MV-Performer: Taming Video Diffusion Model for Faithful and Synchronized Multi-view Performer Synthesis

YIHAO ZHI, SSE, CUHKSZ, China
CHENGHONG LI, FNii-Shenzhen, China and SSE, CUHKSZ, China
HONGJIE LIAO, SSE, CUHKSZ, China
XIHE YANG, SSE, CUHKSZ, China
ZHENGWENTAI SUN, SSE, CUHKSZ, China
JIAHAO CHANG, SSE, CUHKSZ, China
XIAODONG CUN, Great Bay University, China
WENSEN FENG, Shenzhen University, China
XIAOGUANG HAN*, SSE, CUHKSZ, China, FNii-Shenzhen, China, and Guangdong Provincial Key Laboratory of Future Networks of Intelligence, China

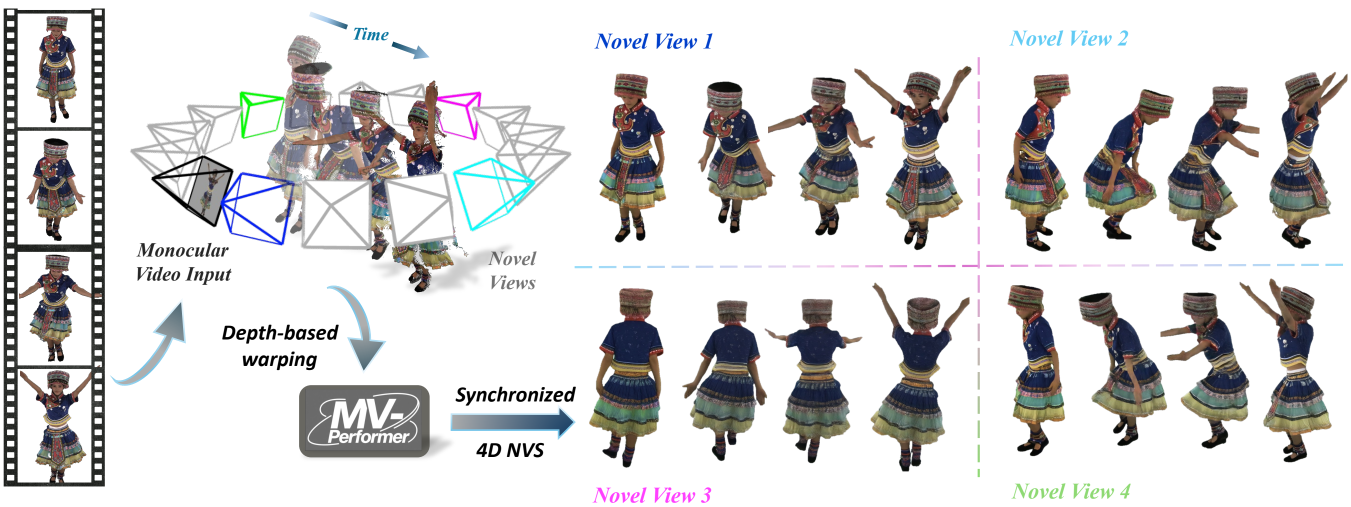

Fig. 1. We propose *MV-Performer* that aims to generate 4D human novel view synthesis from monocular video input. Our method adopts the powerful video diffusion model with the depth-based warping paradigm, enabling 360-degree synchronized multi-view video generation. MV-Performer demonstrates strong capabilities in maintaining both view and temporal consistency for 4D human novel view synthesis.

Recent breakthroughs in video generation, powered by large-scale datasets and diffusion techniques, have shown that video diffusion models can function as implicit 4D novel view synthesizers. Nevertheless, current methods primarily concentrate on redirecting camera trajectory within the front view while struggling to generate 360-degree viewpoint changes. In this paper, we focus on human-centric subdomain and present MV-Performer, an innovative framework for creating synchronized novel view videos from monocular full-body captures. To achieve a 360-degree synthesis, we extensively leverage the MVHumanNet dataset and incorporate an informative condition signal. Specifically, we use the camera-dependent normal maps rendered from oriented partial point clouds, which effectively alleviate the ambiguity between seen and unseen observations. To maintain synchronization in the generated videos, we propose a multi-view human-centric video diffusion model that fuses information from the reference video, partial rendering, and different viewpoints. Additionally, we provide a robust inference

*Corresponding author: Xiaoguang Han (hanxiaoguang@cuhk.edu.cn).

Authors' Contact Information: Yihao Zhi, SSE, CUHKSZ, Shenzhen, China, yihaozhi1@link.cuhk.edu.cn; Chenghong Li, FNii-Shenzhen, Shenzhen, China and SSE, CUHKSZ, Shenzhen, China, chenghongli@link.cuhk.edu.cn; Hongjie Liao, SSE, CUHKSZ, Shenzhen, China, hongjieliao@link.cuhk.edu.cn; Xihe Yang, SSE, CUHKSZ, Shenzhen, China, xiheyang1@link.cuhk.edu.cn; Zhengwentai Sun, SSE, CUHKSZ, Shenzhen, China, zhengwentaisun@link.cuhk.edu.cn; Jiahao Chang, SSE, CUHKSZ, Shenzhen, China, jiahaochang@link.cuhk.edu.cn; Xiaodong Cun, Great Bay University, Dongguan, China, cun@gbu.edu.cn; Wensen Feng, Shenzhen University, Shenzhen, China, sanmumuren@126.com; Xiaoguang Han, SSE, CUHKSZ, Shenzhen, China and FNii-Shenzhen, Shenzhen, China and Guangdong Provincial Key Laboratory of Future Networks of Intelligence, Shenzhen, China, hanxiaoguang@cuhk.edu.cn.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/25/12
https://doi.org/10.1145/3757377.3763935

procedure for in-the-wild video cases, which greatly mitigates the artifacts induced by imperfect monocular depth estimation. Extensive experiments on three datasets demonstrate our MV-Performer's state-of-the-art effectiveness and robustness, setting a strong model for human-centric 4D novel view synthesis. Code is available at https://github.com/zyhbili/MV-Performer.

CCS Concepts: • **Computing methodologies → Computer graphics**;

Additional Key Words and Phrases: 4D Novel View Synthesis, Video Diffusion Model

ACM Reference Format:
Yihao Zhi, Chenghong Li, Hongjie Liao, Xihe Yang, Zhengwentai Sun, Jiahao Chang, Xiaodong Cun, Wensen Feng, and Xiaoguang Han. 2025. MV-Performer: Taming Video Diffusion Model for Faithful and Synchronized Multi-view Performer Synthesis. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3757377.3763935

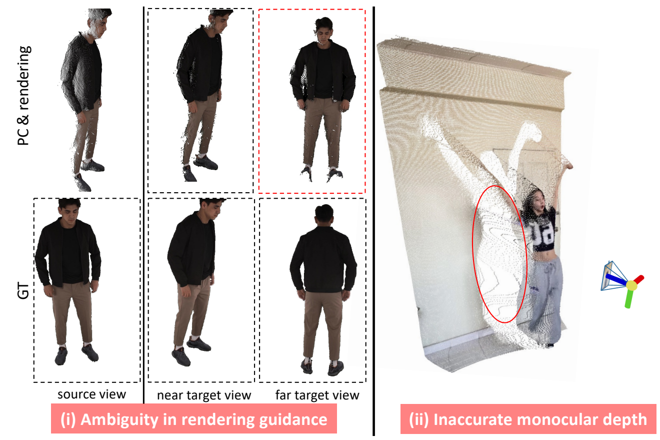

Fig. 2. (i) The depth warping condition at the rear viewpoints presents ambiguity for the model. (ii) Inaccurate monocular depth produce floater-like rendering when there is a significant change in viewpoint.

