References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan, 2021.

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models, 2023.

[3] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation, 2019.

[4] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model, 2022.

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2022.

[6] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models, 2023.

[7] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, 2021.

[8] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing, 2021.

[9] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[10] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models, 2023.

[11] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, 2019.

[12] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild, 2018.

[13] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos, 2022.

[14] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths, 2022.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017.

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

[20] Aleksander Holynski, Brian Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields, 2020.

[21] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections, 2022.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[23] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation, 2021.

[24] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds, 2022.

[25] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. Dynast: Dynamic sparse transformer for exemplar-guided image generation, 2022.

[26] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets, 2022.

[27] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors, 2023.

[28] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[31] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models, 2022.

[32] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion, 2022.

[33] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation, 2020.

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2022.

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J

Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[37] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer, 2018.

[38] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation, 2020.

[39] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. 2021.

[40] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

[41] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, and Yaniv Taigman. Text-to-4d dynamic scene generation, 2023.

[42] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[43] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric challenges, 2018.

[44] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation, 2022.

[45] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[46] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models, 2022.

[47] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo, 2018.

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation, 2022.

[49] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation, 2022.

[50] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation, 2019.

[51] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan, 2021.

[52] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[53] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[54] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation, 2022.

[55] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation, 2019.

Supplementary Material

1. Implementation Details

Our experiments are trained on two NVIDIA A100 GPU's with resolution 512x512. In our first phase of training, we fine-tune our base model UNet on the full training dataset for a total of 5 epochs at a learning rate of $5e$-6. We use an effective batch size of 16 (through 4 gradient accumulation steps). We implement a dropout scheme where null values replace the pose input 5% of the time, the input image 5% of the time, and both input pose and input image 5% of the time during training. We further fine-tune the UNet on a specific sample frame for another 500 steps with a learning rate of $1e$-5 and no dropout. Lastly, we fine-tune the VAE decoder only for 1500 steps with a learning rate of $5e$-5. During inference, we use a PNDM sampler for 100 denoising steps [24].

2. User Studies

We conducted two user studies involving 50 distinct Amazon Mechanical Turk workers to compare our method with state-of-the art image animation approaches [39] [54] and evaluate the quality of our videos. In both surveys, workers evaluated results corresponding to 50 unique input images from the test set of the UBC Fashion dataset [50].

In the first user study, workers were asked their pair-wise preferences between our method and one of the other methods. For each input image, the workers were shown two videos: one containing the input image, our resulting video, and the MRAA resulting video and the other containing the input image, our resulting video, and the TPSMM resulting video. The ordering of our video and other video (MRAA or TPSMM) was randomized for each question. For each videos, workers selected their preference between the videos. The results are shown in Table 3. Overall, the workers had a preference for our method over MRAA and TPSMM.

In the second user study, workers were asked to rate our videos and TPSMM videos on a scale of 0 to 5, where 0 corresponds a video that does not match the input image at all and 5 corresponds to a realistic animation of the input image. During training, workers were shown a video of a different dress for as an example of a ”0” rating and a ground-truth video of the input image as an example of the ”5” rating. The results are shown in Figure 10. Our videos achieved higher scores for image similarity and quality than TPSMM and 85% of users rated the results of our method a 3 or higher.

	# Responses	Total Responses	(%)
Ours $>$ MRAA [39]	1637	2500	(65%)
Ours $>$ TPSMM [54]	1417	2500	(57%)

Table 3: Results of User Study #1: Workers choose between pairs of videos corresponding to input images, either our result vs. MRAA result or our result vs. TPSMM result. Overall, participants preferred our method over both MRAA and TPSMM in terms of quality and similarity to the input image.

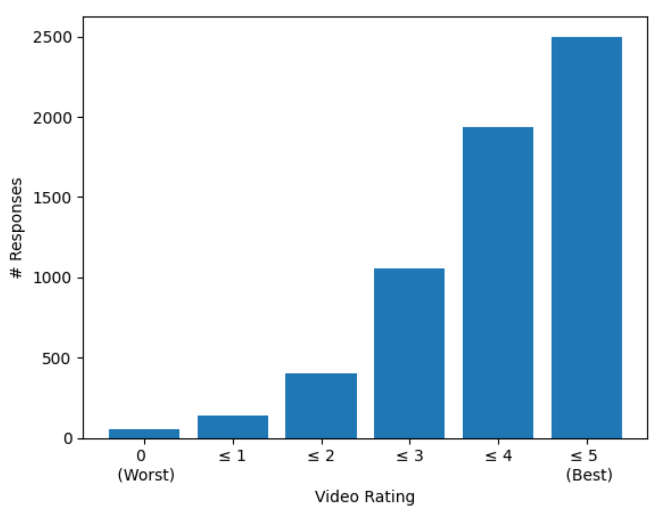

Figure 10: Results of User Study #2: Amazon Mechanical Turk worker ratings of our videos from 0 (video does not match input image) to 5 (video is a realistic animation of the input image). Overall, 85% of workers rated our method a 3 or higher.

