DevilSight: Augmenting Monocular Human Avatar Reconstruction through a Virtual Perspective

Yushuo Chen[1] Ruizhi Shao[1] Youxin Pang[1] Hongwen Zhang[2]
Xinyi Wu[3] Rihui Wu[3] Yebin Liu[1]
[1]Tsinghua University [2]Beijing Normal University [3]Honor Device Co., Ltd

Figure 1. We present DevilSight, which reconstruct 3DGS human avatar with fine-grained dynamic details from monocular video. Our approach utilizes a video generative model to "see in the darkness", generating a human video from alternative perspective. This enables us to capture high-frequency details from the input view while mitigating potential artifacts in unseen regions.

Abstract

We present a novel framework to reconstruct human avatars from monocular videos. Recent approaches have struggled either to capture the fine-grained dynamic details from the input or to generate plausible details at novel viewpoints, which mainly stem from the limited representational capacity of the avatar model and insufficient observational data. To overcome these challenges, we propose to leverage the advanced video generative model, Human4DiT, to generate the human motions from alternative perspective as an additional supervision signal. This approach not only enriches the details in previously unseen regions but also effectively regularizes the avatar representation to mitigate artifacts. Furthermore, we introduce two complementary strategies to enhance video generation: To ensure consistent reproduction of human motion, we inject the physical identity into the model through video fine-tuning. For higher-resolution outputs with finer details, a patch-based denoising algorithm is employed. Experimental results demonstrate that our method outperforms recent state-of-the-art approaches and validate the effectiveness of our proposed strategies.

1. Introduction

Reconstructing high-fidelity and dynamically detailed human avatars holds significant value across a wide range of applications, including gaming, VR/AR technologies, and the movie industry. However, despite notable advancements in human avatar reconstruction methods [6, 32, 33, 36, 49, 65, 90, 91] utilizing multi-view inputs, it still remains a challenge to create an avatar with high-realistic texture details and dynamic fidelity from monocular video.

Previous approaches represent the human avatar with respect to deformable human templates [40, 46], in terms of structured latent codes [50] or implicit 3D representations [10, 14, 16, 17, 25, 75]. By constructing avatars in a canonical pose, these approaches align multi-frame observations into a unified 3D space and leverage pose-conditioned neural networks to interpolate dynamic details through spatial smoothness priors. While this framework ensures multi-view consistency, it struggles to recover high-frequency details due to limited cross-view correlation in the input frames and the required fusion operation. Furthermore, the strong reliance on the naked body template limits its applicability in scenarios involving loose clothing.

Recently, generative models [7, 57, 76] show the poten-

tial to "imagine" previously unseen content from text or images. Moreover, some works [37, 39, 63] achieve the generation and reconstruction of 3D objects by adding camera control. Based on these generative models, many methods [2, 13, 22, 23, 28, 41] leverage generative priors to enhance 3D reconstruction by score distillation sampling (SDS) [52] or explicitly generating multi-view images as pseudo ground truths. While these methods have achieved high-quality 3D static scene reconstruction from sparse-view inputs, reconstructing dynamic sequences remains a significant challenge. This difficulty primarily arises from generating video while maintaining the identity consistency across time and viewpoints. In this paper, we propose leveraging Human4DiT [59] to generate an identical motion sequence from an alternative perspective, thereby providing supplementary supervision signals for reconstruction. Recent advancements have demonstrated the impressive capability of diffusion transformers (DiT) [47] to incorporate control signals from various dimensions. Specifically designed to produce spatio-temporally coherent human videos, Human4DiT supplies the view and temporal consistency priors essential to our approach. In practice, we choose to generate only one video from the back view, for its efficiency and effectiveness to address the majority of unseen regions.

Nevertheless, directly generating the back-view video using Human4DiT presents challenges in accurately reproducing human motions that are physically consistent with the input video. This limitation arises because the generation is conditioned on an identity embedding extracted from a single reference image, which is insufficient to encapsulate the dynamic characteristics inherent to human identity. Consequently, generative models pretrained exclusively on such embeddings also struggle to accurately reproduce the motion. Drawing inspiration from DreamBooth [58], we propose Physical Identity Inversion through model finetuning, enabling video generation that aligns with the physical motion present in the input video. To prevent overfitting and maintain view consistency priors, we have meticulously designed our fine-tuning strategy and carefully selected the parameters involved. Moreover, in order to bridge the resolution gap between captured and generated videos and to mitigate potential artifacts arising from this difference [83], we have developed an innovative algorithm that effectively doubles the resolution of generated videos. Given that DiT-based models are constrained to denoising at fixed resolution, the algorithm involves generating videos by partitioning them into patches. Benefiting from these technical designs, our method achieves superior reconstruction quality compared to previous state-of-the-art methods.

As shown in the experiments, our method achieves high-fidelity human reconstruction with dynamic details and supports dynamically consistent image synthesis at novel views. In summary, our contributions are:

- We propose a novel framework for high-fidelity dynamic human-centric reconstruction from monocular video, which leverage the capability of video generative model to complement the unseen perspective.
- We introduce Physical Identity Inversion through model fine-tuning, effectively embedding personalized visual and physical identity into the model, thereby enabling accurate generation of identical human motions from the rear-view perspective.
- We present a patch-based denoising strategy for super-resolution video generation, effectively bridging the resolution gap between input and generated videos. This approach enables the subsequent reconstructed avatar model to render images with consistent granularity across varying viewpoints.