1 Introduction

Novel view synthesis is a longstanding task in 3D vision and computer graphics, with extensive applications in media content creation, augmented and virtual reality, movie production, etc. Early methods [Avidan and Shashua 1997; Chaurasia et al. 2011; Chen and Williams 2023; Levoy and Hanrahan 2023] attempt to solve it with techniques including multi-view stereo [Furukawa et al. 2015; Seitz et al. 2006] and image warping [Glasbey and Mardia 1998], which explicitly model the stereo, color of each target pixel. With the rise of neural representations and corresponding differentiable rendering techniques [Chen et al. 2022a; Huang et al. 2024b; Jiang et al. 2020; Kerbl et al. 2023; Mildenhall et al. 2021; Park et al. 2019; Shen et al. 2021a; Tewari et al. 2020; Thies et al. 2019], high-fidelity novel view synthesis can be obtained through reconstruction from posed visual observations. However, fine reconstructions often require high capture coverage and density.

Beyond static scenes, a comprehensive 4D human synthesis [Hilsmann et al. 2020; Li et al. 2024c; Orts-Escolano et al. 2016; Xu et al. 2024a], viewable from all angles, is more crucial for enhancing immersive experiences. However, 4D human reconstruction presents unique challenges because of its ill-posedness. For example, a static scene can be thoroughly documented over time using a smartphone; however, when it comes to a person in motion, we are limited to capturing only a partial snapshot at one moment with the same device. Therefore, 4D human novel view synthesis generally demands a synchronized and calibrated multi-view camera system [Cheng et al. 2023; Li et al. 2025; Xiong et al. 2024], which is both costly and sophisticated. Motivated by recent advancements in techniques [Wang et al. 2025a] and datasets [Li et al. 2025; Xiong et al. 2024], we believe it is the opportune moment to make a breakthrough: realizing a 360-degree human-centric dynamic novel view synthesis using only monocular inputs.

Diffusion Probabilistic Models [Ho et al. 2020; Sohl-Dickstein et al. 2015; Song and Ermon 2019] have witnessed huge success in recent years, particularly for image and video generation tasks. Certain diffusion-based models possess the capability to infer and generate the shape and appearance of an object's multiple views from a single frontal image, maintaining high spatial consistency [Kant et al. 2025; Liu et al. 2024, 2023c,b,a; Shi et al. 2023a,b; Voleti et al. 2024; Wang and Shi 2023; Watson et al. 2022]. Building upon these multi-view diffusion models, 4D generation is attainable by additionally enforcing the temporal consistency [Bahmani et al. 2024; Huang et al. 2025; Jiang et al. 2023b; Ling et al. 2024; Ren et al. 2023; Wu et al. 2024c; Zeng et al. 2024] through 4D representations [Fridovich-Keil et al. 2023; Wu et al. 2024b]. Although similar strategies can be directly applied to 4D human scenarios [Pang et al. 2025], their training processes are still expensive, and they remain inadequate for handling large motions and preserving temporal details due to limitations inherent in their foundation models.

Recent rapid evolution of video diffusion model [Blattmann et al. 2023a,b; Chen et al. 2023, 2024b; He et al. 2022; Hong et al. 2022; Lin et al. 2024b; Rombach et al. 2022; Wang et al. 2025a; Xing et al. 2023; Yang et al. 2024a] demonstrates its potential to function as a shader [Gu et al. 2025] and enable camera-controllable video generation [He et al. 2024; Wang et al. 2024b; Wu et al. 2024a]. It is possible to directly infer novel view video content through iteratively sampling and denoising, obviating the need for scene-specific training. Some works [Bai et al. 2025, 2024; Jiang et al. 2024c; Van Hoorick et al. 2024] redirect the camera trajectory via the injection of camera pose embeddings. However, these models generally converge at a relatively slow pace. Moreover, such an implicit condition typically demands a dense array of viewpoints in the training set to guarantee generalizability across arbitrary perspectives. Another line of works [Bian et al. 2025; Liu et al. 2025; Ren et al. 2025; Xiang et al. 2023; YU et al. 2025; Yu et al. 2024] explicitly employ depth geometric priors. They achieve 4D novel view synthesis by first applying depth-based warping and then employing video inpainting. Despite these successes, these works struggle to synthesize at very large viewpoint changes and faithfully preserve multi-view attributes. Apart from the limitations of training data, the reasons are still twofold (Fig. 2): (i) insufficient 3D cues from monocular inputs are provided to the network. (ii) image warping floater at large viewpoints change would be intolerable due to inaccurate monocular depth estimation.

In this paper, we focus on human-centric scenarios and present MV-Performer, a simple yet effective framework that transforms an input monocular video into multi-view synchronized videos. In particular, we extend the pre-trained WAN2.1 [Wang et al. 2025a] to a multi-view video diffusion model that learns the joint distribution of multi-view human-centric videos. To address the aforementioned issues, we devise a network tailored to the data characteristics of MVHumanNet [Xiong et al. 2024]. We contend that using implicit camera embeddings is unsuitable for MVHumanNet [Xiong et al. 2024] due to the limited camera views. To enable a 360-degree novel view synthesis with the explicit depth-based warping paradigm, we excavate additional condition information from the monocular depth. Specifically, we render the camera-dependent normal map from oriented point clouds, which aid the model in distinguishing between observed and unobserved areas. To ensure synchronization within different views and faithfulness toward the reference view, our multi-view video diffusion model adopts the multi-view attention and reference attention mechanisms, which efficiently fuse the information from the reference video, partial rendering, and different viewpoints. Additionally, we provide a robust inference procedure by integrating several state-of-the-art estimation methods [Khirodkar et al. 2024; Li et al. 2024b; Piccinelli et al. 2025], which significantly mitigate the artifacts induced by imperfect monocular depth estimation and provide better guidance to video generation.

Extensive experiments on MVHumanNet [Xiong et al. 2024], DNA-Rendering [Cheng et al. 2023], and collected in-the-wild datasets demonstrate the superior effectiveness and robustness of our proposed MV-Performer. In summary, our contributions are as follows:

- We develop the first generative framework for converting human-centric monocular video to dense multi-view videos, leveraging a cutting-edge video diffusion model and the MVHumanNet dataset.
- We propose a multi-view video diffusion model that learns the joint distribution of multi-view human-centric videos, guided by the normal map rendered from oriented partial point clouds. We show that the depth-based warping paradigm could also enable human appearance and motion synthesis under large viewpoint changes, harnessing the inherent power of the video diffusion model.
- To ensure the generalizability of our framework, we provide a robust inference procedure, which greatly mitigates the artifacts induced by imperfect monocular depth estimation.

2 Related work

2.1 Reconstruction-based 4D Human Modeling

4D novel view synthesis presents significant challenges, which are typically achieved by first reconstructing the dynamic scenes. Numerous highly efficient and expressive 4D representations [Cao and Johnson 2023; Duan et al. 2024; Huang et al. 2024a; Li et al. 2024a; Lin et al. 2024a; Shao et al. 2023; Wang et al. 2025c; Xu et al. 2024b; Yang et al. 2024b] are introduced to improve reconstruction performance. Recently, high-fidelity 4D human reconstruction has been widely investigated to achieve photorealistic digital avatar creation. Multi-view approaches, designed for studio environments with calibrated sensors, leverage diverse scene representations—such as volumetric occupancy fields [Huang et al. 2018], point clouds [Wu et al. 2020], and depth fusion [Yu et al. 2021]—to capture clothed human performances. The success of neural radiance fields (NeRF) [Mildenhall et al. 2020] further advanced this domain, follow-up works [Li et al. 2022, 2023; Liu et al. 2021; Peng et al. 2021a,b; Wang et al. 2022; Zhao et al. 2022a; Zheng et al. 2022, 2023; Zhi et al. 2022] utilize neural rendering techniques to learn a plausible implicit canonical geometry [Pumarola et al. 2021] of clothed humans. while recent work explores 3D Gaussian splatting [Kerbl et al. 2023] for efficient photo-realistic human rendering [Chen et al. 2025, 2024c; Jiang et al. 2024a,b; Li et al. 2024c; Pang et al. 2024; Qian et al. 2024a]. However, these methods rely on specialized hardware, restricting their applicability. In contrast, monocular reconstruction tackles the ill-posed challenge of inferring 3D geometry from single-view inputs [Kocabas et al. 2024; Wang et al. 2024a; Zhao et al. 2025]. NeRF-based works [Guo et al. 2023; Jiang et al. 2022a, 2023a, 2022b; Weng et al. 2022] adopted neural deformation fields to model dynamic humans from monocular videos. Inspired by these methods, recent advances [Hu et al. 2024a; Qian et al. 2024b; Wen et al. 2024; Zhi et al. 2025] optimize 3DGS primitives anchored to explicit [Loper et al. 2015; Pavlakos et al. 2019] or implicit templates [Shen et al. 2021b; Wang et al. 2021; Yariv et al. 2021], achieving articulated avatars with enhanced detail. However, such optimization-based frameworks typically require extensive optimization time to achieve satisfactory performance.

