

# FEED-FORWARD HUMAN PERFORMANCE CAPTURE VIA PROGRESSIVE CANONICAL SPACE UPDATES

**Anonymous authors**

Paper under double-blind review

## A VIDEO RESULTS

We present video comparisons for the novel view synthesis task in the attached .mp4 file. To assess the advantages of our method—leveraging temporal observations aggregated in the canonical space and probabilistic rendering—we compare it against two feed-forward baselines: (1) Neural Human Performer (NHP) Kwon et al. (2021), a temporal deterministic regression model, and (2) Champ Zhu et al. (2024), a per-frame probabilistic method.

We evaluate on two generalization settings using only unseen subjects at test time: (1) *In-domain generalization*, where models are trained on THuman2.1 and 4D-Dress and tested on held-out 4D-Dress subjects, and (2) *Cross-dataset generalization*, where models trained on the same combined dataset (THuman2.1 and 4D-Dress) are tested on MVHumanNet.

## B REPRODUCIBILITY

### B.1 IMPLEMENTATION DETAILS

Rather than focusing on architectural novelty in probabilistic rendering, our contribution lies in demonstrating *how a carefully designed canonical context can significantly improve synthesis quality when paired with off-the-shelf diffusion models*. Therefore, we use a standard pre-trained VAE model and Stable Diffusion model provided by Hugging Face <sup>1</sup>.

**VAE.** For  $U_{\text{vae}}$ , we used the frozen AutoencoderKL VAE (sd-vae-ft-mse) from the Diffusers library with a downscale factor of 8 (e.g., a  $1024 \times 1024 \times 3$  image encodes to a  $128 \times 128 \times 4$  latent).

**Live Frame Feature Extractor.** To extract hierarchical feature maps  $F_t$  from the current live frame  $I_t$ , we use the ResNet-18 model as a feature extractor. The model and its pretrained weights are taken from the torchvision library and kept frozen during training. The extracted feature maps progressively downsample the input resolution to  $1/2$ ,  $1/4$ , and  $1/8$  of the original size, corresponding to the outputs of the conv1, layer1, and layer2 blocks (*i.e.*, first three layers). These features are then upsampled to a common spatial resolution and concatenated along the channel dimension to form the final feature map  $F_t$ .

**Live Frame Encoder and Denoising Network.** Live frame encoder  $U_{\text{live}}$  is implemented using the UNet2DConditionModel from the Hugging Face Diffusers library, initialized from the Stable Diffusion v1.5 U-Net weights provided by Hugging Face. Denoising network  $U_{\text{denoiser}}$  is implemented using UNet3DConditionModel, which inflates the 2D U-Net architecture into a 3D variant. 2D U-Net is initialized from the same Stable Diffusion v1.5 checkpoint. For the augmented module, we load the motion module weights provided by AnimateDiff Guo et al. (2024).

**Context Encoder.** Context Encoder  $U_{\text{enc}}$  comprises a  $3 \times 3 \times 3$  inflated 3D convolution with SiLU activation, followed by two stages: each stage includes a residual 3D convolution, Transformer3DModel block for self-attention, and a downsampling 3D convolution (stride = 2). The final Transformer3DModel refines features before a zero-initialized 3D convolution produces the  $G_{\text{context},t}$ .

<sup>1</sup><https://huggingface.co/docs/diffusers>

**Conditioning of the Canonical Context and the Live Frame During Denoising.**  $U_{\text{denoiser}}$  denoises the novel view image conditioned on the canonical context  $G_{\text{context},t}$  and the current live frame input  $I_t$ . The live frame is conditioned by fusing the intermediate features of the live frame encoder (i.e.,  $G_{\text{live},t}$ ) with those of the denoising network. More specifically, in each transformer block of  $U_{\text{denoiser}}$  and  $U_{\text{live}}$ , intermediate outputs are concatenated along the spatial dimension (axis 1 of the  $[\text{batch}, H \times W, C]$  tensor), followed by self-attention.

