VideoHumanMIB: Unlocking Appearance Decoupling for Video Human Motion In-betweening

Haiwei Xue[1] , **Zhensong Zhang**[2] , **Minglei Li**[3] , **Zonghong Dai**[4] ,
Fei Yu[5] , **Fei Ma**[5*] , **Zhiyong Wu**[1*]
[1]Tsinghua University
[2]Huawei Noah's Ark Lab
[3]Beijing Ruxiaoyi Intelligent Technology Co., Ltd.
[4]Beijing JidianQiyuan InfoTech Co. Ltd
[5]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
WinniyGD@outlook.com, mafei@gml.ac.cn, zywu@sz.tsinghua.edu.cn

Abstract

We propose VideoHumanMIB, a novel framework for Video Human Motion In-betweening that enables seamless transitions between different motion video clips, facilitating the generation of longer and more natural digital human videos. While existing video frame interpolation methods work well for similar motions in adjacent frames, they often struggle with complex human movements, resulting in artifacts and unrealistic transitions. To address these challenges, we introduce a two-stage approach: First, we design an Appearance Reconstruction AutoEncoder to decouple appearance and motion information, extracting robust appearance-invariant features. Second, we develop an enhanced diffusion pretrained network that leverages both motion optical flow and human pose as guidance conditions, enabling the model to learn comprehensive latent distributions of possible motions. Rather than operating directly in pixel space, our model works in a learned latent space, allowing it to better capture the underlying motion dynamics. The framework is optimized with a dual-frame constraint loss and a motion flow loss to ensure temporal consistency and natural movement transitions. Extensive experiments demonstrate that our approach generates highly realistic transition sequences that significantly outperform existing methods, particularly in challenging scenarios with large motion variations. The proposed VideoHumanMIB establishes a new baseline for human motion synthesis and enables more natural and controllable digital human animation.

1 Introduction

Digital humans are widely utilized in the industry, such as health care, online counseling, and public transportation scenarios, which often require 24-hour ready service. Therefore, there is an increasing demand for digital human systems capable of generating continuous, long-term videos with realistic and natural human motion. However, directly generating longer and perfect human motion videos is very challenging, in terms of both quality and efficiency.

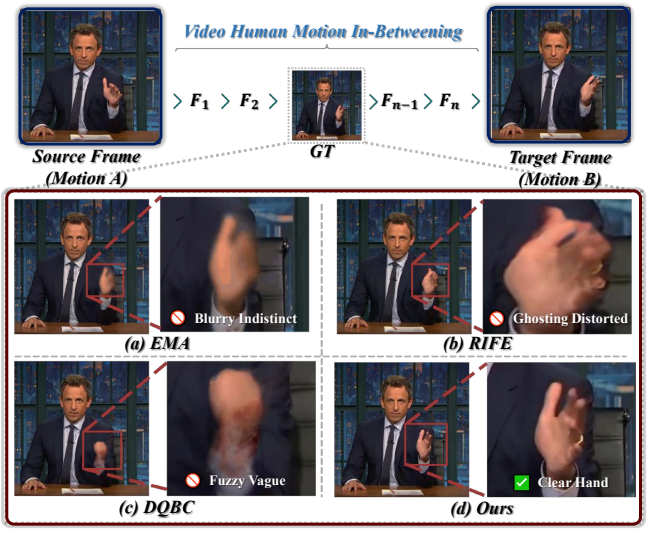

Figure 1: Illustration of the Video Human Motion In-betweening challenge. Given two reference frames, our task is to generate natural transition sequences $(F_1, F_2, ..., F_n)$ without additional motion guidance.

Current approaches [Guo *et al.*, 2023; Lin *et al.*, 2024; Xue *et al.*, 2024; Luo *et al.*, 2024] are only capable of directly generating videos in less than a few minutes with considerable time consumption. Consequently, the majority of digital humans in the industry rely on pre-recorded body motion libraries and manipulate meta motions to produce various digital human videos, which necessitates a smooth and natural transition between different meta motions. Since there has not been a dedicated method for video human motion in-betweening, recent solutions typically employ video frame interpolation (VFI) algorithms [Niklaus *et al.*, 2017; Zhang *et al.*, 2022; Bao *et al.*, 2019a]. Nonetheless, current

*Corresponding author.

VFI algorithm performs worse on large-scale human motion, particularly in handling complex hand movements involving intricate finger articulations and rapid gestures. The algorithm is deficient in the following three aspects:

- Due to *the low percentage of pixels representing hands in the overall video, capturing large-scale and fast-moving hand features becomes challenging*, leading to the intra-frame attention vaguely estimating motion vector [Zhang *et al.*, 2023], such as the blurred body result of EMA [Zhang *et al.*, 2023], shown in Figure 1 (a).
- Because of *the non-rigid hand movement*, incorrect rigid motion fields are generated by methods such as object shift [Huang *et al.*, 2022] or unilateral correlation, resulting in missing hand or erroneous conclusions, as illustrated in Figure 1 (b) and Figure 1 (c).
- Additionally, most VFI algorithms *suffer from a restricted number (1-2 frames) of generated frames*, which result in transition animations that are not natural enough and can easily be noticed by viewers.

To alleviate these issues, the Render In-between approach [Ho *et al.*, 2021] concentrates on human motion sequences. This method introduces a novel motion model that infers nonlinear skeleton movements between frames using a large-scale motion capture dataset. Subsequently, a neural rendering pipeline employs pose prediction with high frame rate to generate full-frame output while maintaining pose and background consistency. However, only similar adjacent frames are input into the neural rendering pipeline based on DAIN [Bao *et al.*, 2019b] interpolation, lacking a global comprehension of a wide range of movement. Consequently, this approach is fundamentally constrained to frame rate enhancement scenarios, where the temporal resolution is increased between adjacent frames with minimal motion variance.

