SinGS: Animatable Single-Image Human Gaussian Splats with Kinematic Priors

Yufan Wu[1*] Xuanhong Chen[1†] Wen Li[1,3] Shunran Jia[2] Hualiang Wei[1*] Kairui Feng[4]
Jialiang Chen[1] Yuhan Li[1] Ang He[1*] Weimin Zhang[1] Bingbing Ni[1] Wenjun Zhang[1]
[1]Shanghai Jiao Tong University, Shanghai, China [2]DreamX Inc [3]Akool Research
[4]The National Key Laboratory of Autonomous Intelligent Unmanned Systems, Shanghai, China
{chen19910528,wmzhang,nibingbing,zhangwenjun}@sjtu.edu.cn, elvinfkr@tongji.edu.cn

Abstract

Despite significant advances in accurately estimating geometry in contemporary single-image 3D human reconstruction, creating a high-quality, efficient, and animatable 3D avatar remains an open challenge. Two key obstacles persist: incomplete observation and inconsistent 3D priors. To address these challenges, we propose ***SinGS****, aiming to achieve high-quality and efficient animatable 3D avatar reconstruction. At the heart of SinGS are two key components: Kinematic Human Diffusion and Geometry-Preserving 3D Gaussain Splatting. The former is a foundational human model that samples within pose space to generate a highly 3D-consistent and high-quality sequence of human images, inferring unseen viewpoints and providing kinematic priors. The latter is a system that reconstructs a compact, high-quality 3D avatar even under imperfect priors, achieved through a novel semantic Laplacian regularization and a geometry-preserving density control strategy that enable precise and compact assembly of 3D primitives. Extensive experiments demonstrate that SinGS enables lifelike, animatable human reconstructions, maintaining both high quality and inference efficiency (up to 70FPS).*

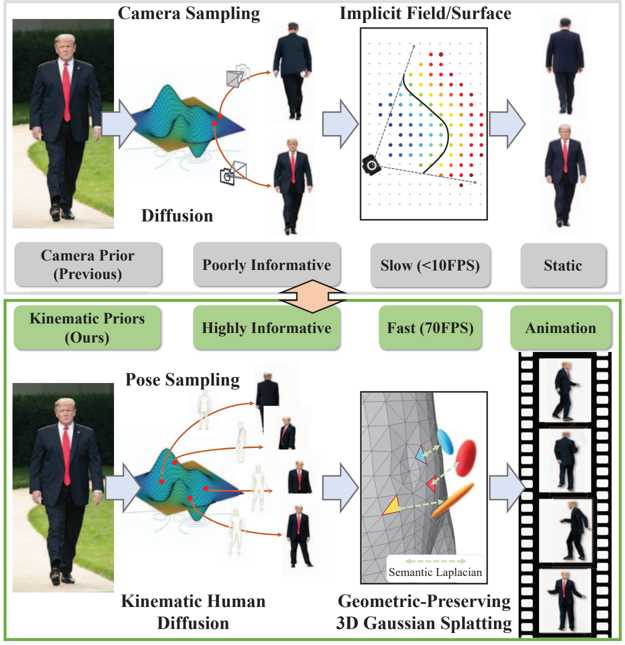

Figure 1. Existing single-image 3D avatar reconstruction methods are limited to sampling prior information from camera space. In contrast, SinGS proposes sampling from pose space, allowing us to obtain both camera and kinematic priors simultaneously. Combined with our Geometry-Preserving 3D Gaussian Splatting, this enables highly efficient animation rendering.

1. Introduction

Single-image 3D human reconstruction (i.e., SIHR) is an emerging modeling technique that aims to create a lifelike 3D avatar from just one image, offering a highly cost-effective and promising approach to 3D modeling. Recently, 3D Gaussian Splatting (3DGS)[14] has made significant breakthroughs in rendering efficiency and quality, establishing itself as a mainstream approach for 3D avatar reconstruction. However, current 3DGS-based methods [23, 26] for single-image human 3D avatar reconstruction suffer from critical limitations in animation, as most are designed primarily for static assets. This paper aims to develop an animation reconstruction framework that achieves a high balance between efficiency and rendering quality.

Currently, the vast majority of SIHR methods [7, 10, 23, 24, 26, 40, 41, 46, 49, 50] focus solely on optimizing geometry while neglecting to enhance animation capabilities. As a result, they are often incapable of animation or require third-party tools [7] to achieve approximate motion. Unlike digital assets, the primary requirement for a human avatar is its ability to animate, making **animation** the core goal of SIHR [9]. Achieving animatable single-image human reconstruction (i.e., Animatable SIHR) is particularly

[*] Work done during an internship at Shanghai Jiao Tong University.
[†] Corresponding author: Xuanhong Chen.

challenging for two main reasons: **Incomplete Observation**: A single image lacks comprehensive viewpoint information, offering an extremely limited, partial observation that cannot address occlusion issues. Current methods [7, 10, 47] mainly rely on fine-tuned diffusion models to synthesize (i.e., 3D hallucination) missing viewpoint information (such as front and back perspectives). However, this approach does not resolve the absence of motion information. While synthesized views can alleviate some occlusion issues, the human body's non-rigid structure requires more than static viewpoints to capture its complex kinematic patterns effectively. **Inconsistent 3D Prior**: Current mainstream methods [5, 7, 10, 47] use Score Distillation Sampling (SDS) [25] to leverage diffusion models as supervisory modules for learning 3D priors. However, the high generative diversity of these models often leads to viewpoint inconsistencies, resulting in implausible imagery (e.g., poor-quality hidden views and inconsistencies with the input image). HumanRef [47] attempts to mitigate the instability of SDS through ReferenceNet [8], but observable inconsistencies remain widespread in the results of ReferenceNet-based methods [8, 33], indicating the difficulty in fully resolving this issue. Consequently, 3D representations should be robust enough to handle imperfect 3D priors, though this perspective has largely been overlooked in current single-image avatar frameworks.

