IDOL: Instant Photorealistic 3D Human Creation from a Single Image

Supplementary Material

In this supplementary material, we provide additional details and visualizations to support the claims made in our main paper. Sec. A provides further details on the *HuGe100K* dataset, including visualizations, important statistics, and the methodology to enhance the 3D consistency and diversity of multi-view images. Sec. B describes the training procedure and setup for our proposed method, *IDOL*. Sec. C presents additional experimental results, including comparison tables and results from the user study.

A. More Details of *HuGe100K* Data

This section provides a more detailed explanation of the *HuGe100K* dataset generation process, along with additional visualizations. Sec A.1 describes our approach to improving the 3D consistency of MVChamp during training, while Sec A.2 presents the prompt template and attribute set used to generate reference images, as well as the generation process with MVChamp. Sec. A.3 demonstrates more visualization of the dataset.

A.1. Improving 3D Consistency of Image Animation

Champ [21] is one of the state-of-the-art Human Image Animation models, enhanced by multiple conditions rendered from DWPose and SMPL. We employ a two-stage training process to enhance the 3D consistency of the Champ model for human multi-view synthetic, referred to as MVChamp.

Fine-tuning Champ on Large-scale Human Videos with Whole-body Conditions To enable Champ to learn more human 3D prior knowledge, we curate a dataset of approximately 100K dance videos for fine-tuning, of which around 20K explicitly contain human turning motions. Full parameter training of MVChamp on such a large dance dataset effectively enhances its understanding of human 3D prior knowledge. Additionally, we employ HaMeR [12], a state-of-the-art model for 3D hand reconstruction, to specifically reconstruct hand poses from images. These reconstructed hand poses are rendered into depth maps and used as an additional pose control signal for precise whole-body reconstruction and animation.

Fine-Tuning Temporal Blocks on 3D Human Dataset We use the open-source scanned dataset THuman 2.1 [19], rendered in Blender, to produce 24 uniformly sampled views along the horizontal dimension to fine-tune the temporal layers of MVChamp using standard diffusion loss.

Improving Temporal Consistency from the First to Last Frames Although the MVChamp model generates highly continuous multi-view images between adjacent frames, significant discrepancies remain between the first and last views, even though these two views are continuous in content. This issue likely arises from the model's emphasis during training on ensuring continuity between adjacent frames while neglecting the larger temporal gap between the first and last frames. Thus, we propose the *Temporal Shift Denoising Strategy* to address this issue. During each denoising step, we shift the current latent inputs and pose condition signals along the temporal axis, moving the latent inputs and pose condition of the last frame to the first frame. This strategy ensures that each frame can access contextual information during most of the denoising steps, effectively eliminating discrepancies between the first and last frames at the same inference cost.

A.2. Generating Balanced and Diverse Images

Balanced, diverse, high-quality, high-resolution, and full-body images are scarce in existing human-centric datasets, and they are challenging to collect on the Internet due to copyright and portrait rights issues. Therefore, we mix the real-life images and generate photorealistic images to obtain the large-scale quantity and high-quality images. Specifically, we extract approximately 10,000 real-life images from the open-source dataset DeepFashion [11] and use Flux [5], a state-of-the-art text-to-image model, to generate balanced and diverse human reference images. We ensure balance and diversity across five dimensions during image generation: *area*, *clothing*, *body shape*, *age* and *gender*. Each dimension value is randomly selected from a large set of options generated by GPT-4 [1], with prompt templates as follows: *Front view, full-body pose of a* {*age*} *old* {*body shape*} {*area*} {*gender*} *wearing* {*clothing*} *and visible hands. He/She stands against a white background, evenly lit.* Ultimately, we collect a total of 100,000 balanced and diverse full-body human reference images.

For each dimension, the possible options are as follows:

1. **Area**: *United States, Canada, Mexico, Guatemala, Cuba, Brazil, Argentina, Colombia, Chile, Peru, United Kingdom, Germany, France, Italy, Spain, Netherlands, Belgium, Switzerland, Poland, Sweden, Nigeria, Egypt, South Africa, Kenya, Morocco, Ghana, Tanzania, Ethiopia, Uganda, Algeria, Saudi Arabia, Iran, Turkey, Israel, United Arab Emirates, Qatar, Kuwait, Jordan, Oman, Lebanon, Kazakhstan, Uzbekistan, Turkmenistan, Kyrgyzstan, Tajikistan, India, Pakistan, Bangladesh, Sri Lanka, Nepal, Bhutan, China,*