2. Related Work

2.1. Neural Rendering for Human Reconstruction

Neural rendering, as it effectively bridging the 3D assets with 2D images, has emerged as a powerful technique for reconstructing human figures directly from images. Many approaches served this problem as reconstructing the human avatar by leveraging shape priors [36, 49, 50, 65, 75] from parametric human body templates [40, 46] or learning the skinning fields according to the skeletons [8, 31, 61, 62]. The human dynamics is then decomposed into rigid motions driven by skeletons and non-rigid deformations predicted by a pose-conditioned neural network. On the other hand, this strategy unifies appearance information from different frames into a common 3D space, facilitating avatar creation and synthesis under novel views and novel poses. However, the quality of the reconstructed avatar is constrained by the ability to translate low-frequency pose parameters into high-frequency dynamic details. PoseVocab [32] decomposed the pose latent vector into per-joint embeddings for richer pose encoding. SLRF [90, 91] divide the entire radiance field into smaller local radiance fields, enabling better representativeness for local body part. Animatable Gaussians [33] and MeshAvatar [6] represented the pose parameters as the corresponding SMPL position map and employed the powerful 2D networks to encode it, achieving the SOTA in reconstructing the details.

Due to the high costs associated with studio-based multi-view data capture, several studies [21, 48, 61, 82] have explored reconstructing humans from monocular images or videos. Single-view scenarios are particularly challenging because inevitable errors in estimated poses can degrade 3D correspondences across frames. To address this issue, some research has focused on error correction networks [25, 75], while others have utilized trainable pose parameters [10, 24]. Recently, 3DGS is incorporated to fur-

ther improve the training and inference efficiency [14, 16, 17, 27, 30, 42, 45, 51, 53, 60, 74, 93]. Nevertheless, in order to stabilize the rendering quality in novel view, which is invisible from input, these methods usually constrained the complexity of the pose-conditioned networks. This limitation may hinder the accurate reconstruction of dynamic details observed in the input images. In contrast, our approach regularize the avatar representation using the priors from generative models, facilitating both reconstruction and novel view synthesis.

2.2. Generative Prior for Human Avatars

With the development of generative models [4, 7, 11, 57, 68, 76], more and more methods [2, 13, 22, 23, 28, 29, 41, 73] seek to inject human priors from large models into human avatar reconstruction. They mainly leverage the power of generative models to iteratively optimize 3D representations. For example, some methods employ score distillation sampling (SDS) loss [52] to optimize 3D representations conditioned on text prompt [18, 20, 22, 28, 29, 35, 41, 69, 78], skeletons [19, 38, 64, 84], and densepose [86], normal maps. However, textual descriptions and 2D pose maps inherently lack precision in representing fine-grained geometric and texture details. Moreover, SDS primarily optimizes 3D parameters by enforcing distributional constraints, which can lead to issues such as oversaturation and excessive smoothing. Although some methods [3] alleviate the above problems by using variational score distillation (VSD) loss [72], the computational cost is higher.

Due to the inherent limitations of SDS loss, subsequent works [1, 2, 5, 13, 23] avoid using it and instead explicitly leverage generative models [57, 63, 88]. These approaches typically generate multi-view images or videos using a pretrained large model, followed by direct 3D reconstruction and optimization. Human-related priors are learned from 2D structures and conditions (e.g., rendered skeletons or surface normals). Similarly, HumanSplat [43] generates multi-view features within latent space and reconstructs the human body through feature aggregation. Recent approaches [34, 44, 77, 80, 87] extends these strategies to video-to-4D generation. However, the absence of explicit 3D structural modeling exacerbates inherent temporal and view inconsistencies in 2D generative models, often leading to flat or blurry results. L4GM [56] and GVFDiffusion [85] directly synthesize 4D content conditioned on the given video frames. Despite this, they primarily handle simple skinning-like motions and fail to capture complex dynamic details. Preserving the human identity while recovering fine-grained motions remains a significant challenge.

Our task formulation is similar to WonderHuman [73], as both leverage generative priors to reconstruct dynamic 3D human avatar from a monocular video. WonderHuman optimize 3D human with SDS loss in both canonical and observation spaces to ensure visual consistency. However, due to the oversaturation and excessive smoothing caused by SDS, the generated invisible regions lack fine details and exhibit noticeable inconsistencies.

2.3. Identity Adaptation

Maintaining identity consistency is crucial in digital humans, particularly when handling complex textures and dynamic movements. Most methods [15, 66, 79, 92] use a single reference image as input and design a reference network to embed the image into the backbone model. However, this approach demands a substantial amount of data for training and is computationally expensive. In addition, IP-Adapter [81] designs an effective and lightweight adapter to enable image prompt for pre-trained text-to-image diffusion models. Specifically, IP-Adapter modifies the cross-attention mechanism by separating cross-attention layers for text and image features. By freezing most parameters, image prompt can be learned by training only the image cross-attention layers. DreamBooth [58] can fine-tune the generative model using a few reference images, ensuring identity consistency. In details, DreamBooth fine-tunes a pretrained diffusion model to link a unique identifier to a specific subject, allowing it to be seamlessly embedded into different scenes. VideoComposer [67] directly concatenates the reference image with noise to achieve identity injection.

