Generative Inbetweening through Frame-wise Conditions-Driven Video Generation

Tianyi Zhu[1] Dongwei Ren[2,✉] Qilong Wang[2] Xiaohe Wu[1] Wangmeng Zuo[1]
[1]Harbin Institute of Technology [2]Tianjin University

Abstract

Generative inbetweening aims to generate intermediate frame sequences by utilizing two key frames as input. Although remarkable progress has been made in video generation models, generative inbetweening still faces challenges in maintaining temporal stability due to the ambiguous interpolation path between two key frames. This issue becomes particularly severe when there is a large motion gap between input frames. In this paper, we propose a straightforward yet highly effective Frame-wise Conditions-driven Video Generation (FCVG) method that significantly enhances the temporal stability of interpolated video frames. Specifically, our FCVG provides an explicit condition for each frame, making it much easier to identify the interpolation path between two input frames and thus ensuring temporally stable production of visually plausible video frames. To achieve this, we suggest extracting matched lines from two input frames that can then be easily interpolated frame by frame, serving as frame-wise conditions seamlessly integrated into existing video generation models. In extensive evaluations covering diverse scenarios such as natural landscapes, complex human poses, camera movements and animations, existing methods often exhibit incoherent transitions across frames. In contrast, our FCVG demonstrates the capability to generate temporally stable videos using both linear and non-linear interpolation curves. Our project page and code are available at https://fcvg-inbetween.github.io/.

1. Introduction

Given the presence of low framerate videos, the generation of high framerate videos has emerged as an active research area with a wide range of applications that demands smooth and stable visual contents. Video interpolation or inbetweening, which focuses on synthesizing intermediate frames between two given frames, has been extensively studied in the literature [7, 22, 23, 38, 61]. However, traditional video interpolation methods are inherently limited in dealing with significant motions due to their reliance on optical flow for modeling frame motion. Recently, generative image-to-video (I2V) models have made remarkable progress in generating coherent videos [4, 31], offering potential solutions to enhance framerates through generative inbetweening [47, 53]. Compared to traditional video interpolation, generative inbetweening leverage the creative ability of generative models and focuses more on handling scenarios with larger time intervals.

When start and end frames are provided, it is a straightforward task to separately generate two coherent videos using an I2V model, but the challenge of inbetweening lies in the ambiguity of the interpolation path caused by large motions. To address this issue, a time reversal strategy [8] is proposed to average the fusion of bidirectional diffusion denoising steps conditioned on start and end frames. Then, temporal attention layers in the I2V model are fine-tuned to enhance motion coherence [47]. More recently, Yang et al. [53] introduced a multi-channel sampling strategy to substitute direct average fusion. However, the ambiguity of interpolation path remains severe and often results in incoherent transitions in generated videos, as shown in Fig. 1.

In this paper, our aim is to mitigate the ambiguity in interpolation path and achieve temporal stability in video generation. We propose a straightforward yet highly effective Frame-wise Conditions-driven Video Generation (FCVG) method. Our FCVG model provides an explicit condition when generating each frame using an I2V model, making it easier to identify the interpolation path between start and end frames, and thus ensuring temporally stable production of visually plausible video frames. Firstly, we extract two conditions from the input key frames to establish robust correspondences between start and end frames using matched lines. In addition, pose skeletons can be incorporated into the conditions to better capture human poses. Subsequently, linear interpolation is employed on a frame-by-frame basis to interpolate the start and end conditions. These frame-wise conditions effectively alleviate ambiguity in determining the interpolation path, and can be seamlessly integrated into the I2V model as control for video frame generation.

The linear frame-wise conditions for generative inbetweening are generally feasible from two perspectives: (*i*)

Figure 1. Video results of GI [47] and our FCVG. More videos are provided in https://fcvg-inbetween.github.io/.

In pioneering video interpolation methods [2, 19, 25], the linear assumption is commonly adopted, which may not truly align with ground-truth temporal consistency but can lead to temporally stable videos for most scenes. As depicted in Fig. 1, our FCVG demonstrates a significantly higher level of video stability compared to existing methods; (*ii*) As illustrated in Fig. 2(b), our frame-wise conditions provide a control path for inbetweening, while still allowing some flexibility for video generation models, whose influence can be further adjusted by tuning the fusion weight between features of condition and video generation branches. Moreover, our FCVG allows users to specify a non-linear interpolation path for frame-wise conditions to generate desired video frames, providing more flexibility in determining the interpolation path.

Extensive experiments are conducted on the collected diverse testing samples including natural landscapes, complex human poses, camera movements and animations, where our FCVG is compared with both video interpolation methods and generative inbetweening methods. Regarding several evaluation metrics, our method outperforms existing methods in terms of frame textures and temporal stability. Particularly when dealing with large motions, unlike other methods that suffer from incoherent transitions across frames, our FCVG exhibits significantly enhanced temporal stability in the generated videos.

2. Related Work

Optical Flow-based Frame Interpolation. Video frame interpolation aims to synthesize intermediate frames between two given input frames [3, 28]. Previous methods primarily rely on optical flow-based approaches[14, 22, 23], which estimate optical flow and apply forward [27] or backward warping to generate intermediate frames. Some techniques incorporate flow reversal techniques to estimate the intermediate flows [19, 40, 50], while others concentrate on directly predicting the intermediate flows [15, 21, 55, 60]. Although these methods yield stable interpolation results in real-world scenes, they exhibit significant artifacts when confronted with large motion or complex scenarios, such as human motion. Moreover, they struggle with the impact of optical flow estimation accuracy [6] in varying data distributions, such as animation or line art. While some approaches have developed models tailored for animation [6, 41] or line art [37, 42, 64], restricting them for specific data types.