## B.2 TRAINING DETAILS

**Selection of Input and Novel View.** For THuman2.1, we randomly select 10 out of 20 available views to initialize the canonical space, then randomly choose 1 as the live frame and 1 novel view from the remaining 10 views. For 4D-Dress, we randomly select 1 input view and 1 novel view from the 4 available views. From the selected input view, 10 frames are sampled at 10-frame intervals to initialize the canonical space.

**Training of the Comparison Methods** For a fair comparison, we train all comparison methods using the same training setup as ours, including the same dataset and the same input/novel view selection strategy.

**Memory.** During training, we use a resolution of  $1024 \times 1024$  with a batch size of 1 on a single NVIDIA L40S GPU, which consumes approximately 40 GB of GPU memory. While this is non-trivial, further optimization such as mixed-precision training or the use of smaller VAE latents with a more effective VAE (e.g., CogVideoX Yang et al. (2024), Wan2.1 Wan et al. (2025)) can be explored in future work.

## B.3 INFERENCE DETAILS

**Selection of Input and Novel View.** For all baselines and our method, we use the frontal view as input (4D-Dress: 0004, MVHumanNet: CC32871A038). The test views are back, left, and right: [0028, 0052, 0076] for 4D-Dress and [CC32871A008, CC32871A037, CC32871A015] for MVHumanNet.

**Runtime and Memory.** While efficiency is not the focus of our work, we report inference runtime and memory usage for reference. All experiments are conducted using a single NVIDIA L40 GPU, rendering images at a resolution of  $1024 \times 1024$  with 10 denoising steps. We generate 20 images per inference pass, which takes approximately 20 seconds (i.e., 1 frame per second). The generation of 20 images at once consumes around 40 GB of GPU memory.

While the runtime and memory usage are nontrivial, they reflect a trade-off for achieving high-quality generation at  $1024 \times 1024$  resolution. Further optimization (e.g., model distillation, lower precision, or fewer denoising steps) is a promising direction for future work.

## C DATASET DETAILS

We use the THuman2.1 Yu et al. (2021), 4D-Dress Wang et al. (2024), MVHumanNet Xiong et al. (2024), and TikTok Jafarian & Park (2021) datasets, all of which are publicly available for research purposes.

### C.1 THUMAN2.1

THuman2.1 consists of 3D scans of clothed human subjects along with their corresponding SMPL-X parameters. Out of approximately 2,500 available scans, we randomly selected 1,846 subjects for training. Each subject is rendered from 20 uniformly distributed camera viewpoints, with the subject positioned at the center. For each view, we render both an RGB image and a corresponding foreground mask. Rendering is performed using the official script provided by the GPS-Gaussian repository<sup>2</sup>.

<sup>2</sup>[https://github.com/aipixel/GPS-Gaussian/blob/main/prepare\\_data/render\\_data.py](https://github.com/aipixel/GPS-Gaussian/blob/main/prepare_data/render_data.py)

## C.2 4D-DRESS

4D-Dress is a multi-view video dataset consisting of 64 human subjects wearing a variety of real-world outfits, including challenging loose garments and jackets. Each subject performs eight different motion sequences, captured by four uniformly distributed cameras. Each sequence contains approximately 200 frames. To evaluate in-domain generalizability, we randomly reserve 30 subjects for testing. The dataset provides RGB images, foreground masks, camera parameters, and corresponding SMPL-X parameters for each frame.

## C.3 MVHUMANNET

MVHumanNet is a multi-view video dataset comprising 4,500 subjects dressed in everyday clothing, captured using 48 synchronized cameras. To evaluate our method’s ability to aggregate temporal features in the canonical feature space, we carefully selected 30 representative motion sequences. These sequences feature subjects with diverse clothing styles, including logos, patterns, and asymmetric designs (e.g., different visuals on the front and back). We ensured that each selected sequence includes clear front, side, and back views, with cameras directly facing the subject. The dataset provides RGB images, foreground masks, camera parameters, and SMPL-X annotations for each frame.

## C.4 TIKTOK

The TikTok dataset is a collection of diverse in-the-wild dance videos from TikTok. We removed background with RemBG (Gatis, 2022). SMPL-X parameters were fitted with the off-the-shelf monocular estimator SMPLest-X (Yin et al., 2025).

