GAS: Generative Avatar Synthesis from a Single Image
Supplementary Material

A Limitations and future works

Although our method achieves state-of-the-art performance in both novel view and pose synthesis, it is not free from limitations. (1) Due to the constraints of current off-the-shelf single-image human mesh recovery methods, we use SMPL to establish geometry and appearance conditions for scalable training. However, SMPL lacks expressiveness in regions such as the face and hands, resulting in artifacts in these areas. Future works could explore scalable solutions inspired by recent efforts, such as regional supervision [8]. (2) While we utilize Stable Video Diffusion to model view and pose synthesis, achieving strong consistency across both spatial and temporal dimensions, some challenges remain. Specifically, we occasionally observe degradation in fine-grained detail quality and difficulties in accurately generating complex clothing textures, particularly during significant clothing deformations. Addressing the former may involve increasing the image resolution and exploring advanced video generative models, such as CogVideoX [9]. For the latter, leveraging synthetic human datasets with intricate textures presents an exciting avenue for future research.

B Ethical considerations

Human subjects data. We adhere to strict ethical guidelines in the collection and use of data involving human subjects. Below, we provide details on how we obtained each dataset utilized in our work:

- THuman2.1 [10] is a publicly available dataset. We signed the necessary agreement with the dataset authors to obtain access via an official download link.
- 2K2K [3] and MVHumanNet [7] follow the same procedure as THuman2.1, involving an agreement with the authors to obtain official download links.
- TikTok Dataset [6] is publicly available and was downloaded in its preprocessed form from MagicPose [1].
- Additional Real-World Data was manually selected and filtered from the publicly released portions of Champ [11] training data.

Broader Impact. Our proposed method enables affordable and accessible solutions for a wide range of applications. By leveraging generative AI, we transform content generation, making it possible to generate novel views and poses from just a casually-captured single image. It has the potential to revolutionize fields such as virtual reality, gaming, and digital content creation by significantly lowering the barriers to high-quality multi-view synthesis and animation.

However, along with these benefits come potential risks that warrant careful consideration. The misuse of generative AI in creating synthetic content raises ethical concerns, such as the possibility of producing misleading or harmful media. Additionally, privacy concerns may arise when real-world data is used for training, especially when the data involves human subjects. Ensuring robust safeguards, transparent practices, and compliance with ethical standards is crucial to mitigate these risks while maximizing the positive impact of our method.

C Implementation details

C.1 Model architecture

We detail the processing of each conditioning input and their integration into the video diffusion model, as illustrated in Figure 1. Starting with a single reference image $I_{ref} \in \mathbb{R}^{H \times W \times C}$, we extract its VAE latent representation, repeat it T times, and concatenate it with the input noise latent along the channel dimension. This combined representation is passed through a convolutional layer to generate $C_{\text{vae}} \in \mathbb{R}^{T \times H_1 \times W_1 \times C_1}$.

For the geometry cue, represented as a sequence of SMPL normal maps, we use a 2D ConvNet $\boldsymbol{\epsilon}_{\text{geo}}$ to extract features $C_{\text{smpl}} \in \mathbb{R}^{T \times H_1 \times W_1 \times C_1}$. Similarly, for the appearance cue, which consists of corresponding NeRF renderings, we pass the sequence through a VAE and subsequently a 2D ConvNet $\boldsymbol{\epsilon}_{\text{appr}}$ to obtain features $C_{\text{nerf}} \in \mathbb{R}^{T \times H_1 \times W_1 \times C_1}$.

The feature representations C_{vae}, C_{smpl}, and C_{nerf} are element-wise added and fed into the diffusion UNet $\mathcal{U}_\Theta$ to predict noise. Additionally, the CLIP embedding of the reference image, $\boldsymbol{h}_{\text{clip}} \in \mathbb{R}^d$, is repeated T times to match the frame sequence and injected into $\mathcal{U}_\Theta$ via cross-attention.

