PoseCrafter: One-Shot Personalized Video Synthesis Following Flexible Pose Control

Yong Zhong*[1], Min Zhao*[1], Zebin You[1], Xiaofeng Yu[2], Changwang Zhang[2], and Chongxuan Li**[1]

[1] Gaoling School of AI, Renmin University of China, Beijing, China
[2] Huawei Technologies Co., Ltd
yongzhong@ruc.edu.cn, gracezhao1997@gmail.com, zebin@ruc.edu.cn, yuxiaofeng16@huawei.com, changwangzhang@foxmail.com, chongxuanli@ruc.edu.cn

Supplementary Material

A Experimental Settings

In this section, we present more experimental details about datasets in Appendix A.1, implementation details in Appendix A.2, and baselines in Appendix A.3. We also detail the inference cost of PoseCrafter in Appendix A.4 and discuss its potential negative impact in Appendix A.5.

A.1 Datasets

In order to fully, reasonably, and comprehensively evaluate methods, we collect 10 high-quality videos in the open domain from YouTube, encompassing a variety of scenes such as interviews, movie clips, and talk shows. The collected videos were already publicly available online. We emphasize that our intention is purely academic and no offensive edits or alterations have been made to the original content. Additionally, we extend our sincere acknowledgments and respect to the original video producers.

For each video on TED, we take 8 frames uniformly from the first 36 frames for training. For inference, we extract pose information of the consecutive 100 frames, beginning with the 46-th frame, thereby setting $N = 8$ and $M = 100$. For each video on TikTok and our collected dataset, due to the need to vary the number of training frames, we designate the final 100 frames as test frames (i.e. $M = 100$) and uniformly select N frames from their preceding frames as training frames.

---

* Equal contribution.
** Correspondence to Chongxuan Li.

A.2 Implementation Details

We initialize Pose ControlNet[3] and VAE[4] using their provided public checkpoints, without any additional fine-tuning on other data. The guidance scale of free-classifier guidance is set to 1 and increased to 3 for attribution editing. It is important to note that a larger guidance scale results in the generated video being more aligned with the target prompt, but potentially less faithful to the training video. We use the default control scale of ControlNet, which is set at 1. Additionally, we adopt the 50 DDIM sampling steps for inference.

We set the source prompt p_s and target prompt p_t as "a person is speaking" for TED videos and "a person is dancing" for TikTok videos. For other datasets, we use the default prompt "a person". For attribute editing, we append relevant text to the source prompt. For example, to modify the hair color of the generated character to red, we use the prompt "a person, red hair" corresponding to its source prompt "a person".

We use the fixed learning rate of 0.003 and a fixed minimal batch size of 8 for all experiments. We set the max training step as 100 for 8 training frames and 2000 for 100 training frames. For the number of training frames N other than 8 or 100, we calculate the max training step using the following formula:

$$\text{round}(\frac{2000-100}{100-8}N + 100 - \frac{2000-100}{100-8}8) = \text{round}(\frac{475}{23}N - \frac{1500}{23}), \quad (1)$$

where round$(\cdot)$ denotes the function that rounds a value to the nearest integer.

It is worth noting that the aforementioned hyperparameters may not represent the optimal settings, but we empirically find that they can yield good results as defaults.

A.3 Baselines

We use only the commercial application GEN-2 for qualitative comparison, adopting its default parameters in all experiments. We utilize the prompt "a person is speaking" on TED and "a person is dancing" on TikTok.

We find that the default learning rate 1e-3 of Disco for human-specific fine-tuning tends to lead to overfitting, thus we adjust it to 1e-4 which yields better results. On TED, we use the default guidance scale of 3 for Disco but 1.5 for fine-tuned Disco to achieve better outcomes. On TikTok, we employ an optimal scale of 1.5, reported in [5], for both Disco and fine-tuned Disco.

Regarding the selection of the reference image for image-to-video methods, we employ every frame from 8 training frames as a reference image for inference on TED, and report average quantitative results. On TikTok, since the training frames vary but share the same first frame, we designate this first frame as the reference image. Moreover, applying image-to-video methods across all training frames incurs substantial budgetary and time expenses, notably when $N = 32$.

