StyleUNet and convolutional networks. First, StyleUNet paints the invisible regions, then the convolutional networks predict each Gaussian attribute.

3.3. Animation and Rendering

Animation. After reconstructing the upper body avatar in the canonical space, we can animate the reconstructed avatar with the new tracking parameters to present new expressions and motions. The template Gaussians inherit their positions from the EHM vertices, so we use the deformed EHM vertex positions (Eq. (1)) as the template Gaussian positions in the pose space. Other Gaussian attributes remain the same, except for the rotation, which undergoes a consistency rotation: $r' = R_{lbs} r$, where R_{lbs} represents the weighted blend skinning rotation matrix. For the UV Gaussians, each Gaussian remains fixed in the local coordinate system relative to its rigged parent triangle, but moves in world coordinates as the triangle deforms. During animation, after obtaining the deformed EHM mesh, we compute the orientation R_t and average edge length σ for each triangle. Then, we use Eq. (3) to transform the UV Gaussians into world coordinates.

Rendering. Although we increase the number of Gaussians by adding UV Gaussians, the number of valid Gaussians after reconstruction may be less than 150000. For complex upper body models, the sparsity of Gaussians may limit the expressiveness. To address this and enhance rendering quality, following [48, 99], we assign each Gaussian with a latent feature c. During rendering, we first obtain a coarse feature map I_F through splatting where the first three dimensions represent the coarse RGB image I_c. Then, the image features are passed through a StyleUNet-based refiner [14], which decodes them into a refined image I_r of the same resolution. Compared to direct RGB rendering, this method strengthens the implicit representation of the Gaussian and improves detail capture.

3.4. Training Strategy and Losses

Similar to the training strategy of [15, 23], we randomly select two images from each image sequence: one as the source image to reconstruct the avatar and the other as the target image to drive the avatar. The loss between the driven result and the target image is then used to train the model. For image loss, we use $\mathcal{L}_1$ and perceptual loss LPIPS $\mathcal{L}_{lpips}$ [108]. To enhance the model's focus on local details, especially the face and hands, we also crop these regions from the image and include them in the loss calculation:

$$\begin{aligned}\mathcal{L}_{image} =& \mathcal{L}_c(I_t, I_r) + \mathcal{L}_c(I_t, I_c) + \lambda_f \cdot \mathcal{L}_c(I_t^{face}, \\ & I_r^{face}) + \lambda_h \cdot \mathcal{L}_c(I_t^{hand}, I_r^{hand}),\ \text{where} \\ \mathcal{L}_c(I_t, I_r) =& \lambda_1 \cdot \mathcal{L}_1(I_t, I_r) + \lambda_{lpips} \cdot \mathcal{L}_{lpips}(I_t, I_r).\end{aligned} \tag{4}$$

Following [70], we introduce a regularization loss to ensure that the Gaussians rigged on the mesh stay close to the thier parent triangles: $\mathcal{L}_{pos} = ||\max(\Delta\mu, \epsilon_{pos})||_2$, where ϵ_{pos} allowed offset threshold relative to the parent triangle's scale. We also regularize the scaling of UV Gaussians: $\mathcal{L}_{sca} = ||\max(s, \epsilon_{sca})||_2$, where ϵ_{sca} represents the allowed scaling threshold. The total loss function is shown as Eq. (5), where λ denotes the weight coefficient.

$$\mathcal{L} = \mathcal{L}_{image} + \lambda_p \cdot \mathcal{L}_{pos} + \lambda_s \cdot \mathcal{L}_{sca}. \tag{5}$$

4. Experiment

4.1. Experimental Setup

Implementation details. Our model is built with PyTorch [64] and trained using the Adam optimizer [44]. We train the model on NVIDIA RTX A6000 GPUs for 200,000 iterations, consuming about 156 GPU hours, with a total batch size of 12. For data processing, we use PIXIE [25], TEASER [55], and HaMeR [66] to perform rough estimations of SMPLX, FLAME, and MANO parameters, respectively. Body masks are extracted using [12], and the background is set to black. Moreover, we extract human and facial keypoints using [5, 17, 57, 101].