Finally, the switcher, represented as a one-hot vector, is

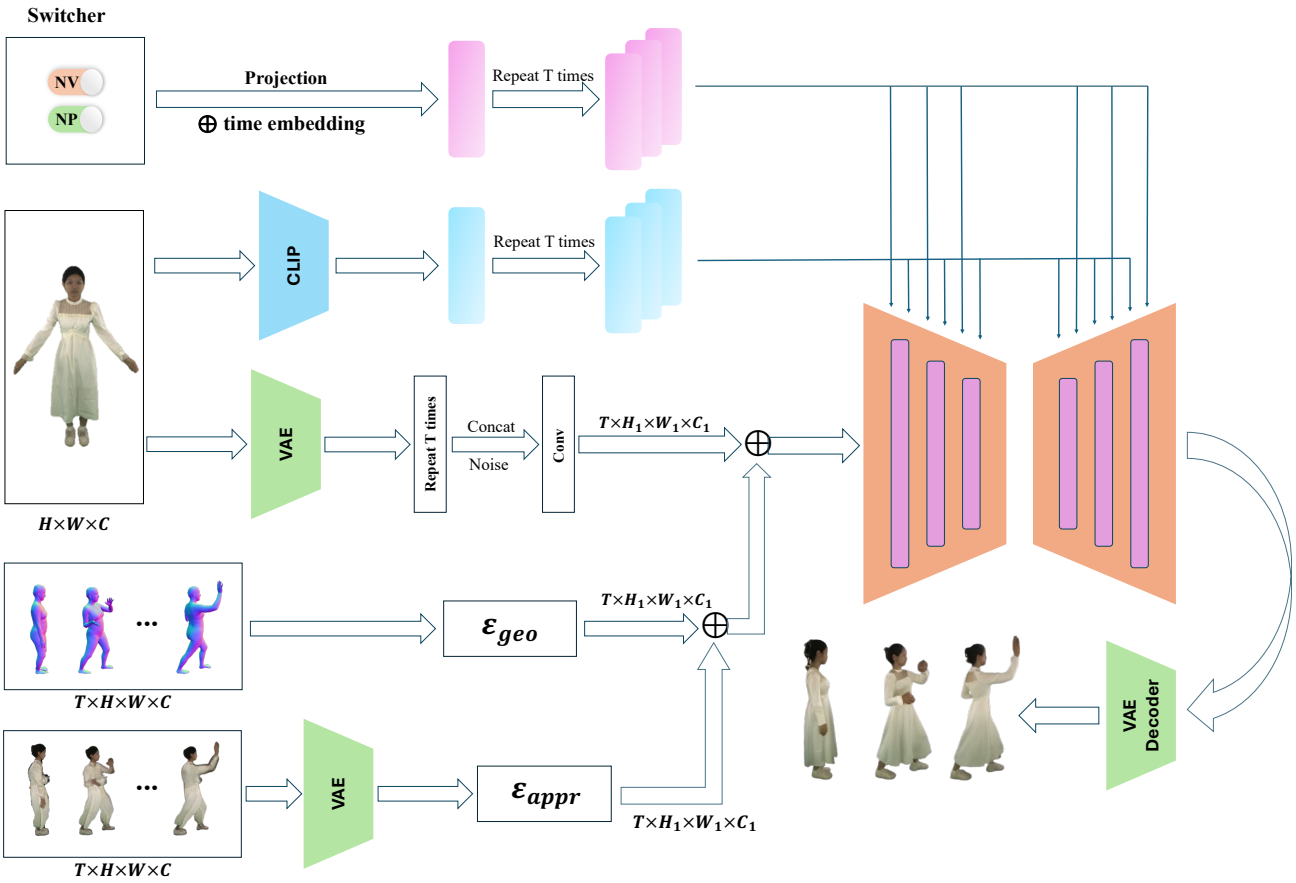

Figure 1. **Network architecture for processing the conditions of the video diffusion model.** ϵ_{geo} and ϵ_{appr} denote geometry encoder and appearance encoder, respectively. $\oplus$ denotes element-wise addition. The switcher embedding and time embedding are injected into the diffusion model in the ResNet layers; the CLIP embedding is injected through the cross attention mechanism.

embedded and element-wise added to the time embedding. This is also repeated T times to align with the frame sequence and injected into $\mathcal{U}_\Theta$ within the ResNet layers.