[3] `https://huggingface.co/lllyasviel/ControlNet-v1-1`
[4] `https://huggingface.co/stabilityai/sd-vae-ft-mse`

A.4 Inference Cost

PoseCrafter, with 1.48 billion parameters and around 1.747×10^{14} FLOPs in total, takes 2.75 GPU minutes to generate 100 frames, utilizing 19.28 GB of memory on a single RTX 3090. In our experiments, using a single RTX 3090 with 24 GB of memory, we successfully generate videos up to a maximum length of 180 frames with good quality. PoseCrafter can generate longer videos with GPUs of larger memory, memory reduction techniques, and long-range video generation strategies.

A.5 Potential Negative Impact

A major concern in human video generation is the risk of creating hyper-realistic videos that may impersonate real individuals. This technology allows for the production of avatars that closely resemble real people, often without their consent. Such convincing "DeepFakes" spark fears of identity theft, fraud, reputational harm, and regulatory challenges.

B Additional Results

In this section, we present more related baselines in Appendix B.1, more qualitative results in Appendix B.2, more quantitative results in Appendix B.3, and failure cases of PoseCrafter in Appendix B.4.

B.1 More Baselines

Tab. 1 and Tab. 2 introduce additional image-to-video baselines, which are marked in gray, for TikTok and TED, respectively. TPS [9] and MRAA [4] represents the state-of-the-art GAN-based methods and rely on ground truth videos. Consequently, [6] develops an alternative version that requires only DensePose sequences. Moreover, [6] establishes a related baseline, PA+CtrlN-V, integrating existing state-of-the-art image and video generation models. This includes the image-to-image method IP-Adapter [7], which maintains the identity of the reference image, Pose ControlNet [8], controlling the pose in generated videos, and the text-to-video model AnimateDiff [2], ensuring temporal consistency in results. We also report the original quantitative results of Disco and MagicAnimate.

It should be noted that these additional baselines cannot be directly compared with our method due to different settings in image resolution, the number of test frames, and evaluation codes. The original Disco trains and evaluates on 256-pixel resolution images, whereas we focus on a higher resolution of 512 pixels. Furthermore, while these image-to-video baselines evaluate using the full frames of test videos excluding reference images on TikTok and TED, we limit our testing to 100 frames for each video, as we require preceding frames for training. Regarding evaluation, MagicAnimate does not provide its implementation codes, and the Disco implementation contains errors[5], especially for PSNR.

[5] `https://github.com/Wangt-CN/DisCo/issues/86`

Table 1: Quantitative results on TikTok test dataset. NFs represent the number of available frames for the corresponding method. Image-GT signifies that these image-level metrics are calculated between each generated image and its corresponding ground truth image. † represents we fine-tune corresponding methods using their tailored strategies for specific subjects. ⋆ indicates the results are directly cited from the [6]. ‡ represents that the results are cited directly from the [5] where it additionally collects 250 internal TikTok-style videos for training.

Method	NFs	Image-GT				Image		Video	
		SSIM ↑	PSNR ↑	LPIPS ↓	MSE-P ↓	FID ↓	CLIP-I ↑	FVD ↓	CLIP-T ↑
TPS⋆ [9]	1	0.560	28.17	0.449		140.37		800.77	
MRAA⋆ [4]	1	0.646	28.39	0.337		85.49		468.66	
IPA [7]+CtrlN [8]-V⋆	1	0.479	28.00	0.461		66.81		666.27	
Disco‡ [5]	1	0.674	29.15	0.285		28.31		267.75	
MagicAnimate⋆ [6]	1	0.714	29.16	0.239		32.09		179.07	
DisCo [5]	1	0.704	15.16	0.358	9.11E-3	76.40	0.827	689.23	0.908
MagicAnimate [6]	1	0.756	17.95	0.265	5.48E-3	57.60	0.846	374.40	0.918
Disco† [5]	8	0.683	15.33	0.371	8.86E-3	65.73	0.813	807.35	0.886
ControlVideo [10]	8	0.738	16.93	0.311	6.78E-3	56.50	0.827	489.30	0.912
PoseCrafter (**ours**)	8	0.765	17.36	0.275	5.76E-3	48.09	0.840	440.49	0.921
PoseCrafter (**ours**)	16	0.776	17.87	0.252	5.23E-3	42.09	**0.864**	397.19	0.919
PoseCrafter (**ours**)	32	**0.786**	**18.56**	**0.233**	**4.05E-3**	**39.65**	0.854	**362.09**	**0.922**

Consequently, we have standardized these elements to ensure a fair comparison in the main text and detail them in the appendix.

