PoseGen: In-Context LoRA Finetuning for Pose-Controllable Long Human Video Generation

Supplementary Material

In this supplementary material, we present more details of our training data, implementation, additional qualitative results, ablation studies, robustness of hand normals, inference efficiency, and limitations.

1. Training Data

1.1. Data Construction

We collect human-centric videos that capture the movement of a single person, such as dancing or live broadcasting, from various online platforms. To ensure the quality of these videos for training, we employ a comprehensive filtering process. Initially, the collected videos undergo a manual screening to discard samples containing undesirable elements such as abrupt scene changes, background jitter, or superimposed special effects. Following this, we apply an automated filtering pipeline to only preserve samples that simultaneously meet the following criteria. (1) **Visual Quality.** We adopt the image and video quality scorers provided by Q-Align [9] to assess the visual quality of a collected video. A threshold of 0.7 is applied for both scorers to exclude low-quality samples. (2) **Human Presence.** We uniformly sample a set of frames from a video and extract human bounding boxes using RTMDet [4]. We then compute the human area ratio according to the estimated bounding boxes. Videos with an average ratio below 0.3 are discarded to eliminate those with insufficient human presence. Meanwhile, this moderate threshold effectively retains videos where humans are captured in close-up, medium, or full shots. (3) **Facial Confidence.** We apply YOLO [5] to extract bounding box annotations for the categories "face" and "head" in each sampled frame. Videos with an average confidence score exceeding 0.75 are retained to ensure clear visibility of human face and head regions. This filtering choice helps avoid significant occlusions from masks, hats, or similar accessories, which can detrimentally affect the performance of identity-preserving human video generation. Finally, we obtain a training dataset that comprises 7,005 high-quality video clips for LoRA finetuning.

1.2. Data Preprocessing

We leverage Sapiens [6], a family of human vision models, to predict pose skeleton maps and hand normal maps for training videos. Regarding pose estimation, the score threshold and the Intersection-over-Union (IoU) threshold for human bounding box detection are both set to 0.3. We render OpenPose [3]-style body and hand keypoints while discarding facial keypoints, in order to avoid identity leakage from the driving video to the generated subject. To balance precision and recall, we empirically set the confidence threshold to 0.3 for body keypoints and 0.7 for hand keypoints to ensure accurate pose skeletons. For hand normal prediction, we first apply the surface normal prediction model of Sapiens to generate normal maps for the entire frame, and then use its body-part segmentation model to localize the desired hand regions. Considering that the quality of hand regions in a training video is often affected by motion blurs, a dropout rate of 0.1 is applied to both hand skeletons and surface normals in order to decouple the strong reliance on these conditions and promote model robustness. We adopt Qwen2.5-7B-Instruct [10] as our text encoder to generate text prompts for reference images. The structure of a prompt typically begins with the image style, followed by a detailed description of the subject and the background, and ends with the type of the camera shot, which is similar to the prompt distribution of Wan2.1 post-training captions.

2. Additional Implementation Details

Our model is built upon the checkpoint of Wan2.1 Image-to-Video (I2V) 14B model [7]. The learnable parameters in the reference image patchifier are initialized by copying those in the noisy video patchifier. To prevent color-related artifacts that we observed in early experiments when concatenating hand normal maps with pose skeletons, we devised an alternative integration strategy. Instead of a simple concatenation, we process the hand normal maps through a separate patchifier and then add its output to the pose skeleton features. This convolution-addition scheme effectively fuses geometric information without introducing color spillover. We then append a zero-initialized linear projection layer after the 3D convolutional layer in the separated hand patchifier to suppress its influence at the beginning of the training process, inspired by the design of ControlNet [11]. During training, the learning rate is linearly warmed up to $1e-4$ over the first 10% of the total training steps, followed by a cosine annealing scheduler for the remaining steps. The LoRA rank and alpha are both set to 16.