Figure 3. Qualitative comparison results on self-reenactment. Compared to others, our method better preserves ID consistency during animation while capturing more detailed facial expressions and hand gestures.

Dataset. We collect a dataset consisting of videos from YouTube, OSX [51], and HowToSign [24], with a focus on human upper body videos. During preprocessing, 5 to 75 frames are sampled at equal intervals from each video depending on the video length. The final training set contains over 26,000 video clips and 620,511 frames. The frames with low hand keypoints confidence due to occlusion are discarded. For the test set, 58 randomly selected IDs are used, with one video per ID, each averaging 15 seconds, totaling 28,287 frames.

Metrics. *Self-reenactment.* We use the first frame of each video as the source image and generate animations based on the motion from the entire video. We evaluate image quality between animated results and original video using PSNR, L1, SSIM, and LPIPS. *Cross-reenactment.* When driving the current ID with motion from another video, no ground truth exists. We use ArcFace [18] to measure the Identity Preservation Score (IPS) which computes the cosine similarity of identity features.

Baseline. *2D-based methods.* We compare GUAVA with several SOTA methods: MagicPose [7], Champ [115], and MimicMotion [110], which enable controllable human animation video synthesis from a single image. *3D-based methods.* We also compare GUAVA with SOTA 3D-based methods like GART [45], GaussianAvatar [34], and ExAvatar [61]. These methods reconstruct 3D human avatars from monocular videos. For comparison, we use the first half of each video for their training and the remaining half for self-reenactment evaluation, while GUAVA uses the first frame as the source image.

	PSNR↑	$\mathcal{L}_1$↓	SSIM↑	LPIPS↓	FPS↑
GUAVA (Ours)	**25.87**	**0.0162**	**0.9000**	**0.0813**	**52.21**
MimicMotion	24.46	0.0200	0.8768	0.0879	0.21
Champ	22.01	0.0258	0.8643	0.1000	0.53
MagicPose	21.25	0.0333	0.8661	0.0913	0.12

Table 1. Quantitative results on the self-reenactment against 2D-based methods. FPS denotes the animation and rendering speed.

	PSNR↑	$\mathcal{L}_1$↓	SSIM↑	LPIPS↓	Recon. input	Recon. time
GUAVA (Ours)	**25.70**	**0.0168**	**0.8976**	**0.0836**	first frame	≈ **98 ms**
ExAvatar	24.09	0.0207	0.8783	0.1064	half video	≈ 2.4 h
GaussianAvatar	23.62	0.0199	0.8780	0.1085	half video	≈ 1.3 h
GART	24.46	0.0195	0.8805	0.1016	half video	≈ 7 min

Table 2. Quantitative results on the self-reenactment against 3D-based methods. Recon. denotes reconstruction.

4.2. Evaluation

Quantitative results. *Self-reenactment.* For 2D-based methods, we evaluate all video frames, while for 3D-based

