Video Motion Graphs

Haiyang Liu[1,2*] Zhan Xu[2] Fa-Ting Hong[2] Hsin-Ping Huang[2] Yi Zhou[2] Yang Zhou[2]
[1]The University of Tokyo [2]Adobe Research

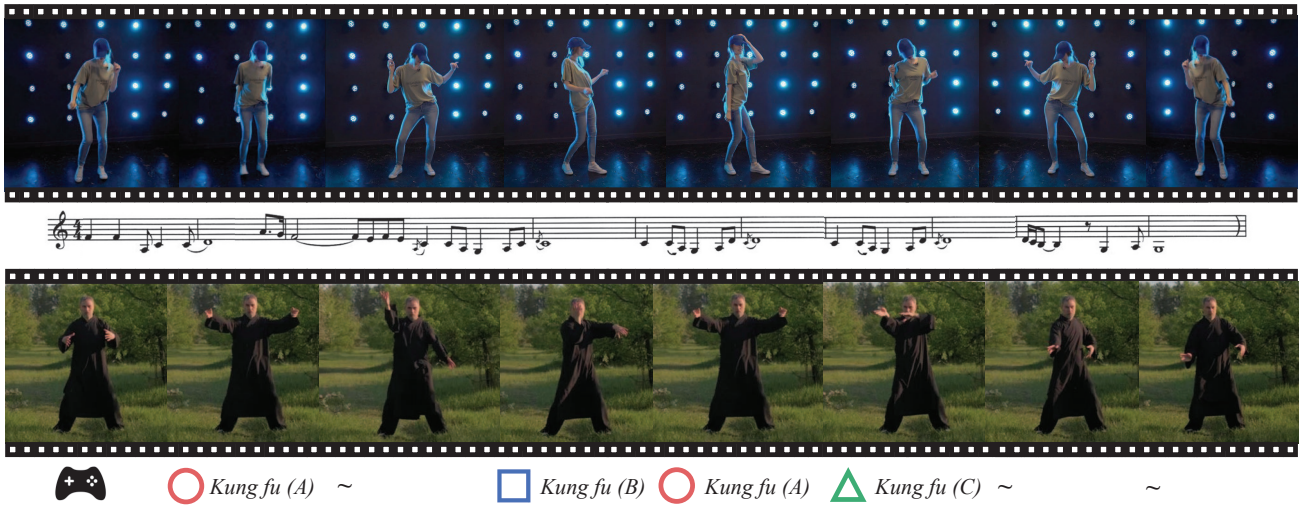

Figure 1. **Video Motion Graphs** is a system to generate human motion videos from a reference video and conditional signals such as music, action tags and sparse keyframes. The video is generated by first retrieving matched video clips from reference video and then generating interpolation frames between clips to smooth the transition boundaries.

Abstract

We present Video Motion Graphs, a system designed to generate realistic human motion videos. Using a reference video and conditional signals such as music or motion tags, the system synthesizes new videos by first retrieving video clips with gestures matching the conditions and then generating interpolation frames to seamlessly connect clip boundaries. The core of our approach is HMInterp, a robust Video Frame Interpolation (VFI) model that enables seamless interpolation of discontinuous frames, even for complex motion scenarios like dancing. HMInterp i) employs a dual-branch interpolation approach, combining a Motion Diffusion Model for human skeleton motion interpolation with a diffusion-based video frame interpolation model for final frame generation. ii) adopts condition progressive training to effectively leverage identity strong and weak conditions, such as images and pose. These designs ensure both high video texture quality and accurate motion trajectory. Results show that our Video Motion Graphs outperforms existing generative- and retrieval-based methods for multi-modal conditioned human motion video generation. Project page can be found here.

1. Introduction

Human motion videos play a vital role across numerous industries, including entertainment, virtual reality, and interactive media. However, capturing high-quality, realistic motion videos can be labour-intensive and costly. Recent advancements in video generation offer solutions in generating human motion videos based on inputs like skeletal animation, action labels, and speech audios, making production more efficient and customizable.

Human motion video generation has two primary approaches: generative- and retrieval-based methods. Generative-based models [7, 16, 21, 48] synthesize all frame pixels from conditional inputs, offering flexibility in generating diverse motions. However, they often produce artifacts for complex contents, such as human distorted limbs, fingers, etc [1, 55]. Retrieval-based models utilize key frames from given reference videos, and generate interpolation frames to ensure smooth transitions [30, 61]. While they require reference material, they typically deliver

*Work done during Haiyang, Fa-Ting, and Hsin-Ping's internship at Adobe Research from 24/06 to 24/09.

higher video quality and maintain the actor's identity. Motivated by the video quality, we focus on the retrieval-based method for real-world applications. Existing retrieval-based methods, such as, Gesture Video Graph (GVR) [61] and TANGO [30], are specifically designed for co-speech gesture video generation and are not applicable to general human motion animation, *e.g.*, dancing, kung fu, etc. In particular, these methods retrieve video frames based on input audio using a motion graph framework [24], and then apply a Video Frame Interpolation (VFI) model to generate the interpolated frames between the retrieved frames. Extending these methods to a comprehensive system to accommodate diverse conditions beyond speech audio presents one main challenges: Their VFI model leverages the linear blended motion guidance and thus limits its ability to handle complex, dynamic motions like human dancing. In this paper, we proposed an improved diffusion-based VFI model, *HMInterp*, to complete the essential functionality of the general video motion graphs system.

The motivation of *HMInterp* is to seamlessly connect retrieved frames. Unlike the original VFI module in GVR, which uses a simple linearly interpolated motion as guidance in a flow-warp-based VFI model, HMInterp addresses the limitations of linear interpolation for complex, dynamic motions. For instance, approximately 78% of mild speech gestures can be reasonably approximated through linear blending, while only 17% of dance gestures which complex motion can, highlighting the need for a more advanced approach for human dynamic motion sequences (see supplemental for details).