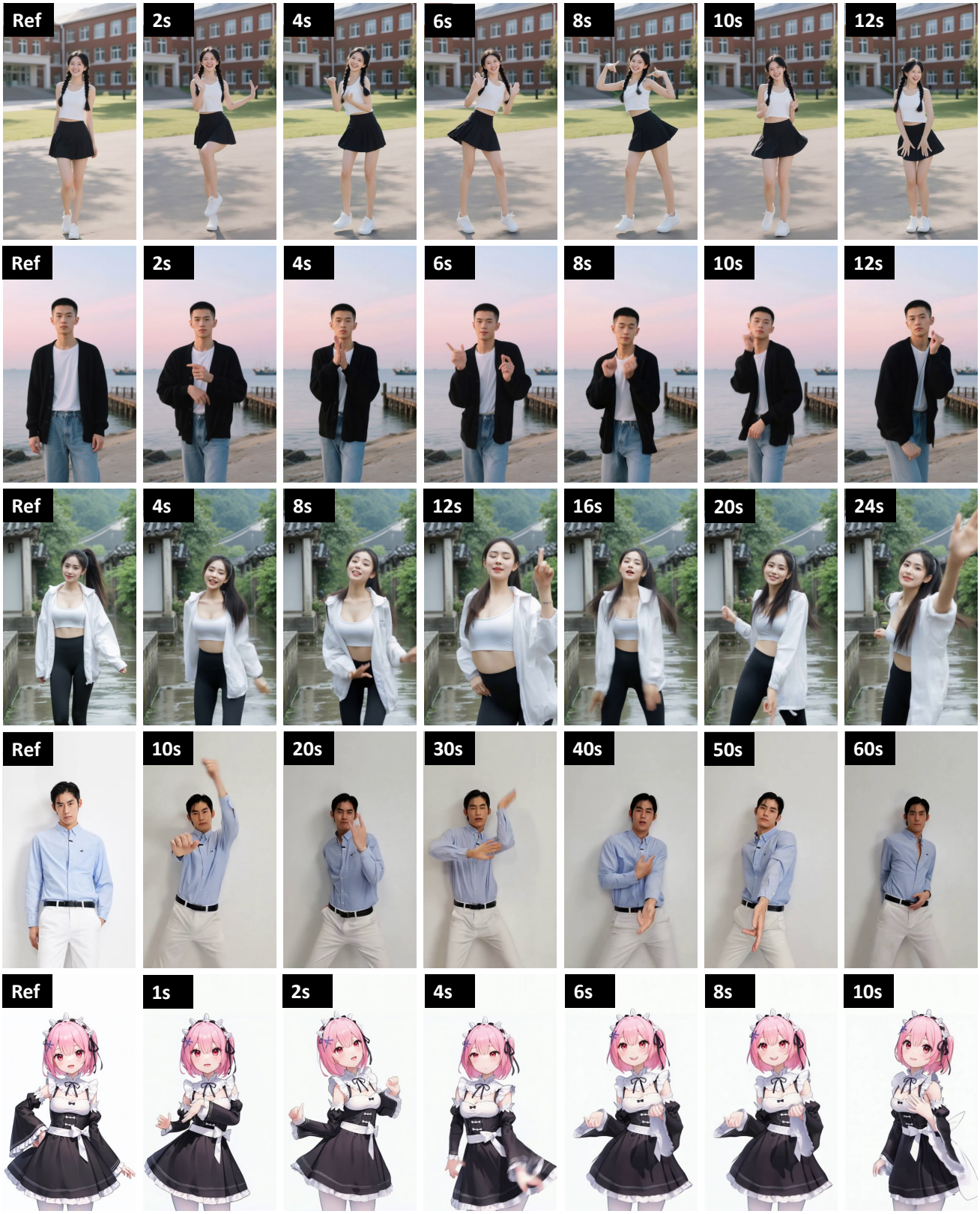

Figure 1. Additional qualitative results of pose-controllable human video generation. The reference images are displayed at the first column. PoseGen is capable of animating cartoon-style images as depicted in the last row.