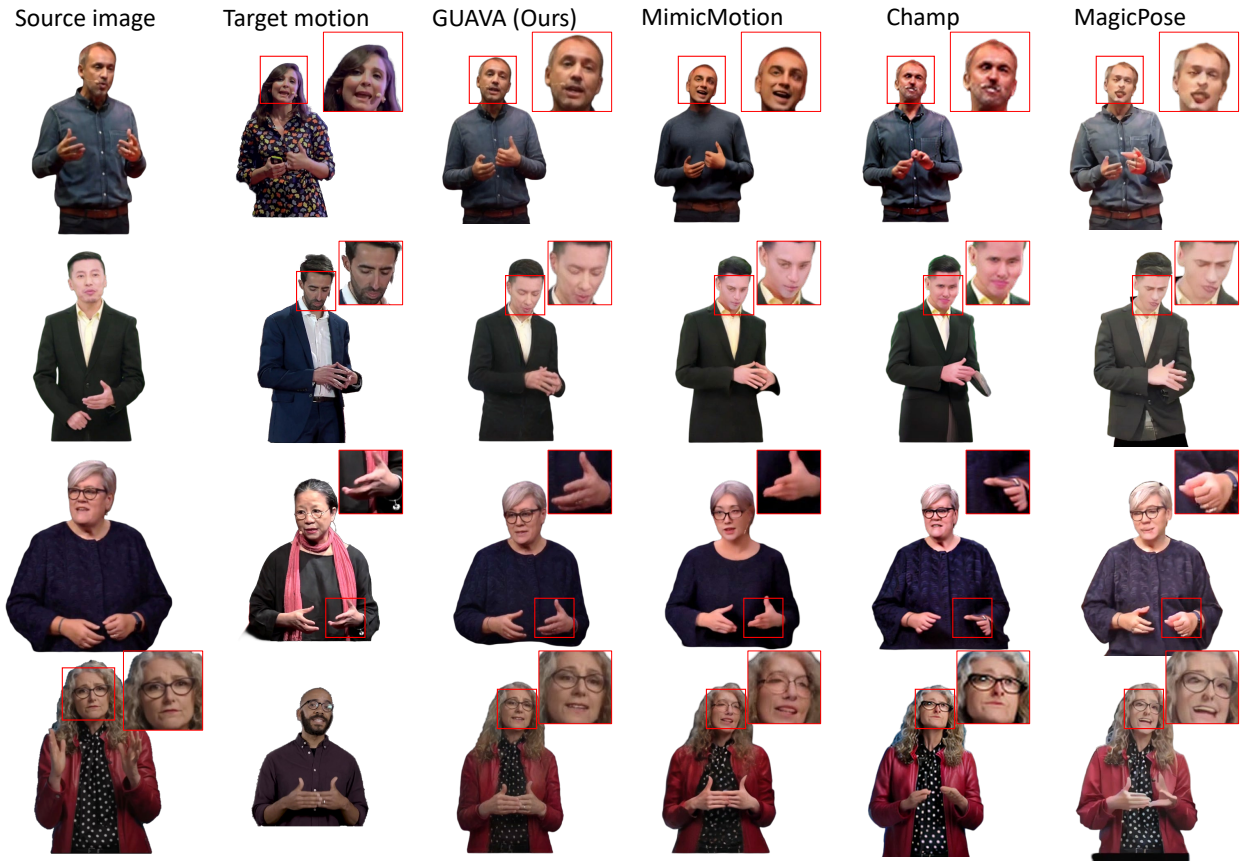

Figure 4. Qualitative results on cross-reenactment against 2D-based methods. Our method demonstrates superior performance in preserving ID consistency across various poses, as well as more accurately capturing the facial expressions and hand gestures of the target motion.

	GUAVA (Ours)	MimicMotion	Champ	MagicPose
IPS↑	**0.5554**	0.1310	0.3677	0.3277

Table 3. ID Preservation Score on cross-reenactment against 2D-based methods. Our method maintains consistent identity.

methods, we assess the latter half. As shown in Tab. 1 and Tab. 2, our method outperforms all others across all metrics, demonstrating high-quality avatar reconstruction with accurate motion and photorealistic rendering. *Cross-reenactment.* We use 10 source images from the test set and ViCo-X [114], driven by 8 videos (5079 frames). Tab. 3 shows that our method significantly outperforms others in IPS, proving its superior ability to preserve identity consistency with the source ID under different poses. *Efficiency.* We evaluate all methods on an NVIDIA RTX 3090 GPU, with results shown in Tab. 1 and Tab. 2. Our method achieves around 50 FPS in animation and rendering, while other 2D-based methods take several seconds per frame. 3D-based methods also support real-time rendering but require minutes to hours for reconstruction, whereas our method completes it from a tracked image in just 0.1s.