In this work, we first introduce a novel task, named **Video Human Motion In-Betweening**, which is different from existing video frame interpolation that aims at increasing frame rate, and that the two frames to interpolate are similar, with no significant variations. Our goal is to generate a sequence of coherent transition frames depicting human motion between two arbitrary motions, accommodating potentially large pose variations. Towards this goal, we propose a two-stage training framework based on diffusion model to generate complete human bodies and realistic human motion transitions of arbitrary length between two given actions instead of simple linear interpolation. Hypothesizing that the information in the human video can be decoupled into human motion information and human appearance information, the proposed model can fit possible human movements as much as possible, including turning and hand movements.

In the first stage, we decouple pixel-wise motion and appearance information from the human videos to preserve physical details. An encoder-decoder reconstruction model similar to Variational AutoEncoder (VAE) is used to convert the human body frames into optical flow information and latent space features of appearance. The reconstruction model can also recover frames from optical flow information and human appearance features. Building upon this foundation, the second stage employs a diffusion-based generation process that takes the optical flow information from both source and target frames as input, guided by extracted pose features. This design enables the framework to explore and synthesize plausible motion trajectories between given poses, with a dual-frame constraint loss specifically formulated to ensure temporal consistency and motion naturalness. Through extensive analysis of motion distributions and model generalization capabilities across diverse datasets, our experiments demonstrate the framework's effectiveness in capturing and reproducing complex human movements, suggesting promising directions for large-scale digital human motion synthesis.

We conduct extensive experiments to validate our hypothesis on several datasets, including MHAD, NATOPS, and the additional HD video collection dataset. The empirical results and ablative studies show our method consistently achieves significant improvements over most VFI methods. The diffusion model generator built from potential motion distribution also significantly improves more natural and human-likeness transition, which makes our method learned from mass video data competitive or superior compared with most VFI models. In summary, our work mainly contributes in three aspects:

- We propose a novel video synthesis framework for the task of video human motion in-betweening. A diffusion framework method based on optical flow and pose features is designed for generating the transition of arbitrary two motions.
- We present a dual-frame constraint loss for training video human motion in-betweening task to achieve a natural and imperceptible human body motion transition.
- The extensive experiments demonstrate the effectiveness of our approach. Our experiments on examining the relationship between the distribution of actions and model generalization provide guidance for further study of large-scale pre-training video generation models.

2 Related Work

Motion In-Betweening The main goal of Human Motion In-Betweening is to generate transitional movements for digital humans from the current action to the next one. These transitions should be realistic and in accordance with human behavior. There has been significant progress in 3D digital humans [Gopinath *et al.*, 2022; Harvey *et al.*, 2020; Li *et al.*, 2022; Kim *et al.*, 2022; Qin *et al.*, 2022; Oreshkin *et al.*, 2023; Starke *et al.*, 2023; Ren *et al.*, 2023; Sridhar *et al.*, 2022; Kaufmann *et al.*, 2020; Tang *et al.*, 2022]. Making transitions by combining body movements, text, and other information. But the majority of these methods rely on joint information and lack texture, resulting in less realistic effects and necessitating extensive time for retargeting and rendering.

However, research is lacking in the context of 2D video simulations illustrating human motion. An associated project, Render In-between [Ho *et al.*, 2021], focuses on enhancing from lower frame rates to higher frame rates to depict intricate movements of digital human figures in videos. Nonetheless,

it is restricted to combining low-frame rate video into seamless high-frame rate video and does not have the capability to define transition animations between arbitrary two actions.

Video Frame Interpolation The insertion of frames in videos is a traditional and important task. Currently, there's been a lot of relevant work [Li *et al.*, 2023; Kong *et al.*, 2023; Reda *et al.*, 2022a; Yu *et al.*, 2023; Huang *et al.*, 2022; Zhang *et al.*, 2023; Lu *et al.*, 2022; Park *et al.*, 2023; Plack *et al.*, 2023] that has made some good progress. For instance, FILM [Reda *et al.*, 2022a] effectively predicts intermediate frames by utilizing a two-stage framework that incorporates a coarse-to-fine strategy. The DFI-WD [Kong *et al.*, 2023] technique utilizes a lightweight motion perception network to estimate intermediate optical flow, and subsequently, a wavelet synthesis network employs flow-aligned context features to predict multiscale wavelet coefficients with sparse convolution for effective target frame reconstruction. The reduction in computation can be as much as 40% while maintaining the same accuracy. Besides the methods mentioned above, Transformers have also been introduced into the video frame interpolation (VFI) task. With its powerful self-attention mechanism, the Transformer model can better capture inter-frame correlations, thereby improving the accuracy of interpolation. EMA [Zhang *et al.*, 2023] uses attention mechanisms between frames to extract both motion and appearance information, effectively capturing key features in the video and generating interpolated frames with high-quality details. VFIT [Lu *et al.*, 2022] utilizes Cross-Scale Window-based Attention to effectively deal with the receptive field and multiscale information. The introduction of the Transformer architecture provides a novel solution to the video frame interpolation problem. However, these methods upgrade generic videos from low frame rate to high frame rate without smoothing the motion of the virtual digital person from the current action to the next action, lacking a focused study of the distribution of human behavior.