B.2 More Qualitative Results

Fig. 1 and Fig. 2 show more qualitative results with the same individual poses and with poses from different individuals, respectively. Although using pose sequences with diverse motions from various humans, PoseCrafter can still maintain the human identity. In summary, PoseCrafter is capable of producing high-quality videos while allowing for flexible pose control.

B.3 More Quantitative Results

We also conduct quantitative experiments on open-domain videos. As shown in Tab. 3, consistent with conclusions on TED and TikTok, PoseCrafter excels in all quantitative metrics than all baselines on our collected open-domain dataset, demonstrating its effectiveness and robustness.

B.4 Failure Cases

PoseCrafter encounters several limitations in generating videos. Constrained by the capabilities of ControlNet [8] and the latent diffusion model [3], PoseCrafter produces mismatched poses (Fig. 3a) and low-quality results for complex poses (Fig. 3b). A significant discrepancy between training poses and inference poses

Table 2: Quantitative results on TED test dataset. NFs represent the number of available frames for the corresponding method. Image-GT signifies that these image-level metrics are calculated between each generated image and its corresponding ground truth image. † represents we fine-tune corresponding methods using their tailored strategies for specific subjects. ⋆ indicates the results are directly cited from the [6].

Method	NFs	Image-GT				Image		Video	
		SSIM ↑	PSNR ↑	LPIPS ↓	MSE-P ↓	FID ↓	CLIP-I ↑	FVD ↓	CLIP-T ↑
TPS⋆ [9]	1					86.65		457.02	
MRAA⋆ [4]	1					35.75		182.78	
IPA [7]+CtrlN [8]-V⋆	1					49.21		281.42	
Disco⋆ [5]	1					27.51		195.00	
MagicAnimate⋆ [6]	1					22.78		131.51	
DisCo [5]	1	0.550	17.22	0.327	3.20E-3	51.57	0.794	309.62	0.909
MagicAnimate [6]	1	0.498	13.32	0.338	1.97E-3	46.96	0.801	194.03	0.912
Disco† [5]	8	0.551	17.98	0.373	3.93E-3	60.56	0.827	244.62	0.909
ControlVideo [10]	8	0.771	22.32	0.196	1.62E-3	29.01	0.866	109.36	0.940
PoseCrafter (**ours**)	8	**0.810**	**23.92**	**0.142**	**1.56E-3**	**20.40**	**0.906**	**80.01**	**0.954**

causes PoseCrafter to yield low-faithfulness videos (Fig. 3c), especially regarding body proportion variation. Moreover, when the training video lacks motion diversity, PoseCrafter is prone to overfitting and struggles with learning temporal consistency.

C Ablation

In this section, we present a quantitative ablation study for the key designs of PoseCrafter in Appendix C.1, explore the impact of training frame quantity in Appendix C.2, and examine the role of sample timestep choice in latent editing in Appendix C.3.

C.1 Key Designs of Inference Framework

We quantitatively investigate the significance of reference frame selection, integration, and latent editing, as demonstrated in Tab. 4. Specifically, reference frame selection improves all metrics compared to ControlVideo, particularly in reconstitution and quality measures, highlighting the importance of selecting an appropriate frame from the training video to DDIM inversion. Furthermore, the integration of the reference frame further enhances all metrics, confirming that putting the pose of the reference frame as the inference pose encourages similarity across it and generated frames, thereby improving video quality. In addition to enhancing the quality of hands and faces, as indicated by SSIM and PSNR, latent editing also improves temporal consistency, i.e. better CLIP-T scores.