To address these challenges, this paper introduces a novel single-image 3D human avatar reconstruction framework, named **SinGS**. Unlike existing methods [23, 26], SinGS focuses on directly constructing an animatable 3DGS avatar. At the core of our SinGS framework are two key components: **Kinematic Human Diffusion** and a highly **Geometry-Preserving 3D Gaussian Splatting**. Specifically, Kinematic Human Diffusion (KHD) is designed to sample a highly 3D-consistent human sequence within the **pose space**, based solely on a single input image prompt. This approach enables high-quality and animatable 3D avatar reconstruction. To train KHD, we constructed a dataset with over ten thousand rotation videos, allowing KHD to learn a rich variety of human structures and texture distributions from different angles, while also gaining extensive knowledge of human kinematics. Unlike existing frameworks [5, 7, 10, 47] that rely on costly and scarce multi-view human data, our approach requires only more affordable and abundant video data, significantly reducing KHD's training costs and better aligning with scaling laws. Geometry-Preserving 3D Gaussian Splatting (i.e., GPGS) tackles artifact and hole issues under imperfect human priors (e.g., minor jitter and 3D inconsistencies) through two key designs: Semantic Laplacian Regularization addresses floating Gaussian spheres caused by 3D inconsistencies. It adjusts the constraint strength based on body parts, allowing more flexibility for areas like hair and tighter constraints for facial details, reducing floating Gaussian spheres while maintaining rendering quality. Geometry-Preserving Density Control resolves over-deletion and incorrect deletion of spheres in the training process. By managing the Gaussian sphere density via mesh edges, it ensures the spheres remain around the mesh, preserving geometry and preventing large holes. Extensive experiments have validated that our SinGS can generate high-quality and animatable 3D avatars at ultra-fast inference speeds (up to 70 FPS) with efficient resource usage (requiring only 3.8GB of GPU memory).

2. Related Work

2.1. Diffusion-based Human Generative Models

Recent advances in diffusion-based human generative models, especially those conditioned on pose, have improved model controllability. Frameworks like ControlNet [48], IP Adapter [44], and ReferenceNet [8], along with better human attribute detection (2D/3D pose estimation, SMPL parameters [18]), enable more precise and consistent results. In image synthesis, models like PIDM [1] and LFDM [22] enhance pose transfer and image adaptation, while LEO [38] uses flow maps for motion representation. For video, DreamPose [13] and DisCo [35] improve pose-to-image translation and control separation. Models like AnimateAnyone [8], MagicAnimate [42], and MagicPose [4] refine pose conditioning with ReferenceNet. Despite these improvements, these models remain resource-heavy and struggle with stability when generating images of complex movements, such as head turns or rotations.

2.2. Single-Image Human Reconstruction

Monocular video-based human reconstruction has reduced the need for specialized equipment, but single-image reconstruction remains challenging due to incomplete data. Early methods like PIFu [29] and SiCloPe [21] used pixel-aligned features and image-to-image translation, with PIFuHD [30] enhancing detail via normal maps. Later methods, such as ARCH [11] and ICON [40], incorporated body priors for animatable avatars, and ECON [41] improved model completeness by separately reconstructing front and back surfaces. Recent approaches like SiTH [7] and HumanRef [47] use generative models to fill in missing data, while SIFU [50] and Diffusion-FOF approximate multi-view features. R-Cyclic Diffuser [3] addresses limitations in Zero-1-to-3 [17], even with human-specific datasets, advancing single-image human reconstruction through generative priors and viewpoint conditioning.

3. Methodology

Given a target image I_r, the objective of the SinGS architecture is to reconstruct a fully animatable 3D avatar $\mathbf{A}_t$ based on I_r. To address the challenge of limited viewpoint and

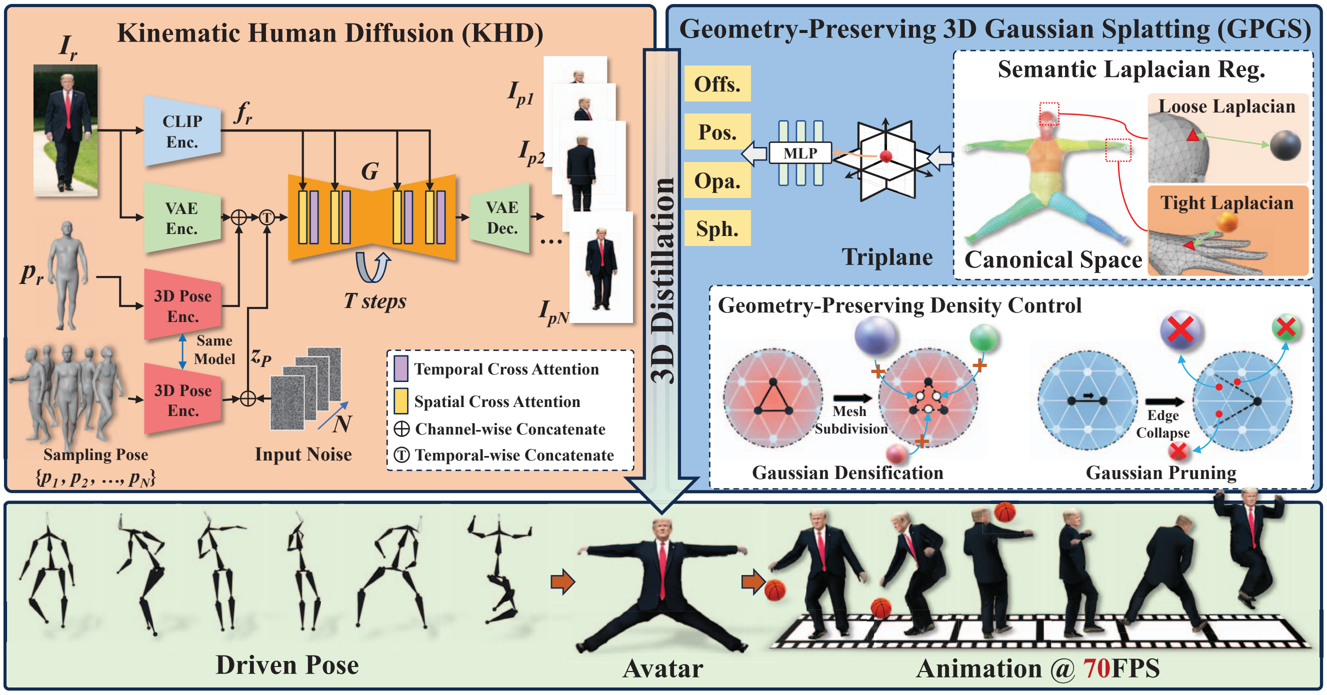

Figure 2. Overall framework of our method.