2.2 Generalizable 4D Human Novel View Synthesis

Neural rendering technologies [Mildenhall et al. 2020; Tewari et al. 2020] have demonstrated strong capabilities in generating high-fidelity renderings across multiple views. However, these methods are typically optimized for a single scene and require densely sampled input views for training. For general scenes, some representative works [Chen et al. 2021, 2024a; Xu et al. 2022] follow the multi-view stereo fashion and propose generic deep neural networks to directly regress neural parameters. To extend their applicability to new human performers and handle sparse-view inputs, later works [Chen et al. 2022b; Hu et al. 2023; Kwon et al. 2021; Mihajlovic et al. 2022; Zhao et al. 2022b] use 3D human prior to anchor the pixel-aligned features accurately on the human template. Although these techniques achieve good results, their rendering speed is slow due to the heavy computations in volume rendering. Recent methods [Hu et al. 2024b; Kwon et al. 2024; Zheng et al. 2024; Zhuang et al. 2024] utilize GPU-accelerated 3DGS rasterization [Kerbl et al. 2023] to achieve both high-speed and photorealistic human rendering from sparse observations. Nevertheless, these methods can only generate promising results for observed viewpoints and still struggle to synthesize fine details in unseen regions.

2.3 4D View Extrapolation via Video Diffusion Models

Diffusion models [Ho et al. 2020; Rombach et al. 2022; Song et al. 2020] have demonstrated remarkable promise in generating novel views from posed sparse view videos [Jin et al. 2025] or even from a monocular video. One line of works [Bai et al. 2025; He et al. 2024] encoding camera pose parameters into the video diffusion models for controlling the viewpoint of the output video. In another line, GEN3C [Ren et al. 2025], TrajectoryCrafter [YU et al. 2025], and others [Bian et al. 2025; Hu et al. 2025] converge on the concept of employing depth-based warping information as prior conditions. However, these models cannot effectively generate synchronized multi-view videos consistent with each other. Recent studies have extended beyond single-camera scenarios, focusing on multi-view video generation. SV4D [Xie et al. 2024] and CAT4D [Wu et al. 2024a] combine 3D shape and motion information from multi-view video diffusion to optimize implicit 4D representations. SynCamMaster [Bai et al. 2024] introduces a multi-view synchronization module to synthesize open-world multi-view videos from a single text prompt and desired viewpoints. For multi-view human video generation, Human4DiT [Shao et al. 2024] introduces a 4D diffusion transformer that disentangles image, viewpoint, and temporal learning. GAS [Lu et al. 2025] employs video diffusion models to enhance novel-view and pose synthesis results from Human NeRF reconstruction. However, these models primarily focus on pose-conditioned human animation from single-image inputs rather than 4D novel view synthesis from monocular videos, and some codes are not publicly available.

3 Method

Given a reference frontal full-body monocular video V^{ref}, comprising f frames, our goal is to synthesize m synchronized novel view human videos $\{V^1, V^2, ..., V^m\}$. These videos should accurately maintain consistency across different views. We tackle this problem by taming the power of MVHumanNet [Xiong et al. 2024] and pre-trained Wan2.1-T2V-1.3B [Wang et al. 2025a]. In this section, we first introduce our synchronized multi-view video diffusion model (Sec. 3.1). Then, we illustrate our camera-dependent normal map designed to handle large viewpoint changes (Sec. 3.2). Finally, we present the inference procedures for in-the-wild scenarios(Sec. 3.3).

3.1 Multi-View Video Diffusion Model with Depth-based Geometric Condition

The overview of our pipeline is shown in Fig. 3. One primary focus of our design is selecting an appropriate condition according to the dataset characteristics.

Depth-based warping. As mentioned in Sec. 1, the open-source multi-view datasets typically comprise 32 to 60 camera views, with the cameras fixed on capture cages. This setup results in a limited training view distribution. Therefore, instead of utilizing Plücker ray as the camera embedding [Bai et al. 2025, 2024; He et al. [n. d.]], we incorporate explicit 3D geometric priors for the precise control of camera viewpoint changes, following the depth-based warping paradigm used in [Bian et al. 2025; Ren et al. 2025; YU et al. 2025; Yu et al. 2024]. To construct the training pairs, we perform RGBD-warping with known camera parameters $\left\{Cam^{ref}, Cam^1, ..., Cam^m\right\}$. Specifically, given a frontal RGB image with its metric depth D, and corresponding camera parameter Cam^{ref} consisting of intrinsics K and extrinsics R, we first unproject the 2D pixels u into the colored partial point cloud X_{color} in the world coordinate:

$$X(u) = R^{-1}D(u)K^{-1}u \tag{1}$$

Subsequently, given new viewpoints $\{Cam^1, Cam^2, ..., Cam^m\}$, we render the per-frame colored point cloud into partial rendering $\mathcal{R}(X_{color}, Cam^i)$ for viewpoint i. In this way, we produce the partial rendering geometric cues for m target views $\{P^1, P^2, ..., P^m\}$. We feed the partial rendering $\{P^1, P^2, ..., P^m\}$ and normal $\{N^1, N^2, ..., N^m\}$ (see Sec. 3.2) geometric condition to the 3D-VAE of Wan2.1 separately and concatenate their output along the channel dimension, resulting in latent features Z^i_{cond}. We further concatenate them with the input noise latents Z_{noise} along the channel dimension.

For network finetuning, we adhere to the principle of simplicity. To achieve faithful and synchronized generation, we specifically modify the pre-trained Wan2.1-T2V-1.3B model [Wang et al. 2025a] by incorporating two primary components in each DiT block:

Ref Attention. The partial rendering explicitly represents the camera transformation and effectively provides the denoising network with known observations. However, some information will inevitably be lost due to occlusion. Inspired by [YU et al. 2025], we implement cross-attention mechanisms between Z_{in} and reference latents Z^{ref}, where Z_{in} denotes the hidden latents in each Dit block. We use Z_{in} as queries and the Z^{ref} as keys and values. The reference latents Z_{ref} are derived from V^{ref} via the VAE encoder in conjunction with a reference patch embedder.

$$Z_{out} = Z_{in} + proj(cross\_attn(Z_{in}, Z_{ref})) \tag{2}$$

The aggregated features are projected back to the original dimension with a zero-initialized linear layer and residual connection. Unlike YU et al. [2025], which incorporates additional attention layers, we reuse the textual cross attention layer for simplicity.

Sync Attention. Despite the consistent underlying 3D geometry P_{color}, challenges persist in maintaining consistency across various camera viewpoints. This issue is particularly pronounced when considering views from the rear. To aggregate information from the hidden latents $Z_{in} = concat(Z^{ref}_{in}, Z^1_{in}, ..., Z^m_{in})$ across different viewpoints, where m is the target number of views. We employ a frame-level spatial self-attention mechanism that functions as synchronized attention:

$$Z_{out} = Z_{in} + proj(self\_attn(Z^{ref}_{in}, Z^1_{in}, ..., Z^m_{in})) \tag{3}$$

The synchronized attention mechanism effectively aggregates per-frame information from multiple views and integrates it into the video diffusion model. Unlike Bai et al. [2024], we do not incorporate camera pose embedding into our model.