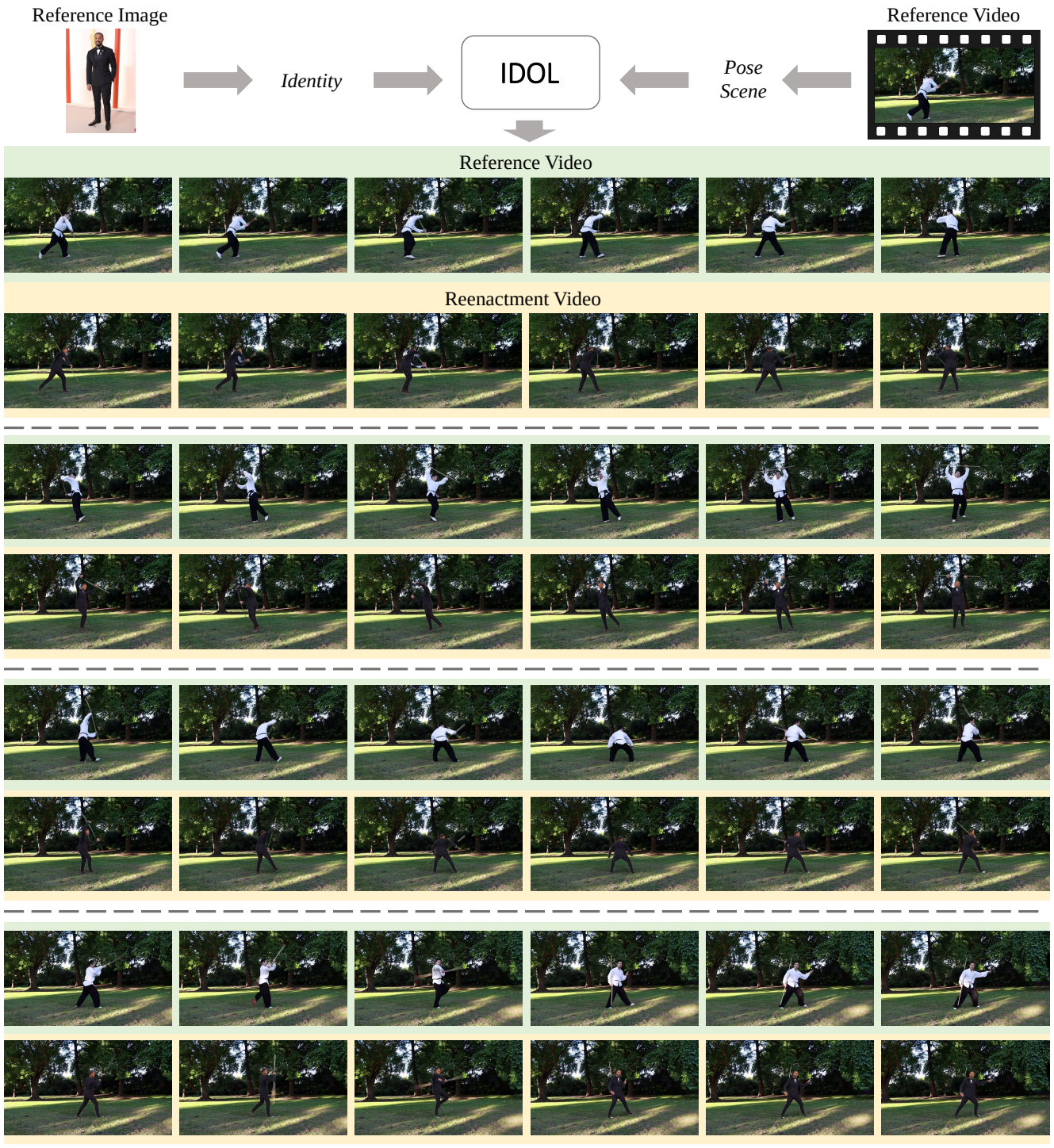

Figure 1. The visualization of the reenactment.

Japan, South Korea, Mongolia, North Korea, Indonesia, Thailand, Vietnam, Malaysia, Philippines, Singapore, Myanmar, Cambodia, Laos, Brunei, Australia, New Zealand, Papua New Guinea, Fiji, Solomon Islands, Jamaica, Haiti, Dominican Republic, Puerto Rico, Trinidad and Tobago, Panama, Costa Rica, Nicaragua, Honduras, El Salvador, Belize, etc.