3. Method

3.1. Overview

Given a monocular sequence of human motions with the corresponding human poses, which could be estimated using existing tools [9, 70], our task is to reconstruct a high-fidelity 3DGS avatar of this subject. High-frequency dynamic details often necessitate a trade-off between accurate input-view reconstruction and generalizable novel view synthesis. To address this, we propose leveraging advanced video generative models to generate additional videos from alternate perspectives as pseudo-supervision. This approach could effectively regularize the 3D representation from overfitting. However, generating a video that replicates human motions from an alternative viewpoint presents significant challenges. It requires preserving the individual's identity both visually and physically, ensuring that distinctive features and movement characteristics remain consistent across perspectives. Additionally, video generative models often face computational memory constraints that limit their output to low-resolution videos, typically below 1080p. It is crucial to bridge resolution gap between generated video and input video, to maintain visual fidelity and meet quality expectations. We propose physical identity inversion (Sec. 3.3) through model finetuning and super-resolution generation (Sec. 3.4) to tackle these

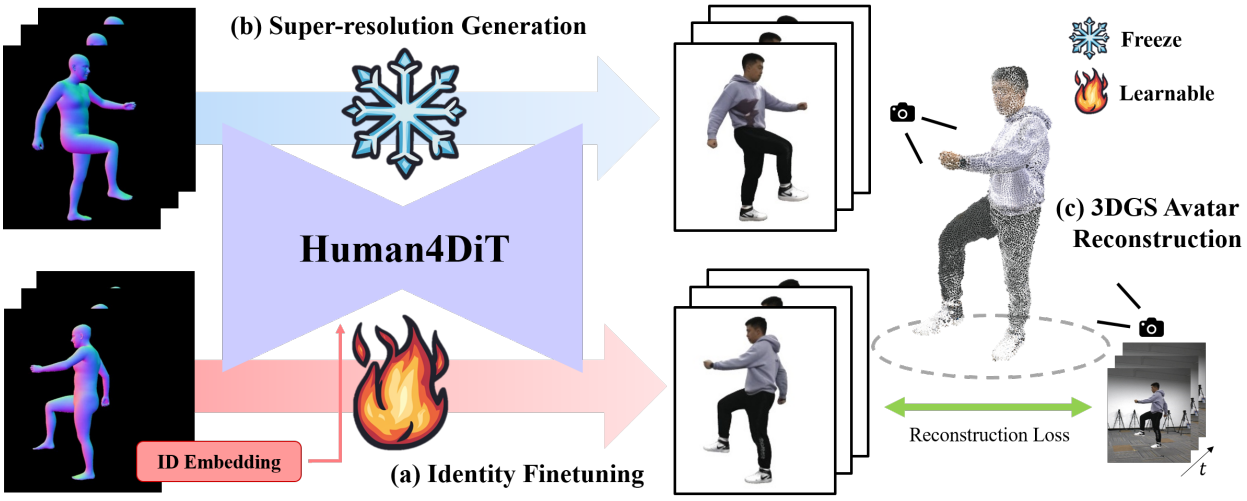

Figure 2. **Overview of our method.** Our method leverages priors from Human4DiT to enable robust monocular avatar reconstruction. It comprises three key components: (a) fine-tuning the model for personalized content generation, (b) generating consistent rear-view motion with superresolution, and (c) reconstructing 3DGS avatars using pseudo multi-view data.

challenges, respectively. An overview of our method is presented in Fig. 2.

3.2. Preliminaries

Expressive 3DGS Avatar Representation. 3D Gaussian Splatting [26] is a 3D point-based representation for efficient and realistic rendering. It's represented by a set of 3D Gaussians, each of which is parameterized by its 3D center position $\boldsymbol{\mu}$, a covariance matrix $\boldsymbol{\Sigma}$, opacity α and color $\boldsymbol{c}$, and distributed as:

$$f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad (1)$$

e-() where the covariance matrix is further parameterized by a rotation quaternion $\mathbf{q}$ and a 3D scaling vector $\mathbf{s}$ practically.

The rendering is conducted by splatting these 3D Gaussians onto the 2D plane, where the color of each pixel is computed by composition of the Gaussians overlapping on it. Following recent advancements on human avatars [16, 33], we represent the avatar as 2D Gaussian map, and utilize UNet $\mathcal{U}$ to produce the Gaussian maps under different poses. Specifically, the Gaussian points are anchored onto the SMPL template by orthogonally projecting the canonical mesh to the front plane and back plane. Each active pixel corresponds to a 3D Gaussian point. Given any human pose $\boldsymbol{\Theta} = \{\theta_i\}_{i=1}^{J}$, these points could be deformed using the transformation matrix from LBS, and other Gaussian attributes are acquired by feeding the front-back projected posed position map [6, 33] $\mathcal{P}_f, \mathcal{P}_b$ to UNet:

$$\mathcal{G}_f(\boldsymbol{\Theta}), \mathcal{G}_b(\boldsymbol{\Theta}) \leftarrow \mathcal{U}(\mathcal{P}_f(\boldsymbol{\Theta}), \mathcal{P}_b(\boldsymbol{\Theta})), \quad (2)$$

where $\mathcal{G} = (\Delta\boldsymbol{\mu}, \mathbf{q}, \mathbf{s}, \alpha, \mathbf{c})$ consists of a pose-dependent point offset $\Delta\boldsymbol{\mu}$ to model non-rigid deformations.

Human4DiT. Human4DiT [59] is a DiT-based [47] approach for generating high-quality, 360-degree human videos with spatio-temporal coherence from a single image. By leveraging a hierarchical structure and the strength of DiT in capturing global features, it enables the synthesis of videos with strong 3D consistency. Specifically, Human4DiT proposes a hierarchical 4D transformer architecture that factorizes self-attention across views, time steps, and spatial dimensions. To enable precise control, dedicated modules are designed to embed camera parameters, temporal information, human motion, and identity. Additionally, Human4DiT collects large-scale multi-dimensional datasets, including 2D videos, multi-view videos, 3D data, and 4D data. By employing a multi-dimensional training strategy, the model effectively leverages these diverse data to enhance the generalization ability.