Specifically, *HMInterp* starts from UNet-based pretrained text-to-video model, AnimateDiff[12], and further consists of a diffusion-based VFI model with motion guidance from a dedicated Motion Diffusion Model (MDM). HMInterp generates smooth interpolated frames while preserving accurate human structure and motion, guided by the MDM module. Based on [44], the MDM is trained within a validated human skeleton space, ensuring both structural integrity and continuous motion of body parts. Moreover, we observed straightforward multi-condition (image and pose) joint training yields identity inconsistent results. To solve this, we introduce condition progressive training. It adopts different training orders and iterations for strong and weak conditions such as pose and image, respectively. Finally, to enhance the performance, we incorporate the ReferenceNet from [16, 53] and reference decoder from [51]. As a consequence, these components enable *HMInterp* to produce realistic interpolated frames.

The quality of *HMInterp* enable us to implement the *Video Motion Graphs* for general human video representation (in Figure 1). From an engineering perspective, we complete the audio-only GVR baseline [61] into a more comprehensive system. In particular, we define a standard four-stage pipeline for graph initialization, searching, frame interpolation and background reorganization. Besides, we adopt task-specific rule-based matching for searching, and introduce keyframe-based editing. This allows the system to retrieve relevant video clips to align with conditions like music beats. It also enables users to replace or customize specific frames by manually editing the output frames.

Overall, our contributions can be concluded as follows:

- We propose *Video Motion Graphs*, a comprehensive retrieval + generation system for general human motion videos. This system enables both sequence and keyframe retrieval, supporting applications such as real-time video generation and keyframe editing. It achieves state-of-the-art performance in generating high-quality, customizable human motion videos.
- We propose *HMInterp*, a high-quality motion-aware, video frame interpolation module with the proposal that i) utilizes Motion Diffusion Model for generative and controllable human structure guidance. ii) adopts condition progressive training to effectively leverage identity strong and weak conditions.

2. Related Work

Generative Human Motion Video Synthesis. Generative methods generate all frames directly from networks with conditions. To the best of our knowledge, there is no unified model to accept flexible conditions and output general human motions. There are task-specific models, such as generating motion video from text [21], music [48], speech [7, 13, 26, 28, 29, 31], and pose [3, 4, 16, 18, 59, 62]. These methods are flexible to generate novel poses, but the video quality is sub-optimal. The video quality depends on the video generation backbone [42]. Even though the backbone has moved from UNet [2, 5, 6, 42, 57] to DiT [10, 15, 55, 58] the state-of-the-art video generation model such as Open-Source Stable Video Diffusion [1] still has broken hands and faces. This makes a video full of generated frames appear unnatural. Different from them, we only apply the generative model to a few frames and retrieve from reference video for the task to ensure higher video quality.

Retrieval Human Motion Video Synthesis. GVR [61] and TANGO [30] are previous works for Retrieval Motion Generation. They generate motion in three steps: (i) creating a motion graph based on 3D motion and 2D image domain distances, (ii) retrieval of the optimal path within this graph for the motion on the path is best matched to target speech, (iii) blending the discontinues frames by an interpolation network based on linear blended motion guidance. However, their system is only designed for co-speech

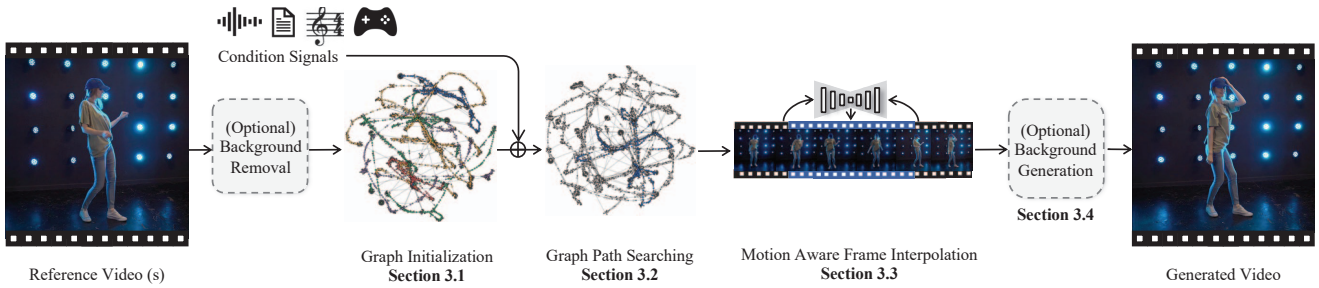

Figure 2. **System Pipeline of Video Motion Graphs.** Given input reference video(s) and a condition signal like music, Video Motion Graphs generates a new video in four steps: (i) representing the video as a directed graph, where nodes are RGB frames and edges indicate valid playback transitions, (ii) retrieving a frame playback path in the graph to match conditions based on task-specific rules, such as beat alignment, (iii) blending discontinuous frame transitions with a Motion-Aware Frame Interpolation Model, and (iv) optionally changing the video background through background removal and generation models.

talkshow videos with stable and static backgrounds. We imporve the VFI to make it work on general human motions.

Video Frame Interpolation. VFI is a classical low-level problem aimed at generating intermediate frames from beginning and end frames. The problem shifts from reconstruction to generative when motion dynamics increase. The fully end2end methods [8, 9, 17, 20, 23, 25, 32, 34, 35, 37, 38, 45, 50, 54], directly estimating middle frames using an end-to-end model trained on VFI tasks like VFI Diffusion [19] or leverage pre-trained Text2Video models [11], such as DCInterp [52], which leverage the spatial and temporal patterns from VideoCrafter [5]. On the other hand, the solutions that rely on flow-warping or intermediate guidance show more promising results, FILM [41] and VFIFormer [33] are repetitive works with CNN and transformer backbones. However, as content differences increase, the estimated flow is not accurate enough and often leads to hands disappearing. To solve this, Pose-Aware Neural Blending [61] and ACInterp [30] introduce explicit linear blended motion guidance with CNN and Diffusion. However the linear blending could not handle complex motions such as dance. Unlike these, we leverage both VFI models and generative motion guidance to maintain video texture and motion correctness.