kinematic information inherent in single-image reconstruction, I_t is first input into the Kinematic Human Diffusion (i.e., KHD) module to complete missing information, generating an augmented sequence of rendered images $S_t = \{I_{p_1}, \cdots, I_{p_i}, \cdots, I_{p_N}\}$, where N represents the total number of poses generated, and p_i indicates each corresponding 3D human pose. KHD reframes the less-constrained, underdetermined 3D reconstruction problem as a generative task, thus producing a fully-constrained, fully-observed 3D reconstruction formulation. Finally, S_t is fed into our compact 3DGS module to reconstruct an animatable 3D human avatar. The compact 3DGS module is capable of reconstructing a watertight, high-quality human model even with imperfect 3D priors. The distilled avatar is high-quality, computationally efficient, and well-suited for promising applications in AR, VR, and similar fields.

3.1. Kinematic Priors

3.1.1 Kinematic Human Diffusion

Unlike existing hallucination architectures [7, 10, 50], our Kinematic Human Diffusion does not rely on camera input, which is prone to estimation errors and has the added challenges of multi-view data scarcity and high costs (e.g., multi-view human data typically requires capture with specialized setups like a light stage). Our approach instead posits that sampling within the motion space provides complete priors for viewpoint and kinematic features. Additionally, our KHD does not adopt a ReferenceNet structure [8]; while ReferenceNet has been highly successful, it requires maintaining two UNet models, significantly increasing the framework's resource demands and preventing integration of longer contexts (e.g., 24 frames). A shorter contextual window cannot provide an image sequence with a sufficient diversity of motion samples (e.g., 150 frames). Considering the above two aspects, our KHD is designed as a diffusion structure, as illustrated in Figure 2. KHD $\mathcal{G}$ takes a single image I_r as input and, constrained by a randomly sampled sequence of poses $\{p_1, p_2, ..., p_N\}$, augments the input image into a sequence $S_t = \{I_{p_1}, I_{p_2}, \ldots, I_{p_N}\}$. This process can be formulated as follows:

$$S_t = \mathcal{G}(I_r; p_1, p_2, \ldots, p_N). \tag{1}$$

Model Architecture. Inspired by UniAnimate [36], our KHD also removes the ReferenceNet and directly inputs the reference image I_r and the driving sequence into the UNet. The reference image I_r is encoded by the VAE encoder of the diffusion model into a latent code z_r, serving as an appearance reference throughout KHD. Additionally, I_r is encoded by the CLIP image encoder [27] into features f_r, which are injected into the network through cross-attention by replacing the original text encoder's features in the diffusion model. This enables high-fidelity preservation of the subject's appearance characteristics. Existing works typically use 2D poses (e.g., OpenPose [2], DW-Pose [43]) as the driving signal. However, our experiments reveal that 2D poses fail to guide the diffusion model accurately in scenarios with extensive occlusions, such as when the subject is turning. Moreover, 2D poses cannot reliably differentiate between different viewpoints of the human body. Therefore, we use SMPL refined by ScoreHMR [31] as the driving signal $\{p_1, p_2, ..., p_N\}$. It is worth noting that Champ [51] also uses SMPL as guidance; however, their SMPL estimations lack accuracy, and they rely on overly strict guidance inputs, such as depth and normal vectors. This approach leads to suboptimal results, performing even worse than 2D poses in practice, for instance, by generating inconsistent body shapes. The SMPL parameters are converted into a mesh model in the specified pose using Linear Blend Skinning (LBS) [18]. This mesh is rendered as an RGB image based on the fixed camera pose maintained throughout the process. The rendered image is then encoded by a **3D Pose Encoder**, a network composed of several convolutional layers, into a latent code z_{p_i} with the same dimensionality as z_r. All z_{p_i} codes are first concatenated with the input noise along the channel dimension, and then sequentially along the temporal dimension to obtain the final representation $z_P \in \mathbb{R}^{T\times C\times H\times W}$. Simultaneously, I_r is processed to estimate and encode its pose latent code z_{p_r}, which is then concatenated with z_P along the temporal dimension before being fed into the diffusion model's backbone. Following most video generation models [36, 37], the UNet component of the diffusion model is also designed as a 3D UNet to ensure 3D consistency in the generated output.

Training Dataset and Strategy. Human image animation [8, 36] primarily focuses on generating diverse frontal human movements (e.g., dancing). As a result, their training data (e.g., TikTok [12]) is largely composed of performance videos of dancers, typically featuring frontal views and wide movement ranges. In our experiments, we found that models trained on this type of data provide limited support for our 3D reconstruction tasks. These models often fail to generate results with high consistency, exhibit significant flickering in fine details, and display abnormal motion in clothing dynamics. In response to this, we collected over 6,000 high-definition videos (each around 20 seconds long) from YouTube, featuring activities such as model showcases and dancing, with a focus on segments that include turning motions. Additionally, we incorporated 600 clips from the UBC Fashion dataset [45], primarily showcasing model runway turns. As a key part of our dataset, we also integrated frontal-view data from MVHumanNet [39]. We used SAM2 [28] to segment human figures from the background in MVHumanNet, then merged the results with a white background to create videos with a consistent background, effectively eliminating the background variability caused by bounding boxes in MVHumanNet. These three sources together form a dataset of over 10,000 training videos, which we call **RotationHuman**. Notably, compared to multi-view data, this type of data collection is significantly more cost-effective and thus better suited for building large-scale datasets. During training, our KHD generates 64 pose-driven images in a single pass, setting the context window to 64 frames (i.e., $N = 64$). In the inference stage, we directly generate 128 frames in one pass. The advantage of one-pass inference is that it avoids consistency jitter in overlapping transition frames. KHD training requires only one phase, where the objective is to reconstruct the ground truth images corresponding to each pose, with the training loss formulated as follows:

