Blur2Sharp: Human Novel Pose and View Synthesis with Generative Prior Refinement

Supplementary Material

1. Boarder Impacts

Our model might raise serious concerns regarding privacy and potential misuse. As these models can generate realistic and manipulable representations of individuals without their consent, they may be exploited to create non-consensual deepfakes, impersonate identities, or produce misleading visual content. This causes a high risk for public figures and vulnerable individuals because a single publicly available photo could be repurposed for malicious intent. We strongly encourage readers to use this work responsibly and strictly within the bounds of legal and ethical guidelines.

2. Training Pipeline and Network Learning Details

Our training pipeline consists of two main stages. Since we evaluate our method on two datasets, we train separate models for each dataset independently. Notably, the two datasets differ in camera viewpoints, which introduces a domain gap. In the first stage, we train the Human NeRF model for 40,000 iterations on the MVHumanNet [14] dataset and 33,000 iterations on the HuMMan [1] dataset using a batch size of 1. Input images are first center-cropped to a resolution of 1024×1024 pixels, then resized and normalized to 512×512. Once trained, the Human NeRF model generates coarse renderings for each sample, which are saved for use in the subsequent refinement stage. To filter out background regions from the coarse images, we apply a threshold of 0.5 to the human region mask predicted by the Human NeRF model.

The second stage of our method refines the coarse outputs using a generative diffusion model trained in two phases. To initialize the model, we adopt pre-trained weights from Stable Diffusion 1.5 [12] for both the multi-view Denoising UNet and the Reference Network, while the 1D attention module is initialized with weights from AnimateDiff V3 [4]. In the first training phase, we train a base version of the diffusion model without multi-view and 1D attention for 80,000 iterations on MVHumanNet and 50,000 iterations on HuMMan, using a learning rate of 1×10^{-5} and a batch size of 2. In the second phase, we freeze the MLGF module, the RGB and normal encoders, and the Reference Network. We then enable both multi-view and 1D attention mechanisms, and fine-tune only the multi-view Denoising UNet for 30,000 iterations on MVHumanNet and 15,000 iterations on HuMMan. The entire training process takes approximately five days on a single NVIDIA A6000 GPU.

3. Training Objective

For the Human NeRF training stage, following [7], we optimize a composite objective that integrates both photometric and geometric constraints. The primary reconstruction loss $\mathcal{L}_{\text{recon}}$ measures pixel-level discrepancies between the rendered and ground truth images. To ensure silhouette consistency, we include a mask loss $\mathcal{L}_{\text{mask}}$, implemented as the binary cross-entropy between predicted and ground truth masks. To further enhance perceptual quality, we incorporate a structural similarity loss $\mathcal{L}_{\text{SSIM}}$ and a perceptual feature loss $\mathcal{L}_{\text{LPIPS}}$, computed using a pretrained VGG network. The complete objective for Human NeRF training is defined as:

$$\mathcal{L}_{\text{NeRF}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}, \tag{1}$$

where λ_{mask}, $\mathcal{L}_{\text{mask}}$, and $\mathcal{L}_{\text{SSIM}}$ are loss weights.

The diffusion refinement stage employs a dual-domain velocity-based prediction (v-prediction) objective [13] that jointly optimizes RGB appearance and surface normal geometry. The RGB branch is conditioned on three inputs: the reference image I_{ref}, RGB features F_{rgb} derived from Human NeRF renderings, and geometry features F_{geo}, which provide structural pose guidance. Similarly, the normal branch is conditioned on the reference normal map N_{ref}, normal features F_{normal}, and the shared geometry features F_{geo}. The v-prediction objective is formally defined as:

$$\begin{aligned} \mathcal{L}_{\text{v-pred}} = \mathbb{E}. \Big[&\left\| \hat{v}_\theta(I_t; I_{\text{ref}}, F_{\text{rgb}}, F_{\text{geo}}) - v_t^I \right\|_2^2 \\ &+ \left\| \hat{v}_\theta(N_t; N_{\text{ref}}, F_{\text{normal}}, F_{\text{geo}}) - v_t^N \right\|_2^2, \end{aligned} \tag{2}$$

where $I_t = \alpha_t I + \sigma_t \epsilon_I$ and $N_t = \alpha_t N + \sigma_t \epsilon_N$ denote the noisy RGB and normal latents at timestep t, respectively. v_t^I and v_t^N represent the target velocities for the RGB and normal branches.

4. Supplementary Network Architecture Specifications

Details of Coarse Image Conditioning. In Stage 1, We utilize SHERF [7] to generate coarse renderings I_{coarse} from a single reference image, capturing both novel poses and viewpoints. These initial renderings preserve the subject's

