MOTION PRIOR DISTILLATION IN TIME REVERSAL SAMPLING FOR GENERATIVE INBETWEENING

Anonymous authors
Paper under double-blind review

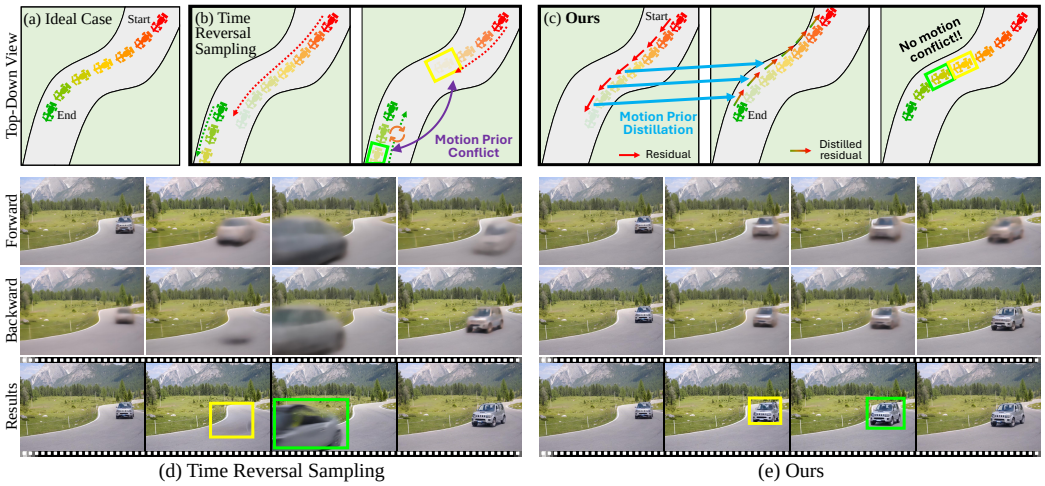

Figure 1: **An overview of our motion prior distillation.** (a) Ideal case of inbetweening task. (b) Motion prior conflict coming from existing time reversal sampling. (c) The proposed motion prior distillation. (d) A real example of time reversal sampling. We can observe ghosting artifact (yellow) and reverse play (green) caused by the motion conflict. (e) A result from the proposed method.

ABSTRACT

Recent progress in image-to-video (I2V) diffusion models has significantly advanced the field of *generative inbetweening*, which aims to generate semantically plausible frames between two keyframes. In particular, inference-time sampling strategies, which leverage the generative priors of large-scale pre-trained I2V models without additional training, have become increasingly popular. However, existing inference-time sampling, either fusing forward and backward paths in parallel or alternating them sequentially, often suffers from temporal discontinuities and undesirable visual artifacts due to the misalignment between the two generated paths. This is because each path follows the motion prior induced by its own conditioning frame. We thus propose Motion Prior Distillation (MPD), a simple yet effective inference-time distillation technique that suppresses bidirectional mismatch by distilling the motion residual of the forward path into the backward path. MPD alleviates the misalignment by reconstructing the denoised estimate of the backward path from distilled forward motion residual. With our method, we can deliberately avoid denoising the end-conditioned path which causes the ambiguity of the path, and yield more temporally coherent inbetweening results with the forward motion prior. Our method can be applied to off-the-shelf inbetweening works without any modification of model parameters. We not only perform quantitative evaluations on standard benchmarks, but also conduct extensive user studies to demonstrate the effectiveness of our approach in practical scenarios.

1 INTRODUCTION

Recent advances in diffusion models have significantly improved the performance of image and video generation tasks. In particular, image-to-video (I2V) diffusion models (Blattmann et al.,

2023a; Xing et al., 2024b; Bar-Tal et al., 2024; Yang et al., 2025b) demonstrate strong capabilities across diverse applications, as they can generate temporally coherent videos from a single conditioning frame. From a generative perspective, this progress has extended video frame interpolation to *generative inbetweening*, which aims to generate natural intermediate frames between two keyframes. However, I2V diffusion models are not directly applicable to bounded generation where both start and end frames serve as a dual-constraint.

To address this, recent studies have explored time reversal sampling, which employs the temporally forward/backward denoising paths conditioned on the start/end frames during the iterative reverse denoising process. This can be categorized into two approaches, namely parallel and sequential, according to how these two paths are integrated. In the parallel approach (Feng et al., 2024; Wang et al., 2025b; Zhu et al., 2025), samples from the forward and backward paths are denoised simultaneously at each denoising step, and then linearly interpolated to form the input for the next denoising step. In contrast, the sequential approach (Yang et al., 2025a) samples two denoising paths sequentially by inserting a single re-noising step between them.