2. **Clothing**: *T-shirts, Jeans, Casual pants, Dresses, Shorts, Tank tops, Sweaters, Cardigans, Jumpsuits, Hoodies, Suits, Business shirts, Formal skirts, Dress*

pants, Blazers, Tie, Waistcoats, Formal shoes, Briefcases, Leather belts, Sport shirts, Fitness clothes, Sports shoes, Tracksuits, Gym shorts, Leggings, Swimwear, Cycling gear, Compression wear, Evening gowns, Tuxedos, Long dresses, Tailcoats, Cocktail dresses, Party wear, Ceremonial suits, Ball gowns, Dress shoes, Fine jewelry, Hiking clothes, Waterproof jackets, Thermal wear, Camping gear, Fishing vests, Hunting apparel, Snowboarding pants, Rain boots, Cotton shirts, Linen dresses, Chiffon blouses, Sandals, Sunglasses, Short sleeves, Beachwear, Crop tops, Wool coats, Thick cotton sweaters, Fur jackets, Beanies, Boots, Gloves, Scarves, Thermal leggings, Padded parkas, Insulated boots, Hanfu, Kimono, Sari, African tribal dresses, Scottish kilts, Bavarian lederhosen, Moroccan kaftans, Hawaiian shirts, Russian ushankas, Streetwear, Avant-garde designs, Fusion wear, Boho chic, Minimalist styles, High fashion, Urban outfits, Eco-friendly clothing, Techwear, Nurse uniforms, Firefighter gear, Construction vests, Police uniforms, Military boots, Lab coats, Coveralls, Military uniforms, Academic gowns, Judicial robes, Clerical vestments, Diplomatic suits, Regalia, etc.

3. **Body shape**: *Slight, Lean, Petite, Athletic, Fit, Average, Built, Buff, Bodybuilder, Full-figured, Stocky, Large.*
4. **Age**: *20–30 years, 30–40 years, 40–50 years, 50–60 years, 60–70 years, 70–80 years, 80–90 years.*
5. **Gender**: *Female and male.*

A.3. Additional Visualization

Fig. 3 shows the diversity of reference images generated using our prompt template and attribute set. Fig.4 and Fig. 5 illustrate the multi-view images under diverse poses generated by our MVChamp.

A.4. Application: Human Video Reenactment

The goal of this application is to replace a person in a reference video with a new identity while preserving the background and pose. Given a reference image that provides the target identity, and a reference video that provides the pose and background of the original person, the task is to seamlessly swap the person in the video while maintaining the integrity of the scene. We visualize the results in Fig. 1.

To achieve this, we follow a multi-step process:

Identity Reconstruction: The *IDOL* model is used to reconstruct an animatable 3D human from the reference image. This model generates a highly detailed and realistic representation of the target identity, allowing us to manipulate the avatar to match various poses.

Background Inpainting: The video inpainting process restores the regions of the video frame where the original person has been replaced, ensuring a seamless background. It involves detecting and tracking the target area using a segmentation method, which is initialized and refined by the widely used zero-shot segmentation model, Segment Anything Model (SAM)[8]. Once the target area is segmented and tracked, the remaining regions are completed using the video inpainting method, ProPainter[20], ensuring the background is seamlessly restored with no traces of the replaced identity.

Pose Animation: The target pose is extracted from the reference video [3, 9], and the reconstructed human model is animated to match this pose. The IDOL model provides precise control over the 3D human's pose, including fine details such as **finger movements**, allowing it to adapt dynamically to the reference video's actions. After animating the 3D human, we render it into the target view and seamlessly blend it with the background.

Utilizing *IDOL*, our process offers an efficient and high-quality solution for identity replacement in videos, providing greater stability and lower computational cost compared to 2D-based approaches [6, 21]. This opens up new possibilities for digital content creation and interactive media applications.

A.5. Representation Comparisons

To further illustrate the differences between our method and previous approaches, we provide a comparison in Fig. 2. Below, we explain the key differences:

Comparison to PIFU: PIFu predicts the 3D human shape directly from a given image without leveraging a parametric model prior. While effective for simple cases, it often lacks robustness and precision, particularly when handling challenging poses or incomplete observations [16].