3.2 Camera-dependent Normal Map Condition

The previous warping-based method can only handle small viewpoint changes [Xiang et al. 2023; YU et al. 2025]. We attribute this limitation to the ambiguity between front and back perspectives under larger viewpoint changes. To address this issue, we propose to leverage camera-dependent normal map condition to facilitate 360-degree synthesis. As illustrated in Fig. 3, we adopt a view-dependent rendering strategy to provide an intuitive representation of surface orientation. Specifically, given the point cloud normal vector $\vec{n}$ and the camera viewing direction $\vec{d}$, both defined in the world coordinate system (with $\vec{d}$ derived from the camera's rotation matrix), we

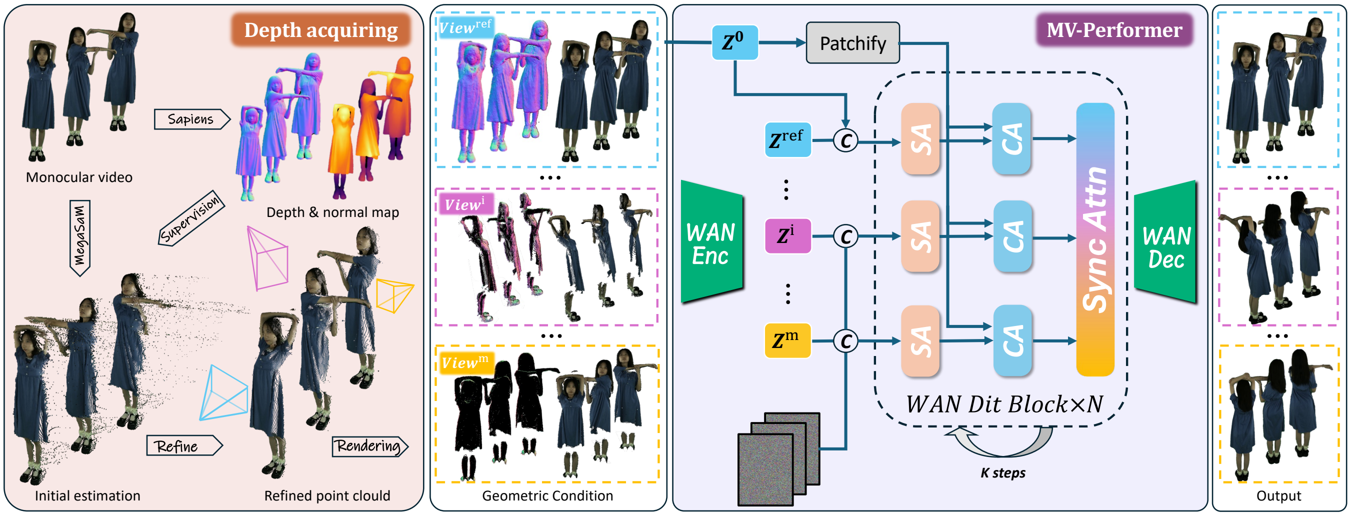

Fig. 3. The overview of our MV-Performer. "SA"and "CA" are abbreviations for self-attention and cross-attention, respectively. We first estimate the depth and normal from Sapiens [Khirodkar et al. 2024] and then use these estimates to refine the noisy point cloud output from MegaSaM [Li et al. 2024b]. Next, we render the refined point cloud with corresponding colors to novel views as geometric conditions. Finally, we feed them into MV-Performer to synthesize a 4D human video from novel viewpoints.

compute the dot product $o = \vec{n} \cdot \vec{d}$ for each point. The value of o indicates the surface orientation: $o > 0$ implies the surface is facing the camera, while $o < 0$ denotes it is facing away. For visualization, we map the normal vectors from the $[-1, 1]$ range to the RGB color space $[0, 1]$, and assign black to surfaces where $o < 0$, effectively masking back-facing areas. This strategy not only highlights the geometric structure of the point cloud but also conveys precise orientation cues, which are critical for accurate multi-view synthesis. We denote the camera-dependent normal map rendering videos as $\{N^1, N^2, ..., N^m\}$ for m target views.

3.3 Inference with Refined Monocular Depth

For in-the-wild inference, we need to perform the depth-based warping to get the partial rendering of the novel view, necessitating a metric depth estimation method. However, existing approaches [Piccinelli et al. 2025] continue to face challenges in producing high-fidelity depth outputs. Specifically, the depth drift toward the background significantly degrades the generation quality for large viewpoint changes, mainly due to the domain gap. To tackle this issue, we propose a depth refinement process by integrating several state-of-the-art estimation methods. Specifically, given a monocular video input $V^{ref} = \{I_0, I_1, ..., I_f\}$ comprising f frames, we first estimate the per-frame unified metric depth $\hat{D}_i$ and camera parameters using MegaSaM [Li et al. 2024b], and the high-quality relative depth $\tilde{D}_i$ normal map $\tilde{N}$ using Sapiens [Khirodkar et al. 2024]. Then, we align the relative depth $\tilde{D}_i$ to the coarse metric depth $\hat{D}_i$:

$$\arg\min_{\alpha,\beta} = ||(\alpha \cdot \tilde{D}_i + \beta) - \hat{D}_i||_2 \tag{4}$$

This can be effectively solved for scale and shift with a least-squares criterion which has a closed-form solution [Yu et al. 2022]. Finally, we further optimize the aligned depth using normal map $\tilde{N}$ with normal consistency supervision in 2DGS [Huang et al. 2024b] for around 100 iterations.

4 Experiments

4.1 Datasets

To access quantitative metrics, we conduct experiments on two extensively used multi-view human modeling datasets, MVHumanNet [Xiong et al. 2024] and DNA-Rendering [Cheng et al. 2023]. **We only use the training part of MVHumanNet [Xiong et al. 2024] as training set.** Additionally, we collect 5 monocular videos from Bilibili and TikTok to demonstrate generalizability.

MVHumanNet. MVHumanNet [Xiong et al. 2024] is a multi-view video dataset with over 9000 identities in everyday clothing. MVHumanNet++ [Li et al. 2025], an expanded version of MVHumanNet, offers additional depth, normal estimations, and more robust mask segmentation and SMPLX fitting. We utilized 16-view videos from a training set comprising 5,400 subjects for our training process. For evaluation purposes, we selected a test set consisting of 10 subjects. In this test set, we employed even-numbered views to conduct the assessment.

DNA-Rendering. DNA-Rendering [Cheng et al. 2023], another multi-view video dataset, features some professional actors and complicated clothing. In alignment with the MVHumanNet evaluation setup, we sampled 10 subjects from the 8 camera views subset. This dataset is utilized for evaluation purposes.

4.2 Baselines

To the best of our knowledge, we are among the first to concentrate on the subdomain of 360-degree, human-centric 4D novel

Table 1. Quantitative results on MVHumanNet and DNA-Rendering. ↓ indicates lower is better while ↑ indicates higher is better. ReCamMaster* is the finetuned version using MVHumanNet.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
MVHumanNet [Xiong et al. 2024]					
Champ	11.23	0.813	0.328	55.92	5.54
ReCamMaster	6.97	0.600	0.620	154.03	10.78
ReCamMaster*	11.62	0.817	0.287	26.44	2.17
TrajectoryCrafter	4.18	0.493	0.722	154.00	17.25
Ours	**24.35**	**0.926**	**0.066**	**24.47**	**0.12**
DNA-Rendering [Cheng et al. 2023]					
Champ	9.08	0.750	0.399	58.59	4.73
ReCamMaster	6.46	0.595	0.602	138.25	7.80
ReCamMaster*	10.02	0.769	0.342	36.78	4.28
TrajectoryCrafter	4.72	0.498	0.758	154.66	15.52
Ours	**15.63**	**0.861**	**0.152**	**30.05**	**0.73**

view synthesis from monocular input. As a result, there are limited established methods available for direct benchmarking.