C.2 Training details

We leverage a combination of 3D human scans, multi-view videos and monocular videos to train our diffusion UNet $\mathcal{U}_\Theta$, geometry cue encoder ϵ_{geo} and appearance cue encoder ϵ_{appr}.

For 3D scans, we utilize 20 renderings for training the novel view synthesis task. A view is randomly selected as the reference, and the model is tasked with predicting all 20 consecutive novel views. Note that the reference view and the predicted starting view do not need to be the same. Furthermore, the order of the predicted views is randomly determined, *i.e.*, either clockwise or counterclockwise.

For multi-view videos, we train the model for the novel pose synthesis task. A frame is randomly selected as the reference image, and the model is tasked with predicting a 20-frame video clip. For a single pose, we also experiment with the novel view synthesis task. However, due to the fluctuating camera trajectory and sparse camera setup, we observe suboptimal outcomes.

For monocular videos, each video is split into a sequence of images at 30 frames per second. We sample one image every four consecutive frames for training. Similarly, a frame is randomly chosen as the reference, and the model is tasked with predicting 20 consecutive frames.

D Additional results

D.1 Runtime at inference

Given a single in-the-wild image, we report the runtime of our proposed method, in terms of both novel view and pose synthesis. The detailed runtime breakdown is as follows. (1) geometry cue rendering; (2) appearance cue rendering; (3) video diffusion inference. The runtime is reported on a single NVIDIA A800 GPU and measured in seconds.

Novel view synthesis. We report the runtime for generating 20 novel views given a single input image, as shown in

Table 1.

	20 frames	Per frame
Geo. cue rendering	5.81	0.29
Appr. cue rendering	28.6	1.43
Diffusion inference	15.88	0.79
Total runtime	50.29	2.51

Table 1. **Runtime at inference for generating 20 novel views of a single human image.**

Novel pose synthesis. We report the runtime for generating 100 consecutive novel poses from a single input image, as shown in Table 2. The additional video diffusion inference time arises due to the 6-frame overlap between consecutive video segment windows.

	100 frames	Per frame
Geo. cue rendering	29	0.29
Appr. cue rendering	143	1.43
Diffusion inference	106.49	1.06
Total runtime	278.49	2.78

Table 2. **Runtime at inference for generating 50 consucutive novel poses from a single human image.**

Efficiency comparison. We provide a comparison of runtime and VRAM usage with the strongest baseline Champ [11], as shown in Table 3.

	fps	VRAM (GB)
Champ	0.57	9.88
Ours	0.40	5.32

Table 3. **Efficiency comparison with baseline method.**

D.2 Ablation on merging novel view & pose tasks

We provide quantitative ablation analysis on merging both novel view and novel pose tasks into one model. From the static novel view synthesis perspective, the unified framework allow us to train on the abundant internet videos - which leads to the improved generalization, as shown in the main paper. It does not hurt the in-domain dataset performance according to Table 6.

From the pose animation perspective, training on additional 3D datasets can improve the quality when we animate the avatar from a novel view, as shown in Table 4.

Training data	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
Monocular videos	28.67	**0.946**	0.041	208.3
Full	**28.74**	0.945	**0.040**	**188.5**

Table 4. **Quantitative ablation on merging view synthesis and pose animation into one model.** Results reported for novel view animation task on MVHumanNet dataset.

D.3 Ablation on switcher

In addition to the qualitative ablation for the switcher presented in the main paper, we also conduct quantitative ablation analysis on the switcher. With the switcher, our method effectively supports both novel view and pose synthesis, while also providing comparable or even better quantitative results, as shown in Table 5.