Comparison to GTA/SIFU: GTA and SIFU utilize loop optimization [16, 17] to align the reconstructed output with SMPL models. While this alignment step is crucial for pixel-aligned operations [13], it introduces several significant drawbacks:

- High computational cost: Loop optimization requires multiple iterations, adding several minutes of processing time. Additionally, it depends on the estimation of intermediate representations such as masks, normals, and skeletons.

- Error accumulation: Misalignments during optimization can accumulate over iterations, degrading the quality of the final 3D human reconstruction.

Our Approach: In contrast, our method adopts a direct and efficient pipeline: We extract image features using a large-scale encoder [7], which captures rich and detailed visual information. We then predict the 3D human shape and appearance in a uniform space, directly providing the 3D human reconstruction along with the estimated SMPL-X parameters.

By decoupling feature extraction from SMPL-X-based 3D prediction, our approach avoids the error accumulation inherent in optimization-based methods. When pose information is unnecessary, our method relies primarily on body

shape estimation, reducing the dependency on precise pose alignment. Furthermore, our method supports direct animation and editing (*e.g.*, shape and texture), unlocking additional applications and expanding its potential value in digital content creation.

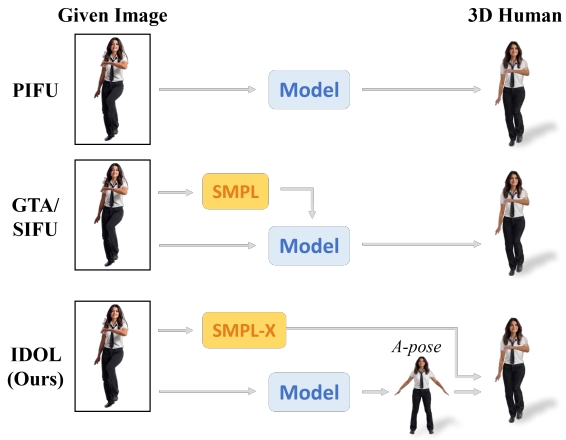

Figure 2. Visualization of different approaches for 3D human reconstruction. Unlike PIFu, which directly predicts the 3D human without a parametric prior, and GTA/SIFU, which relies on computationally expensive loop optimization for SMPL alignment, our IDOL method leverages SMPL-X as a prior. This enables more robust and accurate reconstruction while avoiding the pitfalls of error accumulation. Furthermore, our method supports direct animation and editing, enabling additional applications in digital content creation.

B. More Details of *IDOL*

In this section, we describe the training setup and methodology for our proposed method, *IDOL*.

B.1. Implement Details

Our models are trained on a cluster of 32 NVIDIA H100 GPUs for approximately 1 day, with a batch size of 32. The optimization is performed using the Adam optimizer with a learning rate of $5e-4$. A warm-up schedule of $3,000$ steps is employed to stabilize training in the initial stages.

The training loss function is a weighted combination of VGG perceptual loss and Mean Squared Error, balanced with a $1:1$ ratio. This loss formulation ensures both perceptual quality and pixel-wise accuracy.

B.2. Network Architecture

The proposed network consists of a multi-stage structure designed for high-dimensional feature extraction and reconstruction tasks. The primary components include the pre-trained encoder, UV-Alignment Transformer, and UV decoder. For the encoder, we utilize the large-scale model Sapiens [7] to extract and tokenize human features from the input image.

UV-Alignment Transformer. The neck module employs a hierarchical design inspired by recent advancements in vision transformer architectures [7], featuring a decoder embedding layer with a width of 1536 and 16 transformer layers. Each transformer encoder layer consists of the following components:

1. A layer normalization operation for input stabilization, enhancing training dynamics, and preventing gradient instability.
2. A multi-head self-attention mechanism that maps inputs into query, key, and value representations, followed by a linear projection layer to integrate attention outputs. This process is regularized through dropout for improved generalization and further normalized to ensure consistent feature scales.
3. A feed-forward network (FFN) composed of two dense layers with a GeLU activation function applied between them. The FFN architecture is complemented by intermediate normalization layers to enhance stability and improve optimization convergence.

UV Decoder. The decoder begins by reshaping tokens into a 2D feature map of 64×64 resolution. It employs a hierarchical upsampling and convolutional strategy to progressively refine and synthesize outputs. The upsampling mechanism uses transposed convolutional layers to increase spatial resolution, with each stage incorporating normalization and non-linear activation for stable feature transformations. Specifically:

1. Upsampling Blocks: The decoder incorporates multiple transposed convolutional layers, which double the spatial resolution at each stage. Instance normalization and `SiLU` activations provide stable scaling and enable non-linear feature transformations.
2. Convolution Block: Three convolutional layers with output channels $\{128, 128, 32\}$ further process the features, applying instance normalization and activation functions to improve feature quality and representation.