Diffusion-based Frame Interpolation. In recent years, diffusion models have demonstrated remarkable capabilities in generating high-quality images and videos [1, 5, 12, 34, 43, 46]. Several studies have explored the effectiveness of diffusion models for video frame interpolation [7, 17, 38, 45], particularly in addressing complex motions that pose significant challenges for optical flow-based methods. Some approaches treat the input frames as conditions and utilize large-scale data to train a diffusion model for video frame interpolation [18, 51, 52]. Other approaches leverage pre-trained image-to-video diffusion models, incorporating new sampling strategies to achieve frame interpolation [8, 47, 53]. TRF [8] proposes a time reversal sampling strategy that fuses bidirectional motion from two parallel diffusion denoise steps conditioned on the start and end frame. Based on the time reversal strategy, Generative Inbetweening [47] fine-tunes diffusion models by utilizing temporal attention information to maintain motion consistency, while VIBIDSampler [53] introduces a bidirectional sampling approach rather than direct average fusion ensuring on-manifold generation of intermediate frames.

Controllable Video Generation. Recent works try to introduce controllable conditions to video generation models [59], such as camera motion [9, 49] and human pose [13, 58, 63]. Among these methods, the approach utilizing lightweight adapters [26, 56] is preferred by many researchers due to its elimination of the need for pre-training the large diffusion model. ControlNet [56] uses zero convolution to connect certain trainable layers copied from pre-trained large models to the original layers. To reduce the additional computational cost, ControlNeXt [31] introduces a lightweight module and fine-tunes several parameters in the diffusion model, aligning them using cross normalization.

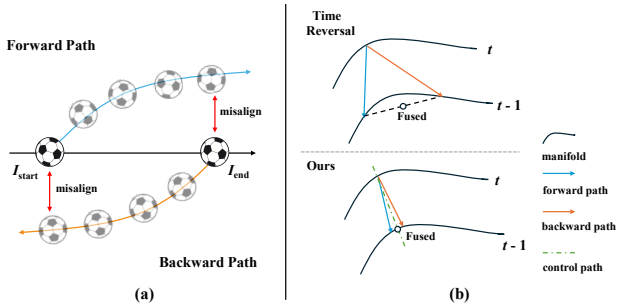

Figure 2. (a) The ambiguity in interpolation path, where the stochastic nature of forward and backward paths leads to misalignment or even generating substantially different contents in two paths. (b) Our frame-wise conditions serve as a control path, enabling rough alignment of the forward and backward paths, thereby confining the fusion process closer to the manifold.

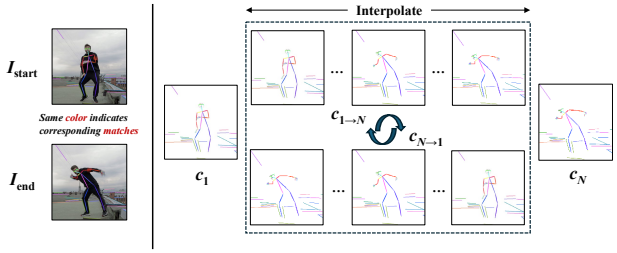

Figure 3. The process for acquiring forward and backward frame-wise conditions $\boldsymbol{c}_{1\to N}$ and $\boldsymbol{c}_{N\to 1}$. Two initial conditions $\boldsymbol{c}_1$ and $\boldsymbol{c}_N$ can be obtained by establishing correspondence between start frame $\boldsymbol{I}_{\text{start}}$ and end frame $\boldsymbol{I}_{\text{end}}$, where the same color indicates corresponding matches. Then, frame-wise conditions are obtained by interpolating $\boldsymbol{c}_1$ and $\boldsymbol{c}_N$.

3. Method

3.1. Preliminaries

Diffusion model [12] is a type of generative model that uses a denoising network to iteratively denoise random Gaussian noise to generate a high-quality image or video. Specifically, for I2V diffusion models, such as Stable Video Diffusion (SVD) [4], given a video $\boldsymbol{x} \in \mathbb{R}^{N\times 3\times H\times W}$ containing N frames, SVD first encode $\boldsymbol{x}$ using an autoencoder $\mathcal{E}(\cdot)$ to get latent representation $\boldsymbol{z} \in \mathbb{R}^{N\times C\times H\times W}$ [20]. The forward process of diffuse gradually adds noise to latent representation $\boldsymbol{z}$ as follows

$$\boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sigma_t \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, α_t and σ_t represent the noise level for denoising time t, defined by the noise schedule. As for backward process, a 3D UNet [35] denoiser f_θ is used for iteratively denoising under the image condition $\boldsymbol{c}_{\text{image}}$. The optimization objective is formulated as

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{c}_{\text{image}},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),t}\left[\|\boldsymbol{v} - f_\theta(\boldsymbol{z}_t, \boldsymbol{c}_{\text{image}}, t)\|_2^2\right], \tag{2}$$

where $\boldsymbol{v} = \alpha_t\boldsymbol{\epsilon}_t - \sigma_t\boldsymbol{z}_t$ is referred as v-prediction [36]. Finally, the generated videos are obtained through the VAE decoder $\hat{\boldsymbol{x}} = \mathcal{D}(\boldsymbol{z}_0)$.

Inspired by the strong capability of diffusion models in video generation, several studies have endeavored to integrate diffusion models into video inbetweening to tackle challenges posed by intricate and extensive motions that are difficult to address using optical flow-based methods [7, 8, 17]. A direct approach for applying diffusion models in video inbetweening involves utilizing key frames as conditions for retraining the diffusion model. However, this often necessitates substantial data and computational resources and may also be influenced by discrepancies in data distribution. Consequently, some approaches leverage pre-trained diffusion models for video inbetweening [8, 47, 53]. The fundamental concept behind these methods is to fuse the temporal forward path and backward path after each denoising step conditioned on start and end frames, respectively, i.e., a process referred to "time reversal" [8, 47].