$$\mathcal{L} = \mathbb{E}_\theta\left[\|\epsilon - \epsilon_\theta(z_t, f_r, t)\|^2\right], \tag{2}$$

where, t denotes the denoising step, and z_t represents the noise at step t. Note that z_t in KHD contains diffusion states for N different driving poses. The feature f_r is the CLIP image encoding of the reference image I_r, used to ensure ID consistency in the generated results. θ represents the parameters of the 3D UNet and the 3D Pose Encoder. The overall training process involves learning to reconstruct the target video based on I_r as input and the pose set $\{p_1, p_2, ..., p_N\}$ as conditioning. The generated sequence S_t represents the sampling results of the subject in the input image I_r within the pose space, and it can be directly used for 3D avatar reconstruction.

3.2. 3D Distillation with Geometry-Preserving 3D Gaussian Splatting

Once we obtain the sequence of pose images sampled from KHD, the goal is to distill a animatable and real-time rendered human avatar using Geometry-Preserving 3D Gaussian Splatting (GPGS). This challenge can be addressed based on mainstream approaches [16, 32, 52] in monocular 3D reconstruction. As illustrated in the Figure 2 on the right, we initialize the positions of Gaussians using the human template mesh that has undergone loop subdivision. The initial positions are set to the vertex locations, with a default of 2 upsampling iterations. This typically results in approximately $110.6K$ Gaussians. To model the avatar representation, all vertices are fed into a learnable triplane [52] to extract vertex-wise features. These features are subsequently passed to MLPs to predict the geometry and appearance attributes of the Gaussians, respectively. The entire modeling process occurs in canonical space, and we finally map this canonical space into posed space using Linear Blend Skinning (LBS) based on the SMPL framework.

3.2.1 Semantic Laplacian Regularization

Even during the reconstruction of real video, it is challenging to recover perfect static 3D consistency. Therefore, while the generated videos may appear visually plausible, every tiny ambiguity during lifting the 2D outcomes into 3D space potentially leads to catastrophic issues. Early experimental results indicated that common reconstruction meth-

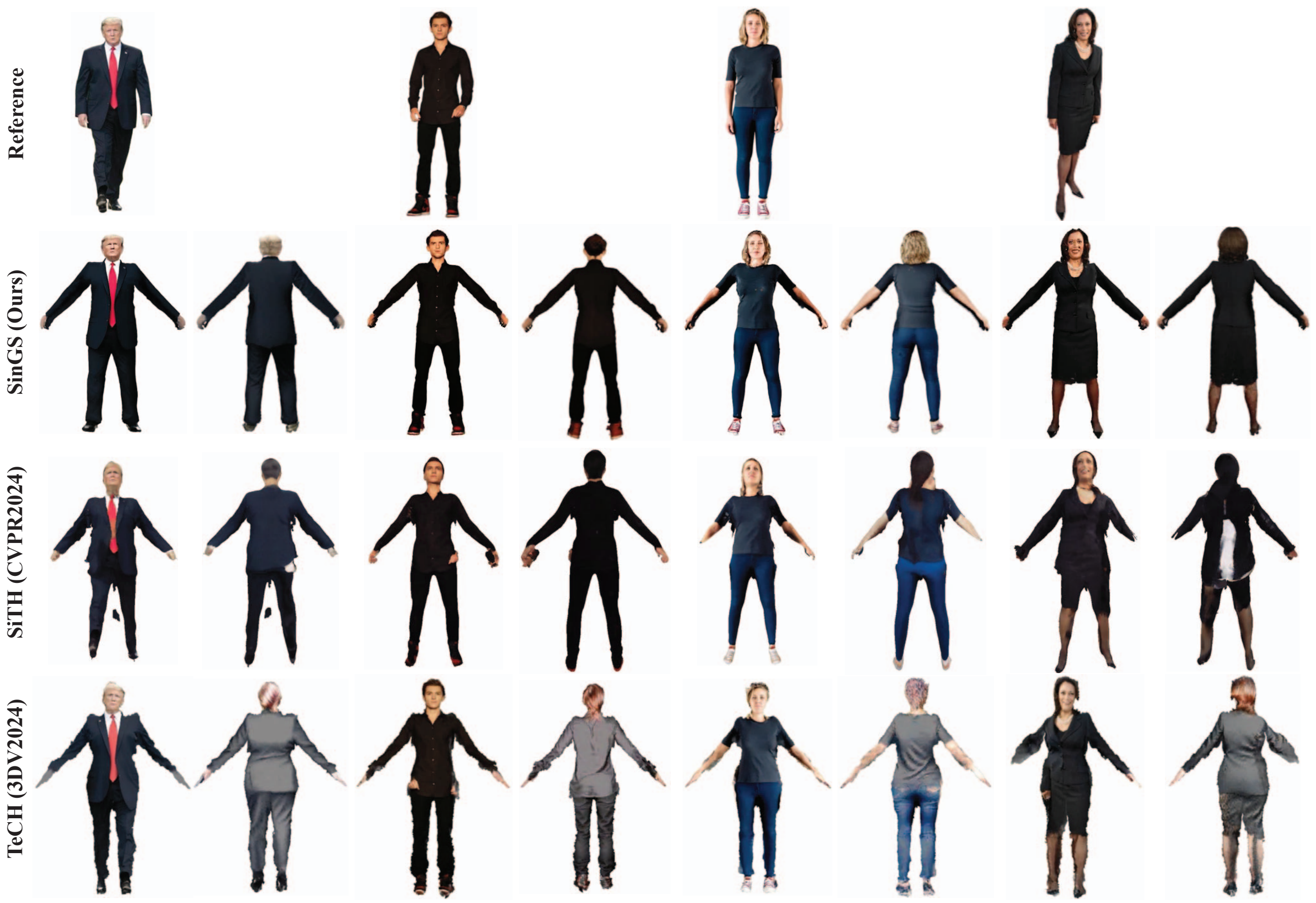

Figure 3. Visual comparison of our method with TeCH [10] and SiTH [7] for single-image reconstruction. We reposed the reconstructed results to A Pose, showcasing both animation and reconstruction capabilities. We highly recommend checking the suppl. for more results.We have provided a large number of animation result images and videos in the suppl.