Head Module. Following [18], we construct two distinct convolutional networks for decoding geometry and color separately. These networks progressively process feature channels, transitioning from an initial channel size of 32 to the target parameters $\delta_{\mu_k}, \delta_{\mathbf{s}_k}, \delta_{\mathbf{r}_k}$ for geometry and $\mathbf{c}_k$ for color.

C. Experiment

In this section, we present additional experimental results, including comparison tables and the user study. We show additional visual comparisons in Fig. 6 and Fig. 7. We compare with the reported results by Weng et al. [14] and AlBahar et al. [2].

Figure 3. Visualization of diverse images generated by Flux [5].

Figure 4. Visualization of examples from *HuGe100K*, where the images are generated by Flux and used to generate multi-view images.

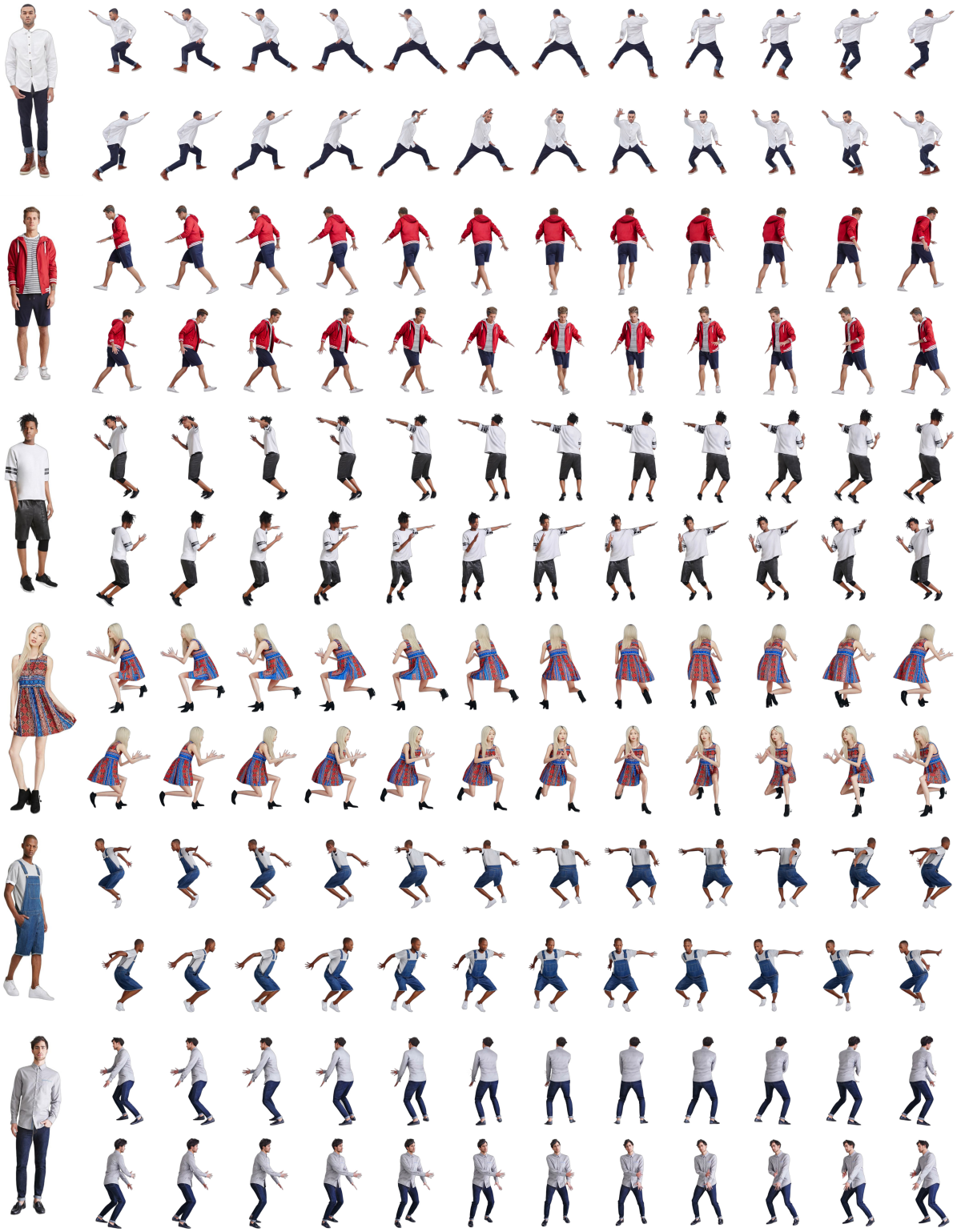

Figure 5. Visualization of examples from *HuGe100K*, where the images are derived from the DeepFashion [11] dataset and used to generate multi-view images.