	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
w.o. switcher	26.58	**0.944**	0.042	198.9
w. switcher	**26.77**	0.943	**0.041**	**194.8**

Table 5. **Ablation on switcher for novel view synthesis task on THuman.**

D.4 Novel pose results on scan datasets

We show the novel pose animation results for subjects in THuman and 2K2K in Figure 2, where the reference images are animated through pose sequences derived from disparate videos.

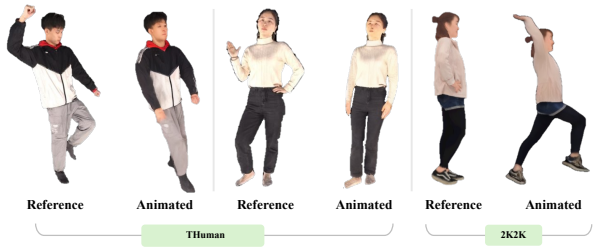

Figure 2. **Novel pose results on THuman and 2K2K.** The reference images are animated by pose sequences derived from MVHumanNet dataset.

D.5 Robustness to input view angles

We train our model using an arbitrary view as input. Thus, we are interested in the novel view synthesis robustness from different input view angles. To this end, we present our quantitative results in Table 7. We find that our pipeline demonstrates robustness with different views of input. Visualized results are shown in Figure 6, where our proposed method can maintain the faithful appearance near the reference view and generate reasonable appearances in unseen regions.

D.6 Ablation on video diffusion model

Fig. 3 (left) presents quantitative ablations on both novel view and pose synthesis tasks, showing consistent improvements over the direct output of the generalizable NeRF results, denoted as before diff. Fig. 3 (right) shows qualitative results on MVHumanNet dataset: NeRF helps the diffusion

Training data	PSNR ↑		SSIM ↑		LPIPS ↓		FVD ↓	
	THuman	2K2K	THuman	2K2K	THuman	2K2K	THuman	2K2K
3D scans only	**26.82**	28.76	0.936	0.953	**0.040**	0.040	**189.5**	**187.9**
3D scans+dynamic videos	26.77	**28.82**	**0.943**	**0.954**	0.041	**0.039**	194.8	191.3

Table 6. **Quantitative ablation on merging view synthesis and pose animation into one model.** Results reported for novel view synthesis task on in-domain THuman and 2K2K testset.

Angle	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
Front	28.46	0.952	0.04	178.3
Back	28.34	0.952	0.04	211.5
Left side	29.18	0.956	0.039	199.9
Right side	29.29	0.955	0.038	175.2
Mean	28.82	0.954	0.039	191.3
Std	0.487	0.002	0.001	17.42

Table 7. **Quantitative results on 2K2K dataset for novel view synthesis from different input view angles.**

model preserve identity on nearby views, while the diffusion model refines distant views, producing sharper results with fewer artifacts.

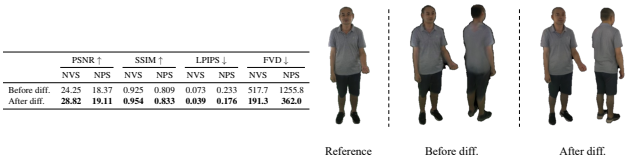

Figure 3. Quantitative and qualitative results before/after diff.

D.7 Video results

We provide additional video results, including in-the-wild avatar synthesis, comparison with baselines, ablations and additional results. Details explained below:

- In-the-wild avatar synthesis: We demonstrate novel view synthesis, pose animation and interactive 4D video synthesis on in-the-wild avatars.
- Comparison with baselines: We compare our results with Animate Anyone [4] and Champ [11] on THuman [10] for novel view synthesis and TikTok [6] for pose animation. The video results demonstrate our method outperforms the two baseline methods in terms of consistency and quality.
- Ablations: We highlight the importance of the dense appearance cue and switcher through the ablation videos.
- Additional results: We show novel view synthesis results on MVHumanNet [7]. We also show free-view animation results, where the human subject and the camera are both moving.