ods produced results that lacked smoothness and compactness, resulting in existing disconnected and floating gaussians. Following [20], we found that applying Laplacian regularization on the canonical mesh, both for geometry and appearance, effectively mitigated the presence of gaussians out of control. However, it also constrained offsets. In other words, isometrical Laplacian constraint on the whole body overly relied on the accuracy of the SMPL mesh, also making it difficult to fit areas such as hair and clothing that deviate significantly from the SMPL surface. Considering the complexity of the human body and the impossibility to obtain extremely accurate SMPL estimation, we repartition all gaussian vertices into $L = 15$ distinct semantic parts and dynamically adjust the smoothing strength. Besides, we impose the constraint on the anchors points which are closer to the rendering surface than gaussians center. The Semantic Laplacian Regularization (SLR) is defined as follows:

$$\mathcal{L}_{SLR} = \sum_{u \in \Omega} \lambda_u \left\| a_{u,v} - \frac{1}{|M(u,v)|} \sum_{v \in M(u,v)} a_{u,v} \right\|, \qquad (3)$$

where Ω represents the set of semantic regions, a denotes the surface anchor point of the v-th Gaussian vertex in the u-th semantic region, and $M(u, v)$ refers to the set of adjacent points of the Gaussian vertex $a_{u,v}$ located within the u-th semantic region. Experiments demonstrate that the Semantic Laplacian Regularization significantly impacts the handling of more intricate situations, particularly when dealing with errors between the SMPL mesh and the actual character, which can arise from both inaccurate estimations and complex clothing. We apply Laplacians in both position and color spaces; the former enhances the smoothness of the rendering surface, while the latter ensures a more continuous variation in color space.

3.2.2 Geometry-Preserving Density Control Strategy

The strategies in vanilla 3DGS were not originally build for complex, dynamic structures like the human body. While works [20] have adapted the approach, such as initializing Gaussians based on SMPL meshes, these efforts pri-

Table 1. Quantitative comparisons with existing methods on UBC fashion dataset [45]. Note that metrics in the table are measure with the background-removed video. Some methods are unable to perform animation, and therefore, these metrics are not available for those methods. **Bold** indicates the best results, while Underlined represents the second-best results.

Method	Prior Type	PSNR ↑	LPIPS ↓	FVD ↓	Contex. ↓	CLIP Score ↑	Speed↑
DisCo [35] (CVPR24)	-	30.98	0.259	167.7	2.134	82.4%	1FPS
Animate Anyone [8] (CVPR24)	-	31.33	0.238	148.8	1.901	86.6%	2FPS
Champ [51] (ECCV24)	-	31.21	0.251	155.1	2.038	84.3%	1FPS
UniAnimate [36] (ArXiv24)	-	32.31	**0.211**	141.3	1.811	87.1%	2FPS
SHERF [9] (ICCV23)	-	27.39	0.301	202.3	2.998	70.1%	
SIFU [50] (CVPR24)	Camera	-	-	-	2.891	72.1%	5FPS
GST [26] (ArXiv24)	Camera	-	-	-	3.011	71.7%	50FPS
TeCH [10] (3DV24)	Camera	29.01	0.271	188.1	2.331	79.3%	6FPS
SiTH [7] (CVPR24)	Camera	-	-	-	2.776	74.7%	<1FPS
HumanRef [47] (CVPR24)	Camera	-	-	-	2.167	81.5%	3FPS
SHERT [46] (CVPR2024)	Inpainting	29.31	0.266	179.8	2.319	78.8%	10FPS
UniAnimate + GPGS (ours)	Kinematics	29.97	0.251	168.7	1.992	82.1%	**70**FPS
KHD (ours)	-	**32.81**	0.221	**134.4**	**1.768**	**92.3%**	2FPS
SinGS (ours)	Kinematics	31.33	0.240	156.3	1.847	88.2%	**70**FPS

marily lay a solid foundation for the optimization process but fail to maintain this advantages throughout. Incompatibilities still arise in optimization stage, prompting us to propose a novel density control scheme that is inherently compatible with hybrid representations, designed for adaptive adjustment body-wide density. This scheme draws on mesh operation algorithms [6] to better address the unique challenges of human body. For primitives that locate in under-expression mesh areas, we subdivide these regions by adding gaussians on every edge of the selected geometric triangles, which creates much smaller triangles. And we skip the trivial post-smoothing for vertices as the position of gaussians will continue to be optimized in the network. For candidate Gaussians that are selected due to excessively large scales, as well as all newly added Gaussians, we perform rescaling to adjust them to an average size. In the case of anomalous gaussians, especially those are almost transparent, we collapse relevant edges and replace them with single vertex, resulting in decimated adjacent faces. And the priority of selected edges is decided by gaussian attributes and edge length. This entire pruning process preserves the principal structure of the human body. Our strategy result in better reconstruction and animation as the optimization was conducted under connectivity and compactness of the entire human body rather than producing holes and stacks. Moreover, this approach synergistically enhances the effectiveness of our hybrid representation.