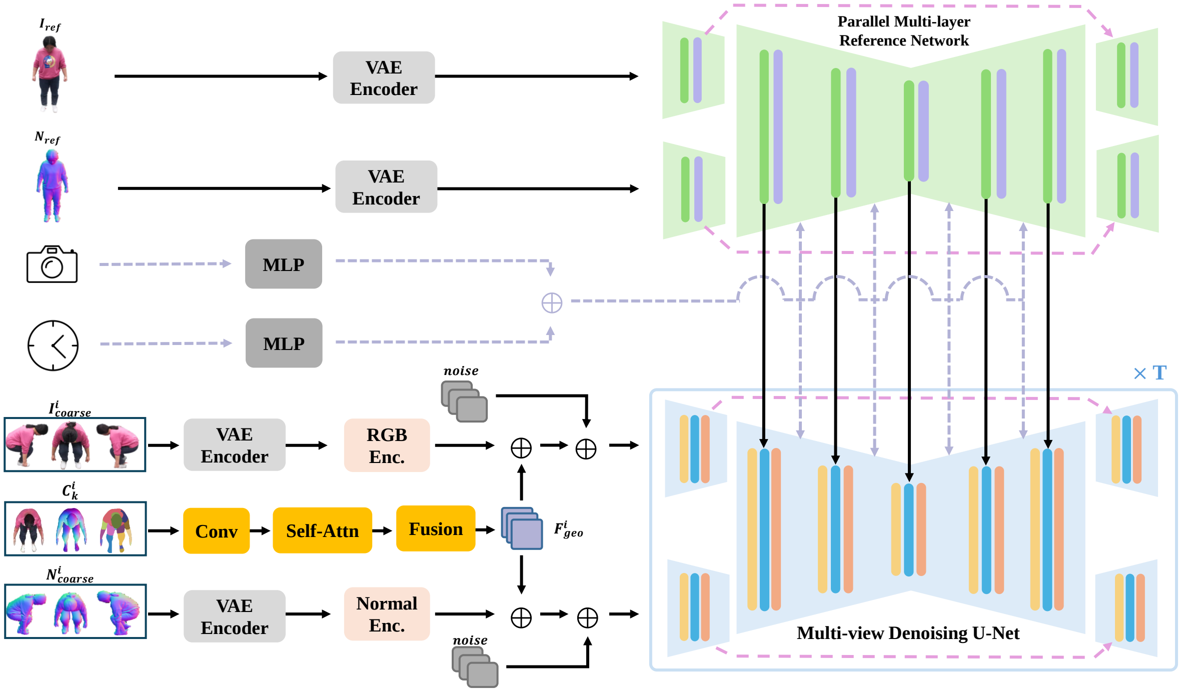

Figure 1. Architecture of Generative Prior Refinement. The network processes coarse NeRF renderings I_{coarse} and predicted normal maps N_{coarse} through RGB encoder and normal encoder. Extracted features are combined with geometry features F_{geo} via element-wise addition ($\oplus$) before injection into our multi-view diffusion model. Note that $\mathbf{C}^i_k$ denotes the three condition maps (i.e., texture, normal, and semantic) used in MLGF module.

appearance while approximating the target geometry. We then predict the corresponding normal maps N_{coarse} using a pretrained off-the-shelf normal estimator [9]. For conditioning the generative model, both the coarse RGB images and normal maps are encoded into latent representations using VAE and then processed through separate encoder networks E_{rgb} and E_{normal} as shown in Fig.1, each network is consisting of four convolutional layers as structured in Tab. 1. The resulting latent features are combined with geometry features F_{geo} from MLGF via element-wise addition, followed by convolutional processing of the noise distributions in both the RGB and normal domains.

Conditioning on Camera View and Time Step. Unlike prior methods such as MagicMan [5] that only leverage camera rotations for viewpoint conditioning, we incorporate both the rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$ and the translation vector $\mathbf{t} \in \mathbb{R}^{3\times1}$ to facilitate more precise viewpoint control. These extrinsic parameters together define the full camera-to-world transformation. We flatten the rotation and translation matrix into a 12-dimensional vector $\mathbf{v} \in \mathbb{R}^{12}$ and pass it through a feed-forward network to obtain the camera embeddings $\mathbf{f}_{\text{cam}} \in \mathbb{R}^{1024}$. These embeddings are then combined with the denoising time step embeddings $\mathbf{f}_{\text{time}} \in \mathbb{R}^{1024}$ of the diffusion model via element-wise addition to provide viewpoint-aware conditioning during the diffusion process. As shown in Tab. 3, this design leads to consistently better results compared to using rotations alone.

Architecture	Channels
Conv2d (kernel: 3×3, stride: 1) SiLU	4 → 16 -
Conv2d (kernel: 3×3, stride: 1) SiLU Conv2d (kernel: 3×3, stride: 1) SiLU	16 → 16 - 16 → 32 -
Conv2d (kernel: 3×3, stride: 1) SiLU Conv2d (kernel: 3×3, stride: 1) SiLU	32 → 32 - 32 → 96 -
Conv2d (kernel: 3×3, stride: 1) SiLU Conv2d (kernel: 3×3, stride: 1) SiLU	96 → 96 - 96 → 256 -
Conv2d (kernel: 3×3, stride: 1), zero-initialized	256 → 320

Table 1. Architecture of RGB and normal encoders

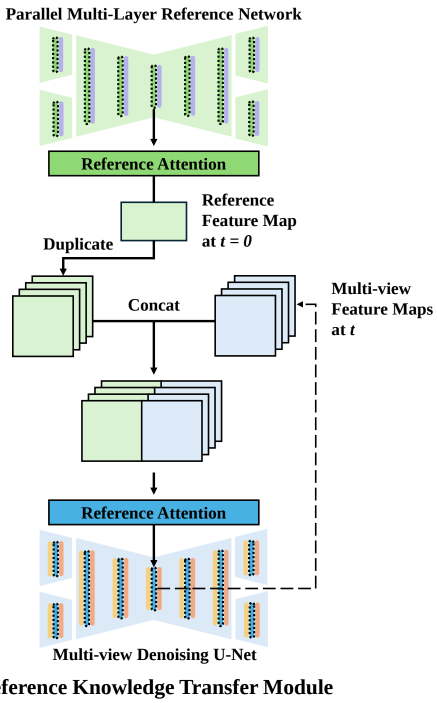
Parallel Multi-Layer Reference Network
Reference Attention
Reference Feature Map at *t = 0*
Duplicate
Concat
Multi-view Feature Maps at *t*
Reference Attention
Multi-view Denoising U-Net
Reference Knowledge Transfer Module

Figure 2. Architecture of Reference Knowledge Transfer Module. The reference feature map is extracted via the reference attention mechanism in the Parallel Multi-Layer Reference Network and duplicated to match the dimension of the multi-view feature map. The duplicated reference feature map is then concatenated with the multi-view feature maps and passed into the reference attention of the Multi-view Denoising UNet for knowledge transfer.