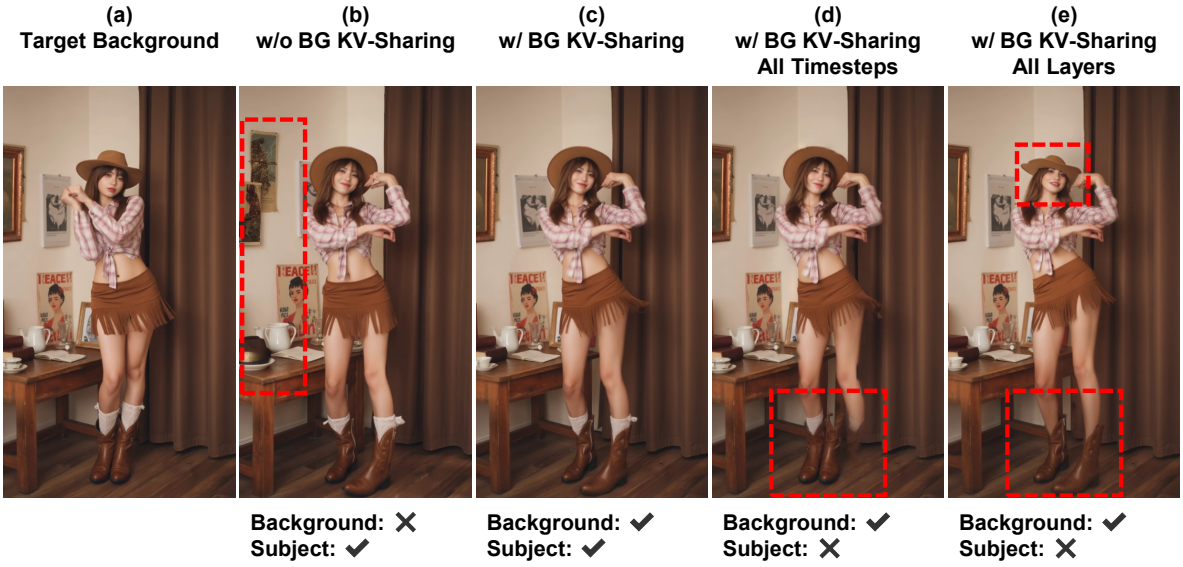

Figure 2. Ablation study on background KV-sharing in terms of denoising timesteps and transformer layers. (a) The source segment with the target background. (b) Background KV-sharing is not applied. (c) Background KV-sharing is applied to the first five timesteps and the last ten layers. (d) Background KV-sharing is applied to all timesteps and the last ten layers. (e) Background KV-sharing is applied to the first five timesteps and all layers.

3. Additional Qualitative Results

We present more qualitative results of pose-driven human video generation in Fig. 1. These visualizations showcase that our model effectively synthesizes realistic human videos while preserving identity consistency across diverse scenes and complex motion patterns. Notably, PoseGen exhibits impressive generalization capabilities beyond the real-human videos it was trained on. As demonstrated in the bottom row of Fig. 1, by first applying a pose retargeting process [8] to the driving motion, our model can successfully animate out-of-domain subjects like cartoon characters, faithfully transferring the motion while preserving the distinct artistic style.

4. Additional Ablation Study

Following the settings in DiTCtrl [1], we perform background KV-sharing at early denoising timesteps and at deep transformer layers. To further evaluate the effect of denoising timesteps and transformer layers, we conducted an ablation on the background KV-sharing mechanism, and the results are presented in Fig. 2. As shown in Fig. 2, background KV-sharing not only preserves the desired background from the source segment when generating the current segment, but also successfully drives the subject according to the pose signals without introducing noticeable artifacts. Specifically, applying background KV-sharing across all denoising timesteps causes inadvertent propagation of human pose information from the source to the current segment, resulting in overlapping poses and reduced motion fidelity. In contrast, implementing background KV-sharing across all transformer layers allows low-level visual features to leak from the source segment, leading to visual artifacts on the subject. These observations corroborate the findings in MasaCtrl [2], which indicate that source background is primarily conveyed during initial denoising steps and in deeper network layers. Our approach retains the first denoising step, a minor deviation from DiTCtrl [1] and MasaCtrl [2], to ensure consistency for the identical background expected across video segments.