3. Video Motion Graphs

The core idea behind the Video Motion Graphs is to represent an input video as a motion graph structure and synthesize output videos through motion graph search [30, 61]. Specifically, given a reference video and target conditional signals (*e.g.*, speech audio, music, motion tags), the system generates video in four steps: (1) representing the reference video as a graph, with nodes as video frames and edges as valid transitions (Section 3.1); (2) framing conditional video generation as a graph path-searching problem, where path costs are guided by conditional signals (Section 3.2); (3) employing a Video Frame Interpolation (VFI) model to smooth discontinuous boundaries (Section 3.3); and (4) joining all searched and interpolated frames to construct the final output video. Optionally, background removal and generation models can be applied to handle reference videos with highly dynamic backgrounds (Section 3.4).

3.1. Graph Initialization

We represent the reference video as a graph $\mathcal{G} = \{\mathbf{V}, \mathbf{E}\}$, where each vertex $v \in \mathbf{V}$ denotes one video frame, and $e \in \mathbf{E}$ indicates if two frames (vertices) can be concatenated temporally with smooth transition. All original consecutive frames in the reference video are naturally connected by an edge. To measure the smoothness of the transition for other pairs of frames, we calculate the difference in human poses captured in frames. We followed GVR [61] and TANGO [30] to compute the difference for 3D pose as $d^l_{i,j} = ||J^l_i - J^l_j||_2$. Slightly different from TANGO and GVR, we replace 2D difference with $d^g_{i,j} = ||J^g_i - J^g_j||_2$, where J^i are 3D poses from any 3D pose detector. We then follow [61] to pick a hyperparameter threshold τ based on the average nearest-neighbour distance calculated on the entire graph. To this end, an edge $e = (v_i, v_j)$ exists if $d^l_{i,j} + d^g_{i,j}$ is smaller than the threshold τ. Finally, we follow the graph pruning in TANGO [30] to remove dead-end nodes and obtain a final valid graph (an illustration can be found in Figure 2).

3.2. Path Searching

We formalize video generation from $\mathcal{G}$ as a graph search problem. Given a target signal, such as speech audio, music, or motion tags, we search for a valid graph path that minimizes the total path cost. This cost function combines (1) the intrinsic edge distance defined in Section 3.1 with (2) task-specific scores based on input conditions. In this paper,

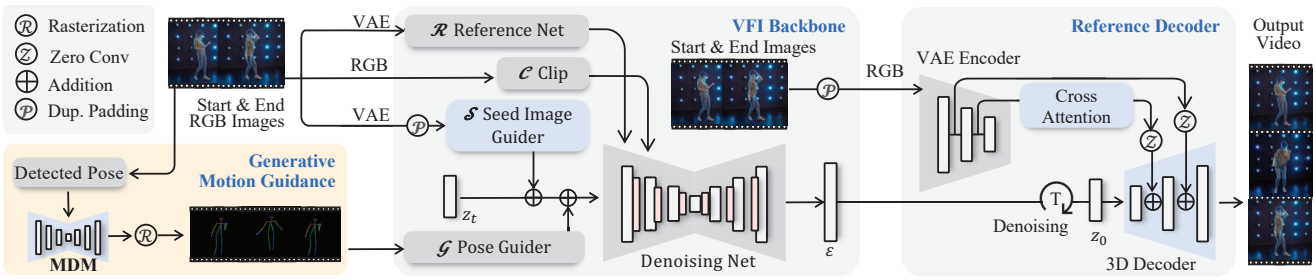

Figure 3. **HMInterp.** During inference, HMInterp takes start and end RGB images to generate interpolated video frames. It consists of three modules: (i) Generative Motion Guidance, which leverages our proposed MDM module to generate interpolated 2D poses and render them as RGB videos, (ii) the proposed VFI module that uses input images and poses to generate interpolation frames in latent space, and (iii) a Reference Decoder that combines denoised latent representations with input images to decode the final output video in pixel space through low-level feature injection.

we define four video generation tasks suited to the proposed Video Motion Graphs: (1) human action generation via motion tags, (2) text-to-motion generation, (3) music-driven dance video generation, and (4) talking-avatar generation with gesture animation from speech audio. We have four dedicated heuristic-based searching solution for each of the four tasks (please check supp. for details.) It will define the per-frame path cost functions perform the sequence-level path searching with dynamic programming (DP) for offline processing or with efficient Beam Search [47] for real-time applications. After searching, we can get a path with all necessary frames aligning well with the conditional signal, but it has discontinuous frames.

3.3. HMInterp for Video Frame Interpolation

In this section, we introduce *HMInterp*, the core enabling technology that allows Video Motion Graphs to seamlessly connect discontinuous frames, ensuring smooth motion transitions. As shown in Figure 3, HMInterp takes start and end frames as input and generates 12 interpolated frames (equivalent to 0.5 seconds at 24 FPS). Unlike existing frame interpolation methods [19, 51, 52], we propose a dual-branch interpolation approach that leverages both a Motion Diffusion Model (MDM) [44] for human skeletal motion interpolation and a diffusion-based Video Frame Interpolation (VFI) for inbetweening frame generation. By fusing motion guidance from the MDM into the VFI with condition progressive training, HMInterp enhances both motion accuracy and high video texture quality. We first introduce the entire VFI Backbone for ensuring video textural quality, and then show how we generate and fuse the motion condition to allow correct motion trajectory.