Details of Reference Knowledge Transfer Module. Similar to MagicMan [5], We implement a Reference Knowledge Transfer Module to effectively propagate informative features from a single reference view to all target views, thereby enhancing appearance consistency between the generated multi-view images and the reference image. As illustrated in Fig. 2, the module first extracts a reference feature map from the reference attention block of the Reference Network. This feature map encodes high-level appearance information from the reference image. To enable alignment across views, the reference feature map is spatially and dimensionally duplicated to match the shape of the multi-view feature maps. These duplicated reference features are then concatenated with the multi-view features and passed to the reference attention layers of the denoising UNet, where attention computations integrate reference-guided appearance cues into the multi-view generation process.

Details of Dual-Branch RGB-Normal Conditional Diffusion Model. As shown in Fig.1, to enable the joint generation of RGB images and normal maps, we first extract features separately from the RGB and normal latents using their respective downsampling blocks. The resulting intermediate features are then stacked and averaged to form a unified representation that captures both appearance and geometric information. This unified representation is passed through the shared intermediate layers of the UNet, while the residuals from the early modality-specific blocks are preserved and routed through their corresponding upsampling layers using the skip connections. This design ensures that each modality retains its unique characteristics throughout the denoising process.

5. Applications

Our method can be directly applied to realistic multi-view and multi-pose virtual try-on, enabling photorealistic visualization of how a garment appears from various angles and body poses using only a single reference image. Given a reference image of a person and a target garment, we first apply a pretrained 2D virtual try-on model [2] to generate a 2D try-on result aligned with the reference pose. We then apply **Blur2Sharp** to produce high-quality images under both novel poses and novel viewpoints, conditioned on the 2D try-on image. As shown in Fig. 3, **Blur2Sharp** is capable of generating realistic and consistent try-on results across diverse viewpoints and human poses.

5.1. Additional Ablation Studies

Evaluating Coarse RGB-Normal Conditioning strategies. To better understand the effectiveness of our proposed dual-branch conditioning mechanism, we conduct an ablation study comparing different methods of injecting coarse RGB renderings and predicted normal maps into the diffusion model. As an alternative baseline, we experiment with a simple conditioning approach in which the RGB and normal latents are concatenated with the noise input along the channel dimension. To accommodate this change, we modify the input convolutional block of the diffusion network to accept an 8-channel input and initialize the weights accordingly.

As shown in Fig. ??, this naive concatenation strategy fails to recover high-frequency texture details, despite maintaining accurate pose geometry. The approach yields significantly worse perceptual quality metrics, with higher LPIPS and FID scores reported in Tab. 2. We hypothesize that the simple concatenation mechanism indiscriminately mixes all conditioning signals—including blurry features—without effective feature decomposition or selective enhancement, allowing low-frequency artifacts from the input RGB and normal maps to persist throughout the denoising process.

Evaluating Different Inputs of Camera Conditioning. To assess the contribution of our camera conditioning approach, we perform an ablation study by removing the

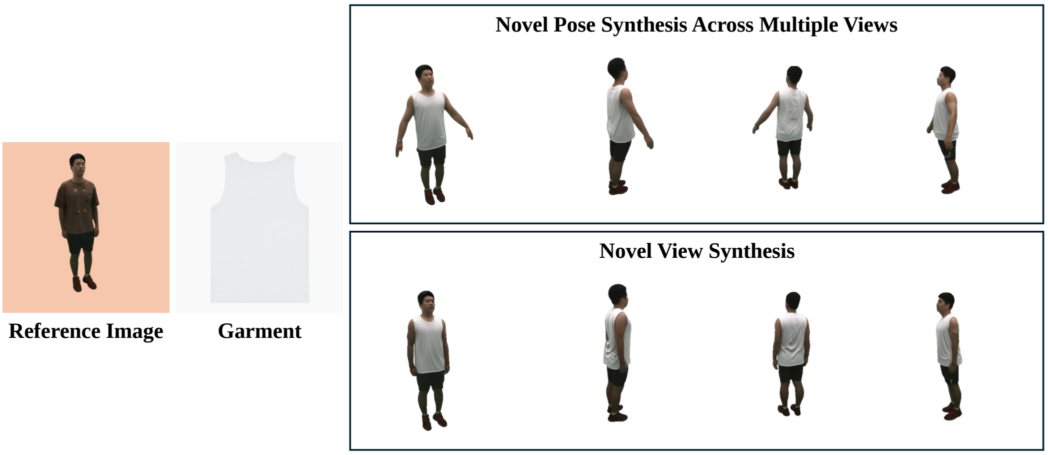

Figure 3. Given a reference image and a target garment, we first employ a 2D virtual try-on model to generate a try-on image. Blur2Sharp then synthesizes multi-view images under novel poses and viewpoints based on this try-on result.

graphics[scale=0.65]pic/ablation$_{c}oncat.pdf$

Figure 4. Ablation studies and qualitative comparisons of different RGB-Normal conditioning strategies. Red boxes indicate enlarged regions.

Table 2. Quantitative comparison of different coarse RGB-Normal conditioning strategies for novel pose and view synthesis.

Method	Novel Pose Synthesis				Novel View Synthesis			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Concat	22.31	0.940	0.047	34.04	22.56	0.942	0.046	34.04
Conv+Add	**23.31**	**0.946**	**0.039**	**24.38**	**23.82**	**0.948**	**0.036**	**23.50**

translation component from the extrinsic matrix and conditioning the model using only the 3×3 rotation matrix. Our full method uses the complete 3×4 extrinsic matrix, which includes both rotation and translation. Quantitative results in Tab. 3 show that removing translation leads to a drop in performance across all metrics, indicating that the translation matrix provides valuable information for improving generation quality.

Evaluating Different SMPL Conditions. As shown in Tab. 4, incorporating all three SMPL priors (texture, normal, and semantic maps) leads to notable improvements in both structural alignment and perceptual quality for novel pose and view synthesis. Specifically, the inclusion of SMPL normal and semantic maps (w/o. texture) enhances structural alignment (higher PSNR) across views and poses by providing geometric cues, while the variant model with SMPL texture priors only (w/o. normal & semantic) contributes to improved perceptual realism (lower FID) by preserving fine-grained surface appearance. Our full model achieves the lowest FID and reduced LPIPS, indicating enhanced visual fidelity, while maintaining strong PSNR and SSIM values that reflect better structural consistency compared to ablated variants. We provide the qualitative results in Fig. 5.