Video Generation Diffusion Models Recently, VDM [Ho *et al.*, 2022b] has expanded diffusion models into the realm of videos, kick-starting exploration of diffusion models in video generation. Specifically, they've transformed a 2D UNet into a spatio-temporal factorized 3D network and further introduced joint image-video training and gradient-based video extension techniques. Make-A-Video [Singer *et al.*, 2022] and Imagen Video [Ho *et al.*, 2022a] employ a series of extensive diffusion models to synthesize large-scale videos based on given text prompts. However, previous video-based video generation methods have performed diffusion and denoising processes in pixel space, requiring significant computational resources. Inspired by LDM [Rombach *et al.*, 2022] and LFDM [Ni *et al.*, 2023], these methods learn from a latent feature space instead of learning directly from image pixels. In a similar vein, we employ an enhanced diffusion model for video human motion in-betweening tasks, where instead of pixel-level features, the diffusion model only needs to be fed with lower dimensional optical flow and pose feature information. More recent text-to-video models, such as OpenSora, OpenSoraPlan, and EasyAnimate [Lin *et al.*, 2024; Zheng *et al.*, 2024; Xu *et al.*, 2024], have demonstrated enhanced controllability and quality in short video generation. These approaches introduce novel architectures and training strategies, improving the fidelity and temporal consistency.

3 Method

Our approach aims to create human-like transitional animations between two motions of a digital human. For ease of presentation, we start this section by presenting the definition of the used symbols in our approach. We then introduce the Appearance Reconstruction AutoEncoder (ARAE), which translates human video frames to human appearance and motion optical flow features (3.1). Thereafter, we elaborate the improved diffusion model based on motion optical flow and pose information, involving a forward noisy process and a backward denoising process. Moreover, we introduce a dual-frame constraint loss to guide the model seamlessly connect with the source and target frames as much as possible (3.2). Finally, during the inference stage, the ARAE decoder will generate the complete transitional video from the reference frame and the motion optical flow sequence generated by the diffusion model (3.3).

Given a sequence of real consecutive action frames $F = \{F_1, F_2, \ldots, F_i\}$ $(1 \leq i \leq n)$, with n being the training frame number. For brief, we define the end frame of the current action as the source frame F_1, the start frame of the next action as the target frame F_n, the predicted results frames as $\hat{F} = \{\hat{F_1}, \hat{F_2}, \ldots, \hat{F_n}\}$, the pose information of the source frame and target frame as P_1 and P_n, the motion optical flow as $\mathbf{f} = \{f_1, f_2, \ldots, f_n\}$, where f_1 represents the motion optical flow information of the source frame, and f_n denotes the motion optical flow feature of the target frame. Moreover, we define the noise sequence of the diffusion model as X_t, guide condition as C, the denoise optical flow result as $\hat{f}$, and the timestamp as t.

3.1 Stage1: Appearance Reconstruction AutoEncoder

Generally, 2D human poses in videos can be extracted using pose estimators [Jiang *et al.*, 2023; Contributors, 2020] to obtain coarse-grained motion features. Nevertheless, the precision of 2D pose estimators is influenced by the handcrafted structural design. And pose estimation has lost the association information between the human body and the video background, which brings challenges to conditional guide video synthesis. According to Siarohin et al. [Siarohin *et al.*, 2019; Ni *et al.*, 2023], AutoEncoder is able to extract appearance information and motion optical flow information in the image. Our idea is to extract high-dimensional action features from the latent space with an AutoEncoder, based on the observation that the human video can be decomposed into human action information and human appearance information. To this end, we introduce a pre-trained Appearance Reconstruction AutoEncoder (ARAE) to establish the mapping from human video to human appearance and motion optical flow. The ARAE is depicted in the pink background portion of Figure 2, where the ARAE Encoder takes the source frame F_1 as input to extract the appearance feature A, the Optical Flow Predictor extracts the motion optical flow information f, and