3. Different Videos for Source Person and Driving Pose Sequence

We show in Figure 11 that DreamPose can animate an input image using motion from a video containing a different person and garment identity. As such, our method is applicable in practice when ground-truth motion is unavailable.

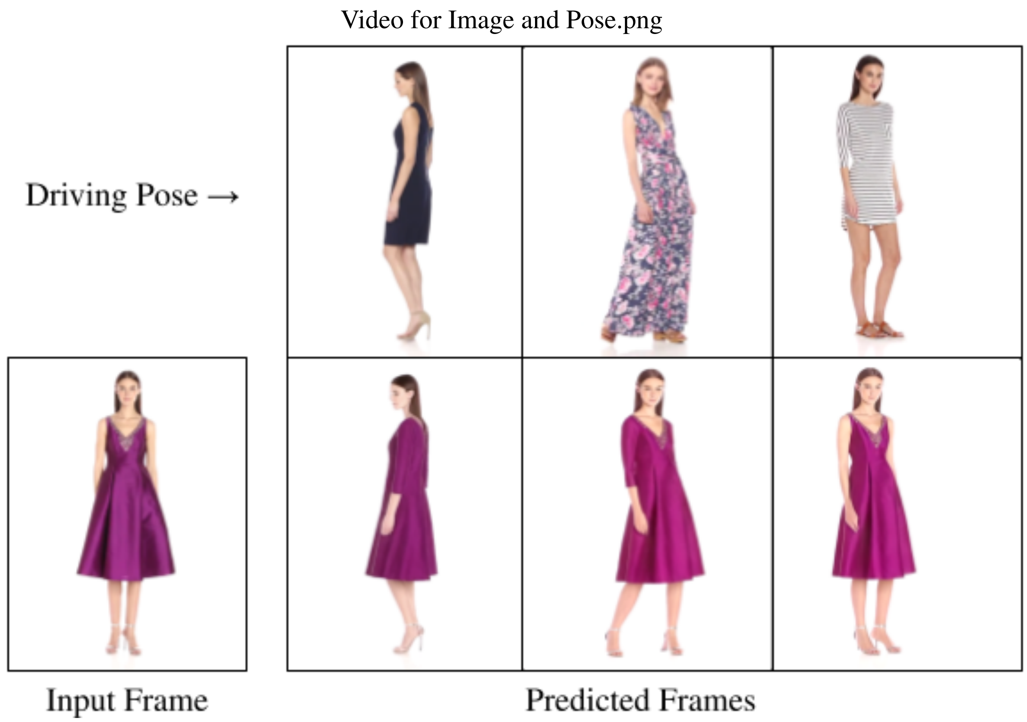

Figure 11: Qualitative results for conditioning on subject and pose from different videos.

4. Multiple Input Frames

While DreamPose demonstrates high-quality results with only a single input image, DreamPose can also be fine-tuned with an arbitrary number of input images of a subject. We showcase the results of training with multiple input images in Figure 12. We find that additional input images of a subject increase the quality and viewpoint consistency.

5. Deep Fashion Results

We demonstrate the effectiveness of our method on a popular dataset, DeepFashion, in Figure 13 [4, 11]. Although trained exclusively on the UBC Fashion video dataset, DreamPose performs well on unseen retail images, even to new backgrounds, model identities, accessories, and patterns.

6. Application to Pose Transfer

While adapted for image-to-video synthesis, DreamPose is also an effective pose transfer tool. In Figure 14, we compare DreamPose to two state-of-the-art pose transfer models: DynaST [25] and PIDM [4]. Our method is better able to preserve fine-details, such as shoe appearance, hemline, and face identity, than DynaST or PIDM.

Figure 12: Results after training with 1, 3, 5, and 7 input images. Increasing the number of input frames improves fidelity of pose, facial identity, and color.

Figure 13: DreamPose results on unseen samples from the DeepFashion dataset [11]. Despite being trained exclusively on the UBC Fashion Dataset, our method generalizes to new garments and model identities after subject-specific finetuning of the base model.

Figure 14: Comparison of Pose Transfer Results. We compare our method to two state-of-the-art pose transfer methods, DynaST [25] and PIDM [4].