Qualitative results. *2D-based methods.* The visual comparison results for self-reenactment and cross-reenactment are shown in Fig. 3 and Fig. 4. Leveraging strong diffusion model priors, 2D-based methods can generate high-quality images. However, the animations from Champ struggle to accurately recover gestures and facial expressions, with blurred hands. MagicPose faces similar issues, with noticeable color distortions. MicmicMotion exhibits better gesture and facial expression performance but fails to maintain identity consistency. In contrast, our approach not only maintains identity consistency with the source image but also restores complex gestures, head poses, and detailed facial expressions like blinking and talking. *3D-based methods.* The qualitative results under the self-reenactment setting are shown in Fig. 5. GaussianAvatar and GART struggle with fine finger and facial expression driving. ExAvatar performs better in these areas, but these methods lack generalization, producing incomplete results for unseen regions and large artifacts in extreme poses. In contrast, our method generates reasonable results for unseen areas, shows better robustness in extreme poses, and provides more accurate

	PSNR↑	$\mathcal{L}_1$↓	SSIM↑	LPIPS↓
Full (Ours)	**25.87**	**0.0162**	**0.9000**	**0.0813**
w/o refiner	24.93	0.0188	0.8851	0.1060
w/o inverse	25.65	0.0168	0.8977	0.0864
w/o UV Gaussians	25.82	0.0164	0.8971	0.0877
w/o EHM	25.60	0.0168	0.8950	0.0846

Table 4. Ablation results under the self-reenactment setting.

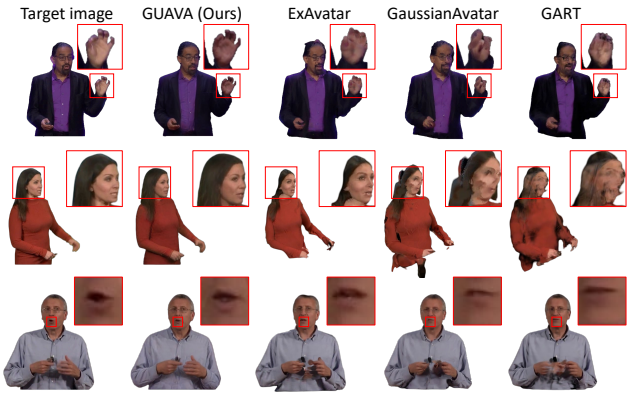

Figure 5. Visual results on self-reenactment against 3D-based methods. Our method reasonably generates unseen regions while capturing more detailed facial expressions and hand gestures.

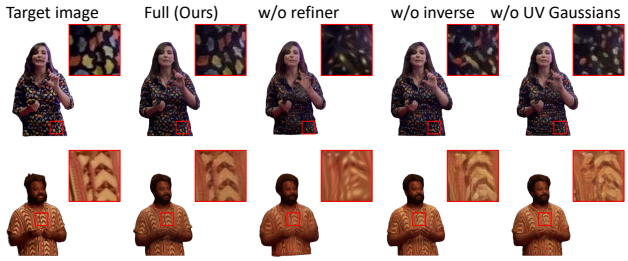

Figure 6. Qualitative results of ablation. Our full method more accurately captures texture details from the source image.

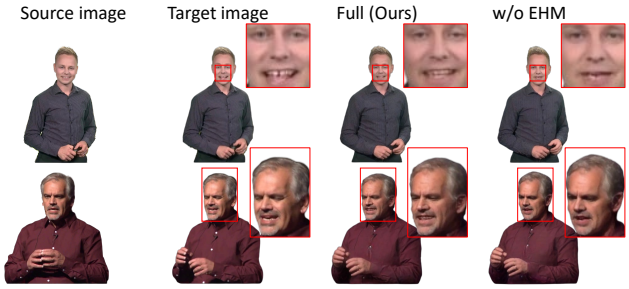

Figure 7. Ablation results of EHM. Using EHM, our method more finely recovers facial expressions and shapes.

and detailed hand and facial expressions.