E Verification of design choices

We found that generative human novel view synthesis remains relatively under-explored, primarily due to the challenge of synthesizing consistent appearances across multiple novel views, especially in unseen regions. We have shown the importance of generalizable appearance cue and geometry cue in the main paper. Here, we aim to additionally validate the choice of leveraging video diffusion models and training with smooth camera orbits.

E.1 Image v.s. video diffusion model

Image diffusion model is believed to be good at image-to-image translation tasks. Specifically, given a single reference image and the target appearance cue, we can train a image diffusion model to synthesize the target image. The model architecture can be a variant of Animate Anyone [4] and Champ [11], where we adopt two ReferenceNets to inject the rich information from the reference image and appearance cue respectively. The results are shown in Figure 4. To enable multi-view consistent synthesis for a single subject, we have tried adding an 1D temporal-axis attention layers [2] and only fine-tune these new added layers. We have also attempted to leverage human geometric prior to construct 3D correspondence across different views [5].

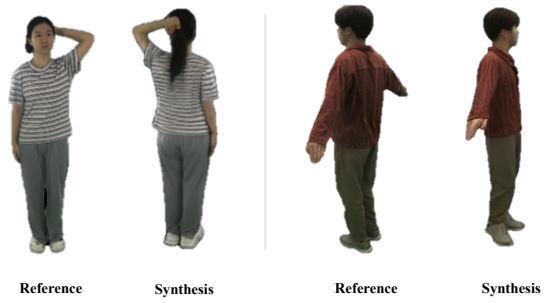

Figure 4. **Image-to-image novel view synthesis on MVHumanNet dataset by using image diffusion model and appearance cue.**

Consistency comparisons. Compared to directly using a pretrained video diffusion model, these image-based diffusion methods exhibit significantly inferior performance. Their results are shown in Figure 5.

Free-view interpolation. Due to GPU memory limitations, we are restricted to training with approximately 20 novel views per subject per batch. During inference, we also test and compare the ability of both models to generate a dense trajectory of novel views (*e.g.,* 100 views). However,

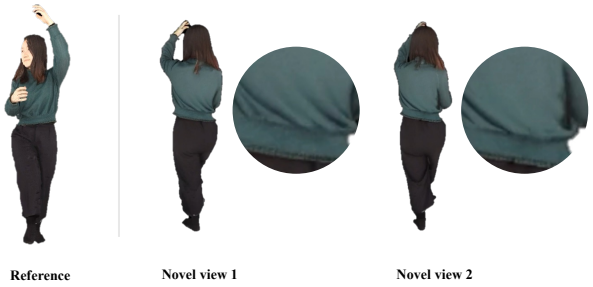

Figure 5. **Novel view synthesis on THuman2.1 by a multi-view image-based diffusion model.** Inconsistent clothing wrinkles appear between two adjacent novel view generations.

neither approach achieves satisfactory results: image-based diffusion models show significant inconsistencies, while video diffusion models produce blurry frames.

E.2 Novel view camera trajectories

On our 3D scan dataset, we render a smooth camera trajectory with the same elevation and evenly distributed azimuth. We also explore the possibility of leveraging the fluctuant novel views in MVHumanNet dataset to learn the multi-view consistency. However, we empirically find that the diffusion models are not able to capture the consistency even with the aid of appearance cues, causing the textures to flow across views.

References

[1] 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 *Forty-first International Conference on Machine Learning*, 2023. 1

[2] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 4

[3] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2023)*, 2023. 1

[4] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 4

[5] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 4

[6] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 1, 4

[7] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. 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*, pages 19801–19811, 2024. 1, 4

[8] Zhongcong Xu, Chaoyue Song, Guoxian Song, Jianfeng Zhang, Jun Hao Liew, Hongyi Xu, You Xie, Linjie Luo, Guosheng Lin, Jiashi Feng, et al. High quality human image animation using regional supervision and motion blur condition. *arXiv preprint arXiv:2409.19580*, 2024. 1

[9] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[10] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 1, 4