5.2. Efficiency Analysis

Tab. 5 presents the memory usage and computation time required for each stage of the pipeline.

5.3. Additional Experimental Results

Novel Pose Synthesis Across Multiple Views. Fig. 9 and 10 present additional qualitative results for novel pose synthesis across multiple views in comparison with state-of-the-art approaches.

Novel View Synthesis Fig. 11 and 12 present additional qualitative results for novel view synthesis in comparison with state-of-the-art approaches.

Our method demonstrates superior performance, exhibiting enhanced multi-view consistency and photorealism.

Comparison with 3DGS-based Methods. To further investigate, we examined IDOL [16], a state-of-the-art 3DGS-based [8] method that constructs an explicit 3D reconstruction to generate novel poses and views. We trained our purely 2D method on a subset of the HuGe100K

Table 3. Quantitative ablation study of camera inputs on MVHumanNet dataset.

Method	Novel Pose Synthesis				Novel View Synthesis			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Rotation only	23.27	0.946	0.039	25.21	23.79	0.948	0.036	24.79
Rotation and Translation	**23.31**	**0.946**	**0.039**	**24.38**	**23.82**	**0.948**	**0.036**	**23.50**

Table 4. Quantitative ablation study on different input configurations of the MLGF module. ”w/o. texture” denotes the model variant without SMPL texture maps, while ”w/o. normal & semantic” refers to the configuration excluding both SMPL normal and semantic maps. ”Ours” represents the proposed method that integrates SMPL texture, normal, and semantic priors for comprehensive geometric guidance. The top two methods are highlighted in **bold** and underline.

Method	Novel Pose Synthesis				Novel View Synthesis			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
w/o. texture	23.28	0.945	0.039	28.08	23.81	0.948	0.037	27.81
w/o. normal & semantic	23.18	0.945	0.039	27.44	23.63	0.947	0.037	26.98
Ours	**23.31**	**0.946**	**0.039**	**24.38**	**23.82**	**0.948**	**0.036**	**23.50**

Table 5. Memory and Runtime Efficiency of each stage of our proposed method.

Method	Memory (GB)	Time (sec)
Rendering SMPL Condition Maps	2.47	1.58
Human NeRF	6.38	4.51
Generative Prior Refinement	6.87	14.12

dataset [16], which was originally introduced by IDOL [16], to evaluate its performance under similar conditions. As shown in Tab. 8, our method achieves competitive perceptual quality without relying on explicit 3D reconstruction. While IDOL [16] generally preserves the reference image's geometry and appearance, even small deviations in pose or viewpoint can lead to floating artifacts, a common limitation of 3DGS-based approaches. Notably, these methods were developed around the same period, demonstrating that high-quality novel pose and view synthesis can be achieved using a simpler 2D approach.

Generalizability Analysis. To evaluate the generalizability of our method beyond controlled settings, we compare it with SHERF [7], Animate Anyone [6], Champ [15], and IDOL [16] on in-the-wild datasets, DeepFashion [11] and SHHQ [3], using camera poses and SMPL parameters estimated by CLIFF [10]. All methods are trained on the HuMMan dataset [1], except for IDOL [16], which uses its official pre-trained weights. Qualitative results in Fig. 6, 7, and 8 show that diffusion-based methods like Animate Anyone [6] and Champ [15] often produce inconsistencies, such as misplaced faces, while IDOL [16] tends to generate body shapes that deviate from the reference image and introduces floating artifacts. In contrast, our method preserves identity, maintains high-fidelity details, and generates coherent novel views and poses under real-world conditions, demonstrating strong robustness and versatility. A user study further confirms that our method effectively preserves the reference appearance while aligning with user preferences, with details provided in the following section.

User Study. To validate the perceptual advantages of our method, we conducted a user study with 31 participants, consisting of 20 questions across four aspects: **Identity & Appearance Preservation**, **Novel View Synthesis**, **Novel Pose Synthesis**, and **Cross-View Consistency**. Each aspect contained five questions with different subjects, and participants evaluated visual quality, identity preservation, and consistency according to the task in each category. For comparison, IDOL [16] was evaluated using its official pre-trained model without additional retraining on our dataset. Its rendered images may have slightly different camera viewpoints, but participants were instructed to focus on visual fidelity and detail preservation, ensuring the user study results remain meaningful. The results of this study are summarized in Tab. 6.

We further conducted an in-the-wild user study with 16 participants, consisting of five questions designed to evaluate the robustness of each method under unconstrained scenarios. The study was performed on in-the-wild datasets, including DeepFashion [11] and SHHQ [3]. Participants compared methods in terms of identity preservation, visual realism, and overall perceptual quality. The results, summarized in Tab. 7, demonstrate our method's strong generalizability across diverse, real-world images.