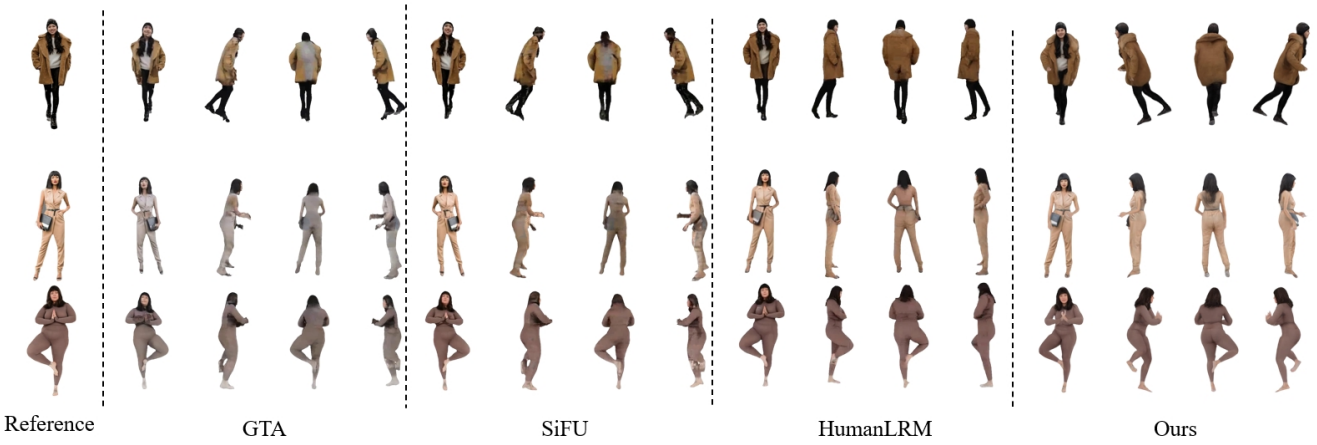

Figure 6. More visualization for comparison in the in-the-wild cases. We compare with the reported results by HumanLRM[14].