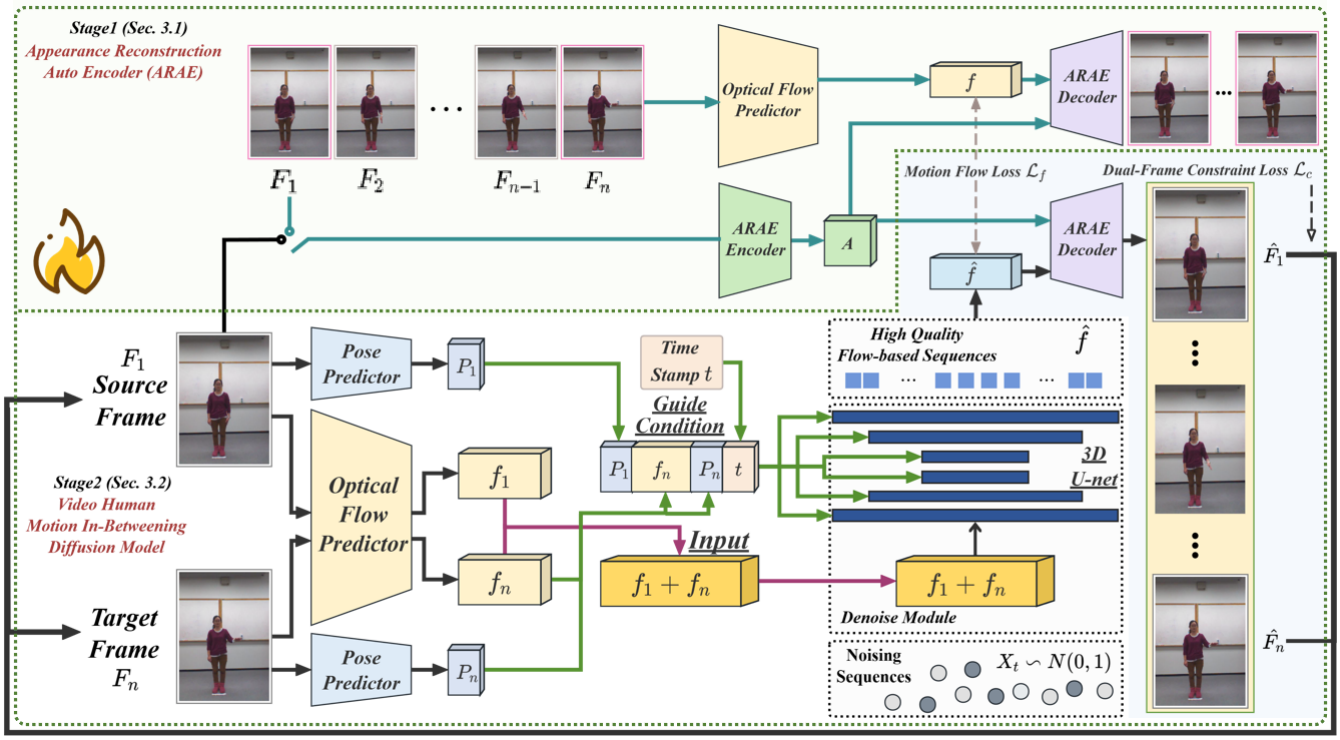

Figure 2: **Overview of Video Human Motion In-Betweening Baseline, VideoHumanMIB**. Firstly, we pre-train the Appearance Reconstruction AutoEncoder (ARAE) to decouple the human appearance information and the motion optical flow features from the video frames (the pink background portion). Second, we extract the latent optical flow and pose information from the source and target frames through optical flow and pose predictors separately. Then, using the latent features of the source and target frames as input, pose information and target frame latent optical flow as guide conditions, we train the flow-based diffusion model for generating the motion optical flow of transition frames (the green background portion). Finally, we use the decoder from the pre-trained ARAE to generate a series of transitional frames (the blue background portion). During training, the diffusion model is not directly trained on image pixels. Instead, it learns latent features in a low-dimensional space and is trained jointly with the dual-frame constraint loss and the motion flow loss.

the ARAE Decoder reconstructs the frame $\hat{F}$. The ARAE can be pre-trained end-to-end with self-supervised training.

3.2 Stage2: Video Human Motion In-Betweening Diffusion Model

Now that we have mapped the optical flow features from human video, we aim to synthesize optical flow features by gradually denoising from pure Gaussian noise, employing a DDPM [Ho *et al.*, 2020] in the latent space for generation, and be prepared for reconstructing transitional action clips. The enhanced diffusion model [Ho *et al.*, 2020] consists of two main stages: the forward noise addition stage and the reverse noise reduction stage.

Forward Process Due to computational power and time constraints, we propose to conduct diffusion and denoising the latent space of the video. We train a diffusion model to generate motion optical flow samples starting from a Gaussian noise $X_T \sim \mathbb{N}(X_T; 0, I)$ in T timestep. The diffusion process adds noise to X_0 according to variance schedule $\beta_1, \beta_2, ..., \beta_T$. Eventually, the sampling point will be similar to Gaussian noise.

$$q(X_1 : T|X_0) := \prod_{t=1}^{T} q(X_t|X_{t-1}) \quad (1)$$

$$q(X_t : T|X_{t-1}) := \mathbf{N}(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I) \quad (2)$$

Thus, to make X_t recover to X_0, the diffusion model needs to be trained to learn the denoising process. Typically, 3D-UNet is commonly used in image synthesis for the denoise module.

Reverse Process As mentioned before, when generating intermediate frames, we must consider the alteration in optical flow between the source frame and the target frame while ensuring the structural integrity of the human body during the synthesis process. Finally, the transition frames should be naturally and imperceptibly rendered from the source frame to the target frame. To keep the alteration in optical flow between the source frame and the target frame, we employ the pre-train optical flow predictor to extract the motion optical flow features f_1, f_n from the source frame F_1 and the target frame F_n. We combine the latent feature $f_1 + f_n$ as the input for denoise module. Moreover, we customize a video frame sampling method to ensure that the model can create transition animations bilaterally from one motion to the others.

The method allows the data to be randomly reversed, accelerated, or decelerated with some probability p of the video. To guarantee the human-like structural integrity during the synthesis process, we use the RTMPose [Jiang *et al.*, 2023;

Contributors, 2020] method to extract the human pose information P_1, P_n.

We leverage the target frame's motion optical flow representation f_n alongside pose information as conditioning signals to ensure human-like motion synthesis and limb integrity. The diffusion model then learns to predict a motion optical flow sequence $\hat{f}$ by denoising X_t conditioned on $C = [f_n, P_1, P_n, t]$, where t denotes the noise timestamp. This denoising process is implemented through residual connections across the Unet's downsampling, middle, and upsampling layers. To achieve seamless motion transitions, we enforce temporal consistency by ensuring that the generated sequence precisely aligns with both source and target frames at its endpoints. Specifically, we introduce a dual-frame constraint loss $\mathcal{L}_c$ that optimizes the synthesized motion sequence to maintain coherent connections with the source and target frames, effectively minimizing visual discontinuities in the transition process.

$$\mathcal{L}_c(X_0) = \sum_{i=0}^{B} \|\hat{F}_1 - F_1\|_2^2 + \|\hat{F}_n - F_n\|_2^2 \tag{3}$$

where B denotes batch size, $\hat{F}_1, \hat{F}_n$ is the first and last frames generated by the diffusion model. However, we do not use the consistency loss between the adjacent frames because it is proven to cause video ghosting by experiments and increase the complexity of training.