However, the existing time reversal sampling methods cannot guarantee that the two temporal paths converge into a single coherent motion during the sampling process because each sample is obtained with the motion prior of its respective conditioning frame. This limitation arises from the nature of I2V models which are trained to predict forward consecutive frames. As shown in Fig. 1 (b) and (d), a backward path initialized at the end frame tends to generate forward-looking sequences instead of faithfully reconstructing historical frames. This highlights that the fundamental challenge lies not merely in how to connect the forward and backward paths, but in how to align them.

To this end, we aim to overcome this fundamental misalignment between two temporal paths by proposing a novel inference-time distillation approach, called **Motion Prior Distillation (MPD)**. Our key intuition is that the residual of the denoised estimates contains motion information induced by a given start frame. Inspired by this, during early denoising steps, our method distills the motion residual induced by the start frame into the backward path. Since our approach deliberately avoids denoising the end-conditioned path, we can drive the backward path to follow the time reversed motion residual of the forward path, thereby achieving bidirectional path alignment (See Fig. 1 (c) and (e)). This single path design effectively removes conflicting motion priors while preserving endpoint consistency, allowing two temporal paths to converge into coherent motions.

Through extensive evaluations, we demonstrate that our method consistently outperforms relevant methods including existing time reversal sampling strategies. In addition, since conventional metrics are not fully capable of evaluating temporal coherence and human preference, we further conduct user studies to validate its robustness under practical scenarios in the presence of complex motion patterns and large temporal displacements.

2 Related Works

Video Frame Interpolation. Video frame interpolation (VFI) aims to synthesize intermediate frames between two input frames while maintaining spatial and temporal coherence (Lyu et al., 2024; Kye et al., 2025). Supervised methods (Bao et al., 2019; Niklaus & Liu, 2018; Park et al., 2020; Lei et al., 2023; Kong et al., 2022; Li et al., 2023; Huang et al., 2022; Lu et al., 2022; Reda et al., 2022) that rely on estimating optical flows have been practically adopted due to their robust performance and interpretable motion trajectories. However, errors in estimated flows often lead to failures particularly under the scenes with occlusion or non-linear motion (Long et al., 2024). Recently, diffusion-based VFI methods (Danier et al., 2024; Voleti et al., 2022) attempt to leverage generative capabilities of diffusion models to improve the perceptual quality of interpolated frames. While these methods enhance perceptual fidelity, their performance still degrades under large temporal displacements between two frames.

Generative Video Inbetweening. With the advancement of video diffusion models (Ho et al., 2022b;a; Blattmann et al., 2023b;a), VFI has broadened into *generative inbetweening*, which is interested in the set of semantically plausible interpolations. Some approaches (Jain et al., 2024; Xing et al., 2024a;b; Wang et al., 2025a; Zhang et al., 2025) train diffusion models to condition on two input frames for interpolation, yielding greater robustness to ambiguous and large motion where traditional methods have struggled. While effective, they typically require substantial train-

ing resources. Other approaches leverage pre-trained large-scale I2V diffusion models and achieve remarkable performances by incorporating new sampling techniques. TRF (Feng et al., 2024) proposes a time reversal sampling strategy that fuses forward and backward denoising paths in parallel, each conditioned on the start and end frames. Building on this strategy, GI (Wang et al., 2025b) enhances reverse motion fidelity by fine-tuning a diffusion model through rotation of temporal self-attention maps to generate temporally reversed frames. Similarly, FCVG (Zhu et al., 2025) proposes a method that injects explicit line correspondences as frame-wise conditions to alleviate the ambiguity of inbetweening path. Meanwhile, ViBiDSampler (Yang et al., 2025a) introduces a new time reversal strategy that employs sequential sampling along forward and backward paths to achieve on-manifold generation of intermediate frames. However, all of them still operate with two independent motion priors from the start and end frames, so convergence to a single coherent trajectory is not guaranteed.