We mainly compare MV-Performer with three baselines: TrajectoryCrafter [YU et al. 2025], ReCamMaster [Bai et al. 2025], and Champ [Zhu et al. 2024], where the first two methods are the state-of-the-art, open-sourced camera-controlled video diffusion models, and the last one is the human image animation method. We finetuned ReCamMaster [Bai et al. 2025] on MVHumanNet [Xiong et al. 2024] for 20 epochs to make a fairer comparison.

We do not compare to Human4Dit [Shao et al. 2024], and Disco4D [Pang et al. 2025] because they primarily focus on animation rather than 4D novel view synthesis. Moreover, they have not provided open-source code, and we face difficulties affording the training costs for reproducing Human4Dit [Shao et al. 2024].

4.3 Implementation Details

As noted by Bai et al. [2024], we also encounter challenges in directly optimizing our full pipeline. To address this, we implement a progressive training strategy. Our formulation allows for a natural decoupling of the pipeline into two distinct stages: first, video inpainting, followed by synchronization. In the initial stage, we refrain from incorporating the synchronization module and train all other parameters for 5 epochs. In the subsequent stage, our focus shifts to synchronization; thus, we freeze all other modules and exclusively train the synchronization module for an additional 5 epochs. Throughout both training phases, we utilize the AdamW [Loshchilov and Hutter 2017] optimizer set the learning rate at 1×10^{-4} and gradually decrease it to 2×10^{-5}. All experiments are conducted with an effective batch size of 6×12 on 6 NVIDIA A100. We perform $K = 50$ steps sampling for all experiments. Our full pipeline can simultaneously generate around 10 videos with 49 frames on a custom-level GPU with 24G memory like RTX3090.

4.4 Quantitative and Qualitative Results

Tab. 1 presents the quantitative results on two datasets, which show that existing models [Bai et al. 2025; YU et al. 2025; Zhu et al. 2024] are not good at this task. Our method is the first to achieve faithful and 360-degree synchronized multi-view synthesis from human-centric monocular video. We exhibit the qualitative comparisons using two datasets in Fig. 8 Fig. 9, respectively. It can be observed that MV-Performer outperforms all baselines by an order of magnitude. Notably, our generated frontal videos are nearly pixel-aligned with the frontal ground truth, while MV-Performer also produces consistent and reasonable back-view imagination. This is consistent with the reported FVD scores. Moreover, MV-Performer accepts only frontal-view videos as input, while the backside clothing patterns are synthesized by the video diffusion model. Although discrepancies exist between the generated backside textures and the ground truth, the results remain reasonable and acceptable. Visually, both ReCamMaster and TrajectoryCrafter can only produce plausible frontal views while struggling to generate significant viewpoint changes in the video. ReCamMaster*, the finetuned version model, shows improvements across all metrics. However, it remains deficient in fine-grained camera control and struggles with generalizing to out-of-distribution camera poses. This issue of leveraging implicit camera embedding is also highlighted in Tang et al. [2025]. Despite Champ [Zhu et al. 2024], being adapted from an image-based model rather than a native video generation framework, struggles to preserve identity consistency during animation. Besides, these methods fail to maintain consistency across different viewpoints. In contrast, our method is capable of generating coherent and faithful 360-degree multi-view synthesis, even in challenging scenarios involving complex clothing. For additional visual results of in-the-wild performers, please refer to the supplementary video.

4.5 Ablation Studies

We ablate each component in MV-Performer using MVHumanNet [Xiong et al. 2024], DNA-Rendering [Cheng et al. 2023] and in-the-wild dataset.

Table 2. Ablation Studies on the whole framework.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
w/o normal cond (A)	15.61	0.858	0.165	36.60	0.837
w/o sync module (B)	15.38	0.856	0.163	38.96	0.898
w/o (A) & w/o (B)	15.21	0.850	0.169	39.13	1.06
Ours full	**15.63**	**0.861**	**0.152**	**30.05**	**0.73**

Camera-dependent normal condition. To demonstrate the effectiveness of our proposed conditioning signal, we conducted an experiment where this signal was omitted during the finetuning process. As shown in Tab. 2, all metrics exhibit a noticeable degradation without camera-dependent normal condition. Furthermore, Fig. 4 illustrates these results more clearly. It can be observed that without the facilitation of our proposed condition signal, the model produces incorrect results due to the condition ambiguity, which indicates that the normal condition serves as a strong geometric cue, alleviating such errors. Notably, this can be regarded as a finetuned version

Fig. 4. Our proposed camera-dependent normal condition assists the model in distinguishing between observed and unobserved condition information, resulting in a more accurate 360-degree synthesis.

of TrajectoryCrafter [YU et al. 2025] on MVHumanNet [Xiong et al. 2024] dataset. We emphasize that our customized design plays a crucial role in addressing this challenging problem.

Sync module. As discussed in Sec. 3.1, most existing camera-controllable video diffusion models face challenges maintaining consistency across different views. To address this issue, we implement synchronization attention to improve 4D view consistency. As illustrated in Fig. 5, the incorporation of the view-sync module results in a more consistent and visually enhanced appearance. Furthermore, the synchronization operation can also enhance the quantitative performance.

Depth refinement. We evaluate the effectiveness of the depth refinement process on the in-the-wild data by replacing depth with the initial estimation from MegaSaM [Li et al. 2024b]. As exhibited in Fig. 6, it is evident that depth fidelity significantly influences the final results. Inaccurate depth maps result in noisy warping, and this issue intensifies with increasing viewpoint changes (from left to right). The model generates unnatural body appearances due to floaters in the condition signals near the human body. In contrast, our integrated depth refinement process mitigates these floaters caused by inaccurate monocular depth estimations, generating clean point clouds. We achieve high-quality generation outcomes with clean geometric cue conditions.

Sampling steps. We also show the influence of sampling steps in Tab. 3. Reducing the sampling steps leads to poorer performance, particularly in FID. 25-50 denoising steps strike a balance between quality and cost.

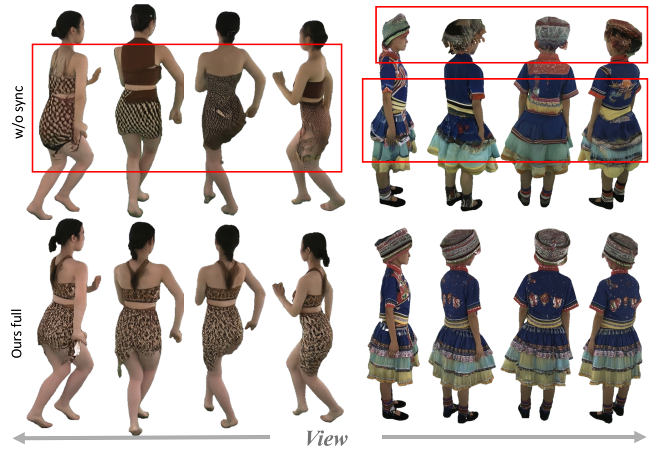

Fig. 5. The syncronization attention largely enhance the generation consistency across views.

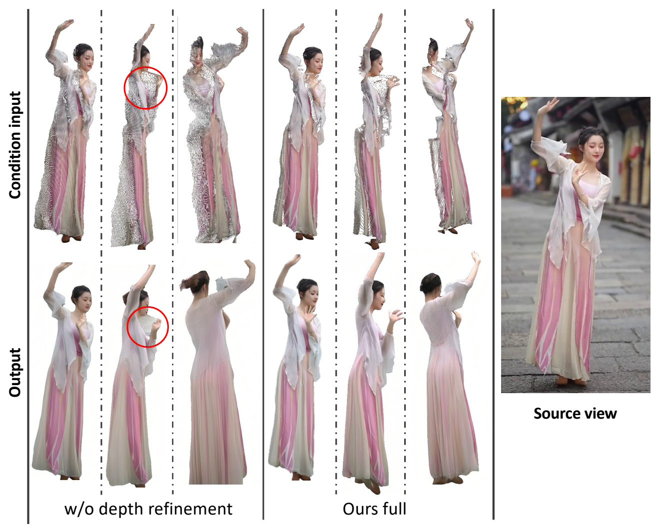

Fig. 6. The initial estimated point clouds contain floaters near the edges of the character, leading to bad guidance to the video diffusion model. In contrast, our method achieves clean estimations and yields pleasing results.