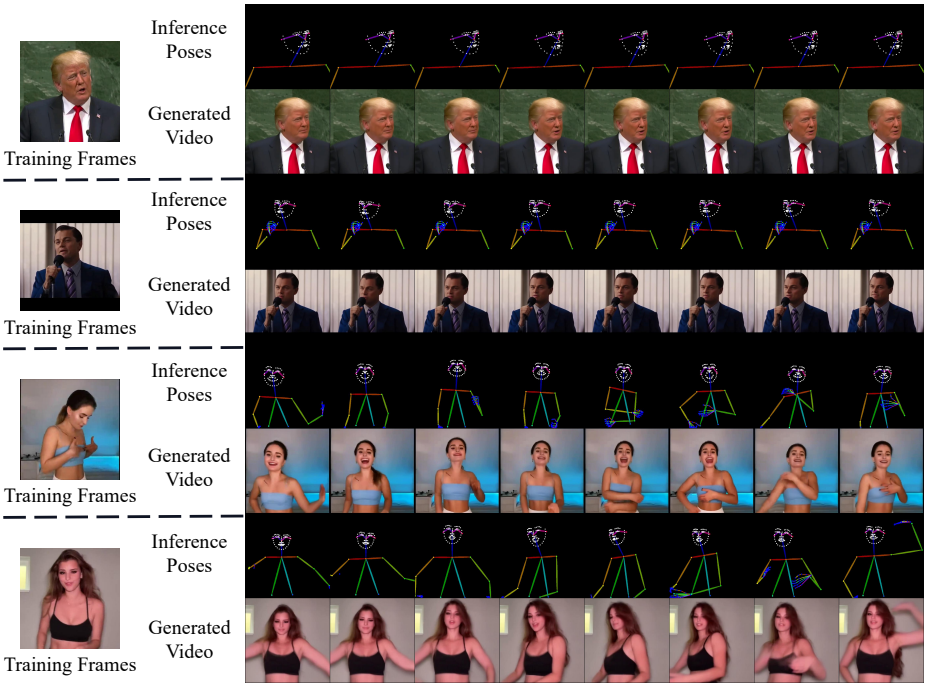


Fig. 1: Inference from poses of the same individual ($N = 100$ and $M = 100$). Time progresses from left to right.

We analyze that the features[6] of generated images corresponding to edited latent diverge more from real images' features than those corresponding to unedited latent, resulting in a decrease in FID and FVD scores. However, reconstruction metrics, which consider the distance between each generated frame and its corresponding ground truth, are more representative of quality than FID and FVD, which merely calculate the overall distance between all images and generated images. Hence, latent editing positively impacts video quality.

It is important to note that our strategy of selecting a reference frame from the training video whose pose coordinates are closest to inference poses may not be optimal. Therefore, exploring more effective methods for constructing a pseudo reference video is a significant area for further research.

C.2 Training on More Frames

We explore the impact of varying the training video length N on generated videos. For this purpose, we present Video-SIM to evaluate the resemblance

[6] FID and FVD use the 2D Inception and 3D Inception to extract image features, respectively.

Fig. 2: Inference poses of other individuals ($N = 100$ and $M = 100$). Time progresses from left to right.

between a training video $\mathbf{X}$ and a test video $\mathbf{X}'$, which is defined as:

$$\sum_{l=1}^{M} \min_{1 \le i \le N} ||\mathbf{p}_l - \mathbf{p}_i||_2^2, \tag{2}$$

where $\mathbf{p}_l$ is a pose of a certain frame in $\mathbf{X}'$, $\{\mathbf{p}_i\}_{i=1}^{N}$ is a pose sequence of $\mathbf{X}$, and the test video length M is 100 in our experiments. Essentially, Video-SIM sums up the shortest distances between each pose in $\mathbf{X}'$ and the nearest pose in $\mathbf{X}$.

Fig. 4 depicts curves of PSNR and Video-SIM where the number of training frames ranges from 8 to 180 uniformly selected from our collected open-domain videos. With the increase in training frames, better Video-SIM enhances the similarity of generated videos to the ground truth, as reflected in higher PSNR. However, in the later phases, despite significant gains in the number of training frames, PSNR has only slight fluctuations limited by marginal improvements in Video-SIM. This suggests that gathering videos encompassing a wider range of motions pertinent to target scenes can enhance the performance of PoseCrafter.

Table 3: Quantitative results on our collected open-domain dataset. NFs represent the number of available frames for the corresponding method. † represents we fine-tune corresponding methods using their tailored strategies for specific subjects. ‡ indicates corresponding methods directly use **ground-truth** frames of target poses as input conditions.

Method	NFs	**SSIM** ↑	**PSNR** ↑	**LPIPS** ↓	**FVD** ↓
PIDM [1]	1	0.358	9.36	0.690	1921.43
MRAA‡ [4]	1	0.654	15.26	0.423	1101.64
TPS‡ [9]	1	0.691	17.22	0.317	723.00
DisCo [5]	1	0.589	13.55	0.434	965.47
MagicAnimate [6]	1	0.621	14.84	0.317	518.23
Disco† [5]	8	0.638	15.51	0.310	583.38
ControlVideo [10]	8	0.774	20.40	0.174	279.90
PoseCrafter (**ours**)	8	**0.808**	**21.70**	**0.143**	**255.61**

Table 4: Ablation study for key designs of the inference framework on TED. We progressively incorporate our key designs into ControlVideo and highlight both the best results and those extremely close to them in bold.