4.3. Ablation Studies

To validate each component's effectiveness, we perform ablation studies under the self-reenactment setting. The quantitative results are summarized in Tab. 4.

w/o refiner. Without the refiner, sparse Gaussians struggle to capture high-frequency texture details accurately. As shown in Fig. 6, this leads to elliptical artifacts when dealing with intricate patterns. The significant drop in the LPIPS metric further confirms this limitation.

w/o inverse. This setting removes inverse UV mapping during UV Gaussians synthesis, directly feeding the screen-space appearance feature map into the UV decoder. While the StyleUnet in the UV decoder can implicitly learn some mappings, inaccuracies occur. As shown in Fig. 6, it struggles to capture structured textures precisely, resulting in blurred outputs and lower performance across all metrics.

w/o UV Gaussians. Relying solely on a limited number of template Gaussians, the model maintains a comparable PSNR but lacks expressiveness, leading to blurry outputs, as reflected in the lower LPIPS score. Fig. 6 further illustrates this issue, where the model struggles to reproduce high-frequency clothing textures.

w/o EHM. Using the original SMPLX for tracking and reconstruction weakens the model's ability to capture fine facial details. As shown in Fig. 7, it sometimes struggles to track facial expressions or head shapes precisely, causing reconstruction and driving errors. The overall performance also drops across all metrics.

5. Discussion

Conclusion. We propose GUAVA, a fast framework for expressive 3D upper-body Gaussian avatar reconstruction from a single image, enabling real-time animation and rendering. To enhance facial expression, shape, and pose tracking, we introduce EHM with an accurate tracking method. We further propose inverse texture mapping and projection sampling to reconstruct Ubody Gaussians, composed of UV Gaussians and template Gaussians, where UV Gaussians enhance texture details. Extensive experiments show that GUAVA provides accurate animation, high-quality rendering, and superior efficiency.

Limitation. Although GUAVA achieves high-quality rendering, its geometric accuracy is limited by training data diversity. Details like clothing wrinkles, loose garments, and hair are difficult to reconstruct. Additionally, limited by the mesh prior, GUAVA struggles to capture regions far from the mesh, making large hairstyles like afros difficult to reconstruct. The lack of clothing deformation modeling limits realistic pose-driven garment changes. Finally, since the dataset focuses on frontal views, GUAVA cannot generate a full 360-degree avatar. These limitations suggest areas for future research and improvements.

Acknowledgement

This research was funded by the National Key Research and Development Program of China (Project No.2022YFB36066) and partially funded by the Shenzhen Science and Technology under Grant (KJZD20240903103210014, JCYJ20220818101001004).

References

[1] *Instant Volumetric Head Avatars*, 2023. 2

[2] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 20364–20373, 2022. 2

[3] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024. 3

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 3, 5

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 3

[7] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023. 3, 6

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 2

[9] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 2

[10] Jinnan Chen, Chen Li, Jianfeng Zhang, Hanlin Chen, Buzhen Huang, and Gim Hee Lee. Generalizable human gaussians from single-view image, 2024. 3

[11] Xingyu Chen, Yufeng Liu, Dong Yajiao, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[12] Sergej Chicherin and Karen Efremyan. Adversarially-guided portrait matting. *arXiv preprint arXiv:2305.02981*, 2023. 5

[13] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[14] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 37:57642–57670, 2025. 3, 5

[15] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*, 2024. 3, 5

[16] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2

[17] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018. 3, 5

[18] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 6

[19] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 3

[20] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *European Conference on Computer Vision*, pages 316–333. Springer, 2024. 3

[21] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. *Advances in neural information processing systems*, 31, 2018. 3

[22] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37:2127–2160, 2025. 2

[23] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 5

[24] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[25] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of

expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 1, 3, 5

[26] 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 (ToG), Proc. SIGGRAPH*, 40(8), 2021. 2

[27] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 1, 2

[28] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024. 3

[29] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14783–14794, 2023. 2

[30] 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

[31] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[32] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 3

