Animating General Image with Large Visual Motion Model Supplementary Materials

Dengsheng Chen Xiaoming Wei Xiaolin Wei
Meituan
Beijing, China
{chendengsheng,weixiaoming,weixiaolin02}@meituan.com

1. Related Work

Advancements in Generative Synthesis for Image Animation The recent advancements in generative models have paved the way for the photorealistic synthesis of images from textual prompts [9, 10, 17, 38–40]. By extending the generated image tensors over time, these models can be adapted to synthesize video sequences [6, 8, 25, 32, 45, 59, 64]. Despite their capability to create plausible video sequences that encapsulate the spatiotemporal characteristics of real footage, these sequences often exhibit artifacts such as incoherent motion, and unrealistic temporal texture variations, and sometimes violate physical constraints like mass preservation. Instead of generating videos purely from text, certain techniques animate a static image. A plethora of contemporary deep learning methods utilize a 3D-Unet architecture to generate video volumes directly from an input image [18, 22, 24, 28, 31, 49]. As these models are essentially the same video generation models (but conditioned on image information instead of text), they exhibit similar artifacts as mentioned earlier. A potential solution to these limitations is to animate an input source image through explicit or implicit image-based rendering, i.e., manipulating the image content according to motion derived from external sources such as a driving video [30, 42–44, 54], motion or 3D geometry priors [7, 20, 27, 33–35, 47, 52, 55–57, 62], user annotations [5, 13, 15, 21, 23, 53, 58, 61], or a physical simulation. Although these methods exhibit enhanced temporal coherence and realism, they necessitate additional guidance signals or user input or otherwise depend on limited motion representations (e.g., optical flow fields, as opposed to full-video dense motion trajectories).

Exploration of Motion Models and Motion Priors Several studies have employed representations of motion beyond two-frame flow fields, in both Eulerian and Lagrangian domains. For instance, Fourier or phase-based motion representations (like ours) have been utilized for motion magnification and visualization [41], or video editing applications [16]. These representations can also be employed in motion prediction, where an image or video informs a deterministic future motion estimate, or a more comprehensive distribution of potential motions. Our work can be viewed as learning priors for motion induced by underlying scene dynamics, where our prior is an image-conditioned distribution over long-range dense trajectories [50, 51]. Recent studies have highlighted the benefits of modeling and predicting motion using generative models in several closed-domain settings such as humans and animals [2, 14, 19, 36, 48, 60].

Significance of Large-Scale Models Large-scale models have made significant strides in various visual tasks in recent years [63]. For instance, OpenAI's DALL-E [4] and CLIP [37] models have demonstrated excellence in image generation and text-to-image translation tasks [37]. Similarly, Google's BigGAN [11] and StyleGAN [1] have made substantial breakthroughs in image generation. By training on copious amounts of data, these large-scale models can learn intricate visual features and complex generation patterns. However, despite their significant accomplishments in static image generation, numerous areas remain unexplored in dynamic visual content, particularly in motion model generation. In this context, our work represents the first attempt to design a large-scale visual motion model. Our model can generate high-quality video sequences and capture complex motion patterns and temporal dependencies, thus excelling in tasks such as video generation and image animation.

2. Implementation and Pretraining Details

2.1. Network Structure

Neural Image Renderer $\mathcal{R}$ We adopt the architecture outlined in [29] to implement $\mathcal{R}$, primarily due to its simplistic design. In this scenario, $\mathcal{R}$ deploys two down-sampling blocks to construct a feature pyramid,

(a) Samples generated through training on the SHM dataset.

(b) Samples generated through training on the FMM dataset.

Figure 1. More examples. We strongly recommend watching the videos on the homepage.

$\{F_{I_0}^{H\times W}, F_{I_0}^{\frac{H}{2}\times\frac{W}{2}}, F_{I_0}^{\frac{H}{4}\times\frac{W}{4}}\}$. Subsequently, $\mathcal{R}$ renders a driving image $\hat{I}_k$ utilizing six residual blocks and two upsampling blocks.

Optical Flow Predictor $\mathcal{P}$ The optical flow predictor $\mathcal{P}$ is implemented using MRAA [44]. This approach is proficient in estimating optical flow δ_x, δ_y, and the motion intention weight ω, based on the identified object parts. To augment the performance, we also integrate the equivariance loss, following the methodology in [35].

Motion Encoder $\mathcal{E}$ and Decoder $\mathcal{D}$ We incorporate a Variational Auto-Encoder (VAE) with a continuous latent space of dimension 4, as initially introduced by Rombach et al. [39]. To enhance the model's performance, we integrate a multi-scale gradient consistency loss and a KL-divergence regularization with a weight of 10^{-6} [35].