Method	NFs	**Image-GT**				**Image**		**Video**	
		SSIM ↑	**PSNR** ↑	**LPIPS** ↓	**MSE-P** ↓	**FID** ↓	**CLIP-I** ↑	**FVD** ↓	**CLIP-T** ↑
ControlVideo [10]	8	0.771	22.32	0.196	1.62E-3	29.01	0.866	109.36	0.940
+ Reference-Frame Selection	8	0.799	23.35	0.158	1.59E-3	22.03	0.892	74.51	0.942
+ Reference-Frame Insertion	8	0.802	23.78	**0.141**	**1.57E-3**	**18.46**	**0.908**	**66.91**	0.949
+ Latent Editing	8	**0.810**	**23.92**	**0.142**	**1.56E-3**	20.40	**0.906**	80.01	**0.954**

C.3 Implementation Time of Latent Editing

We introduce a parameter α to determine the specific sampling time at which the latent editing operation will be executed, and we analyze the effect of α on generated videos in terms of CLIP-I (faithfulness), CLIP-T (temporal consistency), and SSIM (quality). As depicted in Fig. 5, with an increase in α, the faithfulness of generated videos first slightly rises and then decreases after $\alpha = 20$. Simultaneously, temporal consistency and overall video quality consistently decrease for α values ranging from 1 to 50. Therefore, we set $\alpha = 1$ as our default value, indicating that we edit the initial latent $\mathbf{Z}_T$ before the DDIM sampling process.

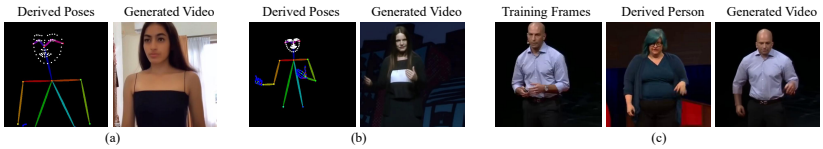
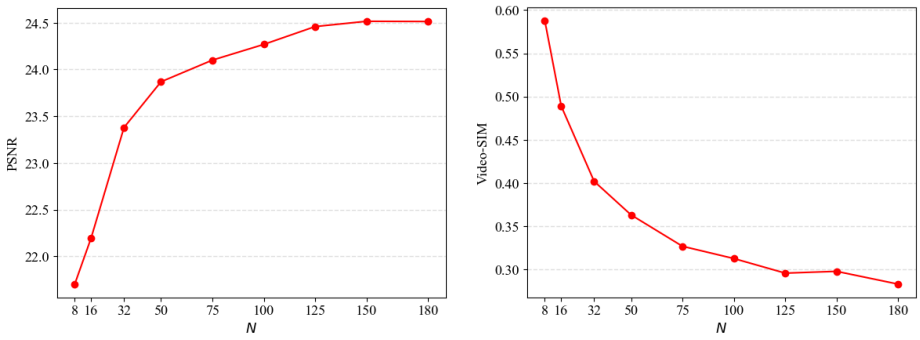


Fig. 3: Failure cases include (a) misalignment of the digital human's eyes with the derived poses (i.e., closed eyes), (b) poor rendering of the right hand, and (c) changes in the facial structure of the target subject.



(a) PSNR **(b)** Video-SIM

Fig. 4: The curves of PSNR and Video-SIM.

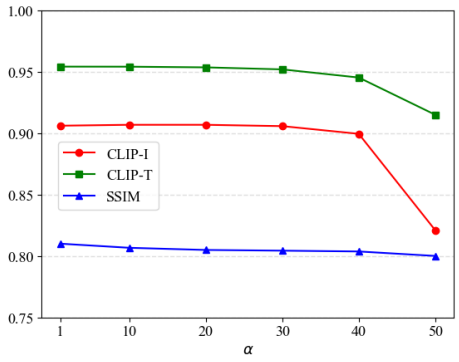


Fig. 5: The curves of CLIP-T, CLIP-I, and SSIM.

References

1. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. In: CVPR (2023)
2. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
4. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13653–13662 (2021)
5. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. arXiv preprint arXiv:2307.00040 (2023)
6. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. arXiv preprint arXiv:2311.16498 (2023)
7. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
8. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
9. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: CVPR (2022)
10. Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J.: Controlvideo: Adding conditional control for one shot text-to-video editing. arXiv preprint arXiv:2305.17098 (2023)