Human4DiT iteratively denoises a gaussian noise ϵ to obtain a clean latent representation z, which is then passed through a decoder to generate the final output. During training, the model is trained to predict the applied noise from the noisy latent z_t:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\mathbf{z}_t, c_{\text{ref}}, c_{\Theta}, \epsilon, t}\left(\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, c_{\text{ref}}, c_{\Theta}, t\right)\|_2^2\right) \quad (3)$$

where ϵ_θ represents the denoising transformer, t represents the timesteps, and c_{ref} and c_{Θ} represent the conditions of identity reference and human poses, respectively.

3.3. Physical Identity Inversion by Finetuning

In Human4DiT, the human identity is injected to the network through a CLIP [54] embedding c_{ref} extracted from a reference image. This approach has several limitations.

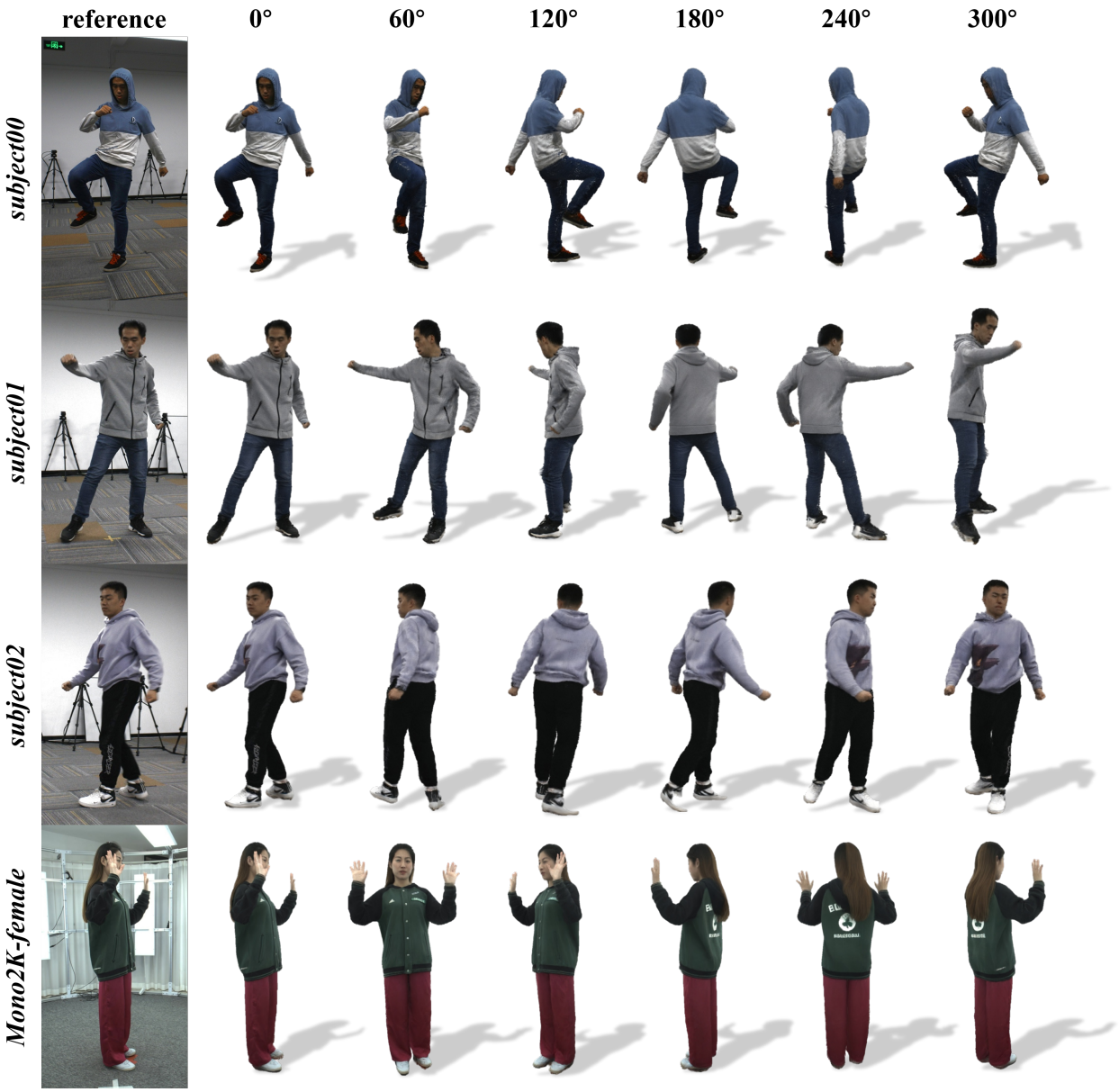

Figure 3. **Freeview Rendering of our reconstructed subjects from *THuman4.0* and *Mono2K*.** By leveraging the priors from Human4DiT, we not only accurately reconstruct dynamic details from the monocular input but also supports novel view synthesis with equivalent quality. Zoom in to see more details.

On the one hand, the embedding conflates visual patterns from both the human subject and the background, leading to insufficient emphasis on the human identity. On the other hand, since the embedding is based on a single image, it provides an incomplete representation of the human identity, inadequately capturing the complex physical properties necessary for high-quality generation.