Motion Diffusion Model For the denoising model ϵ_θ within the motion diffusion framework, we employ the conditional 3D U-Net architecture as proposed by Ho et al. [26], which comprises four down-sampling and four up-sampling 3D convolutional blocks. We encode I_0 and any optional text using ImageBind, subsequently utilizing them as two distinct tokens in the cross-attention module. ImageBind generates a vector of 1024 dimensions for both image and text. The latent motion vector z_0 of image I_0 is transformed to $\{v_1, \cdots, v_K\}$ and is also supplied to ϵ_θ through concatenation with x_t.

2.2. Dataset

During the pre-training phase, we selectively utilized the following three datasets.

- **CelebV-HQ**: The CelebV-HQ dataset is a comprehensive video facial attributes dataset on a large scale. It encompasses 35,666 video clips featuring 15,653 identities and 83 manually annotated facial attributes. These attributes span appearance, action, and emotion, offering a diverse and extensive dataset for facial analysis.

- **WebVid-10M**: The WebVid-10M is a large-scale dataset comprising short videos with textual descriptions obtained from stock footage websites. It includes 10.7M video-caption pairs, amounting to 52K hours of video content.

- **HD-VILA-100M**: The HD-VILA-100M is a large-scale, high-resolution, and diverse video-language dataset, specifically designed to aid multimodal representation learning. The dataset incorporates examples of video clips and transcriptions generated by Automatic Speech Recognition (ASR).

2.3. Pre-training Stage

Primary Training The initial training phase of the Motion Accretion Model was conducted using the WebVid10M dataset, necessitating 224 A100-GPU Days. Due to the presence of watermarks in the WebVid10M dataset, a random selection of approximately 1M videos from the HD-VILA-100M dataset was employed for fine-tuning, requiring an additional 32 A100-GPU Days. This process yielded our first general motion accretion model, denoted as $\mathcal{M}_{general}$. This model exhibits proficiency in predicting motion details across various scenarios and generating high-resolution images.

Augmenting Human Portrait Reconstruction To enhance the model's capacity for human portrait reconstruction, specifically the intricate facial movements, we continued to train $\mathcal{M}_{general}$ using the CelebV-HQ dataset, requiring 64 A100-GPU Days. The resulting model, denoted as $\mathcal{M}_{facial}$, was trained with images randomly selected from video data, with a frame interval not exceeding 150 as I_0 and I_k. Depending on the specific downstream task, we can selectively utilize different Motion Accretion Models for transfer learning.

Preprocessing for the Motion Diffusion Model Prior to the training of the Motion Diffusion Model, we preprocessed the data using the Motion Accretion Model trained in the initial phase to generate the sequence $\{z_0, z_1, \cdots, z_K\}$. This sequence was then supplied as a batch of data to the Motion Diffusion Model. To minimize redundant computations, all video data were converted into latent motion vectors and stored in the TFRecorder format for subsequent retrieval. It is worth noting that the efficacy of the prior knowledge acquired by the Motion Diffusion Model is contingent on the quality of the provided $z_0, z_1, \cdots, z_K$. To mitigate potential errors in estimating scene optical flow and significant reconstruction errors by the Neural Image Renderer during drastic scene changes, we implemented measures to reduce these errors during the pre-training stage.

Data Generation For each video, we selected a continuous segment of approximately 8 frames, i.e., $K = 8$, and evaluated the reconstruction error of I_0 and the final frame I_K. If the reconstruction error $\| \hat{I}_K - I_K \|_2^2$ exceeded 50, the current video sequence was discarded. We selected one frame as I_0 every 4 frames. For the CelebV-HD dataset, we generated approximately 300GB of training data using this method. Given the extensive volume of the WebVid10M dataset, we limited each video to randomly extract 5 video sub-segments for constructing training data, eventually generating over 1TB of training data.

Training the Denoise Model Following this, we trained the denoise model ϵ_θ on these two datasets separately, resulting in $\epsilon_{\theta_{general}}$ and $\epsilon_{\theta_{facial}}$. Both models underwent approximately 64 A100-GPU Days of training. During the training process, all frames were scaled and cropped to a resolution of 512×512. That is, $p \in \mathbb{R}^{128 \times 128}$, $z \in \mathbb{R}^{16 \times 16}$. More training configurations are listed in Tab. 1.