Hence, the training objective includes a motion flow loss $\mathcal{L}_f = \sum_{i=0}^{B} \|\hat{f} - f\|_2^2$ and a dual-frame constraint loss $\mathcal{L}_c$. The motion flow loss $\mathcal{L}_f$ is $L1$ loss between predicted and ground-truth motion optical flow. The dual-frame constraint loss $\mathcal{L}_c$ is used to eliminate the difference in frame reconstruction usually caused by the pixel-level reconstruction loss and further improve the realism of the generation. In summary, the overall training objective is

$$\mathcal{L}_{loss} = \mathcal{L}_f + \lambda \mathcal{L}_c \tag{4}$$

where λ is the weight of the dual-frame constraint loss.

3.3 Inference

As shown in Figure 2, in the inference stage, the ARAE Encoder encodes the source and target frames into latent motion flow features f_1 and f_n. Then $f_1 + f_n$ is the input for the denoising module of the 3D-Unet diffusion model. Simultaneously, the pre-trained RTMPose module [Jiang *et al.*, 2023; Contributors, 2020] is used to infer the Pose feature information P_1, P_n of F_1 and F_n. Thus, $C = [f_n, P_1, P_n, t]$ is used as a condition to obtain a high-quality flow-based frame sequence through inference. The resulting optical flow $\hat{f}$ and the latent feature A of the source frame by ARAE encoder, are fed into the ARAE decoder, which gradually renders the frame images in order, ultimately synthesizing a complete transition video. While we extend it to arbitrary motions transition prediction, we are the first to demonstrate its effectiveness in video human motion in-betweening.

4 Experiments

4.1 Implementation Details

Datasets We perform experiments on the following full-body human video datasets: MHAD [Chen *et al.*, 2015] human action dataset comprises 861 videos showcasing 27 actions executed by 8 individuals. This dataset encompasses various types of human actions, including sports activities, hand gestures, and everyday tasks. Each action is recorded multiple times, typically 3–4 times. From these recordings for each motion, one is selected as a test sample, while the remaining ones are utilized as training data. Since this dataset contains full-body motion and is well suited for the Human Motion In-Betweening task, our experimental analysis is mainly based on the MHAD dataset. This dataset provides nearly 20 minutes of training data, making it suitable for exploring the range of human body movements and modeling generalization capabilities in small datasets.

NATOPS [Song *et al.*, 2011] dataset contains 24 different actions recorded by 20 individuals. These actions are employed in the exchange of information with US Navy pilots and include some common signal processing, such as the swing of the arms, the stopping of the wings and various hand gestures. Due to some actions not being fully recorded on the body, only the visible data of the entire body is retained.

Following the experimental settings by [Ni *et al.*, 2023], we resized all the videos to 128×128 resolution. Since the dataset is relatively small, we enhanced the data with random selection, flipping, video acceleration, and slow-motion effects. Additionally, we also conduct experiments at a high resolution of 512×512 , successfully validating the efficacy of the model. Due to space limitations in the paper, the large resolution comparison results can be found in the Appendix.

Evaluation Metrics We evaluate the generated videos through both objective and subjective metrics. For objective evaluation, we adopt the Fréchet Video Distance following previous works [Zhou *et al.*, 2022; Ni *et al.*, 2023; Singer *et al.*, 2022]. The FVD metric [Unterthiner *et al.*, 2018], analogous to Fréchet Inception Distance [Heusel *et al.*, 2017], quantifies the spatial-temporal similarity between real and synthetic videos by calculating the Fréchet distance between their feature distributions. For subjective assessment, we employ two complementary metrics. The Human Recognition Rate measures the ability of volunteers to distinguish synthetic videos from real ones, with a lower HRR indicating better synthesis quality as the generated videos become more indistinguishable from real ones. The Mean Opinion Score evaluates the overall quality through a five-point scale ranging from terrible to excellent, where volunteers assess multiple aspects including video quality, human body integrity, and motion naturalness.

Experimental Setup Our model is implemented in PyTorch on a 40GB Nvidia RTX A100 GPU, and it takes around 2–3 days for training. Longer N frames settings require more training time. We train using the Adam optimizer with a learning rate of $1e^{-4}$, running for 800 epochs with a batch size of 64.

We invite 30 volunteers to participate in subjective scoring. The participants mainly include graduate students and