To overcome these limitations, we inject the human identity by fine-tuning the model using input video data. Specifically, we define a unique, learnable embedding, c_{id}, initialized from the CLIP embedding corresponding to the human subject, which enables a richer representation of identity beyond the constraints of a single-image reference. The fine-tuning is performed using the reconstruction loss analogous to Equation 3. Additionally, to preserve capability to generate multi-view consistent video, we alternate tuning the model between the input video and the multi-view dataset from Human4DiT. Notably, only the attention layers conditioned on identity are fine-tuned.

Table 1. **Quantitative comparisons** on input view reconstruction and novel view synthesis with HumanNeRF [75], GaussianAvatar [16] and AnimatableGaussians (AG for short) [33]. The best and the second best are highlighted in **bold** and underlined fonts, respectively.

Dataset	Method	Input View			Novel View 1			Novel View 2		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
THuman4.0	HumanNeRF	30.69	0.9515	0.0841	25.84	0.9438	0.1051	24.97	0.9372	0.1163
	GaussianAvatar	31.13	0.9709	0.0820	22.25	0.9533	0.1271	22.02	0.9500	0.1346
	AG	**34.06**	**0.9798**	**0.0554**	26.55	0.9588	0.1071	25.65	0.9542	0.1193
	Ours	32.97	0.9769	0.0558	**26.79**	**0.9605**	**0.0995**	**25.98**	**0.9560**	**0.1111**
Mono2K	HumanNeRF	27.91	0.9315	0.0822	26.02	0.9305	0.0873	25.45	0.9263	0.0941
	GaussianAvatar	29.13	0.9729	0.0734	21.26	0.9549	0.1071	20.83	0.9542	0.1121
	AG	**33.15**	**0.9807**	**0.0533**	**26.29**	0.9654	0.0819	25.60	0.9591	0.0972
	Ours	32.19	0.9768	0.0614	26.26	**0.9670**	**0.0788**	**25.73**	**0.9633**	**0.0898**

3.4. Back-view Generation with Super Resolution

In order to fully capture previously unseen details, we opted to generate the video from a rear perspective. In particular, based on the SMPL parameters of the first frame, we derived the new camera extrinsic matrix by rotating the original camera 180 degree around the axis passing through the root position and directed along the global orientation.

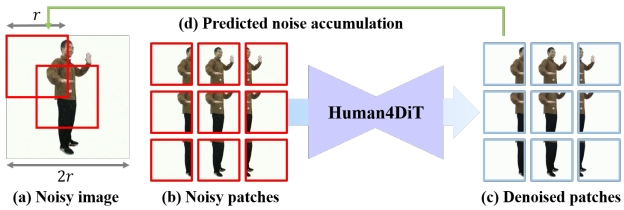

Figure 4. **Illustration of Super-resolution Generation.** At each diffusion timestep, we partition the image into 9 overlapping patches. The noises are then predicted independently by patch and accumulated by weighted sum.

Human4DiT generates images at a standard resolution of 768×768, which is considerably lower than the typical 2K resolution or higher found in most images. This discrepancy creates a pronounced resolution gap when transitioning from the front view to the back view. To address this, we instead generate images at twice the original resolution in both width and height. Enlarging the canvas won't affect the decoding because of the shift invariance from VAE decoder. Consequently, we propose an algorithm that produces larger latent representations using a smaller, fixed-size diffusion denoising approach. Specifically, we employ a similar sliding window strategy to maintain spatial consistency in the generated images, as how other methods preserve temporal consistency in video generation. As illustrated in Fig. 4, at each denoising timestep the latent image is partitioned into nine overlapping patches; noise is predicted independently for each patch and then merged via a weighted sum. Correspondingly, during model fine-tuning (Sec. 3.3), we incorporate randomly cropped videos into the training dataset to enhance patch-based generation.

After generating the back-view video, we directly treat it as a real captured video for pseudo supervision.

4. Experiments

Dataset. We evaluated our method using the THuman4.0 dataset [90], a multi-view dataset with a resolution of 1330×1150, featuring characters with rich textures and dynamic details. To assess our method on higher-resolution images, we collected an additional dataset, named *Mono2K*, comprising images at the resolution of 1500×2048. For evaluation, we selected all three sequences (*subject00*, *subject01* and *subject02*) from THuman4.0 and two typical sequences (*Mono2K-male* and *Mono2K-female*) from *Mono2K*. Each sequence includes manually selected video clips featuring turning motions to ensure full-body visibility. For fair comparison in novel view synthesis, we utilized ground truth SMPL-X poses fitted from multiple views as input. For methods requiring SMPL poses, we converted them manually using the official tool [46]. Foreground masks were obtained using Segment Anything 2 [55].

Baselines. We compared our method with several state-of-the-art approaches, including HumanNeRF [75], GaussianAvatar [16], and Animatable Gaussians [33]. Both GaussianAvatar and Animatable Gaussians are 3DGS-based methods for constructing human avatars, aiming to reconstruct dynamic details through 2D UNet architectures. GaussianAvatar is designed for monocular inputs, while Animatable Gaussians is tailored for multi-view inputs. Due to the training efficiency of NeRF-based methods, when evaluating HumanNeRF on the *Mono2K dataset*, we resized the images to 750×1024.

The comparison is also conducted with state-of-the-art video-to-4D approaches, L4GM [56] and GVFDiffusion [85], both of which are designed for general objects.