Video Frame Interpolation (VFI) Backbone. We build our model on top of an existing UNet based text-to-video model [12], and adapt it for VFI task with additional conditioning signals, *i.e.*, start and end frames. To enhance the performance, we also adopt the ReferenceNet from pose-

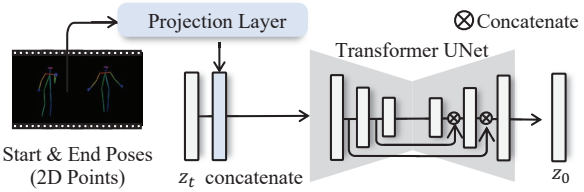

Figure 4. **Details of Motion Diffusion Model.** During training, MDM learns to reconstruct target interpolated 2D joint positions. It fuses the start and end poses through feature concatenation before the Denoising Transformer UNet. The vanilla Transformer in MDM is modified with skip connections and feature concatenation to enhance the details of the generated motion trajectory.

to-video models[16, 53], replace the CLIP text encoder to CLIP image encoder[39], and leverage the Reference Decoder from ToonCrafter[51]. Specifically, the start and end frames are encoded through a pre-trained VAE, duplicated padded to fill intermediate frames, and then fed as Seed Image Guiders in the VFI input (see Figure 3). In addition to CLIP features, we further inject two types of low-level features to improve frame quality. Inspired from [16, 53], we use ReferenceNet to inject hierarchical latent feature guidance from the start and end frames, significantly enhancing identity and appearance fidelity. Additionally, for tasks involving full-body video generation, human details like faces may degrade at lower resolutions (*e.g.*, 256×256) due to VAE decoder limitations. To mitigate this, we implement an improved Reference Decoder for realistic videos, which is originally proposed by ToonCrafter [51] for animation videos. It applies skip-connections from the VAE encoder low-level latents to boost detail retention. Unlike [51], our reference decoder is initialized from the temporal decoder in Stable Video Diffusion [1]. In addition, we propose to use duplicated reference frames as input, which is simple but brings clear performance gain without extra additional inference costs. We compare the effectiveness of our re-implementation against [51] in Figure 8. The above design

Table 1. **Comparison of Human Motion Video Generation.**

	PSNR ↑	LSIPS ↓	MOVIE ↓	FVD ↓
AnimateAnyone [16]	35.55	0.044	54.68	1.369
MagicPose [4]	35.64	0.048	51.97	1.277
UniAnimate [49]	36.75	0.042	49.89	1.090
MimicMotion [60]	36.30	0.047	46.84	1.078
Ours ($f = 32$)	**42.91**	**0.009**	**37.31**	**0.180**
Ours ($f = 64$)	42.75	0.010	37.53	0.213
Ours ($f = 216$)	39.75	0.029	39.89	0.799

allows the backbone to generate videos with high textural quality.

Motion Diffusion Model (MDM). We then introduce the MDM that generates interpolated 2D poses between start and end frames. These interpolated 2D poses serve as the final conditioning input for the VFI, enabling it to produce more accurate human poses with valid structures. The existing MDM [44] often loses motion details when using a vanilla 8-layer transformer. To address this, we implemented a UNet-like transformer architecture (see Figure 4), which fuses features from shallow to deeper layers using concatenation and skip-connection. This design allows our MDM to generate more accurate, non-linear motion interpolation trajectories (see ablation in Table 5). During training, we guide the VDM with ground truth 2D poses, while during inference, we condition it on generated 2D poses from the MDM.

Condition Progressive training. We first train the Reference Decoder and MDM separately and then freeze the VAE, MDM and CLIP. The remaining trainable parameters are ReferenceNet, Seed Image Guider, Pose Guider, and Denoising Net. For Denoising Net we load the pretrained weights from AnimateDiff. The straightforward implementation is training the image conditions (ReferenceNet, Seed Image Guider) together with pose conditions (Pose Guider). However, this yield facial appearance inconsistent between the generated and groundtruth frames (See Figure 9 for details). To solve this, we consider training different conditions progressively: We first conduct Seed Pre-Training, train the VFI module with image condition only for long iterations (100k), to ensure the interpolated frames follow the reference appearance accurately. then, we apply Few-Step Pose Finetuning, combine the image and pose conditions to train VFI module with full conditions for a few iterations (8k). In our experiments, we observed that the other implementations, including swapping the training order, fine-tuning with pose condition only, or fine-tuning with pose guidance for more steps will impact human identity preservation. The condition progressive training could mitigate this effect.

Table 2. **User Study Win Rate.** Subjective comparison between our model and baselines across different motion generation tasks.

Preference of Ours	Dance [48]	Gesture [14]	Action [21]
Texture Quality	82.10%	78.38%	69.12%
Cross-Modal Align.	88.39%	47.63%	45.21%
Overall Preference	84.99%	70.24%	61.05%

Table 3. **Reference Video Length Impact.** Ablation study with FVD, Motion Diversity, and Frame Consistency (LPIPS).

	FVD ↓	Motion Div. ↑	FC (LPIPS) ↓
DanceAnyBeat [48]	1.981	3.669	0.0993
Ours (10s database)	0.611	2.636	0.0466
Ours (100s database)	0.497	5.724	0.0418
Ours (1000s database)	**0.413**	**6.024**	**0.0427**
Real Video	-	5.951	0.0408

Loss Functions. VFI and MDM modules are trained with v-predication and x_0-prediction respectively. The Reference Decoder is trained with MSE and perceptual loss.