[11] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 3, 4

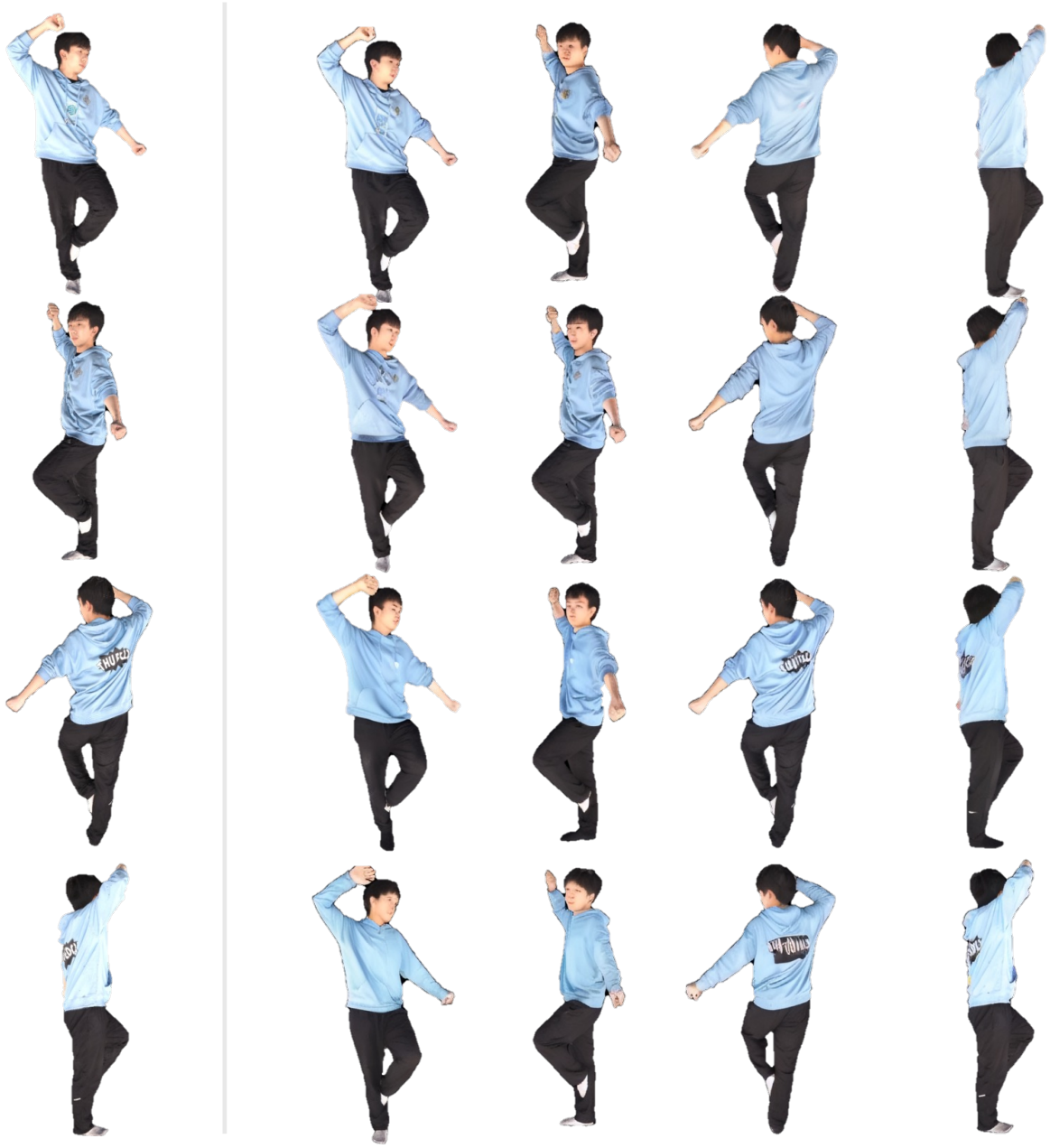

Figure 6. **Qualitative results of the generated novel views for various input views of the same human subject.**