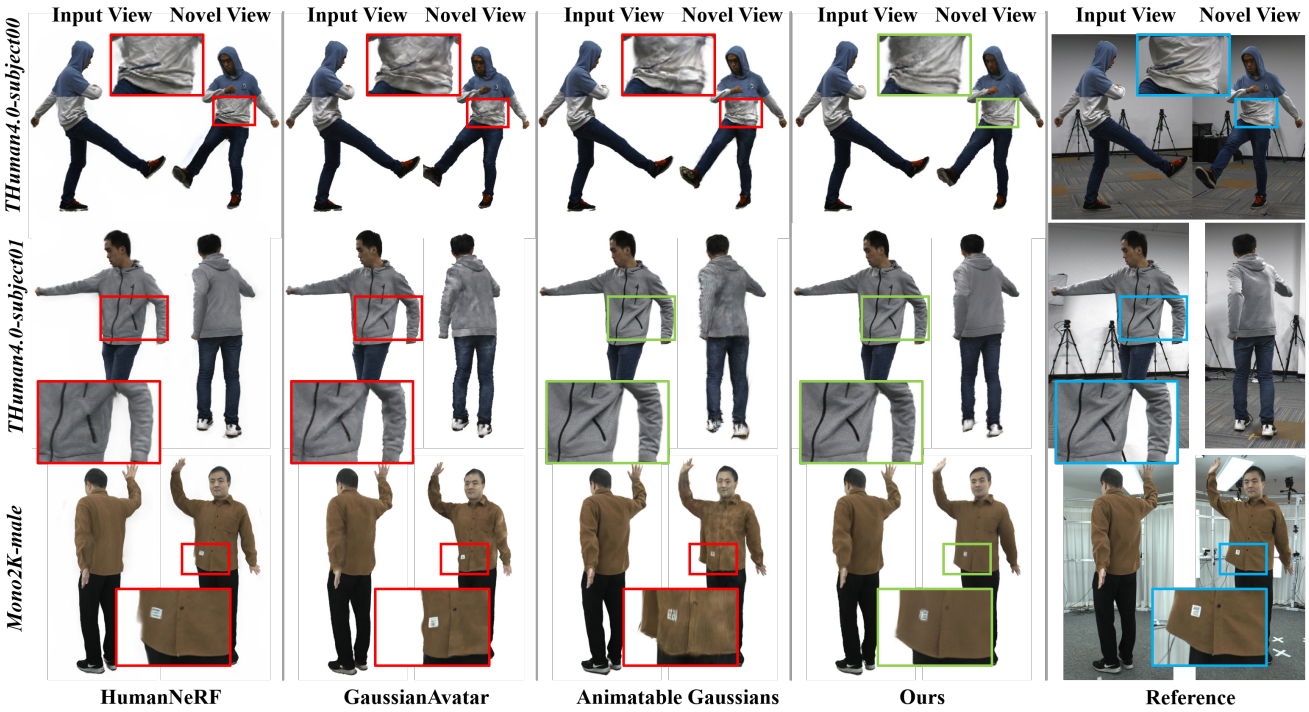

Figure 5. **Qualitative Comparisons** on input view reconstruction and novel view synthesis. Our approach achieves high-fidelity reconstruction by leveraging priors from Human4DiT. Zoom in to see more details.

L4GM leverages the generated multi-view videos, whereas GVFDiffusion adopts a holistic encode–decode paradigm to directly generate 4D content. All methods are reproduced using their publicly available codebases.

Metrics. We conducted our evaluation using established image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM) [71] and Learned Perceptual Image Patch Similarity (LPIPS) [89]. PSNR and SSIM are calculated over the entire image, with backgrounds set to white, while LPIPS is computed within the minimal square bounding box compassing the body.

4.1. Main Results

Fig. 3 showcases our reconstruction results by rendering the reconstructed subject from different angles of viewpoints. Our approach demonstrates the capability not only in recovering the fine-grained details from the input view, but also produce plausible rendering quality at novel views. These results prove the robustness and adaptability of our approach in complementing missing details and regularizing dynamic human representations.

4.2. Comparisons

We assess the effectiveness of our approach through evaluations on both input view reconstruction and novel view synthesis. The quantitative results are presented in Tab. 1. As illustrated in Fig. 5, by leveraging priors from a video generative model, our method significantly outperforms other state-of-the-art approaches. Due to its NeRF-based representation, HumanNeRF cannot accurately recover the fine details from the input, primarily producing dynamics through skinning. In contrast, both GaussianAvatar and Animatable Gaussians demonstrate the superior representational capabilities of 3DGS for precise reconstruction. However, they exhibit artifacts in unseen regions, likely due to the lack of effective regularization for 3DGS, even though GaussianAvatar was specifically designed for monocular input. Complementing the back view not only supplies additional reconstruction details but also regularizes the avatar representation, mitigating potential artifacts and enabling our method to achieve high-fidelity and high-quality reconstruction.

Moreover, as illustrated in the last example in Fig. 5, both HumanNeRF and GaussianAvatar are unable to accurately capture certain dynamic phenomena, such as the fluttering clothes, because their representations are tightly constrained to the human template. Our approach leverages video generation that rigorously preserves the subject's physical identity while ensuring view consistency with the input, enabling the reconstruction of these intricate dynamic details. This capability distinguishes our method from prior approaches, which have struggled to achieve similar results.