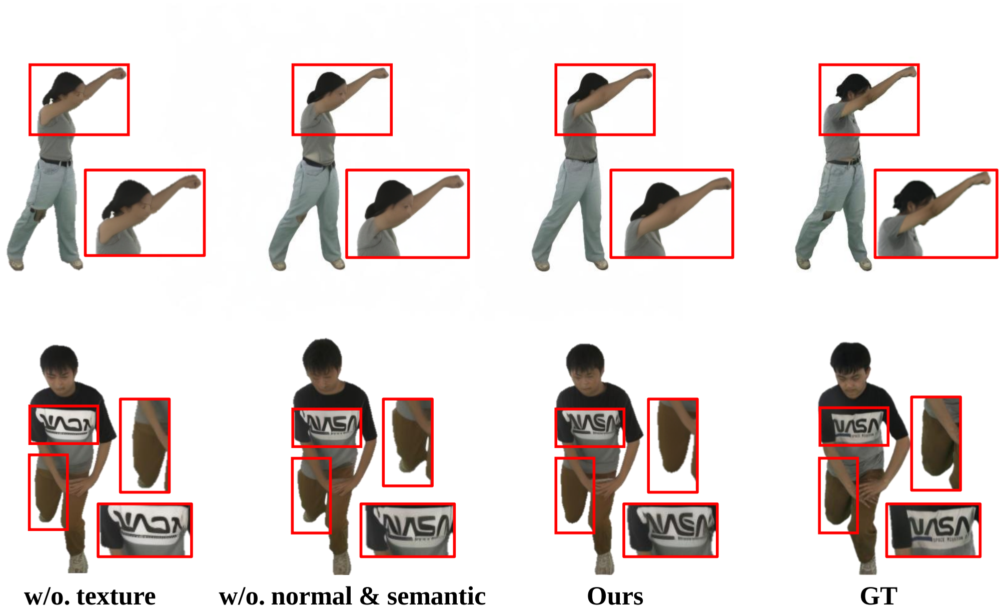

Figure 5. Qualitative results of different SMPL Conditions. Red boxes indicate enlarged regions.

Table 6. User study results across four aspects: **Identity & Appearance Preservation**, **Novel View Synthesis**, **Novel Pose Synthesis**, and **Cross-View Consistency**. Each value indicates the percentage of votes received by each method (higher is better). The best result in each column are highlighted in **bold**.

Method	Identity & Appearance	Novel View	Novel Pose	Cross-View Consistency
SHERF	13.55%	0.00%	5.16%	0.65%
Animate Anyone	5.16%	14.19%	8.39%	16.77%
Champ	12.90%	16.13%	8.39%	9.03%
IDOL	32.90%	7.10%	8.39%	0.65%
Ours (Blur2Sharp)	**35.48%**	**62.58%**	**69.68%**	**72.90%**

Table 7. In-the-wild user study results. Each value indicates the percentage of votes received by each method (higher is better). The best result is highlighted in **bold**.

	SHERF	Animate Anyone	Champ	IDOL	Ours (Blur2Sharp)
User Preference (%)	28.75	7.50	17.50	15.00	**31.25**

Table 8. Quantitative comparison of novel pose synthesis across 4 views on the HuGe100K dataset [16]. The evaluation metrics include PSNR, SSIM, LPIPS, and FID. The best method is highlighted in **bold**.

Method	HuGe100K			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
IDOL	20.67	0.897	0.159	23.3966
Ours (Blur2Sharp)	**24.22**	**0.936**	**0.042**	**7.984**

References

[1] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022. 1, 5, 9, 10

[2] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, 2024. 3

[3] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 5

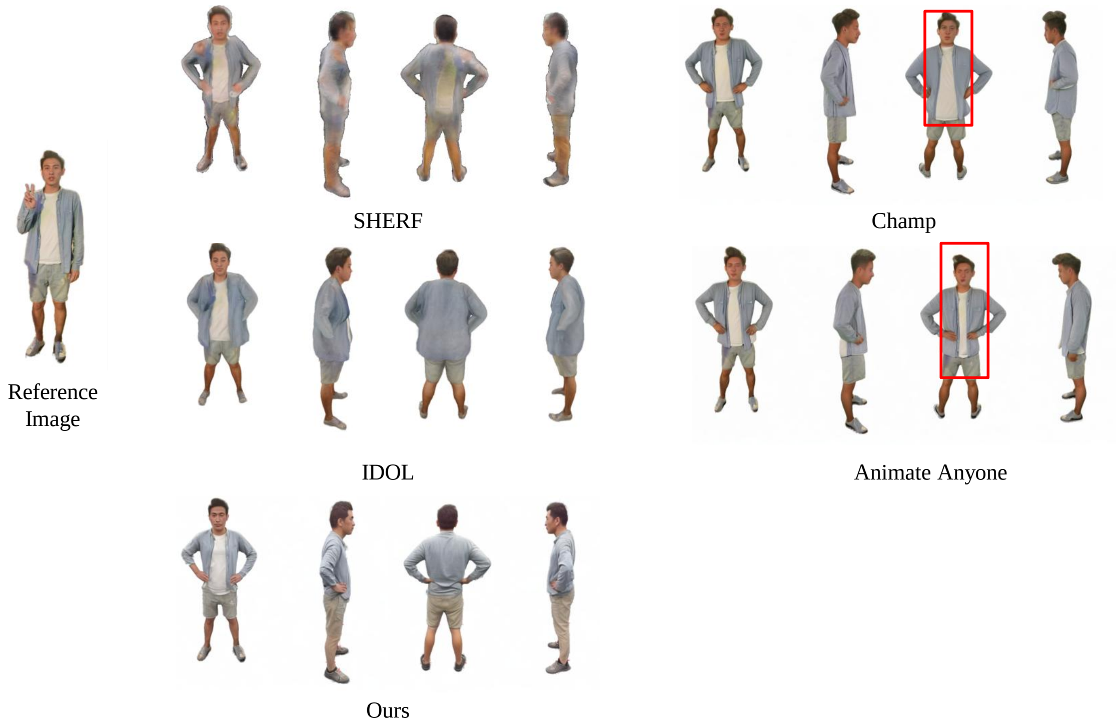

Figure 6. Qualitative results of in-the-wild setting across multiple views on the HuGe100K [16] dataset. We compare our method with IDOL [16]. Red boxes highlight appearance ambiguities or multi-view inconsistencies.