3.2. Motivation of Frame-wise Conditions

Although time reversal provides a means to directly utilize pre-trained video generations for inbetweening, it exhibits certain limitations [8] that are summarized as follows: (***i***) The motion generated by I2V models tends to be diverse rather than stable. While this diversity is advantageous for pure I2V tasks, it introduces significant ambiguity when applying the time reversal strategy for video inbetweening. As depicted in Fig. 2(a), the stochastic nature of generating forward and backward paths leads to misalignment and even substantially different contents in two paths, resulting in an unstable and unrealistic videos. (***ii***) Tedious tuning of hyper-parameters related to temporal conditioning within the I2V model is required for each input pair, such as motion bucket ID and frames per second. (***iii***) Inference efficiency is constrained by certain techniques, e.g., noise re-injection [8], aimed at mitigating the ambiguity but significantly increasing inference time (approximately 1.5 to 3 times longer).

Indeed, if the first fundamental limitation in ambiguity can be addressed, the subsequent two issues can also be readily resolved. To this end, several studies have made significant efforts in mitigating the misalignment between forward and backward paths [47, 53]. Nevertheless, as depicted in Fig. 2(b), there still exists considerable stochasticity between these paths, thereby constraining the effectiveness of these methods in handling scenarios involving large motions such as rapid changes in human poses. The ambiguity in the interpolation path primarily arises from insufficient conditions for intermediate frames, since two input images only provide conditions for start and end frames. Therefore, in this work, we suggest offering an explicit condition for each frame, which significantly alleviates the am-

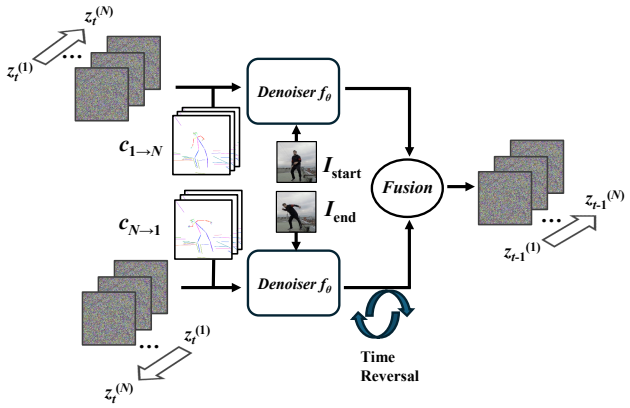

Figure 4. Inference of FCVG at time t.

Algorithm 1: Inference of FCVG

Input: $\boldsymbol{I}_{\text{start}}, \boldsymbol{I}_{\text{end}}, f_\theta, \mathcal{D}, \boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Computing $\boldsymbol{\lambda}$ with $\lambda_i = 1 - \frac{i-1}{N-1}, i \in \{1, ..., N\}$;
Extracting conditions $\boldsymbol{c}_1, \boldsymbol{c}_N$ from $\boldsymbol{I}_{\text{start}}, \boldsymbol{I}_{\text{end}}$;
$\boldsymbol{c}_{1\to N} = \text{interpolate}(c_1, c_N)$;
$\boldsymbol{c}_{N\to 1} = \text{flip}(\boldsymbol{c}_{1\to N})$;
for $t \leftarrow T : 1$ **do**
  $\tilde{\boldsymbol{z}}_t = f_\theta\left(\boldsymbol{z}_t, \boldsymbol{I}_{\text{start}}, \boldsymbol{c}_{1\to N}, t\right)$;
  $\boldsymbol{z}_{t,f} = \text{flip}(\boldsymbol{z}_t)$;
  $\tilde{\boldsymbol{z}}'_t = f_\theta\left(\boldsymbol{z}_{t,f}, \boldsymbol{I}_{\text{end}}, \boldsymbol{c}_{N\to 1}, t\right)$;
  $\tilde{\boldsymbol{z}}'_{t,f} = \text{flip}(\tilde{\boldsymbol{z}}'_t)$;
  $\boldsymbol{z}_{t-1} = \boldsymbol{\lambda} \cdot \tilde{\boldsymbol{z}}_t + (1 - \boldsymbol{\lambda}) \cdot \tilde{\boldsymbol{z}}'_{t,f}$;
end
Return: $\mathcal{D}(\boldsymbol{z}_0)$

biguity of the interpolation path. As shown in Fig. 2(b), frame-wise conditions ensure that the forward and backward paths are relatively aligned during the denoising process, thereby rendering a simple fusion method adequate to confine the fusion process closer to the manifold.

3.3. Proposed FCVG

Given two input frames $\boldsymbol{I}_{\text{start}}$ and $\boldsymbol{I}_{\text{end}}$, our FCVG aims to generate N video frames, whose start and end frames should be consistent with $\boldsymbol{I}_{\text{start}}$ and $\boldsymbol{I}_{\text{end}}$, respectively. As shown in Fig. 4, our FCVG provides frame-wise conditions for video generation model to make the generated frames be temporally stable. Based on the time reversal strategy, our FCVG at time t can be specifically formulated as

$$\tilde{\boldsymbol{z}}_t = f_\theta\left(\boldsymbol{z}_{t+1}, \boldsymbol{I}_{\text{start}}, \boldsymbol{c}_{1\to N}, t\right), \tag{3}$$

$$\tilde{\boldsymbol{z}}'_t = f_\theta\left(\text{flip}(\boldsymbol{z}_{t+1}), \boldsymbol{I}_{\text{end}}, \boldsymbol{c}_{N\to 1}, t\right), \tag{4}$$

$$\boldsymbol{z}_t = \boldsymbol{\lambda} \cdot \tilde{\boldsymbol{z}}_t + (1 - \boldsymbol{\lambda}) \cdot \text{flip}(\tilde{\boldsymbol{z}}'_t), \tag{5}$$

where $\boldsymbol{z}_{t+1} \in \mathbb{R}^{N\times C\times H\times W}$ is diffusion noises from time $t+1$, f_θ is the pre-trained denoiser model with control module, $\boldsymbol{c}_{1\to N}, \boldsymbol{c}_{N\to 1} \in \mathbb{R}^{N\times 3\times H\times W}$ are temporal aligned forward and backward frame-wise conditions, $\text{flip}(\cdot)$ denotes flipping the sample along the time dimension, and $\boldsymbol{\lambda} \in \mathbb{R}^N$ is fusion weights with $\lambda_i = 1 - \frac{i-1}{N-1}, i \in \{1, ..., N\}$. The symbol $\cdot$ denotes frame-wise multiplication, i.e., λ_i is multiplied with noises for i-th frame. By iteratively denoising until time $t = 0$, final inbetweening video frames can be obtained by $\hat{\boldsymbol{x}}_0 = \mathcal{D}(\boldsymbol{z}_0)$.