	UniAnimate-DiT	PoseGen[1]	PoseGen[2]
Time (s/it)	64.9 / 69.5	75.0 / 83.3	71.2 / 77.5
Memory (G)	54.1 / 25.3	57.0 / 27.1	45.3 / 14.7

Table 1. **Inference efficiency**. PoseGen[1] denotes generating non-overlapping segments with KV-sharing. PoseGen[2] denotes generating in-between segments without KV-sharing. Each cell value is presented as "without / with" memory management. We run the test at 1280×720 resolution on a A100 GPU. Our method exhibits comparable runtime and manageable memory usage.

5. Robustness of Hand Normals

As illustrated in Fig. 3, the generation quality is robust to inaccurate hand normals, as our framework jointly leverages pose skeletons, hand normals, and the network's inherent temporal modeling capability to generate videos.

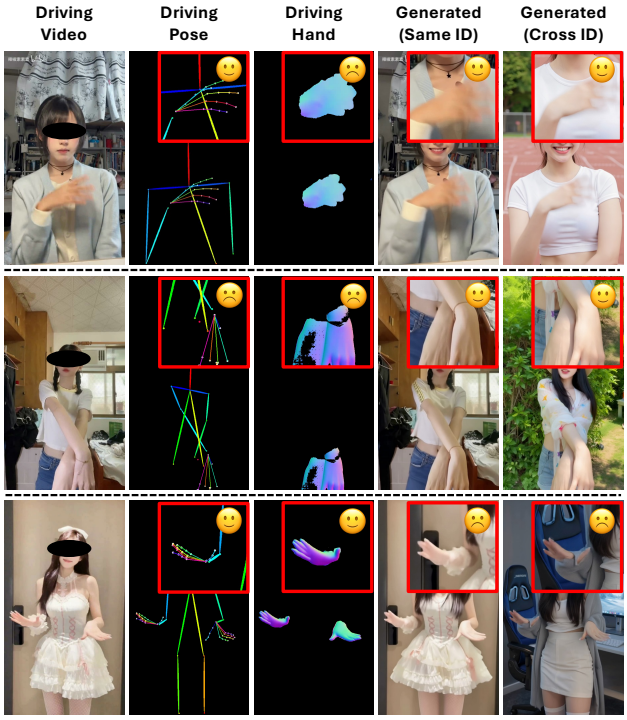

Figure 3. **Illustrations of hand surface normal robustness.** In the 1st row, despite inaccurate normal caused by motion blur, the model generates plausible hands by referring to pose. In the 2nd row, even when both pose and normal are inaccurate, our method can generate reasonable hands thanks to temporal modeling. The 3rd row presents a failure case where complex clothing interferes with hand appearance despite accurate conditions.

6. Inference Efficiency

We compare inference cost with UniAnimate-DiT [8], which adopts a similar architecture, in Tab. 1. For segment-interleaved generation, the runtime overhead mainly comes from 1/4 overlapping between adjacent segments that we designed for better temporal coherence, while the memory overhead stems from caching key-value pairs on GPU for background KV-sharing.

7. Limitations

Despite its strong performance, PoseGen also exhibits a few limitations. First, the interleaved segment-wise generation approach may introduce subtle drift in fine-grained details (such as the subject's delicate accessories) or in occluded regions of the reference image. Second, the model's ability to create seamless long videos hinges on the selected source segment having a static background, as any source variations can lead to unnatural artifacts in the segments to be generated. Third, in terms of controllability, our framework does not support nuanced control of facial expressions, which we leave for future work.

References

[1] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7763–7772, 2025. 3

[2] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 3

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[5] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. ultralytics/yolov5: v7.0 - yolov5 sota realtime instance segmentation. *Zenodo*, 2022. 1

[6] 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. 1

[7] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

[8] Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025. 3, 4

[9] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 1

[10] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[11] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1