[33] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 2

[34] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6

[35] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20418–20431, 2024. 3

[36] Yingdong Hu, Zhening Liu, Jiawei Shao, Zehong Lin, and Jun Zhang. Eva-gaussian: 3d gaussian-based real-time human novel view synthesis under diverse camera settings. *arXiv preprint arXiv:2410.01425*, 2024. 3

[37] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 3

[38] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *IEEE Conference on Computer Vision (ICCV)*, 2023. 3

[39] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 1

[40] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633. IEEE, 2023. 2, 3

[41] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 4

[42] Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, and Wonjun Kim. Sampling is Matter: Point-guided 3d human mesh reconstruction. In *CVPR*, 2023. 2

[43] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[44] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[45] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19876–19887, 2024. 3, 6

[46] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 2

[47] 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. 1, 2, 3

[48] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5

[49] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 3

[50] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 2023. 1, 2

[51] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 1, 6

[52] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5904–5913, 2019. 3

[53] Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1120–1129, 2024. 3

[54] Yuxiao Liu, Zhe Li, Yebin Liu, and Haoqian Wang. Texvocab: texture vocabulary-conditioned human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1715–1725, 2024. 1

[55] Yunfei Liu, Lei Zhu, Lijian Lin, Ye Zhu, Ailing Zhang, and Yu Li. Teaser: Token enhanced spatial modeling for expressions reconstruction. *arXiv preprint arXiv:2502.10982*, 2025. 3, 5

[56] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[57] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 3, 5

[58] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 3

[59] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[60] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 1, 2

[61] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *ECCV*, 2024. 3, 6

[62] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4

[63] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[64] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[65] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2, 3

[66] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 5

[67] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 3

[68] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 1

[69] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9054–9063, 2021. 2

[70] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2, 3, 4, 5

[71] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*, 2018. 3

[72] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 3

[73] Minghan Qin, Yifan Liu, Yuelang Xu, Xiaochen Zhao, Yebin Liu, and Haoqian Wang. High-fidelity 3d head avatars reconstruction through spatially-varying expression conditioned neural radiance field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4569–4577, 2024. 2

[74] 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. 3

[75] George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2490–2501, 2024. 2

[76] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[77] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3

[78] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2

[79] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 130–141, 2024. 3

[80] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. *Advances in Neural Information Processing Systems*, 37: 140032–140065, 2025. 2

[81] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[82] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018. 3

[83] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 3

[84] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[85] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 3

[86] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[87] Yang-Tian Sun, Qian-Cheng Fu, Yue-Ren Jiang, Zitao Liu, Yu-Kun Lai, Hongbo Fu, and Lin Gao. Human motion transfer with 3d constraints and detail enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4682–4693, 2022. 3

[88] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[89] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3

[90] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision*, 2022. 2

[91] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. *arXiv preprint arXiv:2307.00040*, 2023. 2

[92] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 3

[93] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 3

[94] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. Gac-gan: A general method for appearance-controllable human video motion transfer. *IEEE Transactions on Multimedia*, 23:2457–2470, 2020. 3

[95] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

[96] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[97] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[98] Jun Xiang, Yudong Guo, Leipeng Hu, Boyang Guo, Yancheng Yuan, and Juyong Zhang. One shot, one talk: Whole-body talking avatar from a single image. *arXiv preprint arXiv:2412.01106*, 2024. 3

[99] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian

head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[100] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[101] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 3, 5

[102] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 3

[103] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[104] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 896–905, 2024. 3

[105] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *European Conference on Computer Vision*, pages 341–359. Springer, 2024. 3

[106] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Kangjie Chen, Minghan Qin, Yu Li, and Haoqian Wang. Hravatar: High-quality and relightable gaussian head avatar. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26285–26296, 2025. 3

[107] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[108] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[109] Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, and Yu Li. Accurate 3d face reconstruction with facial component tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9033–9042, 2023. 2

[110] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 3, 6

[111] Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Trans. Graph.*, 2023. Just Accepted. 2

[112] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19680–19690, 2024. 3

[113] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[114] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. Interactive conversational head generation, 2023. 7

[115] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 6