4.6 Application

An application of generative novel view synthesizers is to serve as generative priors [Jiang et al. 2024c; Liu et al. 2023b; Shi et al. 2023a; Tang et al. 2025; Yu et al. 2024]. We show that MV-Performer could potentially act as a prior for monocular avatar reconstruction. Without loss of generality, we add the comparison with GauHuman [Hu et al. 2024a] on MVHumanNet [Xiong et al. 2024]. Specifically, we use MV-Performer to generate two side-view and one back-view videos from frontal view videos as priors. We combine them with original frontal view videos to train GauHuman [Hu et al. 2024a]. As shown in Fig. 7 and Tab. 4, due to limited observations, GauHuman [Hu et al. 2024a] produces strong artifacts when viewed from the rear, resulting in poorer results. After incorporating the prior, we

Table 3. Performance under different sampling steps.

Steps	PSNR ↑	SSIM ↑	LPIPS ↑	FID ↓	FVD ↓
MVHumanNet [Xiong et al. 2024]					
5	24.90	0.931	0.078	55.54	0.14
10	24.65	0.929	0.074	43.43	0.13
25	24.40	0.927	0.069	30.26	0.12
50	24.35	0.926	0.066	24.47	0.12
DNA-Rendering [Cheng et al. 2023]					
5	15.72	0.864	0.166	54.85	0.74
10	15.65	0.862	0.161	45.00	0.74
25	15.63	0.861	0.155	34.97	0.74
50	15.63	0.861	0.152	30.05	0.73

observe performance improvements across all metrics, reducing the artifacts behind the performers. Fig. 7 and Tab. 4 also reveal the potential of directly using the video diffusion model to perform 4D novel view synthesis.

Table 4. We validate the effectiveness of prior on MVHumanNet.

Methods	PSNR↑	SSIM↑	LPIPS↑	FID↓	FVD↓
GauHuman	18.63	0.866	0.179	129.35	5.96
GauHuman+Prior	20.97	0.901	0.146	60.02	1.81
MV-Performer	24.35	0.926	0.066	24.47	0.12

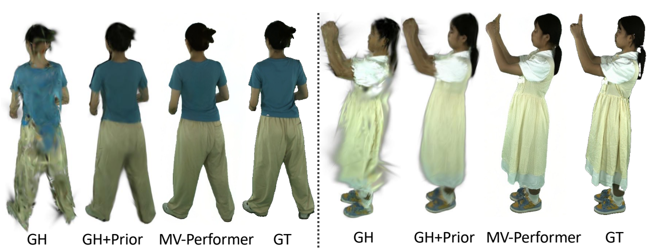

Fig. 7. Using MV-Performer as a generative prior. "GH" means GauHuman [Hu et al. 2024a]

5 Limitations

For the training process, despite the robust VAE offered by WAN2.1 [Wang et al. 2025a], preserving face region details remains challenging due to reconstruction errors, which limit the upper bounds of the human generation quality. For inference, MV-Performer essentially counts on the stability of the depth estimation methods [Li et al. 2024b; Piccinelli et al. 2025]. Our generated results would fail when faced with poor depth estimation. However, this problem could be solved by finetuning the depth estimation model with the metric human depth in MVHumanNet++ [Li et al. 2025]. Moreover, the video diffusion model generally requires multi-step denoising during inference, resulting in relatively high computational overhead and slow inference speed. Distilling MV-Performer into a smaller and one-step denoising version [Wang et al. 2025b] is a promising direction toward practical application. MV-Performer may degrade in quality for untrained origin and certain skin tones, which is limited by the potential bias in WAN2.1 and the existing dataset. Finally, limited by the computational resource, we can only conduct experiments on the 1.3B version of WAN2.1 [Wang et al. 2025a].

6 Conclusion

In this paper, we present MV-Performer, a novel framework for 360-degree human-centric novel view synthesis from monocular full-body videos. To address the limitations of existing warping-based methods, which often struggle with significant viewpoint changes, we introduce a camera-dependent normal map geometric condition signal. This approach effectively resolves the ambiguity between seen and unseen regions of the input human performer. Furthermore, we proposed a robust inference procedure to handle in-the-wild videos, significantly reducing artifacts caused by imperfect monocular depth estimation. Benefiting from the aforementioned design, our multi-view human-centric video diffusion model ensures temporal and geometric consistency across synthesized viewpoints. Extensive experiments on three datasets validate that MV-Performer outperforms the existing camera-controllable video diffusion model, establishing a strong model for 4D human-centric novel view synthesis. Our framework opens new possibilities for immersive VR/AR, free-viewpoint video, and synthetic data generation, which will benefit numerous downstream tasks.

Acknowledgments

The work was supported in part by NSFC with Grant No. 62293482, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Guangdong Provincial Outstanding Youth Fund (No. 2023B1515020055), by Shenzhen Science and Technology Program No. JCYJ20220530143604010, NSFC No. 62172348, the Shenzhen Outstanding Talents Training Fund 202002, the Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. SYSPG20241211173853027).

References

Shai Avidan and Amnon Shashua. 1997. Novel view synthesis in tensor space. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1034–1040.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 2024. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7996–8006.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. 2025. ReCamMaster: Camera-Controlled Generative Rendering from A Single Video. arXiv:2503.11647 [cs.CV] https://arxiv.org/abs/2503.11647

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. 2024. SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints. arXiv:2412.07760 [cs.CV] https://arxiv.org/abs/2412.07760

Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. 2025. GS-DiT: Advancing Video Generation with Pseudo 4D Gaussian Fields through Efficient Dense 3D Point Tracking. *arXiv preprint arXiv:2501.02690* (2025).

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22563–22575.

Ang Cao and Justin Johnson. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 130–141.

Gaurav Chaurasia, Olga Sorkine, and George Drettakis. 2011. Silhouette-Aware Warping for Image-Based Rendering. In *Computer Graphics Forum*, Vol. 30. Wiley Online Library, 1223–1232.

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022a. TensoRF: Tensorial Radiance Fields. In *European Conference on Computer Vision (ECCV)*.

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14124–14133.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. arXiv:2310.19512 [cs.CV]

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024b. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. arXiv:2401.09047 [cs.CV]

Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. 2025. TaoAvatar: Real-Time Lifelike Full-Body Talking Avatars for Augmented Reality via 3D Gaussian Splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 10723–10734.

Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. 2022b. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *European Conference on Computer Vision*. Springer, 222–239.

Shenchang Eric Chen and Lance Williams. 2023. View interpolation for image synthesis. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 423–432.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. 2024a. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*. Springer, 370–386.

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

Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. 2023. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering. *arXiv preprint* arXiv:2307.10173 (2023).

Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 2024. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12479–12488.

Yasutaka Furukawa, Carlos Hernández, et al. 2015. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision* 9, 1-2 (2015), 1–148.

Chris A Glasbey and Kantilal Vardichand Mardia. 1998. A review of image-warping methods. *Journal of applied statistics* 25, 2 (1998), 155–171.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. 2025. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control. *arXiv preprint arXiv:2501.03847* (2025).

Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 2023. 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*. 12858–12868.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. [n. d.]. CameraCtrl: Enabling Camera Control for Video Diffusion Models. In *The Thirteenth International Conference on Learning Representations*.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent Video Diffusion Models for High-Fidelity Long Video Generation. (2022). arXiv:2211.13221 [cs.CV]

Anna Hilsmann, Philipp Fechteler, Wieland Morgenstern, Wolfgang Paier, Ingo Feldmann, Oliver Schreer, and Peter Eisert. 2020. Going beyond free viewpoint: creating animatable volumetric video of human performances. *IET Computer Vision* 14, 6 (2020), 350–358.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022).

Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. 2023. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9352–9364.