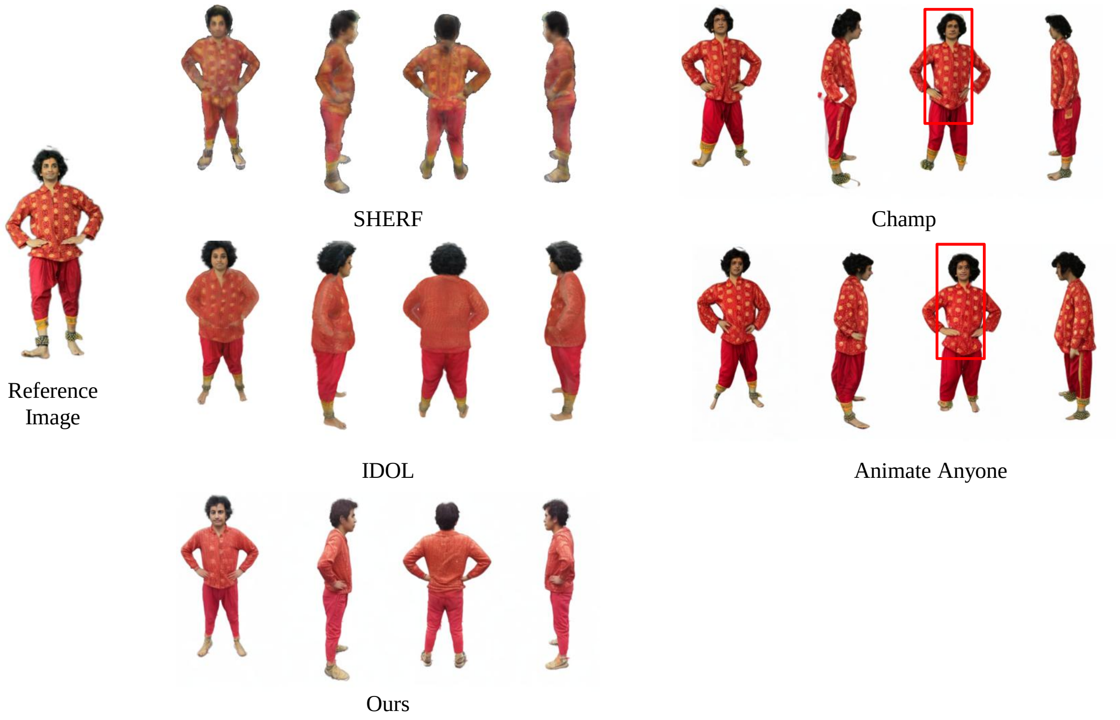

Figure 7. Qualitative results of in-the-wild setting across multiple views on the HuGe100K [16] dataset. We compare our method with IDOL [16]. Red boxes highlight appearance ambiguities or multi-view inconsistencies.

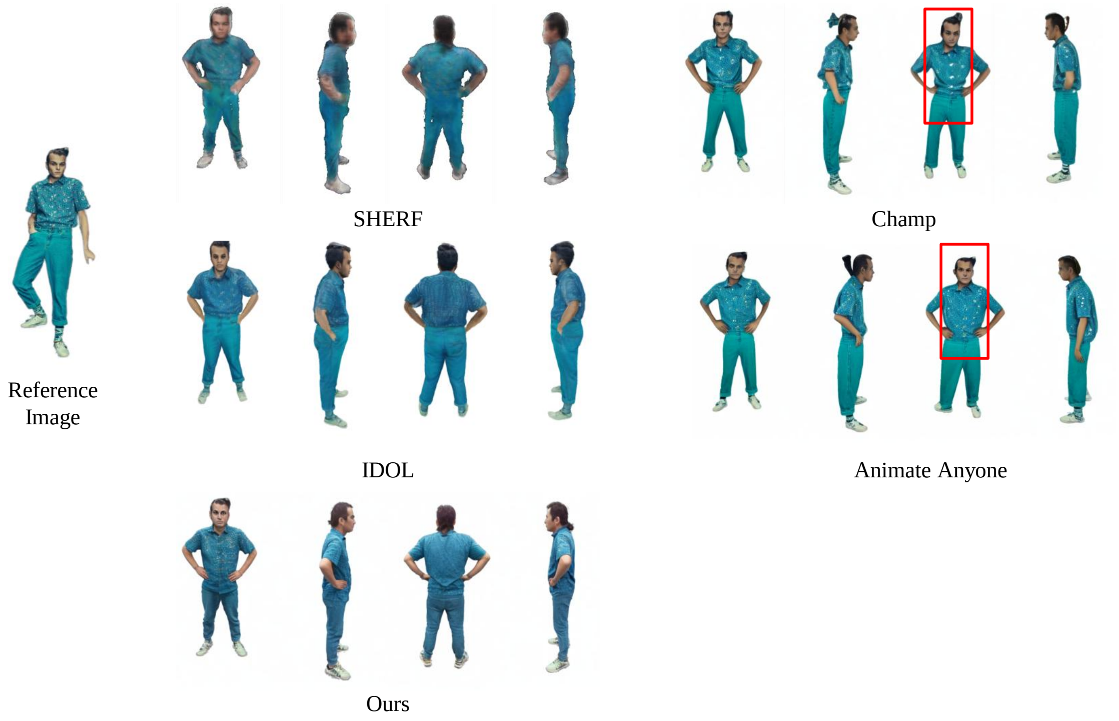

Figure 8. Qualitative results of in-the-wild setting across multiple views on the HuGe100K [16] dataset. We compare our method with IDOL [16]. Red boxes highlight appearance ambiguities or multi-view inconsistencies.