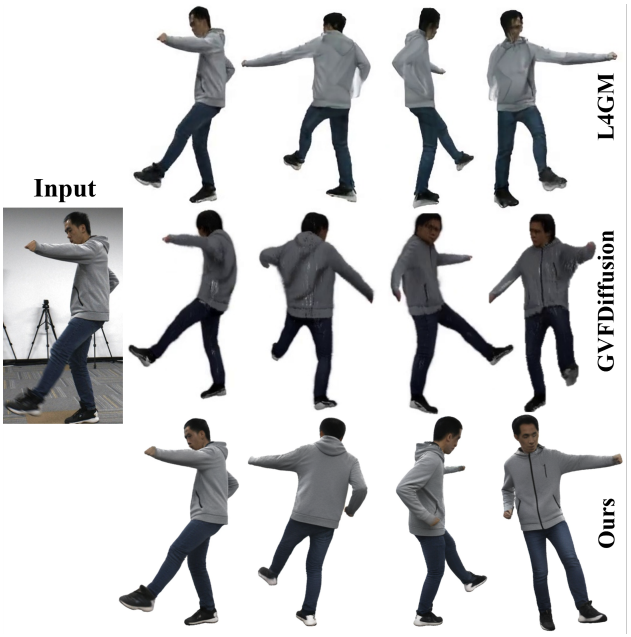

Figure 6. **Qualitative Comparisons** with SOTA video-to-4D methods. By leveraging both human priors and generative priors, our method effectively recovers fine-grained dynamic details.

We further compare our method with SOTA video-to-4D approaches. As shown in Fig. 6, both L4GM and GVFDiffusion struggle to reconstruct complex non-rigid dynamics present in the video. L4GM, due to the absence of human-body priors, fails to maintain consistent human identity across viewpoints and timestamps. GVFDiffusion, on the other hand, fails to encode the full range of motion details in the input video, leading to generated results that deviate substantially from the original content. In contrast, our approach benefits from both 3D human templates and generative priors, achieving high-fidelity dynamic reconstruction.

4.3. Ablation Study

ID Finetuning. Fig. 7 illustrates the generation results of different finetuning strategies. (a) Without finetuning, Human4DiT generates identities with only approximate color similarity. (b) Fine-tuning only the identity embedding yields a generally accurate appearance but lacks precision in details, such as clothing wrinkles, due to insufficient incorporation of the subject's dynamic characteristics. (c) Fine-tuning along with the Human4DiT model implicitly injects the physical properties to generate dynamics. This comprehensive fine-tuning leads to the most accurate reconstruction of the human identity, enhancing the quality of subsequent back-view video generation.

Super-resolution Generation. Fig. 8 presents the back-view generation results at different resolutions. As shown in (b) and (e), the absence of super-resolution leads to blurred

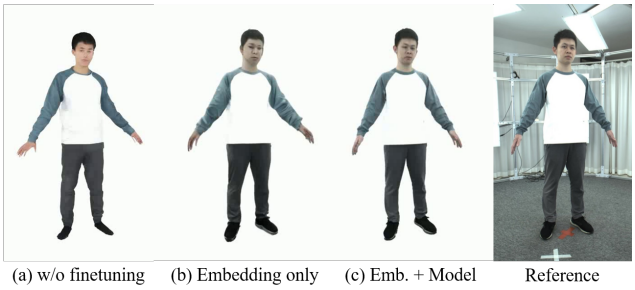

Figure 7. **Ablation study on different ID finetuning strategies.**

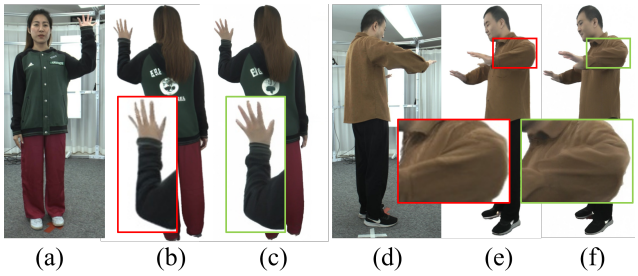

Figure 8. **Ablation study on super-resolution generation.** (a)(d) reference image from input view. (b)(e) standard-resolution generation. (c)(f) super-resolution generation. Zoom in to see details.

clothing wrinkles, as diffusion models struggle to capture fine-grained local details. In contrast, our super-resolution approach not only enhances visual clarity and preserves intricate details but also generates high-resolution images that are crucial for improving the quality and robustness of avatar training.

5. Conclusion

We present a novel framework for high-fidelity dynamic human reconstruction from monocular video. By leveraging video generation to incorporate alternative viewpoints, our approach not only recovers high-frequency dynamic details from the input view but also supports novel view synthesis with equivalent quality. Furthermore, we enhance video generation by employing Physical Identity Inversion through model fine-tuning in conjunction with patch-based super-resolution techniques.

Limitations. Our approach is constrained by the current capabilities of the avatar model. As a pose-conditioned network, the avatar representation faces challenges in reconstructing data with pose-appearance one-to-many issue, a scenario commonly observed with loose clothing.

Acknowledgement. This paper is supported by NationalKey RD Program of China (2022YFF0902200).