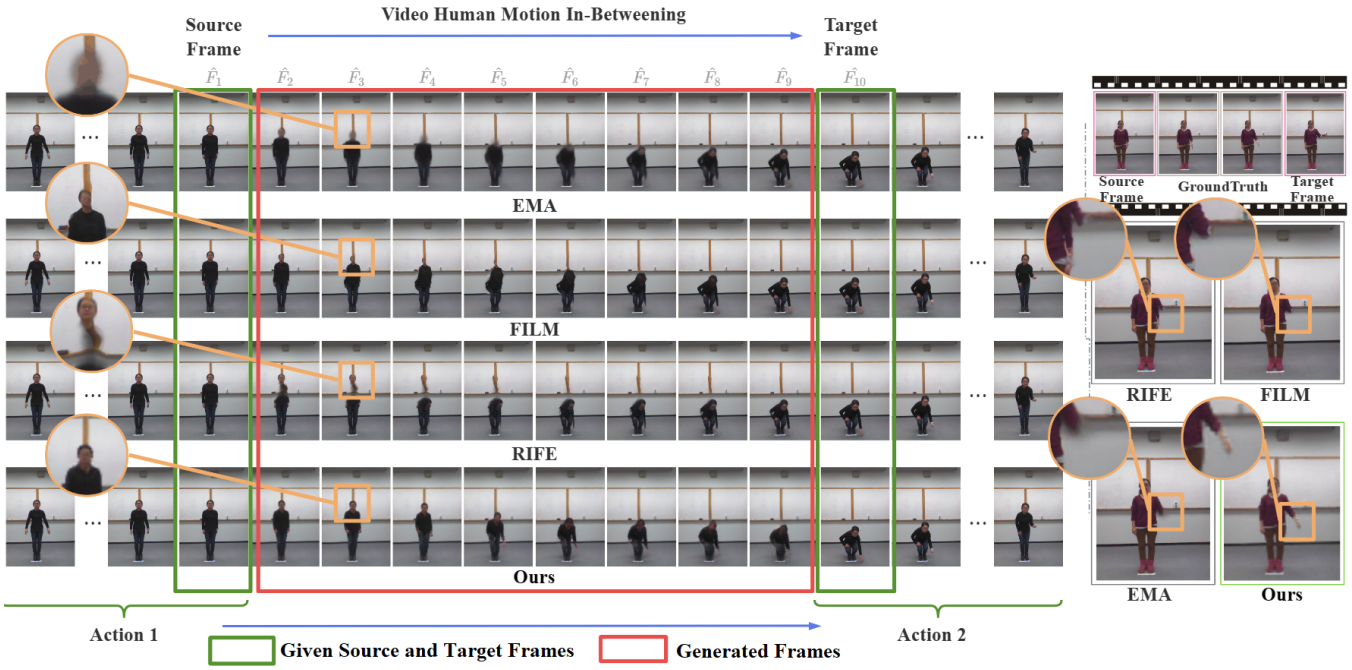

Figure 3: Visualization of our results for human motion in-betweening.

employed professionals, aged between 20 and 40. We take measures to keep the participants uninformed about the particular method variant linked to each video throughout the rating process.

4.2 Experimental Results

Figure 3 shows an illustration of the results of our network for human motion in-betweening frames, where $N = 10$ frames. These test data have not appeared in the training set. Experiments indicate that our proposed method can synthesize transition frames between source frames and target frames more naturally. In particular, we find that when some interaction occurs between hands and body, the model can also have a certain generalization ability and show good performance.

Comparisons As video frame interpolation algorithms currently serve as the primary solution for video human motion in-betweening, we benchmark our approach against several widely adopted VFI methods: EMA [Zhang *et al.*, 2022], FILM [Reda *et al.*, 2022b], and RIFE [Huang *et al.*, 2022]. These methods have demonstrated their effectiveness in industrial applications and provide strong baselines for evaluation. Nonetheless, interpolation algorithms, which aim to convert low-frame rate videos to high-frame rate videos, suffer from a drawback in achieving large-scale digital human motion.

Results Figure 3 shows the synthesis of test data under large-scale motion for the different VFI-based methods and our proposed method. The green box is a reference frame before inputting into the models, which is required to generate a transitional animation for the intermediate orange box. Our method generates transition animations naturally while maintaining the structural integrity of the body. In contrast, the VFI-based methods all indicate varying levels of body defects and perform poorly on tasks such as Human Motion In-Betweening. For example, FILM [Reda *et al.*, 2022b] and RIFE [Huang *et al.*, 2022] show extremely severe distortions and disappearances in the head and upper body as the body motion in the source and target frames transitions from standing to squatting. Although the EMA [Zhang *et al.*, 2022] does not exhibit very severe distortions, it is a noticeable blur in the head. We have observed in numerous test data that the utilization of frame interpolation algorithms leads to the absence of body parts during the generation of transition animations.

We evaluate the VFI-based methods and our method on MHAD [Chen *et al.*, 2015] and NATOPS [Song *et al.*, 2011] testing dataset as shown in Table 1. We observe that: (1) The transition animations generated by EMA and RIFE do not quite deceive the human eye well, with average performance in terms of FVD and MOS. This matches what we have seen in the visualized data. (2) The coarse-to-fine strategy used in FILM makes it excel in frame interpolation for large-scale movements. Especially on the NATOPS dataset, it presents better FVD values. (3) Our proposed method outperforms EMA, FILM, and RIFE in subjective evaluations for motion in-betweening, thus proving the generalization capability of the diffusion model based on latent optical flow in action transition generation. This positively contributes to advancing the practical applications of more natural and humanlike 2D digital humans.

Ablation Study We study the effects of each component in Table 2. We showcase that frame constraint loss $\mathcal{L}_c$, pose features $P_1 + P_n$ and optical flow features $f_1 + f_n$ as inputs to the denoising network are all effective to improve motion in-betweening results.