3.4. Video Background Reorganization

Our *Video Motion Graphs* can generate realistic videos with the modules described above when reference videos have static or nearly stable backgrounds. However, when backgrounds are dynamic (*e.g.*, running scenes on a street), the VFI module struggles to blend such diverse backgrounds, leading to unrealistic artifacts in the final video. To address this, we employ Video Background Removal [22] to remove video dynamic background, then apply the proposed Video Motion Graphs to synthesis human foreground videos, and finally apply Video Background Generation [36] to put back background or generate interesting novel backgrounds.

4. Evaluation

As our system is entirely novel, with no existing open-source methods available for direct comparison on this multi-modal conditioned human video generation task, we compare our method with separately baselines for different sub-tasks, besides, we provide qualitative end-to-end results and showcase various applications in the supplementary material.

Datasets. We combine a series of video datasets focused on human motion to conduct the experiments. The datasets include MotionX [27] for general motion, Show-Oliver [56] TED-Talk for talkshow, and Champ [62] for dance videos. The datasets are around 100 hours for training, we filtered the data based on motion quality (see supplemental for details) for training. The evaluation is conducted on a randomly sampled 367 video test set.

4.1. Evaluation of Video Motion Graphs System

We evaluate video generation quality on human motion videos. The generation quality is measured by objective

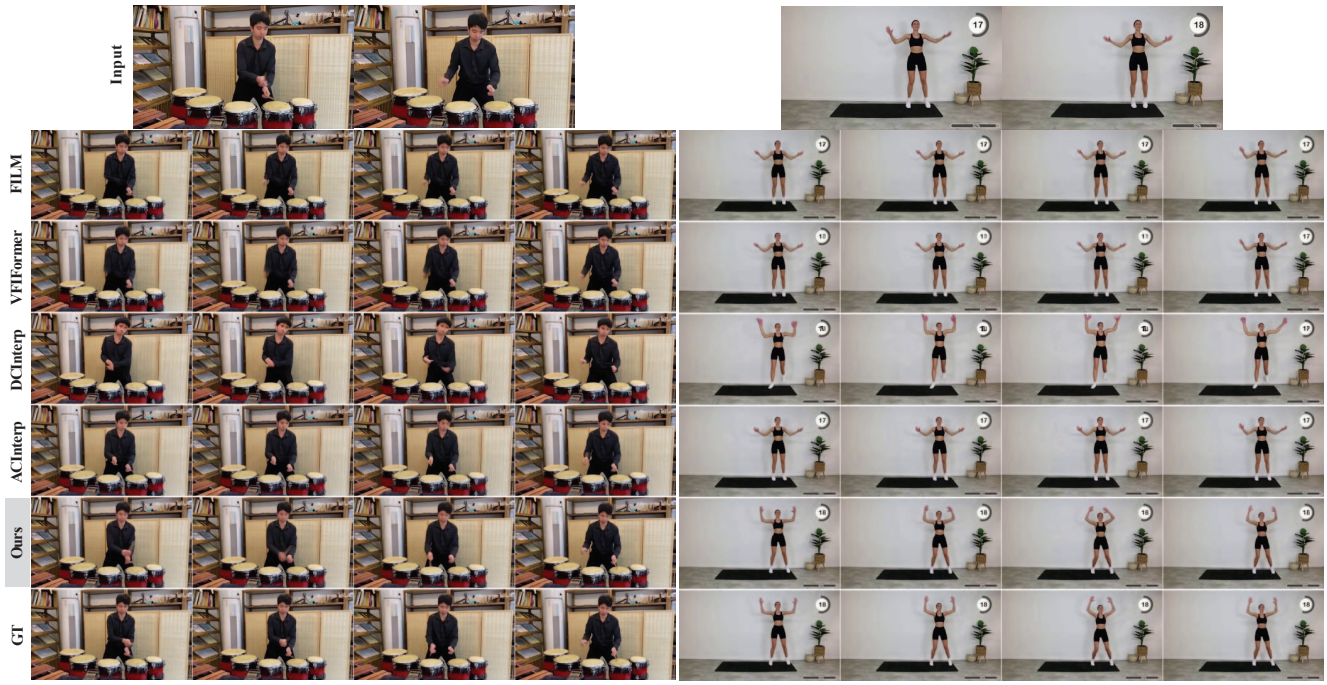

Figure 5. **Subjective Results of HMInterp.** Compared to previous methods, HMInterp generates intermediate frames with accurate motion trajectories, addressing challenges that previous methods with linear motion guidance could not resolve. Top: dynamic motion, such as drumming. Bottom: self-loop motion for fitness activities.

video quality scores and subjective user studies. The video quality evaluation is shown in Table 1. We compared our method with skeleton-pose driven video generation models, including *Animate Anyone* [16], *Magic Pose* [4], *Uni-Animate* [49], and *Mimic Motion* [60]. We select the video length larger than 300 frames in the test set for evaluation. We use the starting frame as the reference frame and GT skeleton poses as input for those methods. For our Video Motion Graphs, we randomly mask out f frames from input videos and recover them by HMInterp interpolation. We evaluated the results when masking out $f = 32, 64, 216$ frames, where higher values indicate a greater challenge for interpolation. The result in Table 1 shows that ours ($f = 32$) significantly outperforms the generated models in all objective terms. Besides, even when it comes to the hardest case with $f = 216$, our method still outperforms existing pose2video models.

We also evaluated the motion alignment and overall performance of generated videos via a subjective user study. We compared different baselines given the specific task. We compared with DanceAnyBeat [48] for music-driven human dance generation, S2G-Diffusion [14] for audio-driven human speech gesture animation, and Text2Performer [21] for action2video (details are in supplement material). We generated 240 videos, 80 per task, as the results and conducted theuser study via Google Forms. For each task, users evaluate 10 videos randomly sampled for each task, 40 videos for all tasks. The results are averaged for all 82 users and shown in Table 2. From the table, we observe that the Video Motion Graphs achieves comparable performance in alignment and gains higher user preference due to notable improvements in video texture quality.