[4] 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. *International Conference on Learning Representations*, 2024. 1

[5] Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement, 2024. 2, 3

[6] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 5, 9, 10

[7] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9352–9364, 2023. 1, 5, 9, 10

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

[9] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 2

[10] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 5

[11] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Large-scale fashion (deepfashion) database. *The Chinese University of Hong Kong, Category and Attribute Prediction Benchmark, Xiaoou TangMultimedia Laboratory*, 2016. 5

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

[13] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 1

[14] 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, 9, 10

[15] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong

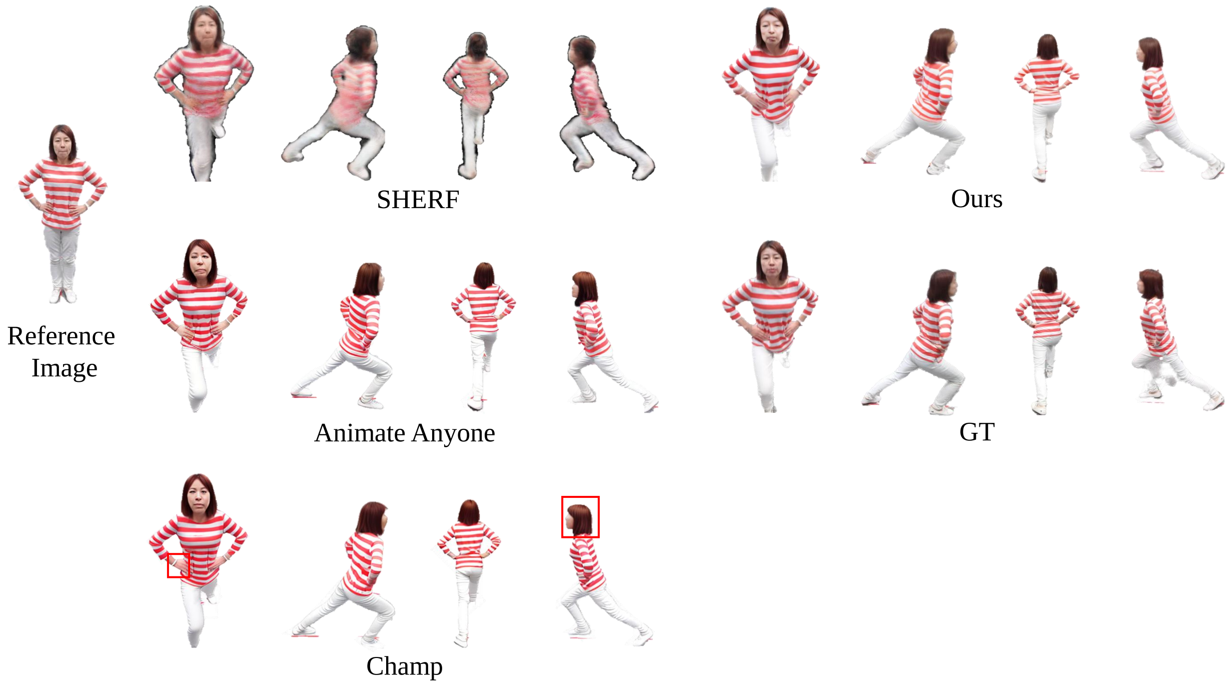

Figure 9. Qualitative results of novel pose synthesis across multiple views on the HuMMan [1] dataset. We compare our method with SHERF [7], Animate Anyone [6], and Champ [15]. Red boxes highlight appearance ambiguities or multi-view inconsistencies.

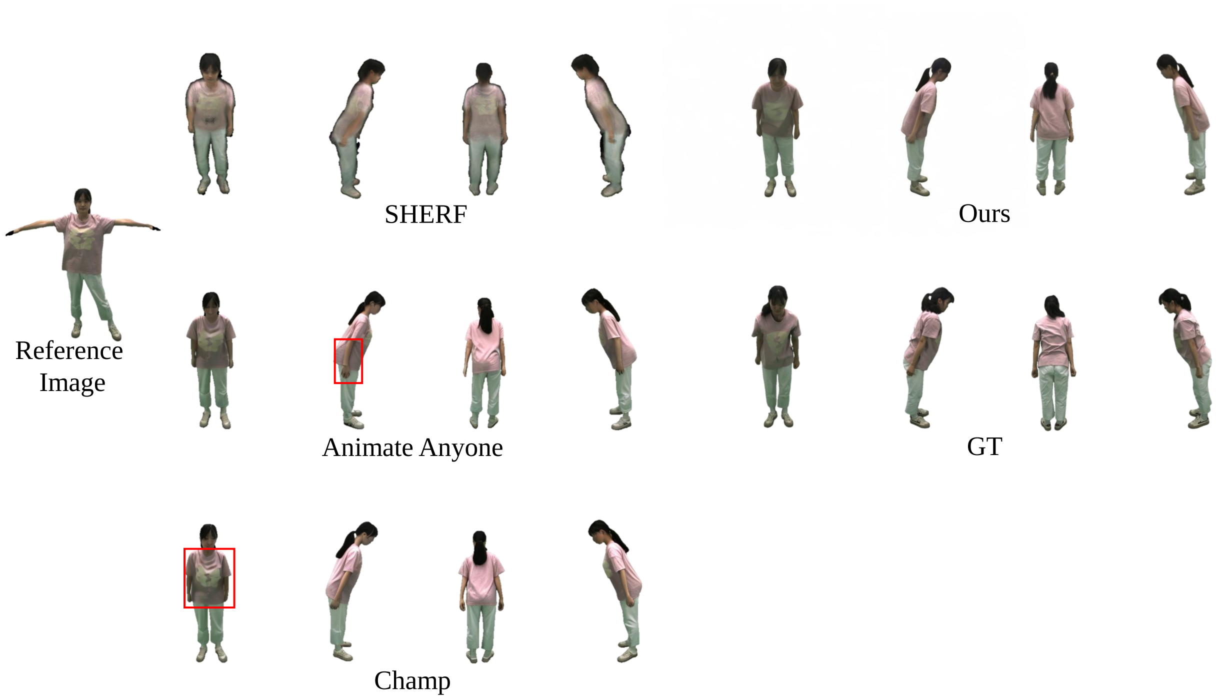

Figure 10. Qualitative results of novel pose synthesis across multiple views on MVHumanNet [14] dataset. We compare our method with SHERF [7], Animate Anyone [6], and Champ [15]. Red boxes highlight appearance ambiguities or multi-view inconsistencies.