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

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. 2025. EX-4D: EXtreme Viewpoint 4D Video Synthesis via Depth Watertight Mesh. arXiv:2506.05554 [cs.CV] https://arxiv.org/abs/2506.05554

Yingdong Hu, Zhening Liu, Jiawei Shao, Zehong Lin, and Jun Zhang. 2024b. Eva-Gaussian: 3D Gaussian-based real-time human novel view synthesis under diverse camera settings. *arXiv preprint arXiv:2410.01425* (2024).

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024b. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery. doi:10.1145/3641519.3657428

Hanzhuo Huang, Yuan Liu, Ge Zheng, Jiepeng Wang, Zhiyang Dou, and Sibei Yang. 2025. Mvtokenflow: High-quality 4d content generation using multiview token flow. *arXiv preprint arXiv:2502.11697* (2025).

Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. 2024a. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4220–4230.

Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. 2018. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 336–354.

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

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

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

Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. 2020. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1251–1261.

Yuheng Jiang, Zhehao Shen, Yu Hong, Chengcheng Guo, Yize Wu, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2024a. Robust dual gaussian splatting for immersive human-centric volumetric videos. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–15.

Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2024b. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19734–19745.

Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. 2024c. Animate3d: Animating any 3d model with multi-view video diffusion. *arXiv preprint arXiv:2407.11398* (2024).

Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. 2023b. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848* (2023).

Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 2025. Diffuman4D: 4D Consistent Human View Synthesis from Sparse-View Videos with Spatio-Temporal Diffusion Models. In *International Conference on Computer Vision (ICCV)*.

Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. 2025. Pippo: High-Resolution Multi-View Humans from a Single Image. (2025).

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

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2024. Sapiens: Foundation for Human Vision Models. *arXiv preprint arXiv:2408.12569* (2024).

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. 2024. HUGS: Human Gaussian Splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2311.17910

Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. 2024. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision*. Springer, 451–468.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. 2021. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems* 34 (2021), 24741–24752.

Marc Levoy and Pat Hanrahan. 2023. Light field rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 441–452.

Chenghong Li, Hongjie Liao, Yihao Zhi, Xihe Yang, Zhengwentai Sun, Jiahao Chang, Shuguang Cui, and Xiaoguang Han. 2025. MVHumanNet++: A Large-scale Dataset of Multi-view Daily Dressing Human Captures with Richer Annotations for 3D Human Digitization. *arXiv preprint arXiv:2505.01838* (2025).

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

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. 2024a. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8508–8520.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. 2024b. MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos. In *arxiv*.

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

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. 2024c. 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*. 19711–19722.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. 2024b. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131* (2024).

Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. 2024a. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21136–21145.

Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. 2024. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8576–8588.

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

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. 2024. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10072–10083.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2023c. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems* 36 (2023), 22226–22246.

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

Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. 2025. Free4D: Tuning-free 4D Scene Generation with Spatial-Temporal Consistency. *arXiv preprint arXiv:2503.20785* (2025).

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023a. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453* (2023).

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

Yixing Lu, Junting Dong, Youngjoong Kwon, Qin Zhao, Bo Dai, and Fernando De la Torre. 2025. GAS: Generative Avatar Synthesis from a Single Image. *arXiv preprint arXiv:2502.06957* (2025).

Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. 2022. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*. Springer, 179–197.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*. 741–754.

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

Hui En Pang, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. 2025. Disco4D: Disentangled 4D Human Generation and Animation from a Single Image. In *CVPR*.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14314–14323.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. 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*. 9054–9063.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. 2025. UniDepthV2: Universal Monocular Metric Depth Estimation Made Simpler. arXiv:2502.20110 [cs.CV] https://arxiv.org/abs/2502.20110

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10318–10327.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2024a. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20299–20309.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 2024b. 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting. In *CVPR*.

Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. 2023. DreamGaussian4D: Generative 4D Gaussian Splatting. *arXiv preprint arXiv:2312.17142* (2023).

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms.

In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 1. IEEE, 519–528.

Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. 2024. Human4DiT: 360-degree Human Video Generation with 4D Diffusion Transformer. *ACM Transactions on Graphics (TOG)* 43, 6 (2024).

Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. 2023. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16632–16642.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021a. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021b. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. 2023a. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110* (2023).

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2023b. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512* (2023).

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. pmlr, 2256–2265.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, and Matthias Niessner. 2025. Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5546–5558.

Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. 2020. State of the art on neural rendering. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 701–727.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. 2024. Generative Camera Dolly: Extreme Monocular Dynamic Novel View Synthesis. *European Conference on Computer Vision (ECCV)* (2024).

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

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025a. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

Hanyang Wang, Fangfu Liu, Jiawei Chi, and Yueqi Duan. 2025b. VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step. *arXiv preprint arXiv:2504.01956* (2025).

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *arXiv preprint arXiv:2106.10689* (2021).

Peng Wang and Yichun Shi. 2023. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201* (2023).

Shaofei Wang, Božidar Antić, Andreas Geiger, and Siyu Tang. 2024a. IntrinsicAvatar: Physically Based Inverse Rendering of Dynamic Humans from Monocular Videos via Explicit Ray Tracing. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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

Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, and Xiaowei Zhou. 2025c. FreeTimeGS: Free Gaussian Primitives at Anytime Anywhere for Dynamic Scene Reconstruction. In *CVPR*. https://zju3dv.github.io/freetimegs

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. 2024b. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. 2022. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628* (2022).

Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. 2024. 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*. 2059–2069.

Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*. 16210–16220.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024b. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20310–20320.

Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. 2020. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1682–1691.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. 2024a. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613* (2024).

Zijie Wu, Chaohui Yu, Yanqin Jiang, Chenjie Cao, Fan Wang, and Xiang Bai. 2024c. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In *European Conference on Computer Vision*. Springer, 361–379.

Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 2023. 3d-aware image generation using 2d diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2383–2393.

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

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. 2023. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. (2023). arXiv:2310.12190 [cs.CV]

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. 2024. MVHumanNet: A Large-scale Dataset of Multi-view Daily Dressing Human Captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. 2022. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5438–5448.

Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 2024a. 4k4d: Real-time 4d view synthesis at 4k resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20029–20040.

Zhen Xu, Yinghao Xu, Zhiyuan Yu, Sida Peng, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. 2024b. Representing Long Volumetric Video with Temporal Gaussian Hierarchy. *ACM Transactions on Graphics* 43, 6 (November 2024). https://zju3dv.github.io/longvolcap

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024a. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. 2024b. Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting. In *International Conference on Learning Representations (ICLR)*.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems* 34 (2021), 4805–4815.

Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. 2025. TrajectoryCrafter: Redirecting Camera Trajectory for Monocular Videos via Diffusion Models. *arXiv preprint arXiv:2503.05638* (2025).

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5746–5756.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. *arXiv preprint arXiv:2409.02048* (2024).

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems* 35 (2022), 25018–25032.

Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. 2024. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*. Springer, 163–179.

Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. 2022a. Human performance modeling and rendering via neural animated mesh. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–17.

Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2022b. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7743–7753.

Yiqun Zhao, Chenming Wu, Binbin Huang, Yihao Zhi, Chen Zhao, Jingdong Wang, and Shenghua Gao. 2025. Surfel-based Gaussian Inverse Rendering for Fast and Relightable Dynamic Human Reconstruction from Monocular Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

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

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

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

Yihao Zhi, Shenhan Qian, Xinhao Yan, and Shenghua Gao. 2022. Dual-space nerf: Learning animatable avatars and scene lighting in separate spaces. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 1–10.

Yihao Zhi, Wanhu Sun, Jiahao Chang, Chongjie Ye, Wensen Feng, and Xiaoguang Han. 2025. StruGauAvatar: Learning Structured 3D Gaussians for Animatable Avatars from Monocular Videos. *IEEE Transactions on Visualization and Computer Graphics* (2025).

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. arXiv:2403.14781 [cs.CV]

Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. 2024. IDOL: Instant Photorealistic 3D Human Creation from a Single Image. *arXiv preprint arXiv:2412.14963* (2024).