One common limitation for all retrieval-based methods that the target should be aligned with the database, such as image retrieval. But in our case, as shown in Table 3, it will outperform current generative models when the database is larger than 100s, which is relatively easy to record. We report the objective scores below via FVD, Motion Div. (2d joints positions diversity), and Frame Consistency (LPIPS frame difference).

4.2. Evaluation of HMInterp

We compare the video frame interpolation performance of HMInterp with previous non-diffusion-based frame interpolation methods FILM [41], and VFIFormer [33], and diffusion-based methods DCInterp [52] and ACnterp [30]. The objective metrics include MOVIE [43], FVD [46] for video quality, and IS, PSNR for image quality (see supplemental for metric details). The results are shown in Table 4. Our HMInterp outperforms baselines on all terms. Some qualitative comparisons are shown in Figure 3. We highlight a limitation of methods like FILM and VFIFormer,

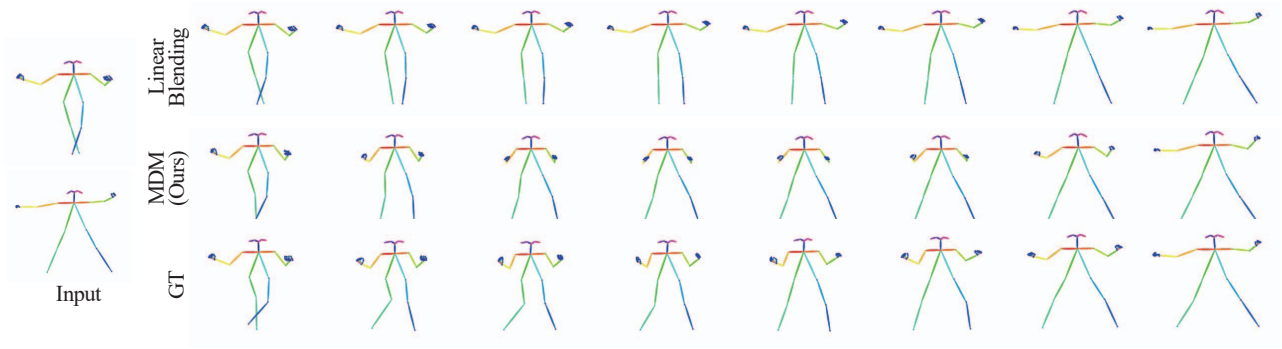

Figure 6. **Ablation of Generative Motion Blending.** Compared to linear blending, our Motion Diffusion Model-based blending generates non-linear intermediate motion for highly dynamic actions, such as dance. Top: linear blending consistently raises the hands above the shoulders. Middle: our blending presents a more complex and natural interpolated dance motion.

Table 4. **Objective comparison of VFI methods.** s denotes the number of start and end frames. Our HMInterp with $s = 1$ outperforms previous non-diffusion-based methods, FILM and VFIFormer, as well as diffusion-based methods, DCInterp and ACInterp. Additionally, HMInterp with $s = 1$, our main model integrated into the Video Motion Graphs, demonstrates high scores at both pixel and feature levels.

	PSNR ↑	LPIPS ↓	MOVIE ↓	FVD ↓
FILM [41]	<u>37.57</u>	0.043	<u>40.64</u>	1.303
VFIFormer [33]	36.29	0.057	52.32	1.390
DCInterp [16]	36.73	0.051	48.62	1.374
ACInterp [61]	37.57	<u>0.040</u>	46.94	<u>1.280</u>
HMInterp ($s = 1$)	**39.53**	**0.034**	**39.18**	**1.210**
HMInterp ($s = 3$)	**40.40**	**0.021**	**37.85**	**0.619**

Table 5. **Ablation Study of HMInterp.** Without motion guidance, performance shows a clear drop in feature-level metrics, such as ISIPS and FVD. Without the low-level reference decoder, pixel-level metrics, such as PSNR and MOVIE, decrease.

	PSNR ↑	LPIPS ↓	MOVIE ↓	FVD ↓
HMInterp ($s = 1$)	**39.53**	**0.034**	**39.18**	**1.210**
w/o motion guidance	39.17	0.048	41.34	1.391
w linear guidance	39.16	0.042	41.06	1.297
w/o reference decoder	37.21	0.039	49.67	1.283
w zero reference decoder	38.13	0.034	40.11	1.221

which tend to miss generating body regions due to the absence of explicit guidance.. Other methods, such as ACInterp with linear interpolation, often produce incorrect motion trajectories. In contrast, our approach maintains both high video quality and accurate motion trajectories.

Ablation of Motion Diffusion Model. As shown in Table 5 and Figure 7, we show the difference that without the MDM, the results degrade on all terms, and the model subjective tends to refer to features from incorrect regions. On the other hand, we demonstrate our MDM is better than linear blending in Figure 4.

Effectiveness of Condition Progressive Training. We compare our condition progressive training with different

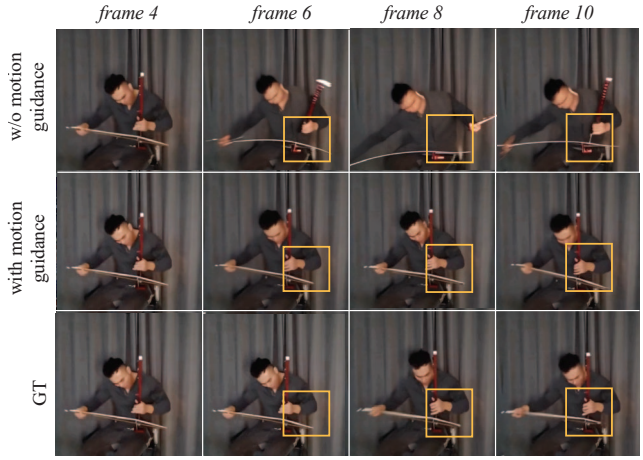

Figure 7. **Additional benefits of motion guidance.** Motion guidance distinguishes regions to obtain the correct interaction with the image content. Top: without motion guidance, BanHu's texture was broken due to out-of-range motion. Middle: with motion guidance, both human motion and BanHu display the correct texture.