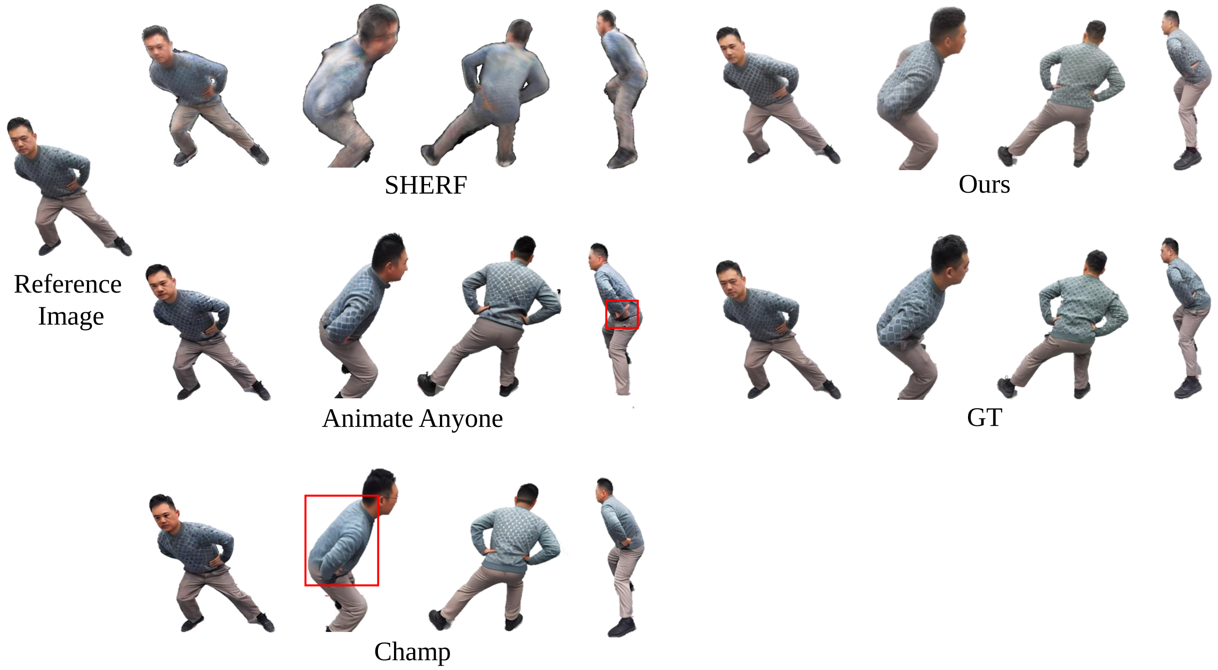

Figure 11. Qualitative results of novel view synthesis on HuMMan [1] dataset. We compare our method with SHERF [7], Animate Anyone [6], and Champ [15]. Red boxes highlight appearance ambiguities or multi-view inconsistencies.

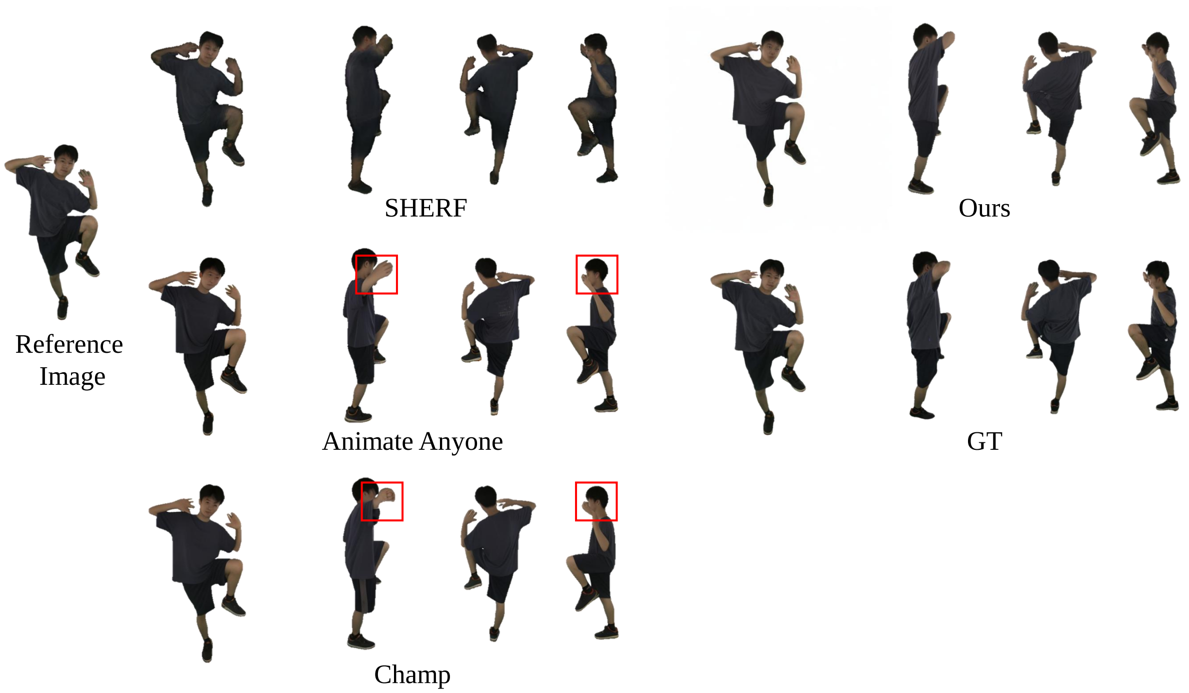

Figure 12. Qualitative results of novel view synthesis on MVHumanNet [14] dataset. We compare our method with SHERF [7], Animate Anyone [6], and Champ [15]. Red boxes highlight appearance ambiguities.

Dong, 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*, pages 145–162. Springer, 2024. 5, 9, 10

[16] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. *arXiv preprint arXiv:2412.14963*, 2024. 4, 5, 6, 7, 8