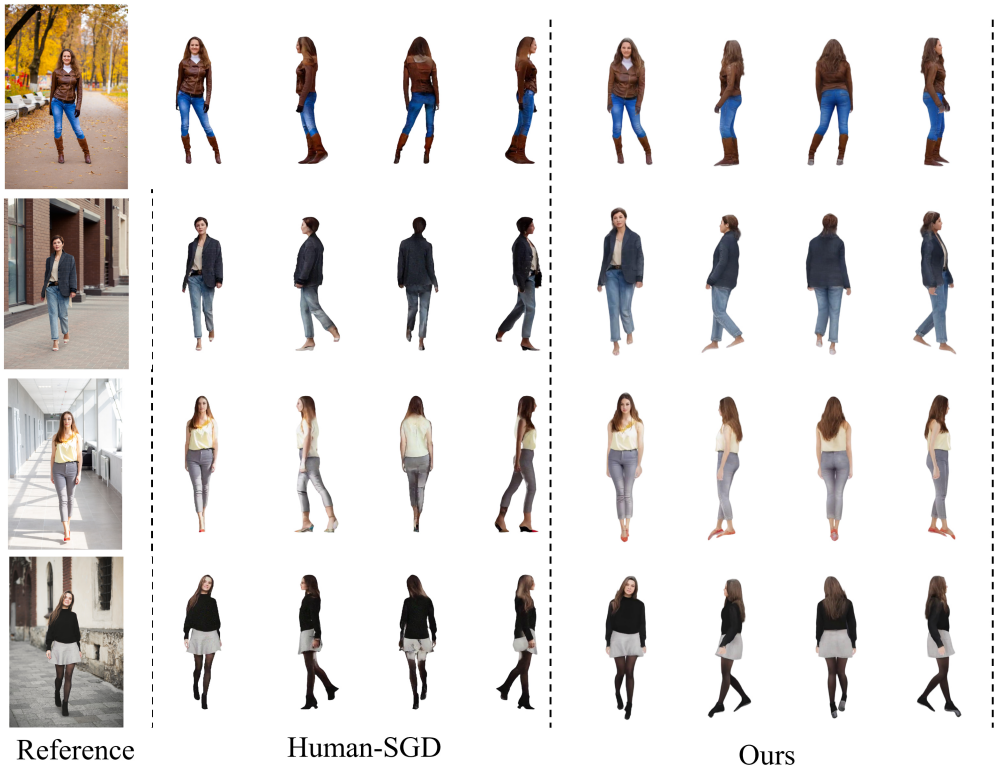

Figure 7. More visualization for comparison in the in-the-wild cases. We compare with the reported results by HumanSGD[2].

More Qualitative Comparisons. We show additional visual comparisons in Fig. 6 and Fig. 7. We compare with the reported results by Weng et al. [14] and AlBahar et al. [2].

Effect of the SMPL-X Parameters on Reconstruction. Although the reconstruction quality remains good with imperfect SMPL-X input, errors such as leaning or bent shapes can occur due to inaccurate pose parameters, as shown in Fig. 6 of the main content. This occurs because the avatar is re-posed based on the estimated SMPL-X parameters. Fig.8b demonstrates that providing accurate pose information resolves this issue.

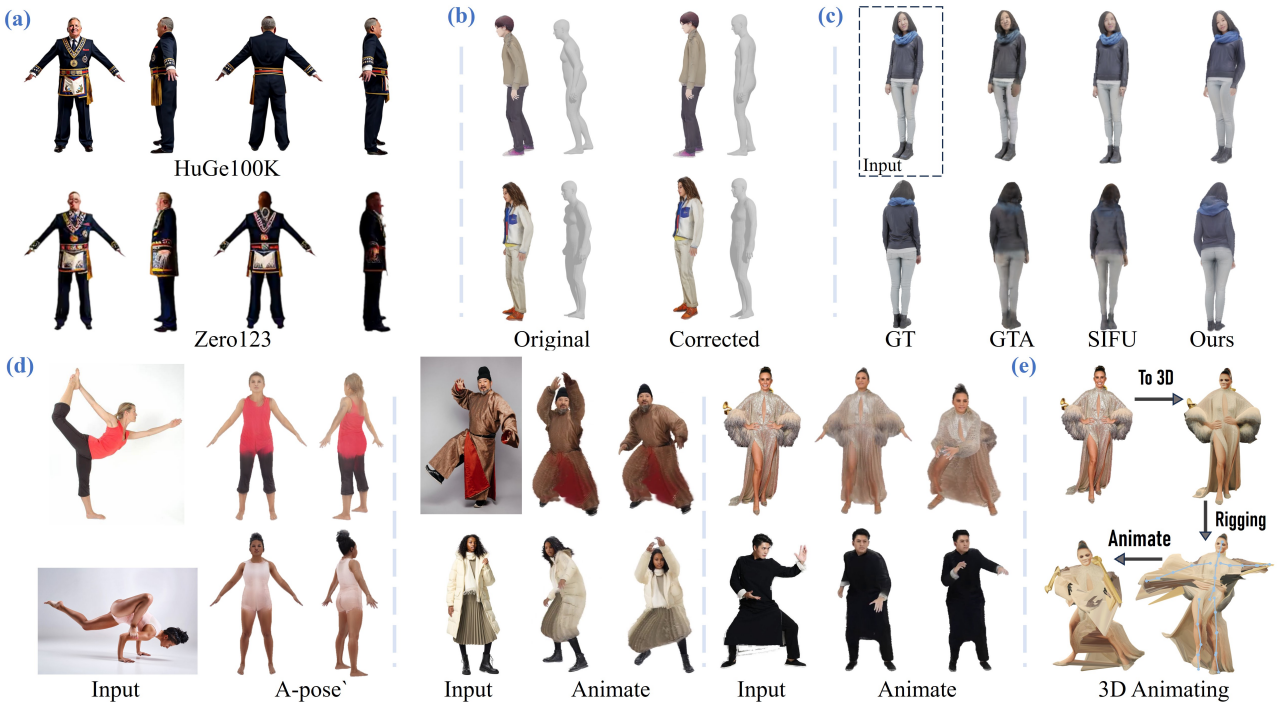

Figure 8. (a) Comparison with Zero123. (b) Original results with leaning/bent poses due to inaccurate SMPL-X, and corrected results with refined SMPL-X. (c) The results on 2K2K. (d) Challenges in large pose and loose cloth animation. (e) 3D animating framework using TRELLIS for image-to-3D and Make-It-Animatable for rigging and animation.