Configs	Values
T	1000
K	8
Betas of AdamW	(0.9,0.999)
Weight decay	0.0
Learning rate	1e-4
Batch size	160
Number of GPUs	32
Image/Text Encoder	ImageBind

Table 1. Hyper-parameters and values in motion diffusion model.

3. Applications

The Large Visual Motion Model (LVMM) is a robust technology that generates dynamic effects from a single static image, enhancing the visual experience across various scenarios. Its primary applications span stochastic generation, conditional animation, and seamless looping.

Stochastic Generation The Stochastic Generation technique facilitates the synthesis of realistic local micro-motions from an initial image, denoted as I_0, leveraging solely its inherent visual features. This capability empowers the Local Visual Motion Model (LVMM) to fabricate dynamic effects that are not merely convincing but also harmonize with the image's visual attributes, thereby enhancing the verisimilitude of the generated motions. This technological innovation is particularly effective in sectors such as entertainment for the production of realistic special effects, virtual reality for the construction of lifelike environments, and digital marketing for the creation of dynamic advertisements.

Our system animates a single static image I_0 by initially predicting a sequence of neural motion vectors $\{z_0, \cdots, z_K\}$, followed by the generation of an animation using our image-based rendering module $\mathcal{R}$ applied to the motion displacement fields obtained via $\mathcal{D}(z_k)$. The explicit modeling of scene motions in our system facilitates the production of slow-motion videos through linear interpolation of the motion displacement fields.

Conditional Animation The Local Visual Motion Model (LVMM), when supplied with an initial image I_0 and auxiliary modal conditions, possesses the ability to modify the image to conform to the stipulated conditions. This feature facilitates the creation of customized animations capable of conveying a specific narrative or complying with a predefined set of instructions, thereby delivering a tailored visual experience. This technological advancement proves particularly effective in domains such as education, where it can be used to create interactive learning resources, storytelling or film-making for the generation of narrative-centric animations, and user interface design for the development of dynamic, interactive components.

Seamless Looping Seamless looping involves generating video clips from an initial image I_0 that can loop continuously without perceptible discontinuities. This functionality is particularly beneficial in creating visually engaging content that can loop seamlessly, offering a continuous and immersive visual experience. This technology finds effective use in digital signage for looping advertisements, in social media for engaging, seamless looping content, and in art and design for captivating, endless visual effects. It is sometimes useful to generate videos with motion that loops seamlessly, meaning that there is no appearance or motion discontinuity between the start and end of the video.

Unfortunately, it is hard to find a large collection of seamlessly looping videos for training diffusion models. Instead, we devise a method to use our motion diffusion model, trained on regular non-looping video clips, to produce seamless looping video. Inspired by recent work on guidance for image editing, our method is a motion self-guidance technique that guides the motion denoising sampling processing using explicit looping constraints. In particular, at each iterative denoising step during the inference stage, we incorporate an additional motion guidance signal alongside standard classifier-free guidance, where we enforce each pixel's position and velocity at the start and end frames to be as similar as possible:

$$\hat{\epsilon}_t = (1+w)\epsilon_\theta(x_t, t, y) - w\epsilon_\theta(x_t, t, \emptyset) + \mu\sigma_t\nabla_{x_t}\mathcal{L}_t$$
$$\mathcal{L}_t = ||z_{K,t} - z_{1,t}||_1 + ||\nabla z_{K,t} - \nabla z_{1,t}||_1$$

where $z_{k,t}$ is the predicted 2D motion displacement field at time k and denosing step t. w is the classifier-free guidance weight, and μ is the motion self-guidance weight.

Qualitative Results We demonstrate more qualitative results in Fig. 1.

4. Dataset used in comparison experiments

Our comprehensive experiments are conducted on an array of publicly available video datasets, as detailed below:

- **MUG Facial Expression Dataset** [3]: This dataset comprises 1,009 videos featuring 52 subjects, each exhibiting one of seven distinct expressions, namely, *anger*, *disgust*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise*. We randomly allocate 26 subjects to the training set, resulting in 465 videos, while the remaining 26 subjects, comprising 544 videos, are assigned to the testing set.

- **MHAD Human Action Dataset** [12]: This dataset contains 861 videos of 27 actions performed by 8 subjects. The actions encompass a wide range of human movements, including sports actions (e.g., *bowling*), hand gestures (e.g., *draw x*), daily activities (e.g., *stand to sit*), and training exercises (e.g., *lunge*). We randomly assign 4 subjects to the training set (431 videos) and the remaining 4 subjects to the testing set (430 videos).