3.3.1 Frame-wise Conditions

Currently, we only have access to two image conditions, i.e., $\boldsymbol{I}_{\text{start}}$ and $\boldsymbol{I}_{\text{end}}$. However, extending them as frame-wise conditions is infeasible. In order to acquire frame-wise conditions, the initial conditions should satisfy two properties that can effectively capture frame motion, and are amenable for extension as frame-wise. Drawing inspiration from prior works [23, 64] that showcase the robustness of global matching in handling large motions and complex scenes, we propose to employ the line matching model to extract the initial conditions from $\boldsymbol{I}_{\text{start}}$ and $\boldsymbol{I}_{\text{end}}$. Compared to conditioning on optical flow, which is often erroneous in certain regions, using line matches as conditions focuses on the most stable regions, avoiding the introduction of erroneous conditions that could disrupt the generation process. Specifically, we utilize the pre-trained GlueStick [30] as our line matching model to establish correspondences between $\boldsymbol{I}_{\text{start}}$ and $\boldsymbol{I}_{\text{end}}$, and subsequently visualize these matches as images with distinct colors representing different line matches, resulting in two initial conditions $\boldsymbol{c}_1$ and $\boldsymbol{c}_N$. Moreover, to further improve human motions, we extract human poses that can be directly added into $\boldsymbol{c}_1$ and $\boldsymbol{c}_N$.

As shown in Fig. 3, the control condition $\boldsymbol{c}_i$ for i-th frame can be obtained by accordingly interpolating $\boldsymbol{c}_1$ and $\boldsymbol{c}_N$, where the forward frame-wise conditions $\boldsymbol{c}_{1\to N}$ are the concatenation of $\{\boldsymbol{c}_i\}_{i=1}^N$ along the time dimension, and backward frame-wise conditions $\boldsymbol{c}_{N\to 1}$ are the flip of $\boldsymbol{c}_{1\to N}$ along the time dimension. We empirically found that linear interpolation is sufficient for most cases to guarantee temporal stability in inbetweening videos, and our method allows users to specify non-linear interpolation paths for generating desired videos.

3.3.2 Injection to Video Generation Model

We follow ControlNeXt [31] to inject frame-wise conditions into the I2V model. We choose SVD [4] as our base I2V model. Compared to other controllable video generation methods, ControlNeXt is light-weight and does not significantly increases inference time. Specifically, the control conditions are first encoded by a lightweight module composed of multiple ResNet [10] blocks. These encoded conditions are first processed using cross normalization to

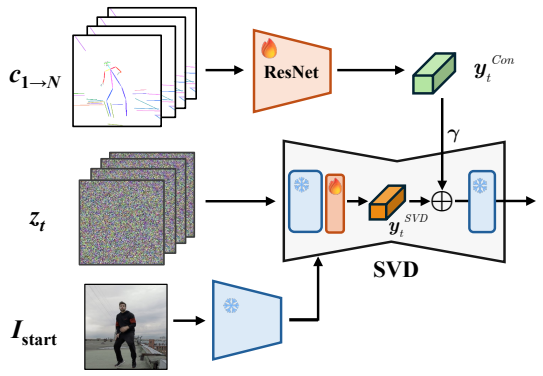

Figure 5. Overview of injecting frame-wise conditions into SVD. To make frame-wise conditions better fit pre-trained SVD, we only need to fine-tune a small set of parameters including the value and output projection matrices within the attention layers, and the lightweight ResNet blocks.

align the distributions of $\boldsymbol{y}^{\text{Con}}$ and $\boldsymbol{y}^{\text{SVD}}$ from the condition and SVD branches. Then at denoising time t, frame-wise conditions can be injected into SVD as follows

$$\hat{\boldsymbol{y}}_t = \boldsymbol{y}_t^{\text{SVD}} + \gamma\, \boldsymbol{y}_t^{\text{Con}}, \tag{6}$$

where γ is a tunable weight for controlling the significance of $\boldsymbol{y}^{\text{Con}}$. The inference of FCVG is detailed in Algorithm 1. Moreover, in order to enhance the compatibility of pre-trained SVD with our frame-wise conditions, we employ a small set of videos for fine-tuning. We freeze the majority of SVD parameters and solely optimize the value and output projection matrices within attention layers, along with lightweight ResNet blocks, as shown in Fig. 5.

The limitations in Sec. 3.2 have been largely resolved in FCVG: (***i***) By explicitly specifying the condition for each frame, the ambiguity between forward and backward paths is significantly alleviated; (***ii***) Only one tunable parameter γ is introduced, and setting it to 1, while keeping hyper-parameters in SVD as default, yields favorable results in most scenarios; (***iii***) A simple average fusion, without noise re-injection, is adequate in FCVG, and the inference steps can be substantially reduced by 50% compared to GI [47].

4. Experiments

4.1. Experimental Setup

Datasets. To verify the performance of our FCVG across diverse scenes, we collect a dataset encompassing a variety of scenes, such as natural environments, indoor/outdor scenes, and human poses, where diverse motion types are encompassed such as camera movements, object motion, human dance actions, and facial expression transitions. Specifically, our dataset consists of 524 video clips, each containing 25 frames, selected from the DAVIS dataset [32] and RealEstate10K dataset [62], supplemented by high-frame-rate videos from Pexels[1]. We randomly split the dataset in a 4:1 ratio for fine-tuning and testing respectively.