4. Experiments

4.1. Experimental Setup

Implementation Details. Our KHD primarily utilizes the RotationHuman dataset we collected, which focuses on human pose image sequences involving turning motions. The specific structure of the 3D Pose Encoder is provided in the supplementary materials. KHD was trained using 8 NVIDIA A100 80GB GPUs, with DeepSpeed and ZeRO-2 employed to reduce memory usage. Similar to Animate Anyone, we resized the video frames to a spatial dimension of 768 × 512. We used a context window size of 64 frames, and to enhance model scalability post-training, we implemented a random context window jitter strategy [36]. During training, the model randomly samples window sizes from 64 to 32 frames in an uniformly distributed manner. The training process utilized the AdamW optimizer [15], with a learning rate set to $5e-5$, and DDPM with 1000 steps for noise sampling. During inference, DDIM with 30 steps was used to reduce inference time. For the 3DGS component, the Adam optimizer [15] was used with a learning rate of $1e-3$, and a cosine annealing learning scheduler was applied to accelerate convergence. The input image sequence size for 3DGS was also 768×512. 3DGS was trained using a single NVIDIA 4090 24GB GPU.

Baseline Models. Our baseline models consist of two parts: the first part includes human image animation models, such as DisCo [35], Animation Anyone [8] [1], Champ [51], and UniAnimate [36]. These models are used to compare our animation capability and the superiority of KHD. The second part consists of single-image 3D avatar reconstruction models, including SHERF [9], SIFU [50], GST [26], TeCH [10], SiTH [7], HumanRef [47], and SHERT [46].

Metrics. We employ both PSNR and LPIPS to assess the quality of video reconstruction. To evaluate the animation capability of our model, we also use FVD [34] to measure the distribution discrepancy between the generated video and the original video, assessing the video generation performance. Furthermore, following HumanRef, we introduce contextual distance [19] and CLIP score [27] to evaluate the texture and semantic similarity between the rendered

[1] https://github.com/MooreThreads/Moore-AnimateAnyone

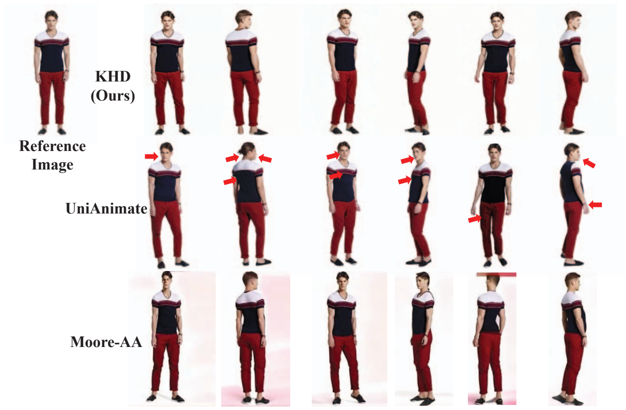

Figure 4. Comparison between KHD and other SOTA human image animation methods. Please zoom in for better observation.

output and the input reference image.

4.2. Qualitative Analysis

Singe-Image Reconstruction. We compare SinGS with two methods from the literature that claim to support animation (TeCH, SiTH). It is worth noting that during our experiments, we found that single-image animation with these methods was extremely difficult to execute, and in some cases, nearly impossible. We had to use third-party code [2] to achieve the reposing for some of these methods. We animate their reconstructed outputs to a common pose (A Pose) for comparison, which allows us to evaluate both reconstruction and animation capabilities. The comparison results are shown in Figure 3, SinGS generates results with better details and maintains high-quality animation, particularly on the in-the-wild data, where SinGS's visual performance significantly outperforms the other methods. This demonstrates the high feasibility and effectiveness of the kinematic prior.

Kinematic Human Diffusion. KHD, as the core component of our pipeline, is compared with state-of-the-art human image animation methods, including Animate Anyone and UniAnimate. As shown in Figure 4, we drive the same character with all three methods, focusing on comparing the quality of their generated results, particularly for the turning pose. From the figure, it is clear that both UniAnimate and our KHD maintain higher consistency than Animate Anyone after the character turns, with KHD surpassing UniAnimate in detail. The details marked by the red arrows in Figure 4 reveals that UniAnimate generates incorrect details, a common issue when using 2D poses, as they often struggle to distinguish body orientation during back-facing poses, leading the model to generate erroneous images. This is detrimental to reconstruction, while KHD generates videos with much higher consistency.

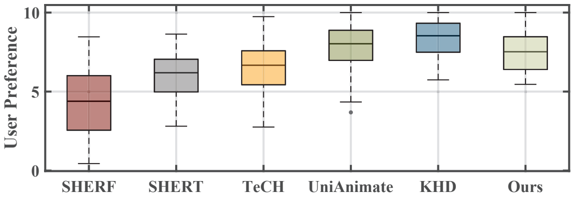

Figure 5. Fifty users rated the video quality, with UniAnimate scoring highest, followed closely by our model, which outperformed others. Our model is more efficient, using only $\mathbf{3.5}$ GB of memory and achieving $\mathbf{70}$ FPS, while UniAnimate requires over 12GB and 0.5 seconds per image.

Geometry-Preserving 3D Gaussian Splatting. Although KHD provides high-consistency pose sampling results, it inevitably has some consistency flaws, meaning it can only offer imperfect prior information. Our geometry-preserving GS method is specifically designed to address this challenge. To validate the superiority of our Geometry-Preserving 3D Gaussian Splatting (GPGS) over general human modeling methods, we compared it with recent state-of-the-art (i.e., SOTA) methods, HAHA [32] and HUGS [16]. We generate a pose sequence using KHD and then performed reconstruction with all 3 methods. The reconstruction results are shown in Figure 7. As seen, both HAHA and HUGS fail to produce a compact 3D avatar, with significant holes or floating artifacts appearing in various areas, particularly during the turning motion. In contrast, our GPGS method is able to reconstruct a compact, watertight 3D avatar model.