- **NATOPS Aircraft Handling Signal Dataset** [46]: This dataset includes 9,600 videos of 20 subjects performing 24 body-and-hand gestures used for communicating with U.S. Navy pilots. The gestures include common handling signals such as *spread wings* and *stop*. We randomly assign 10 subjects to the training set (4,800 videos) and the remaining 10 subjects to the testing set (4,800 videos).

Data Preprocessing All videos are resized to a resolution of 128×128 for compatibility with our models. For the MHAD and NATOPS datasets, we further preprocess the videos by cropping out portions of the background using the provided depth maps. Given that the majority of videos in these datasets are relatively short (averaging no more than 80 frames), we opt for random sampling of 40 frames per training video, rather than uniform sampling. These frames are then sorted chronologically to generate a diverse array of fixed-length video clips for training the stage-two DM.

5. More discussion

Owing to the unified multimodal feature encoding provided by ImageBind, potential applications of LVMM encompass image dynamization guided by text or/and audio. Despite the Motion Accretion Model's proficiency in accurately parsing intricate local micro-dynamics within scenes and synthesizing corresponding video clips, it encounters difficulties in managing large-scale movements. This is particularly evident when the synthesized videos require the generation of substantial content not present in the input frame. Although a generalized Motion Diffusion Model has been trained, it is yet incapable of directly producing high-quality motion trajectories encompassing arbitrary motion forms in a zero-shot manner. To achieve satisfactory results, fine-tuning the motion diffusion model on relevant datasets remains a necessary step.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 1

[2] Hyemin Ahn, Esteve Valls Mascaro, and Dongheui Lee. Can we use diffusion probabilistic models for 3d motion prediction? *arXiv preprint arXiv:2302.14503*, 2023. 1

[3] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE, 2010. 4

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions, 2023. 1

[5] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14707–14717, 2021. 1

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1

[7] Richard Strong Bowen, Richard Tucker, Ramin Zabih, and Noah Snavely. Dimensions of motion: Monocular prediction through flow subspaces. In *2022 International Conference on 3D Vision (3DV)*, pages 454–464. IEEE, 2022. 1

[8] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022. 1

[9] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 1

[10] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 1

[11] Ting-Yun Chang and Chi-Jen Lu. Tinygan: Distilling biggan for conditional image generation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[12] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015. 5

[13] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 1

[14] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1

[15] Yung-Yu Chuang, Dan B Goldman, Ke Colin Zheng, Brian Curless, David H Salesin, and Richard Szeliski. Animating pictures with stochastic motion textures. In *ACM SIGGRAPH 2005 Papers*, pages 853–860. 2005. 1

[16] Ken Dancyger. *The technique of film and video editing: history, theory, and practice*. Routledge, 2018. 1

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[18] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cinns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3742–3753, 2021. 1

[19] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023. 1

[20] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. Animating landscape: self-supervised learning of decoupled motion and appearance for single-image video synthesis. *arXiv preprint arXiv:1910.07192*, 2019. 1

[21] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pages 3233–3246. PMLR, 2020. 1

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[23] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7854–7863, 2018. 1

[24] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 1

[25] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[27] Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5810–5819, 2021. 1

[28] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022. 1

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1

[30] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023. 1

[31] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 1

[32] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 1

[33] Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2022. 1

[34] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35:22438–22450, 2022. 1

[35] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 1, 2, 3

[36] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H. Bermano, and Daniel Cohen-Or. Single motion diffusion, 2023. 1

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[39] 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*, pages 10684–10695, 2022. 1, 3

[40] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[41] Mikio Shinya and Alain Fournier. Stochastic motion—motion under the influence of wind. In *Computer graphics forum*, volume 11, pages 119–128. Wiley Online Library, 1992. 1

[42] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 1

[43] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[44] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 1, 2

[45] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022. 1

[46] Yale Song, David Demirdjian, and Randall Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 500–506. IEEE, 2011. 5

[47] Ryusuke Sugimoto, Mingming He, Jing Liao, and Pedro V Sander. Water simulation and rendering from a still photograph. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1

[48] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1

[49] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. *CoRR*, 2022. 1

[50] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 1

[51] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 835–851. Springer, 2016. 1

[52] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2443–2451, 2015. 1

[53] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 1

[54] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 1

[55] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 1

[56] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5908–5917, 2019. 1

[57] Tianfan Xue, Jiajun Wu, Katherine L Bouman, and William T Freeman. Visual dynamics: Stochastic future generation via layered cross convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2236–2250, 2018. 1

[58] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 1

[59] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023. 1

[60] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023. 1

[61] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 1

[62] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 1

[63] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 1

[64] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1