Evaluation Metrics. Following previous works [8, 47], we adopt LPIPS [57] and Fréchet Inception Distance (FID) [11, 29] to evaluate the quality of individual frames, while employing Fréchet Video Distance (FVD) [44] to assess the overall quality of videos. Additionally, we take two recently proposed metrics VBench [16] and FVMD [24] to assist the evaluation, where VBench assesses videos across multiple dimensions based on pre-trained models, while FVMD refines FVD by emphasizing more on motion consistency. Furthermore, it should be noted that all these metrics are not capable of precisely evaluating temporal stability of generated videos, and thus we highly recommend directly observing more video results provided in our project page.

Implementation Details. For obtaining initial conditions, we utilize the pre-trained GlueStick [30] for line matching and DWPose [54] for estimating human poses. The fine-tuning is performed on ResNet blocks, and the value and output projection matrices within the attention layer of SVD. The fine-tuning process is conducted for 70k iterations using the AdamW optimizer on an NVIDIA A800 GPU, with a learning rate of 1×10^{-6} and $\beta_1 = 0.9, \beta_2 = 0.999$. For fine-tuning, we crop the frames to patches with resolution 512×320. The inference of FCVG is done in $T = 25$ steps without noise re-injection. Without specific clarity, the balancing weight $\gamma = 1$ is set for all the experiments. Due to the robustness of our FCVG, all the hyper-parameters in SVD adopt the default settings.

4.2. Comparison with State-of-the-arts

We compare FCVG with state-of-the-art optical flow-based interpolation method FILM [33], and diffusion-based methods including DynamiCrafter [52], TRF [8] and GI [47].

Quantitative evaluation. To evaluate performance under different motion conditions, we conduct assessments with frame gaps setting to 23 and 12, respectively. As shown in Tab. 1, our method achieves the best performance among four generative approaches across all the metrics. Regarding the LPIPS comparison with FILM, our FCVG is marginally inferior, while demonstrating superior performance in other metrics. Considering the absence of temporal information in LPIPS, it may be more appropriate to prioritize other metrics and visual observation. Moreover, by comparing the results under different frame gaps, FILM may work well when the gap is small, while generative methods are more suitable for large gap. Among these generative methods, our FCVG exhibits significant superiority owing to its explicit frame-wise conditions.

Qualitative evaluation. The visual comparison in Fig. 6 il-

[1]https://www.pexels.com/

Table 1. Quantitative comparison on different interpolation gaps. Our primary focus lies in generative approaches, with FILM being the sole method that employs optical flow. **Bold** refer to the best results. All these metrics are not capable of precisely evaluating temporal stability of generated videos, and thus we highly recommend directly observing video results.

Method	Frame Gap = 23					Frame Gap = 12				
	LPIPS (↓)	FID (↓)	VBench (↑)	FVMD (↓)	FVD (↓)	LPIPS (↓)	FID (↓)	VBench (↑)	FVMD (↓)	FVD (↓)
FILM[33]	**0.1540**	25.00	0.8615	8208.7	543.4	**0.1980**	24.44	0.8667	6975.9	495.4
DynamiCrafter[52]	0.3886	52.66	0.8410	13221.9	978.9	0.3839	37.49	0.8458	11810.7	652.5
TRF[8]	0.3687	42.76	0.8438	10458.0	823.4	0.3742	39.01	0.8478	10076.6	818.4
GI[47]	0.2155	31.39	0.8606	5682.6	524.0	0.2615	32.37	0.8651	4721.0	565.8
Ours	0.1832	**24.05**	**0.8619**	**5607.2**	**437.9**	0.2378	**22.77**	**0.8672**	**4537.4**	**465.6**