Dataset	WE ($\times 10^{-3}$) $\downarrow$
THuman2.1	5.38
HuGe100K (MVChamp)	7.33
Zero123	10.51

Table 1. Warping Error (WE) comparison across datasets and multi-view synthesis methods, evaluating 3D consistency.

Method	MSE $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
SIFU	0.032	15.054	1.303
GTA	0.035	14.833	1.340
Ours	**0.023**	**16.688**	**1.171**

Table 2. Quantitative comparison on the 2K2K dataset.

Experimental Comparison with Other Multi-View Image Generation Models. Here, we compare MVChamp with traditional multi-view image generation models based on text-to-image synthesis, specifically Zero123, in the context of human multi-view generation. We compare WE [10] in Tab. 1, evaluating 50 random cases. The THuman2.1 dataset serves as the upper bound, and HuGe100K shows comparable results; Regarding multi-view generation, MVChamp outperforms Zero123 by 30.1% in 3D consistency. Zero123 generates one novel view at a time, causing multi-view inconsistency. In contrast, MVChamp generates 24 views per batch, ensuring consistency. It also benefits from redundant human priors from dance videos and provides more accurate pose control, enabling well-aligned SMPL-X parameters. See Fig. 8a for a visual comparison.

Additional Cases for Evaluating Generalization to Complex Poses and Loose Clothing. Fig. 8d demonstrates IDOL's capability to handle complex poses and loose clothing. This is made possible by our novel architecture, which extracts global features using Sapiens and the diverse HuGe100K dataset. While loose clothing presents challenges due to significant deviation from the body, HuGe100K provides numerous examples, allowing IDOL to recover animatable 3D avatars effectively and reduce issues like tearing in animations, especially in areas such as skirts. *For more examples and animations, please refer to the introduction video (38s-54s).*

Comparison with 3D Animating Methods. Classical animation methods typically involve image-to-3D conversion, rigging, and animation. Fig.8e demonstrates this pipeline using TRELLIS[15] and Make-It-Animatable [4], which struggles with topology changes, such as detaching the hand from the waist, resulting in artifacts. In contrast,

our approach (left) handles these transitions naturally.

Evaluation on Additional 3D Datasets. We performed the suggested evaluation on 2K2K using the same settings as in the paper. The quantitative and qualitative results are presented in Tab.2 and Fig.8c, offering valuable insights into IDOL's generalizability.

User Study. We conducted a user study with 20 participants via evaluating 50 cases. Participants ranked results based on face, clothing, back-view consistency, and the overall quality. The aggregated results are presented in Tab. 3, showing the superiority of our method.

Method	Face	Clothing	Back	Overall
GTA	0%	2.27%	0%	0%
SIFU	2.27%	4.55%	4.55%	4.55%
HumanLRM	45.45%	43.18%	36.36%	45.45%
Ours	52.28%	50.0%	59.09%	50%

Table 3. The user study. We evaluated IDOL on selected cases reported by HumanLRM[14], designed to highlight their strengths. Despite the selection for HumanLRM, our method achieves slightly superior performance, demonstrating greater robustness and effectiveness under comparable conditions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. 4, 8

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[4] Zhiyang Guo, Jinxu Xiang, Kai Ma, Wengang Zhou, Houqiang Li, and Ran Zhang. Make-it-animatable: An efficient framework for authoring animation-ready 3d characters. *arXiv preprint arXiv:2411.18197*, 2024. 9

[5] https://github.com/black-forest labs/flux. Flux latent rectified flow transformers, 2024. 1, 5

[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. 3

[7] 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, 2025. 3, 4

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[9] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 3

[10] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 9

[11] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 1, 7

[12] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 1

[13] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 3

[14] Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Template-free single-view 3d human digitalization with diffusion-guided lrm. *arXiv preprint arXiv:2401.12175*, 2024. 4, 8, 10

[15] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 9

[16] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 3

[17] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[18] Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. E3gen: Efficient, expressive and editable avatars generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6860–6869, 2024. 4

[19] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 1

[20] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10477–10486, 2023. 3

[21] 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