variants in Table 6 and Figure 9. The straightforward pose-to-video training [16, 53], *i.e.*, directly using the pose condition as the initial stage, generates plausible motion results but leads to notable appearance inconsistencies across frames. Addressing these inconsistencies is essential for achieving artifact-free and production-level video results. Other variants in Table 6, such as sequential training with pose followed by seed image ($P \rightarrow SI$) and simultaneous training ($P + SI$), demonstrate that introducing pose conditions early or simultaneously with identity conditions (seed images) tends to degrade the consistency of generated videos. Specifically, simultaneous condition training ($P+SI$) slightly improves metrics over the pose-only baseline but still falls short compared to a progressive strategy. Our proposed Condition Progressive Training approach, where we first employ identity-strong conditions (seed im-

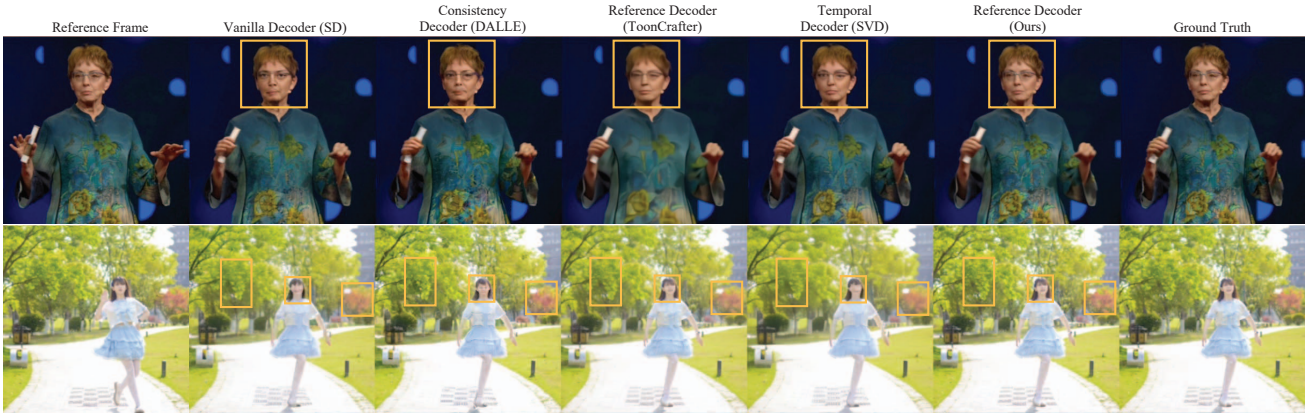

Figure 8. **Ablation of Reference Decoder.** We adopt duplicated padding inputs based on the reference decoder in ToonCrafter. It significantly improves results at lower resolutions, *e.g.*, 256 × 256. Compared to other methods, our implementation provides accurate facial (top) and background (bottom) details.

Table 6. **Ablation study for condition progressive training.** P denotes pose condition (pose-to-video training in AnymateAnyone), SI is Seed Image condition (start and end reference images), *iters* is training iterations. Ours (in gray) progressive training shows best performance. See Figure 9 for subjective results.

Stage 1	Stage 2	PSNR ↑	LPIPS ↓	MOVIE ↓	FVD ↓
P	-	35.55	0.044	54.68	1.369
P	SI	36.81	0.041	51.66	1.330
$P + SI$	-	36.62	0.041	52.41	1.325
SI	-	36.84	0.043	51.89	1.339
SI	P	36.83	0.042	52.03	1.336
SI	$P + SI$ (8k *iters*)	**37.21**	**0.039**	**49.67**	**1.283**
SI	$P + SI$ (30k *iters*)	36.87	0.041	51.69	1.307

ages) and later incorporate identity-weak conditions (pose), achieves the best quantitative performance.

Effectiveness of The Improved Reference Decoder. As shown in Table 5, the reference decoder brings low-level consistency. Due to the memory issue, we keep the single-frame input on vanilla ReferenceNet, but we propose to use duplicated frames instead of zero padding frames on Reference Decoder. Compared with the baseline ToonCrafter [51], this brings a clear benefit with a PSNR improvement over 1.0. The subjective results are shown in Figure 8, compared with baseline decoders. Such as vanilla decoder (Stable Diffusion) [42], Consistency Decoder (DALL-E) [40], Temporal Decoder (Stable Video Diffusion) [1], and Dual-Reference Decoder (ToonCrafter) [51], our Reference Decoder could decode a background consistent video.

4.3. Applications of the System

Real-time video generation. By pre-caching the possible transition results in the memory, our method could generate video in real-time without length limitation. It's recommended to see the real-time kungfu demo to demonstrate the potential applications like video game apps for our method.

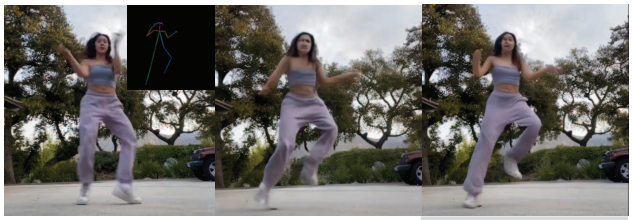

Figure 9. **Subjective comparison for progressive training**. The S1 and S2 denote training stages. After comparing different condition selections for each stage, we observed that training with pose conditions more will decrease the appearance consistency. Our progressive training (in gray) shows highly consistent results with the groundtruth. Refer to Table 6 for objective scores.