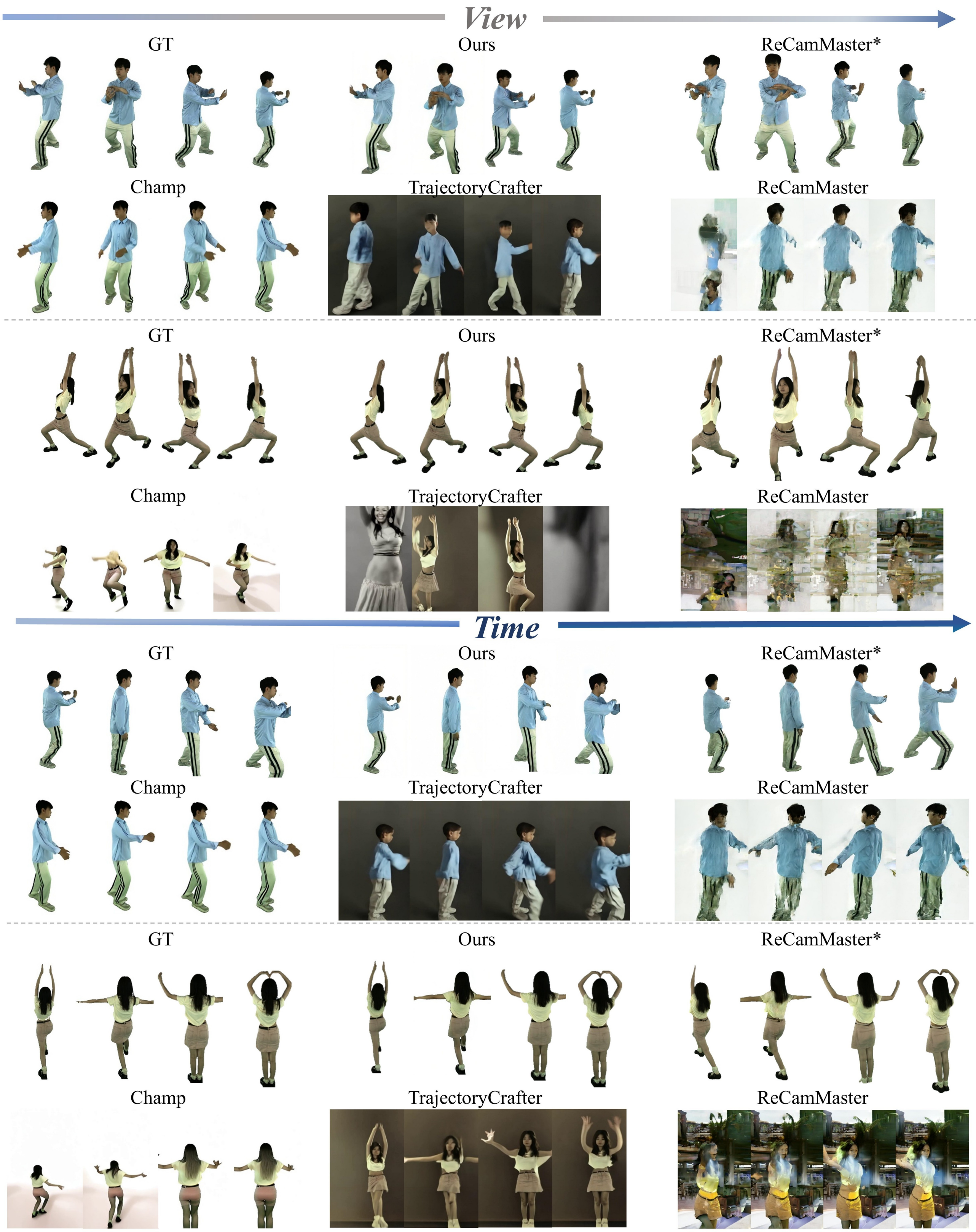

Fig. 8. **Comparison with state-of-the-art methods tested on MVHumanNet dataset. ReCamMaster* is the finetuned version using MVHumanNet.**

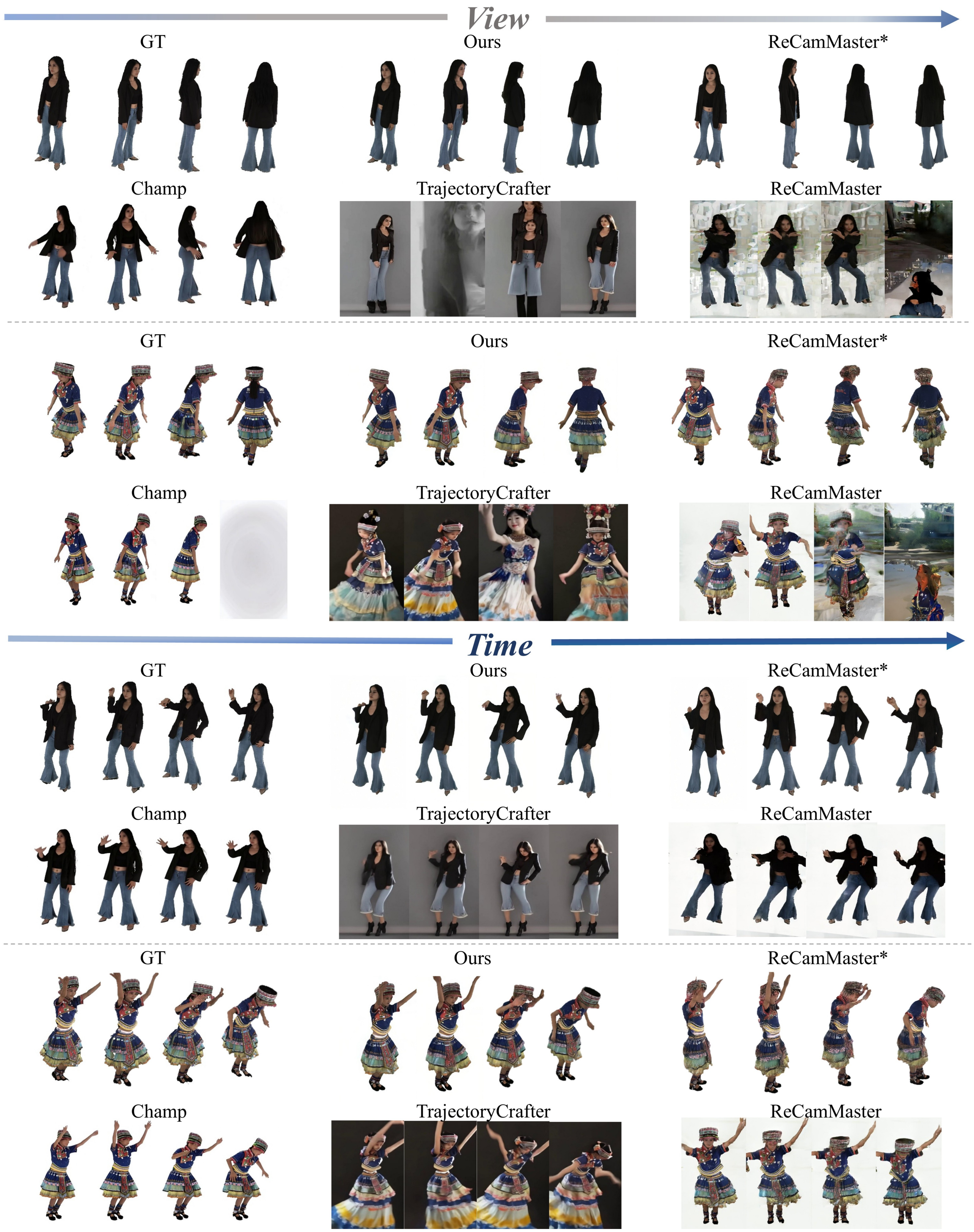

Fig. 9. **Comparison with state-of-the-art methods tested on DNA-rendering dataset. ReCamMaster* is the finetuned version using MVHumanNet.**