	MHAD			NATOPS		
	FVD (↓)	HRR (↓)	MOS (↑)	FVD (↓)	HRR (↓)	MOS (↑)
ASC [Niklaus *et al.*, 2017]	398.45	69.857% ± 6.623	3.652 ± 0.1262	175.82	74.762% ± 6.284	3.642 ± 0.1252
DepthVFI [Bao *et al.*, 2019a]	342.67	68.571% ± 6.714	3.681 ± 0.1285	169.45	72.381% ± 6.425	3.667 ± 0.1264
Phasenet [Meyer *et al.*, 2018]	289.34	67.142% ± 6.825	3.695 ± 0.1296	158.92	70.476% ± 6.587	3.702 ± 0.1268
ACVFI [Niklaus *et al.*, 2017]	271.85	66.190% ± 6.856	3.704 ± 0.1301	152.34	69.047% ± 6.684	3.721 ± 0.1271
EMA [Zhang *et al.*, 2022]	454.99	71.904% ± 6.507	3.619 ± 0.1245	189.74	75.238% ± 6.249	3.638 ± 0.1241
FILM [Reda *et al.*, 2022b]	257.12	65.238% ± 6.894	3.709 ± 0.1307	**145.78**	68.095% ± 6.748	3.738 ± 0.1273
RIFE [Huang *et al.*, 2022]	406.22	70.476% ± 6.604	3.647 ± 0.1255	177.41	76.190% ± 6.166	3.623 ± 0.1244
Ours	**172.34**	**59.523% ± 7.106**	**3.819 ± 0.1144**	147.68	**63.809% ± 6.957**	**3.761 ± 0.1148**

Table 1: Evaluation results of different variant model. Fréchet Video Distance (FVD) (↓), Human Recognition Rate (HRR) (↓) means the smaller the value, the better the performance. HHR and MOS are shown with 95% confidence intervals.

4.3 Effects of Pose Distribution in Motion In-Betweening

To explore the relationship between the pose distribution and the generality of the model, we randomly select 10 actions from 28 actions in the MHAD dataset for visual analysis by t-SNE shown in Figure 4. The color of each slot box on the right of the figure corresponds to the clustering color point.

We discover that: (1) A sequence of actions recorded at once is considered as one sample, and it can be observed that each action category generally presents a certain aggregated distribution. (2) For certain similar actions, the corresponding feature distribution in the latent space distance is also relatively close. For example, the three actions at the bottom of the Figure 4 mainly involve left-handed variations, so their features are close in latent space distance. (3) Combining the observation from Figure 3, it can be noticed that the hand movements in the transition animation are aligned with the prior knowledge of hand movements in the feature space. This proves that the model can successfully model transition animations between two actions even if the transition action does not appear in the training data, if the distribution of actions has been learned.

To further investigate the effect of actions in the training data on the model performance, we evaluate the effect of different numbers of action categories and varying amounts of training data during training. Table 3 demonstrates that

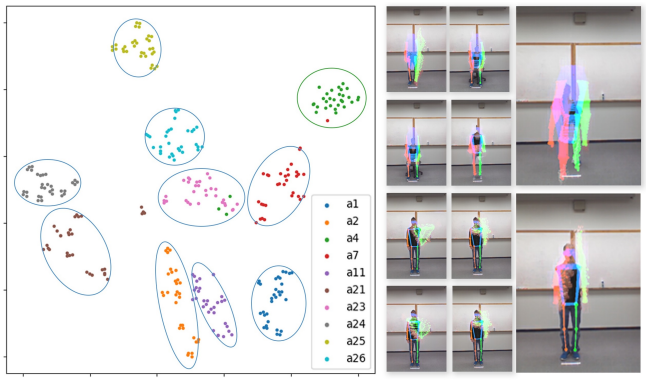

Figure 4: Example action distribution visualization by t-SNE. The colors represent movements in body parts.

	MHAD FVD(↓)	NATOPS FVD(↓)
Ours	172.34	147.68
- $P_1 + P_n$	187.44	150.74
- $f_1 + f_n$	218.12	184.11
- $\mathcal{L}_c$	196.42	176.47

Table 2: Ablation study. '-' is shorthand for 'without'.

Categories	Proportion of Training Usage	FVD (↓)
8 actions	30%	387.49
8 actions	60%	304.14
8 actions	100%	317.11
15 actions	100%	237.65
28 actions	100%	172.34

Table 3: Explore the effects of using different proportions of training data and the categories of actions involved in training.

(1) the inclusion of a more diverse set of action categories in training significantly improves model performance. When the number of action categories in the training set increases from 8 to 15, the FVD decreases significantly. (2) The utilization of a larger dataset for training yields evident advantages up to a certain threshold. Beyond that threshold, however, the gain diminishes in magnitude. In our experiments, increasing the training data from 30% to 60% results in a certain degree of FVD improvement. Still, when the training data goes from 60% to 100%, the FVD results remain similar. We are confident that after expanding the variety of actions, the generated transitional actions will be more human-like and natural. We provided the visualization demo.

5 Conclusion and Future Work

We introduce a novel video synthesis framework for the video human motion in-betweening task and investigate the relationship between the pose distribution and the generality of the model. Our experiments demonstrate the effectiveness of our approach, which can generate the humanlike and natural motion in-betweening with large pose variations. In the future, we will investigate a pre-trained large model for human motion video generation, incorporating human data from online videos to augment the training set.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62076144) and Shenzhen Science and Technology Program (JCYJ20220818101014030).

References

[Bao *et al.*, 2019a] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3703–3712, 2019.

[Bao *et al.*, 2019b] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3703–3712, 2019.

[Chen *et al.*, 2015] 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.

[Contributors, 2020] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[Gopinath *et al.*, 2022] Deepak Gopinath, Hanbyul Joo, and Jungdam Won. Motion in-betweening for physically simulated characters. In *SIGGRAPH Asia 2022 Posters*, pages 1–2. ., 2022.