4.3. Quantitative Analysis

We selected 100 test videos from the UBC dataset for our quantitative analysis. For each video, we chose the first frame as the reference image and used the video's pose sequence as the driving signal. The performance of all methods was assessed by comparing the animation results with the original video to evaluate the differences. Video reconstruction metrics (including PSNR, LPIPS, and FVD) reflect the model's animation capabilities, while Contextual Distance and CLIP score allow for comparison with methods that do not support animation. The quantitative comparison results are shown in Table 1. Human image animation methods (i.e., the second to fifth rows in the table, including our KHD) outperform the reconstruction-based methods across the board. Among them, KHD surpasses the other methods in most metrics, and UniAnimate also demonstrates superior performance, which is expected due to the scale of training data and model size. While SinGS lags behind UniAnimate, it shows a significant advantage over other methods in some metrics. However, SinGS significantly outperforms diffusion-based animation methods in terms of inference efficiency and resource consumption. Our SinGS achieves $\mathbf{70}$ FPS during inference, with only

[2] https://github.com/custom-humans/editable-humans

Table 2. The results of the ablation experiments on Semantic Laplacian Regularization (i.e., Semantic Lap. Reg.) and Geometry-Preserving Density Control (i.e., GPDC) in SinGS.

Setting	PSNR	LPIPS	FVD
with Semantic Lap. Reg.	31.33	0.240	156.3
with KNN Lap. Reg.	30.88	0.250	169.9
w/o Lap. Reg.	30.34	0.272	181.3
with GPDC	31.01	0.249	166.1
naive Density Control [14]	30.22	0.276	190.2

3.8 GB of memory usage (tested on a single NVIDIA 4090 GPU), while UniAnimate, Champ, and Animate Anyone each require more than 0.4 seconds per image and consume over 12 GB of memory. These quantitative results strongly demonstrate that SinGS is highly effective and feasible. We believe that using KHD, a specialized foundational model, will greatly enhance 3D reconstruction performance while significantly reducing the stringent data requirements.

User Study. We invited 50 users to subjectively evaluate the animation results of SinGS. The evaluation was based on a scale from 0 to 10, with 10 being the highest score and 0 being the lowest. The results are visualized in Figure 5.

4.4. Ablation Study

Kinematic Priors Analysis. The number of poses sampled by KHD has a significant impact on the 3D distillation process. We analyzed how the number of poses affects the animation capability of the final reconstructed model, and the results are shown in Figure 6. We incrementally increased the number of poses from 2 (which corresponds to the typical approach of generating front and rear views, as used by most methods) up to 150. We also tested the influence of different prior models (i.e., KHD, UniAnimate, and Animate Anyone) on the results. Additionally, we examined how different 3DGS reconstruction methods handle the use of prior knowledge. As shown, performance improves rapidly with an increase in the number of poses, but the rate of improvement slows significantly after reaching around 100 poses. Moore-AA + HAHA shows the slowest improvement, while KHD + GPGS exhibits the slowest decline in performance. This suggests that increasing the number of poses does not always lead to improved performance for two main reasons: first, the prior models have already provided sufficient information; and second, none of the prior models provide perfectly consistent priors. Excessive sampling can introduce a lot of inconsistent noise, which ultimately harms performance.

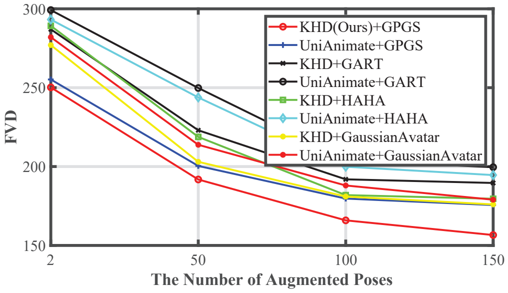

Figure 6. The comparison of the impact of augmented pose numbers on the performance of the 3D avatar models.

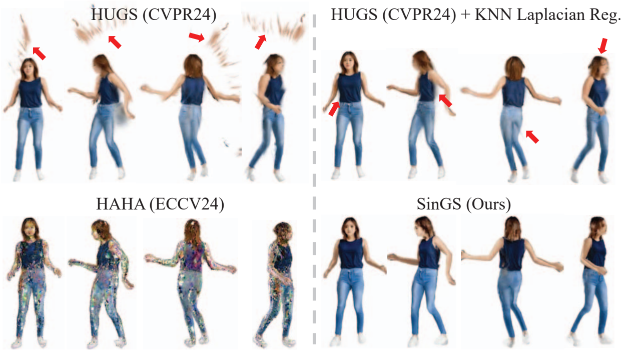

Figure 7. Here is a comparison of our GPGS with two SOTA 3DGS avatar methods, HAHA [32] and HUGS [16]. All methods use images sampled by KHD as body prior.

Geometry-Preserving 3D Gaussian Splatting Analysis. We conducted ablation experiments on the two key designs in GPGS: Semantic Laplacian Regularization and GPDC, with the results shown in Table 2. By replacing and removing the Laplacian regularization, we observed that Semantic Laplacian consistently outperforms in terms of reconstruction metrics. Additionally, GPDC demonstrates a performance advantage over the original density control strategy. The original strategy does not account for dynamic properties and geometric constraints, which makes it ineffective in addressing issues like floating and holes caused by data inconsistencies.

5. Conclusion

In this paper, we propose a new paradigm for single-image 3D avatar reconstruction. Unlike traditional methods that complete missing information in camera space, we sample in human kinematic space, enabling efficient capture of both viewpoints and motion traits. To achieve this, we introduce Kinematic Human Diffusion (KHD) for consistent pose-space generation, supported by our RotationHuman dataset, which focuses on turning motions. Additionally, we propose Geometry-Preserving 3D Gaussian Splatting that reconstructs high-quality, animatable 3D avatars under imperfect human priors. This includes two key innovations: Semantic Laplacian regularization for compact, detailed constraints, and a geometry-preserving density control strategy to maintain stable human geometry. Together, these innovations allow SinGS to achieve high-quality, animatable 3D avatar reconstruction from a single image.

6. Acknowledgment

This work supported by Shanghai Post-doctoral Excellence Programe (2024290). This work was also partly by Shanghai Municipal Bureau of Culture and Tourism and Metaverse Technology Innovation Center of Shanghai Municipal Bureau of Culture and Tourism. Kairui Feng thanks to Xiaomi foundation.