References

[1] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[2] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Guide3d: Create 3d avatars from text and image guidance. *arXiv preprint arXiv:2308.09705*, 2023. 2, 3

[3] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 958–968, 2024. 3

[4] 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. 3

[5] Jinnan Chen, Chen Li, Jianfeng Zhang, Lingting Zhu, Buzhen Huang, Hanlin Chen, and Gim Hee Lee. Generalizable human gaussians from single-view image. *arXiv preprint arXiv:2406.06050*, 2024. 3

[6] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. In *European Conference on Computer Vision*, pages 250–269. Springer, 2024. 1, 2, 4

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 3

[8] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 2

[9] 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*, pages 14783–14794, 2023. 3

[10] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 1, 2

[11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2024. 3

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[13] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024. 2, 3

[14] Hezhen Hu, Zhiwen Fan, Tianhao Wu, Yihan Xi, Seoyoung Lee, Georgios Pavlakos, Zhangyang Wang, et al. Expressive gaussian human avatars from monocular rgb video. *Advances in Neural Information Processing Systems*, 37:5646–5660, 2025. 1, 3

[15] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[16] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 634–644, 2024. 1, 3, 4, 6, 2

[17] 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. 1, 3

[18] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2024. 3

[19] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 36:4566–4584, 2023. 3

[20] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pages 1531–1542. IEEE, 2024. 3

[21] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 2

[22] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14371–14382, 2023. 2, 3

[23] Suyi Jiang, Haimin Luo, Haoran Jiang, Ziyu Wang, Jingyi Yu, and Lan Xu. Mvhuman: tailoring 2d diffusion with multi-view sampling for realistic 3d human generation. *arXiv preprint arXiv:2312.10120*, 2023. 2, 3

[24] 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. 2

[25] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022. 1, 2

[26] 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. 4

[27] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 505–515, 2024. 3

[28] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems*, 36:10516–10529, 2023. 2, 3

[29] Inhee Lee, Byungjun Kim, and Hanbyul Joo. Guess the unseen: Dynamic 3d scene reconstruction from partial 2d glimpses. 2024. 3

[30] 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

[31] 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

[32] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 1, 2

[33] 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*, pages 19711–19722, 2024. 1, 2, 4, 6

[34] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 3

[35] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *2024 International Conference on 3D Vision (3DV)*, pages 1508–1519. IEEE, 2024. 3

[36] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6):1–16, 2021. 1, 2

[37] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[38] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6646–6657, 2024. 3

[39] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 2

[40] 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. 1, 2

[41] Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Kartik Teotia, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Transactions on Graphics (ToG)*, 42(6):1–18, 2023. 2, 3

[42] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 3

[43] 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. *Advances in Neural Information Processing Systems*, 37:74383–74410, 2025. 3

[44] Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Efficient4d: Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742*, 2024. 3

[45] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1165–1175, 2024. 3

[46] 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, 6

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 4

[48] Bo Peng, Jun Hu, Jingtao Zhou, and Juyong Zhang. Selfnerf: Fast training nerf for human from monocular self-rotating video. *arXiv preprint arXiv:2210.01651*, 2022. 2

[49] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Ani-

matable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 1, 2

[50] 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. 1, 2

[51] Sen Peng, Weixing Xie, Zilong Wang, Xiaohu Guo, Zhonggui Chen, Baorong Yang, and Xiao Dong. Rmavatar: Photo-realistic human avatar reconstruction from monocular video based on rectified mesh-embedded gaussians. *arXiv preprint arXiv:2501.07104*, 2025. 3

[52] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3

[53] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5020–5030, 2024. 3

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4

[55] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6

[56] Jiawei Ren, Cheng Xie, Ashkan Mirzaei, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, Huan Ling, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2024. 3, 6

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[58] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 3

[59] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. *arXiv preprint arXiv:2405.17405*, 2024. 2, 4

[60] 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*, pages 1606–1616, 2024. 3

[61] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in neural information processing systems*, 34:12278–12291, 2021. 2

[62] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision*, pages 107–124. Springer, 2022. 2

[63] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 2, 3

[64] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Cheng Lin, Rong Xie, Li Song, Xin Li, and Wenping Wang. Disentangled clothed avatar generation from text descriptions. In *European Conference on Computer Vision*, pages 381–401. Springer, 2024. 3

[65] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European conference on computer vision*, pages 1–19. Springer, 2022. 1, 2

[66] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024. 3

[67] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023. 3

[68] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[69] Yi Wang, Jian Ma, Ruizhi Shao, Qiao Feng, Yu-Kun Lai, and Kun Li. Humancoser: Layered 3d human generation via semantic-aware diffusion model. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 436–445. IEEE, 2024. 3

[70] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024. 3

[71] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[72] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and

diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36:8406–8441, 2023. 3

[73] Zilong Wang, Zhiyang Dou, Yuan Liu, Cheng Lin, Xiao Dong, Yunhui Guo, Chenxu Zhang, Xin Li, Wenping Wang, and Xiaohu Guo. Wonderhuman: Hallucinating unseen parts in dynamic 3d human reconstruction. *arXiv preprint arXiv:2502.01045*, 2025. 3

[74] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2069, 2024. 3

[75] 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*, pages 16210–16220, 2022. 1, 2, 6, 3

[76] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 1, 3

[77] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 3

[78] 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

[79] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3

[80] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion ̂2: Dynamic 3d content generation via score composition of video and multi-view diffusion models. *arXiv preprint arXiv:2404.02148*, 2024. 3

[81] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[82] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[83] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[84] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. 3

[85] Bowen Zhang, Sicheng Xu, Chuxin Wang, Jiaolong Yang, Feng Zhao, Dong Chen, and Baining Guo. Gaussian variation field diffusion for high-fidelity video-to-4d synthesis. *arXiv preprint arXiv:2507.23785*, 2025. 3, 6

[86] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Daniel Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7124–7132, 2024. 3

[87] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *Advances in Neural Information Processing Systems*, 37:15272–15295, 2024. 3

[88] 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. 3

[89] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7, 1

[90] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 1, 2, 6

[91] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4): 1–19, 2023. 1, 2

[92] 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

[93] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. 3