## D SOCIETAL IMPACTS

Our work enables the creation of photorealistic human avatars from monocular RGB video, offering a low-cost alternative to traditional performance capture systems. This capability can broaden access to virtual reality and telepresence, allowing a wider range of users to participate in immersive digital environments. It also holds promise for safety-enhancing applications, such as replacing actors in hazardous scenes or enabling remote physical therapy using only consumer-grade cameras. In creative industries, our method could lower technical barriers for independent creators.

However, this technology also brings important ethical and social concerns. The ability to reconstruct and synthesize human appearances from minimal visual input could be exploited to generate deepfakes or impersonate individuals. Furthermore, without attention to dataset diversity, the system may fail to generalize across body types, skin tones, or clothing styles, reinforcing representational bias. To address these challenges, future work should explore mechanisms for transparent provenance, the development of usage guidelines, and initiatives to educate the public about the capabilities and limitations of synthetic media. We hope that this line of research ultimately contributes to responsible innovation that maximizes societal benefit while minimizing potential misuse.

## E LIMITATIONS AND DISCUSSIONS

**Fine-Grained Detail Reconstruction.** While our method achieves improved quality over the comparison methods, there is still room for enhancing visual fidelity, particularly in fine-grained details. Our method is based on a latent diffusion model (Stable Diffusion Rombach et al. (2022)), where supervision is applied at a  $128 \times 128$  latent resolution. Since fine details occupy only a small portion of this latent space, their reconstruction tends to be underemphasized, which can lead to reduced visual quality in certain localized regions. One promising direction is to adopt more powerful VAEs (e.g., CogVideoX Yang et al. (2024), Wan2.1 Wan et al. (2025)) to better preserve sharp and accurate details.

**Computational Efficiency.** As discussed above, efficiency is not the focus of this work. Rendering high-resolution images ( $1024 \times 1024$ ) requires substantial memory and longer denoising time (see

training and inference details). We leave further optimization (*e.g.*, model distillation, mixed-precision computation, or reducing the number of denoising steps) as promising directions for future work.

**Video Stability.** While our method is able to synthesize details that semantically align with past observations, some video jittering may still occur when the generated frames are viewed sequentially. This is because temporal stability is not explicitly modeled in our framework. In particular, we do not apply frame-to-frame smoothness regularization during training. Addressing this would require dedicated modeling of video stableness regularization or the integration of a video diffusion model Lu et al. (2025), which we leave as a promising direction for future work.

**Limited to a Single Foreground Subject.** Our method is designed to operate on a single foreground human and does not explicitly model the background, multiple subjects, or human–object interactions. Incorporating these components represents a promising direction for future work.

**Limitations of Human Template-Based Alignment.** Our method relies on 4D correspondences defined by a naked human body template to align live-frame features into the canonical space. While this provides a tool for the alignment, it introduces limitations when handling occlusions from fast, non-rigid deformations such as loose clothing in motion. For example, when a blue jacket swings widely during rapid movement, it may temporarily occlude the underlying sweatshirt. Because the template-based correspondence does not account for such occlusions, features from the jacket are projected onto the body surface and fused into the canonical space. This leads to contamination, where appearance information from the occluder (*e.g.*, the jacket) overwrites the features of the occluded region (*e.g.*, the sweatshirt). Consequently, the canonical representation may lose semantic accuracy and visual consistency. Addressing this issue would require occlusion-aware alignment or mechanisms capable of reasoning beyond the visible template surface.

## REFERENCES

- Daniil Gatis. rembg: Background removal tool. <https://pypi.org/project/rembg/2.0.28/>, 2022. Version 2.0.28.
- 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.
- 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 (CVPR)*, pp. 12753–12762, June 2021.
- Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021.
- Yixing Lu, Junting Dong, Youngjoong Kwon, Qin Zhao, Bo Dai, and Fernando De la Torre. Gas: Generative avatar synthesis from a single image. *arXiv preprint arXiv:2502.06957*, 2025.
- 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*, pp. 10684–10695, 2022.
- Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- 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*, pp. 19801–19811, 2024.
- 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.
- Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *arXiv preprint arXiv:2501.09782*, 2025.
- 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)*, June 2021.
- 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.