References

[1] Ankan Kumar Bhunia, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR 2023*, pages 5968–5976. IEEE, 2023. 2

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186, 2021. 3

[3] Kennard Yanting Chan, Fayao Liu, Guosheng Lin, Chuan Sheng Foo, and Weisi Lin. R-cyclic diffuser: Reductive and cyclic latent diffusion for 3d clothed human digitalization. In *CVPR 2024*, pages 10304–10313. IEEE, 2024. 2

[4] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *ICML 2024*. OpenReview.net, 2024. 2

[5] Mingjin Chen, Junhao Chen, Xiaojun Ye, Huan-ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. Ultraman: Single image 3d human reconstruction with ultra speed and detail. *CoRR*, abs/2403.12028, 2024. 2

[6] Michael Garland and Paul S. Heckbert. Surface simplification using quadric error metrics. In *SIGGRAPH 1997*, pages 209–216. ACM, 1997. 6

[7] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR 2024*, pages 538–549. IEEE, 2024. 1, 2, 3, 5, 6

[8] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR 2024*, pages 8153–8163. IEEE, 2024. 2, 3, 4, 6

[9] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. SHERF: generalizable human nerf from a single image. In *ICCV 2023*, pages 9318–9330. IEEE, 2023. 1, 6

[10] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *3DV 2024*, pages 1531–1542. IEEE, 2024. 1, 2, 3, 5, 6

[11] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: animatable reconstruction of clothed humans. In *CVPR 2020*, pages 3090–3099. Computer Vision Foundation / IEEE, 2020. 2

[12] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR 2021*, pages 12753–12762. Computer Vision Foundation / IEEE, 2021. 4

[13] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV 2023*, pages 22623–22633. IEEE, 2023. 2

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 1, 8

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR 2015*, 2015. 6

[16] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: human gaussian splats. In *CVPR 2024*, pages 505–515. IEEE, 2024. 4, 7, 8

[17] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV 2023*, pages 9264–9275. IEEE, 2023. 2

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 2, 4

[19] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV 2018*, pages 800–815. Springer, 2018. 6

[20] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. *CoRR*, abs/2407.21686, 2024. 5

[21] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *CVPR 2019*, pages 4480–4490. Computer Vision Foundation / IEEE, 2019. 2

[22] Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR 2023*, pages 18444–18455. IEEE, 2023. 2

[23] 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. *CoRR*, abs/2406.12459, 2024. 1, 2

[24] Marco Pesavento, Yuanlu Xu, Nikolaos Sarafianos, Robert Maier, Ziyan Wang, Chun-Han Yao, Marco Volino, Edmond Boyer, Adrian Hilton, and Tony Tung. ANIM: accurate neural implicit model for human reconstruction from a single RGB-D image. In *CVPR 2024*, pages 5448–5458. IEEE, 2024. 1

[25] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR 2023*. OpenReview.net, 2023. 2

[26] Lorenza Prospero, Abdullah Hamdi, João F. Henriques, and Christian Rupprecht. GST: precise 3d human body from a single image with gaussian splatting transformers. *CoRR*, abs/2409.04196, 2024. 1, 2, 6

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML 2021*, pages 8748–8763. PMLR, 2021. 3, 6

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *CoRR*, abs/2408.00714, 2024. 4

[29] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV 2019*, pages 2304–2314. IEEE, 2019. 2

[30] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR 2020*, pages 81–90. Computer Vision Foundation / IEEE, 2020. 2

[31] Anastasis Stathopoulos, Ligong Han, and Dimitris N. Metaxas. Score-guided diffusion for 3d human recovery. In *CVPR 2024*, pages 906–915. IEEE, 2024. 4

[32] David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. HAHA: highly articulated gaussian human avatars with textured mesh prior. *CoRR*, abs/2404.01053, 2024. 4, 7, 8

[33] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: emote portrait alive - generating expressive portrait videos with audio2video diffusion model under weak conditions. *CoRR*, abs/2402.17485, 2024. 2

[34] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018. 6

[35] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *CoRR*, abs/2307.00040, 2023. 2, 6

[36] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *CoRR*, abs/2406.01188, 2024. 3, 4, 6

[37] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In *CVPR 2024*, pages 6572–6582. IEEE, 2024. 4

[38] Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. LEO: generative latent image animator for human video synthesis. *CoRR*, abs/2305.03989, 2023. 2

[39] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, Shuguang Cui, and Xiaoguang Han. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *CVPR 2024*, pages 19801–19811. IEEE, 2024. 4

[40] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR 2022*, pages 13296–13306, 2022. 1, 2

[41] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR 2023*, 2023. 1, 2

[42] 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 *CVPR 2024*, pages 1481–1490. IEEE, 2024. 2

[43] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV 2023 - Workshops*, pages 4212–4222. IEEE, 2023. 3

[44] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CoRR*, abs/2308.06721, 2023. 2

[45] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. In *BMVC 2019*, page 51. BMVA Press, 2019. 4, 6

[46] Xiaoyu Zhan, Jianxin Yang, Yuanqi Li, Jie Guo, Yanwen Guo, and Wenping Wang. Semantic human mesh reconstruction with textures. In *CVPR 2024*, pages 142–152. IEEE, 2024. 1, 6

[47] Jingbo Zhang, Xiaoyu Li, Qi Zhang, Yanpei Cao, Ying Shan, and Jing Liao. Humanref: Single image to 3d human generation via reference-guided diffusion. In *CVPR 2024*, pages 1844–1854. IEEE, 2024. 2, 6

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV 2023*, pages 3813–3824. IEEE, 2023. 2

[49] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In *NeurIPS 2023*, 2023. 1

[50] Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR 2024*, pages 9936–9947. IEEE, 2024. 1, 2, 3, 6

[51] 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. *CoRR*, abs/2403.14781, 2024. 4, 6

[52] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *CVPR 2024*, pages 10324–10335. IEEE, 2024. 4