Key-frame editing. Our system supports generating the target path based on keyframes, (see supplement materials for details). The generated video could remain aligned with the predefined keyframes. By replacing the motion guidance generated by MDM with a custom 2D pose video, our HMInterp could generate different motions with the reference video. Make it possible to combine the generative model with the retrieval model for further applications.

5. Conclusion

In this paper, we extend the previous speech gesture Video Motion Graphs system to a more robust and versatile framework for general human motion videos. By introducing HMInterp, a significantly enhanced video frame interpolation model, we enable seamless blending of open-domain human motion videos. It supports a range of applications including keyframe editing, and real-time video generation.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2, 4, 8

[2] 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. 2

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 2

[4] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2023. 2, 5, 6

[5] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2, 3

[6] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2

[7] Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. Vlogger: Multimodal diffusion for embodied avatar synthesis. *arXiv preprint arXiv:2403.08764*, 2024. 1, 2

[8] Duolikun Danier, Fan Zhang, and David R. Bull. St-mfnet: A spatio-temporal multi-flow network for frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3511–3521, 2022. 3

[9] Duolikun Danier, Fan Zhang, and David R. Bull. LDMVFI: video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1472–1480, 2024. 3

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2

[11] Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *European Conference on Computer Vision*, pages 378–395. Springer, 2025. 3

[12] 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. 2, 4

[13] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2263–2273, 2024. 2

[14] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. *arXiv preprint arXiv:2404.01862*, 2024. 5, 6

[15] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[16] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 1, 2, 4, 5, 6, 7

[17] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 3

[18] Ziyao Huang, Fan Tang, Yong Zhang, Xiaodong Cun, Juan Cao, Jintao Li, and Tong-Yee Lee. Make-your-anchor: A diffusion-based 2d avatar generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6997–7006, 2024. 2

[19] Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7341–7351, 2024. 3, 4

[20] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik G. Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9000–9008, 2018. 3

[21] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22747–22757, 2023. 1, 2, 5, 6

[22] Saeed Kamel, Hossein Ebrahimnezhad, and Afshin Ebrahimi. Moving object removal in video sequence and background restoration using kalman filter. In *2008 International Symposium on Telecommunications*, pages 580–585. IEEE, 2008. 5

[23] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang.

Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2022. 3

[24] Lucas Kovar, Michael Gleicher, and Frédéric H. Pighin. Motion graphs. pages 51:1–51:10, 2008. 2

[25] Zhen Li, Zuo-Liang Zhu, Linghao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. AMT: all-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9801–9810, 2023. 3

[26] Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, and Yanbo Zheng. Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. *arXiv preprint arXiv:2409.01876*, 2024. 2

[27] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[28] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3764–3773, 2022. 2

[29] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *arXiv preprint arXiv:2203.05297*, 2022. 2

[30] Haiyang Liu, Xingchao Yang, Tomoya Akiyama, Yuantian Huang, Qiaoge Li, Shigeru Kuriyama, and Takafumi Taketomi. Tango: Co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. *arXiv preprint arXiv:2410.04221*, 2024. 1, 2, 3, 6

[31] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024. 2

[32] Ziwei Liu, Raymond A. Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4473–4481, 2017. 3

[33] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3532–3542, 2022. 3, 6, 7

[34] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1710, 2018. 3

[35] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5436–5445, 2020. 3

[36] Boxiao Pan, Zhan Xu, Chun-Hao Paul Huang, Krishna Kumar Singh, Yang Zhou, Leonidas J Guibas, and Jimei Yang. Actanywhere: Subject-aware video background generation. *arXiv preprint arXiv:2401.10822*, 2024. 5

[37] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. BMBC: bilateral motion estimation with bilateral cost volume for video interpolation. In *European Conference on Computer Vision*, pages 109–125, 2020. 3

[38] Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14519–14528, 2021. 3

[39] 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. 4

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 8

[41] 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. 3, 6, 7

[42] 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. 2, 8

[43] Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE transactions on image processing*, 19(2):335–350, 2009. 6

[44] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2, 4, 5

[45] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: extreme video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14469–14478, 2021. 3

[46] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR, 2018. 6

[47] Volker Steinbiss, Bach-Hiep Tran, and Hermann Ney. Improvements in beam search. In *ICSLP*, pages 2143–2146, 1994. 4

[48] Xuanchen Wang, Heng Wang, Dongnan Liu, and Weidong Cai. Dance any beat: Blending beats with visuals in dance video generation. *arXiv preprint arXiv:2405.09266*, 2024. 1, 2, 5, 6

[49] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong

Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024. 5, 6

[50] Yue Wu, Qiang Wen, and Qifeng Chen. Optimizing video prediction via video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17814–17823, 2022. 3

[51] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *arXiv preprint arXiv:2405.17933*, 2024. 2, 4, 8

[52] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 3, 4, 6

[53] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 2, 4, 7

[54] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8): 1106–1125, 2019. 3

[55] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2

[56] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *CVPR*, 2023. 5

[57] Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: you prompt and we fix the photo. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 40000–40031, 2024. 2

[58] Yongsheng Yu, Haitian Zheng, Zhifei Zhang, Jianming Zhang, Yuqian Zhou, Connelly Barnes, Yuchen Liu, Wei Xiong, Zhe Lin, and Jiebo Luo. Zipir: Latent pyramid diffusion transformer for high-resolution image restoration. *arXiv preprint arXiv:2504.08591*, 2025. 2

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[60] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 5, 6

[61] Yang Zhou, Jimei Yang, Dingzeyu Li, Jun Saito, Deepali Aneja, and Evangelos Kalogerakis. Audio-driven neural gesture reenactment with video motion graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3418–3428, 2022. 1, 2, 3, 7

[62] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*, 2024. 2, 5