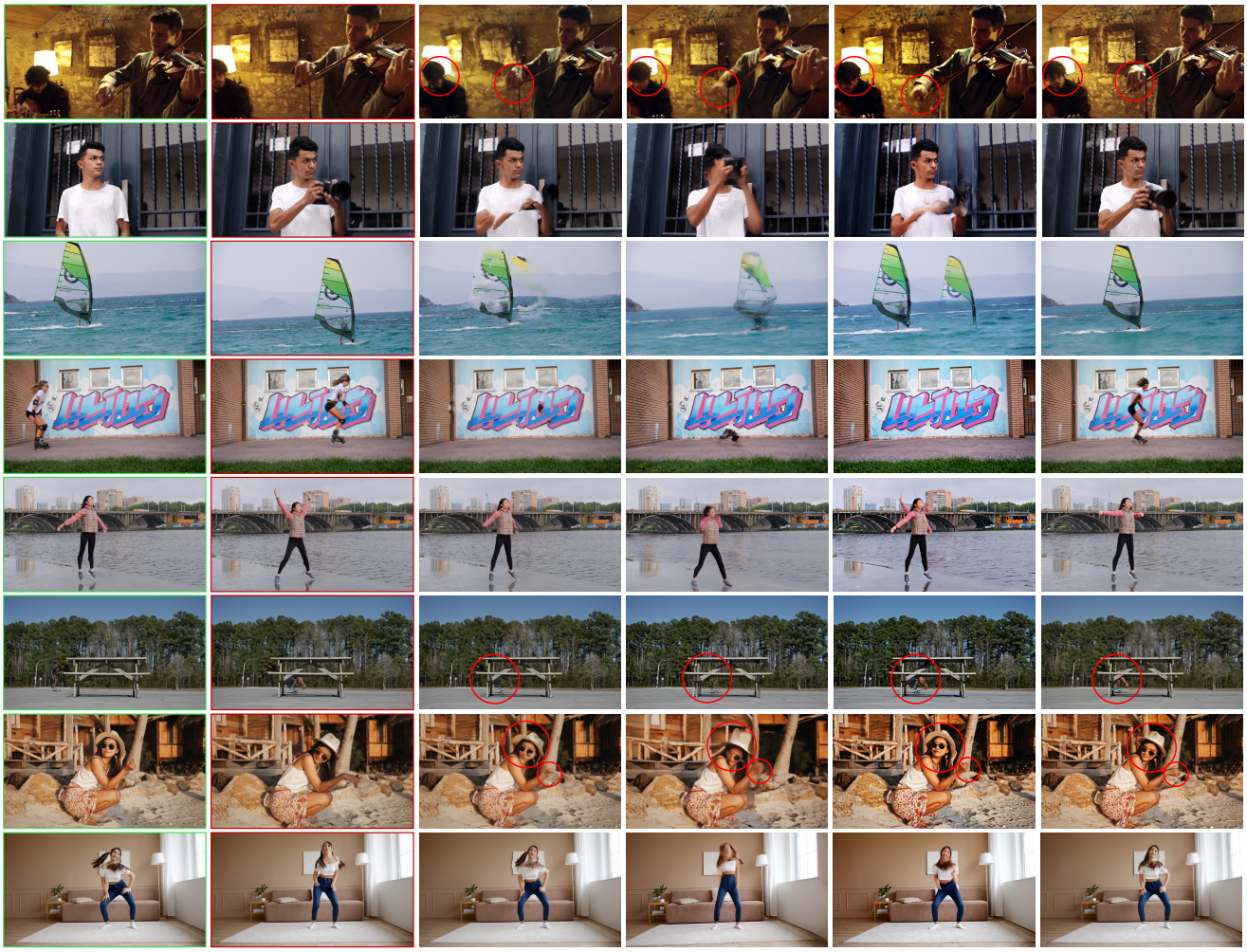

Figure 6. Qualitative evaluation on diverse scenes, where our FCVG is superior in texture details and coherent intermediate motions.

lustrates the superior performance of our method compared to other counterparts in terms of motion stability, consistency, and overall quality. While FILM produces smooth interpolation results for small motion scenarios, it struggles with large scale motion due to inherent limitations of optical flow, resulting in noticeable artifacts such as background and hand movement (in the first case). Generative models like TRF and GI suffer from ambiguities in fusion paths leading to unstable intermediate motion, particularly evident in complex scenes involving human and object motion. In contrast, our method consistently delivers satisfactory results across various scenarios. Even when significant

Figure 7. The effect of control weight γ. Red arrows indicate the movement directions. As the value of γ decreases, the diversity of intermediate motion increases, e.g., up-and-down swinging of the arm, while temporal stability is guaranteed for two cases. The videos are available in the project page.

occlusion is present (in the second case and sixth case), our method can still capture reasonable motion. Furthermore, our approach exhibits robustness for complex human actions (in the last case).

Computational efficiency. In Tab. 2, we evaluate the inference time of diffusion-based methods. DynamiCrafter generates 16 frames at resolution of 512×320, while the other methods generate 25 frames at resolution of 1024×576. DynamiCrafter exhibits advantages in inference time due to its single-pass denoising process. In comparison to TRF and GI, our FCVG facilitates easier alignment between forward and backward paths, ensuring that the results remain within the manifold as in Sec. 3.2. Consequently, although FCVG requires slightly less extra time for conditions extraction, it eliminates noise re-injection step and achieves satisfactory results with only 25 steps, in contrast to the requirement of 50 steps by the other methods. Real-time applications still pose challenges for generative inbetweening due to the high inference cost of pre-trained video generation models.

Table 2. Evaluation of computational efficiency.

Methods	$N \times (H, W)$	T	Time (s)
DynamiCrafter [52]	$16\times(512, 320)$	50	37
TRF [8]	$25\times(1024, 576)$	50	1230
GI [47]	$25\times(1024, 576)$	50	975
Ours	$25\times(1024, 576)$	25	523

4.3. Flexibility of Frame-wise Conditions

Our FCVG allows some flexibility to SVD by setting different values for γ. In Fig. 7, one can see that adjusting γ within a certain range has little impact on video stability but can lead to different directions of movement. Albeit fine-tuned for linear frame-wise conditions, FCVG enables users to specify non-linear motion trajectories, such as ease-in and ease-out motion trajectories. The video examples can be found in our project page.

4.4. Generalization to Animation Videos

We further evaluate the generalization ability of FCVG to hangle animation and lineart videos, whose data types do not appear in the fine-tuning dataset. As depicted in Fig. 8, our FCVG consistently produces visually appealing results, even for challenging scenarios such as head turns, liquid flow, and object deformations. This can be attributed to the robust guidance provided by frame-wise conditions that are beneficial for interpolation involving large motion in linearts and animations [23, 64].

Table 3. Ablation study on condition components.

	LPIPS (↓)	FID (↓)	FVMD (↓)	FVD (↓)
w/o Control	0.2485	27.55	7217.5	536.5
w/o Pose	0.1843	24.70	**5520.9**	446.1
w/o Matching	0.2124	24.17	6546.8	498.8
Full Model	**0.1832**	**24.05**	5607.2	**437.9**

4.5. Ablation Study

We conduct ablations to to discuss the components of conditions and control weight γ.

Condition components. In Tab. 3, 'w/o Control' denotes the exclusion of the entire frame-wise conditions control, while 'w/o Pose' and 'w/o Matching' indicate the removal of human pose and line matching conditions, respectively. The visual results are presented in Fig. 9, from which one can see that the line matching condition governs the overall motion of the scene, and the pose condition benefits details with human movements.

Control weight γ. The impact of γ has been discussed in Sec. 4.3. Based on the findings in Figure 7 and Table 4, FCVG is not very sensitive to the value of γ, and the weight $\gamma = 1$ proves to be suitable for the majority of scenarios.