[Guo *et al.*, 2023] 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.

[Harvey *et al.*, 2020] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[Heusel *et al.*, 2017] 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. *Advances in neural information processing systems*, 30, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Ho *et al.*, 2021] Hsuan-I Ho, Xu Chen, Jie Song, and Otmar Hilliges. Render in-between: Motion guided video synthesis for action interpolation. *arXiv preprint arXiv:2111.01029*, 2021.

[Ho *et al.*, 2022a] 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.

[Ho *et al.*, 2022b] Jonathan Ho, Tim Salimans, Alexey Gritsenko, and OTHERS. Video diffusion models. ., 2022.

[Huang *et al.*, 2022] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[Jiang *et al.*, 2023] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose, 2023.

[Kaufmann *et al.*, 2020] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*, pages 918–927. IEEE, 2020.

[Kim *et al.*, 2022] Jihoon Kim, Taehyun Byun, Seungyoun Shin, Jungdam Won, and Sungjoon Choi. Conditional motion in-betweening. *Pattern Recognition*, 132:108894, 2022.

[Kong *et al.*, 2023] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Ying Tai, Chengjie Wang, and Jie Yang. Dynamic frame interpolation in wavelet domain. *IEEE Transactions on Image Processing*, 2023.

[Li *et al.*, 2022] Yunhao Li, Zhenbo Yu, Yucheng Zhu, Bingbing Ni, Guangtao Zhai, and Wei Shen. Skeleton2humanoid: Animating simulated characters for physically-plausible motion in-betweening. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1493–1502, 2022.

[Li *et al.*, 2023] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *CVPR*, pages 9801–9810, 2023.

[Lin *et al.*, 2024] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[Lu *et al.*, 2022] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *CVPR*, pages 3532–3542, 2022.

[Luo *et al.*, 2024] Xiangyang Luo, Xin Zhang, Yifan Xie, Xinyi Tong, Weijiang Yu, Heng Chang, Fei Ma, and Fei Richard Yu. Codeswap: Symmetrically face swapping based on prior codebook. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6910–6919, 2024.

[Meyer *et al.*, 2018] Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus Gross, and Christopher Schroers. Phasenet for video frame interpolation. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, pages 498–507, 2018.

[Ni *et al.*, 2023] 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.

[Niklaus *et al.*, 2017] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE international conference on computer vision*, pages 261–270, 2017.

[Oreshkin *et al.*, 2023] Boris N Oreshkin, Antonios Valkanas, Félix G Harvey, Louis-Simon Ménard, Florent Bocquelet, and Mark J Coates. Motion in-betweening via deep delta-interpolator. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[Park *et al.*, 2023] Junheum Park, Jintae Kim, and Chang-Su Kim. Biformer: Learning bilateral motion estimation via bilateral transformer for 4k video frame interpolation. In *CVPR*, pages 1568–1577, 2023.

[Plack *et al.*, 2023] Markus Plack, Karlis Martins Briedis, Abdelaziz Djelouah, Matthias B Hullin, Markus Gross, and Christopher Schroers. Frame interpolation transformer and uncertainty guidance. In *CVPR*, pages 9811–9821, 2023.

[Qin *et al.*, 2022] Jia Qin, Youyi Zheng, and Kun Zhou. Motion in-betweening via two-stage transformers. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

[Reda *et al.*, 2022a] Fitsum Reda, Janne Kontkanen, Eric Tabellion, and OTHERS. Film: Frame interpolation for large motion. ., 2022.

[Reda *et al.*, 2022b] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pages 250–266. Springer, 2022.

[Ren *et al.*, 2023] Tianxiang Ren, Jubo Yu, Shihui Guo, Ying Ma, Yutao Ouyang, Zijiao Zeng, Yazhan Zhang, and Yipeng Qin. Diverse motion in-betweening with dual posture stitching. *arXiv preprint arXiv:2303.14457*, 2023.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[Siarohin *et al.*, 2019] 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.

[Singer *et al.*, 2022] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[Song *et al.*, 2011] 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.

[Sridhar *et al.*, 2022] Pavithra Sridhar, Madhav Aggarwal, R Leela Velusamy, et al. Transformer based motion inbetweening. In *Proceedings of the Asian Conference on Computer Vision*, pages 289–302, 2022.

[Starke *et al.*, 2023] Paul Starke, Sebastian Starke, Taku Komura, and Frank Steinicke. Motion in-betweening with phase manifolds. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–17, 2023.

[Tang *et al.*, 2022] Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. Real-time controllable motion transition for characters. *ACM Transactions on Graphics (TOG)*, 41(4):1–10, 2022.

[Unterthiner *et al.*, 2018] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[Xu *et al.*, 2024] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024.

[Xue *et al.*, 2024] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, et al. Human motion video generation: A survey. *Authorea Preprints*, 2024.

[Yu *et al.*, 2023] Zhiyang Yu, Yu Zhang, Dongqing Zou, Xijun Chen, Jimmy S Ren, and Shunqing Ren. Range-nullspace video frame interpolation with focalized motion estimation. In *CVPR*, pages 22159–22168, 2023.

[Zhang *et al.*, 2022] Guozhen Zhang, Yuhan Zhu, Haonan Wang, and OTHERS. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. ., pages 5682–5692, 2022.

[Zhang *et al.*, 2023] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *CVPR*, pages 5682–5692, 2023.

[Zheng *et al.*, 2024] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *Github*, March 2024.

[Zhou *et al.*, 2022] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv e-prints*, pages arXiv–2211, 2022.