5. Limitations and Future Work

Due to the dependence on line matching, incorrect matching results may arise when two input frames exhibit highly similar features, thereby impacting the quality of generated

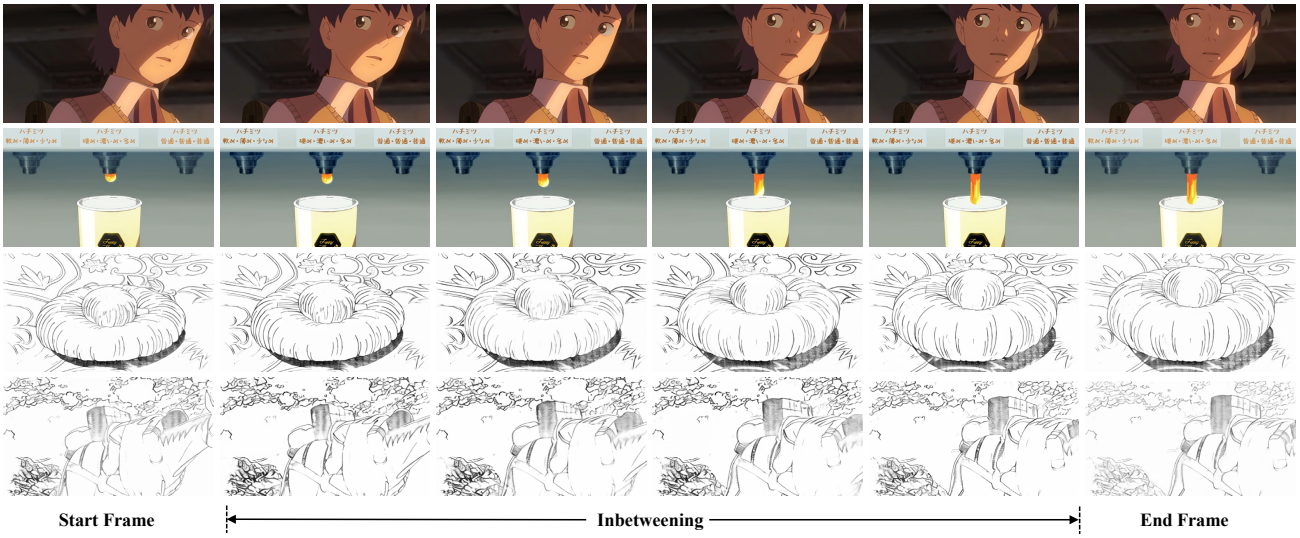

Figure 8. Inbetweening results on animations and linearts. FCVG exibits favorable performance without fine-tuning on these data types.

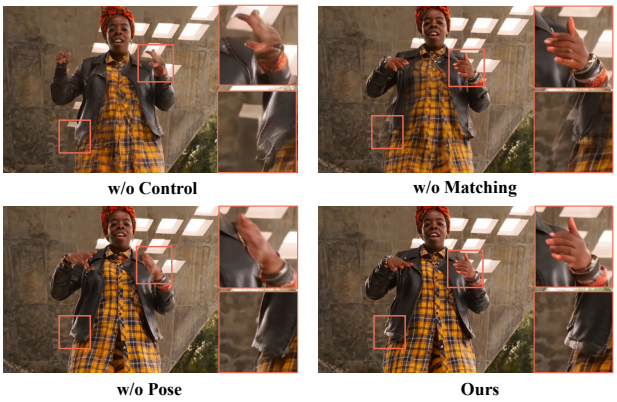

Figure 9. Ablation study on condition components.

Table 4. Quantitative analysis on control weight γ.

	LPIPS (↓)	FID (↓)	FVMD (↓)	FVD (↓)
$\gamma = 0.5$	0.1912	**23.80**	5920.0	**431.4**
$\gamma = 1.0$	**0.1832**	24.05	**5607.2**	437.9
$\gamma = 2.0$	0.1861	24.66	5726.9	441.1

frames. This limitation can be alleviated by manually reducing the control weight, as exemplified in the first instance in Fig. 10. However, when there is a significant difference between the two input frames, matched lines may be sparse, making simple adjustment of the control weight ineffective, as the second example in Fig. 10. This limitation could potentially be addressed in future research by replacing SVD with more robust I2V models. Furthermore, enhancing the interpolation process through diverse control conditions could offer improvements, e.g., incorporating user-specified drag effects like Dragdiffusion [39] or generating control conditions based on generative models using user-defined text inputs [48].

Figure 10. Failure cases. Incorrect matches and significant difference between input frames lead to intermediate artifacts, some of which can be mitigated by adjusting the control weight.

6. Conclusion

In this paper, we propose a frame-wise conditions-driven video generation method, FCVG, for generative inbetweening. To fully exploit the potential of a video generation model in producing temporally stable videos, an explicit condition is provided for each frame. Specifically, frame-wise conditions can be obtained by interpolating two initial conditions that contain matched lines extracted from two input frames, and subsequently injecting them into the video generation model alleviates the ambiguity of inbetweening path. Our FCVG is not very sensitive to the introduced control weight, and thus exhibits robustness across diverse scenes with a fixed weight setting. Moreover, FCVG is able to handle specific interpolation paths defined by users.

Acknowledgements

This work was supported in part by the National Key Research and Development Project (2022YFA1004100), the National Natural Science Foundation of China (62172127 and U22B2035), the Natural Science Foundation of Heilongjiang Province (YQ2022F004).

References

[1] Liu Anan, Su Yuting, Wang Lanjun, Li Bin, Qian Zhenxing, Zhang Weiming, Zhou Linna, Zhang Xinpeng, Zhang Yongdong, Huang Jiwu, and Yu Nenghai. Review on the progress of the aigc visual content generation and traceability. *Journal of Image and Graphics*, 29(06):1535–1554, 2024. DOI:10.11834/jig.240003. 2

[2] Simon Baker, Daniel Scharstein, James P Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92:1–31, 2011. 2

[3] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, pages 3703–3712, 2019. 2

[4] 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, 3, 4

[5] 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 *CVPR*, pages 22563–22575, 2023. 2

[6] Shuhong Chen and Matthias Zwicker. Improving the perceptual quality of 2D animation interpolation. In *ECCV*, pages 271–287, 2022. 2

[7] Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *AAAI*, pages 1472–1480, 2024. 1, 2, 3

[8] Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *ECCV*, pages 378–395, 2024. 1, 2, 3, 5, 6, 7

[9] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. In *ICLR*, 2025. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[11] 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. *NIPS*, 30, 2017. 5

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851, 2020. 2, 3

[13] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pages 8153–8163, 2024. 2

[14] Mengshun Hu, Kui Jiang, Zhihang Zhong, Zheng Wang, and Yinqiang Zheng. Iq-vfi: Implicit quadratic motion estimation for video frame interpolation. In *CVPR*, pages 6410–6419, 2024. 2

[15] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, pages 624–642, 2022. 2

[16] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, 2024. 5

[17] Zhilin Huang, Yijie Yu, Ling Yang, Chujun Qin, Bing Zheng, Xiawu Zheng, Zikun Zhou, Yaowei Wang, and Wenming Yang. Motion-aware latent diffusion models for video frame interpolation. In *ACM MM*, pages 1043–1052, 2024. 2, 3

[18] Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *CVPR*, pages 7341–7351, 2024. 2

[19] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, pages 9000–9008, 2018. 2

[20] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

[21] 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. 2

[22] 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. 1, 2

[23] Chunxu Liu, Guozhen Zhang, Rui Zhao, and Limin Wang. Sparse global matching for video frame interpolation with large motion. In *CVPR*, pages 19125–19134, 2024. 1, 2, 4, 7

[24] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024. 5

[25] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, pages 4463–4471, 2017. 2

[26] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 2

[27] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *CVPR*, pages 5437–5446, 2020. 2

[28] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, pages 261–270, 2017. 2

[29] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, pages 11410–11420, 2022. 5

[30] Rémi Pautrat, Iago Suárez, Yifan Yu, Marc Pollefeys, and Viktor Larsson. Gluestick: Robust image matching by sticking points and lines together. In *ICCV*, pages 9706–9716, 2023. 4, 5

[31] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 1, 2, 4

[32] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5

[33] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *ECCV*, pages 250–266, 2022. 5, 6

[34] 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. 2

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[36] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3

[37] Jiaming Shen, Kun Hu, Wei Bao, Chang Wen Chen, and Zhiyong Wang. Bridging the gap: Sketch-aware interpolation network for high-quality animation sketch inbetweening. In *ACM MM*, 2024. 2

[38] Liao Shen, Tianqi Liu, Huiqiang Sun, Xinyi Ye, Baopu Li, Jianming Zhang, and Zhiguo Cao. Dreammover: Leveraging the prior of diffusion models for image interpolation with large motion. In *ECCV*, pages 336–353, 2024. 1, 2

[39] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *CVPR*, pages 8839–8849, 2024. 8

[40] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *ICCV*, pages 14489–14498, 2021. 2

[41] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. In *CVPR*, pages 6587–6595, 2021. 2

[42] Li Siyao, Tianpei Gu, Weiye Xiao, Henghui Ding, Ziwei Liu, and Chen Change Loy. Deep geometrized cartoon line inbetweening. In *ICCV*, pages 7291–7300, 2023. 2

[43] Zheng Tianpeng, Chen Yanxiang, Wen Xinzhe, Wang Zhiyuan, and Li Yancheng. Research on diffusion model generated video datasets and detection benchmarks. *Journal of Image and Graphics*, 2024. DOI:10.11834/jig.240259. 2

[44] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 5

[45] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *NIPS*, 35:23371–23385, 2022. 2

[46] Wen Wang, Qiuyu Wang, Kecheng Zheng, Hao Ouyang, Zhekai Chen, Biao Gong, Hao Chen, Yujun Shen, and Chunhua Shen. Framer: Interactive frame interpolation. In *ICLR*, 2025. 2

[47] Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steven M Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. In *ICLR*, 2025. 1, 2, 3, 5, 6, 7

[48] Yuan Wang, Zhao Wang, Junhao Gong, Di Huang, Tong He, Wanli Ouyang, Jile Jiao, Xuetao Feng, Qi Dou, Shixiang Tang, et al. Holistic-motion2d: Scalable whole-body human motion generation in 2d space. *arXiv preprint arXiv:2406.11253*, 2024. 8

[49] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, pages 1–11, 2024. 2

[50] Guangyang Wu, Xin Tao, Changlin Li, Wenyi Wang, Xiaohong Liu, and Qingqing Zheng. Perception-oriented video frame interpolation via asymmetric blending. In *CVPR*, pages 2753–2762, 2024. 2

[51] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *ACM TOG*, 43(6):1–11, 2024. 2

[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 *ECCV*, pages 399–417, 2024. 2, 5, 6, 7

[53] Serin Yang, Taesung Kwon, and Jong Chul Ye. Vibidsampler: Enhancing video interpolation using bidirectional diffusion sampler. In *ICLR*, 2025. 1, 2, 3

[54] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, pages 4210–4220, 2023. 5

[55] 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. 2

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[58] 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. 2

[59] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *ICLR*, 2024. 2

[60] Zhihang Zhong, Gurunandan Krishnan, Xiao Sun, Yu Qiao, Sizhuo Ma, and Jian Wang. Clearer frames, anytime: Resolving velocity ambiguity in video frame interpolation. In *ECCV*, pages 346–363, 2024. 2

[61] Kun Zhou, Wenbo Li, Xiaoguang Han, and Jiangbo Lu. Exploring motion ambiguity and alignment for high-quality video frame interpolation. In *CVPR*, pages 22169–22179, 2023. 1

[62] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 5

[63] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, pages 145–162, 2024. 2

[64] Tianyi Zhu, Wei Shang, Dongwei Ren, and Wangmeng Zuo. Thin-plate spline-based interpolation for animation line inbetweening. *arXiv preprint arXiv:2408.09131*, 2024. 2, 4, 7