3 PRELIMINARIES

3.1 STABLE VIDEO DIFFUSION

We base our explanation on Stable Video Diffusion (SVD) (Blattmann et al., 2023a), which is widely adopted in time reversal sampling based methods. Specifically, SVD is a UNet-based latent video diffusion model that is built on EDM framework (Karras et al., 2022). At a reverse denoising step $t \in \{T, ..., 1\}$ with the noise level σ_t, the denoiser $\boldsymbol{D}_\theta$ predicts both the unconditional estimate $\hat{\boldsymbol{x}}_{0,\varnothing}$ and the conditional estimate $\hat{\boldsymbol{x}}_{0,\boldsymbol{c}}$ from the current noisy latent $\boldsymbol{x}_t$:

$$\hat{\boldsymbol{x}}_{0,\varnothing} = \boldsymbol{D}_\theta(\boldsymbol{x}_t; \sigma_t) \quad \text{and} \quad \hat{\boldsymbol{x}}_{0,\boldsymbol{c}} = \boldsymbol{D}_\theta(\boldsymbol{x}_t; \sigma_t, \boldsymbol{c}), \tag{1}$$

where $\boldsymbol{c}$ is the input condition. In EDM framework, the corresponding noise prediction model $\boldsymbol{\epsilon}_\theta$ and score prediction model $\boldsymbol{s}_\theta$ have the following relationship with the denoiser $\boldsymbol{D}_\theta$:

$$\boldsymbol{s}_\theta(\boldsymbol{x}_t; \sigma_t) = -\frac{\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t; \sigma_t)}{\sigma_t} = \frac{\boldsymbol{D}_\theta(\boldsymbol{x}_t; \sigma_t) - \boldsymbol{x}_t}{\sigma_t^2}. \tag{2}$$

To guide the sample toward the condition $\boldsymbol{c}$, the classifier-free guidance (CFG) (Ho & Salimans, 2021) mixes unconditional estimate $\hat{\boldsymbol{x}}_{0,\varnothing}$ with the conditional estimate $\hat{\boldsymbol{x}}_{0,\boldsymbol{c}}$:

$$\hat{\boldsymbol{x}}_{0,\boldsymbol{c}} \leftarrow (1 + w)\hat{\boldsymbol{x}}_{0,\boldsymbol{c}} - w\hat{\boldsymbol{x}}_{0,\varnothing}, \tag{3}$$

where $w \geq 0$ is a guidance strength. At each iteration, we can denoise the sample with Euler step, progressively denoising from Gaussian noise $\boldsymbol{x}_T$ to sample $\boldsymbol{x}_0$:

$$\boldsymbol{x}_{t-1} = \hat{\boldsymbol{x}}_{0,\boldsymbol{c}} + \frac{\sigma_{t-1}}{\sigma_t}\big(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_{0,\boldsymbol{c}}\big). \tag{4}$$

In particular, I2V models take the initial starting frame condition as input and generate videos with its motion prior. To reflect both start and end frame conditions, time reversal sampling process involves denoising two temporal paths with each corresponding frame condition.

3.2 TIME REVERSAL SAMPLING

Parallel Method. Parallel time reversal methods denoise the temporally forward/backward path conditioned on the start/end frame, and then fuse them to produce the intermediate frames (Feng et al., 2024; Wang et al., 2025b; Zhu et al., 2025). Let's denote $\boldsymbol{c}_{\text{start}}$ and $\boldsymbol{c}_{\text{end}}$ as the encoded latent conditions of the start and end frame, respectively. We can express the denoising step as:

$$\boldsymbol{x}_{t-1} = \alpha \boldsymbol{x}_{t-1,\boldsymbol{c}_{\text{start}}} + (1 - \alpha)(\boldsymbol{x}'_{t-1,\boldsymbol{c}_{\text{end}}})' \tag{5}$$

$$\text{s.t.} \quad \boldsymbol{x}_{t-1,\boldsymbol{c}_{\text{start}}} = \hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}} + \frac{\sigma_{t-1}}{\sigma_t}\left(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}\right) \tag{6}$$

$$\text{and} \quad \boldsymbol{x}'_{t-1,\boldsymbol{c}_{\text{end}}} = \hat{\boldsymbol{x}}'_{0,\boldsymbol{c}_{\text{end}}} + \frac{\sigma_{t-1}}{\sigma_t}\left(\boldsymbol{x}'_t - \hat{\boldsymbol{x}}'_{0,\boldsymbol{c}_{\text{end}}}\right), \tag{7}$$

where $(\cdot)'$ indicates a temporal flip along the time dimension and $\alpha \in [0, 1]$ refers to the interpolation weight. However, this method can suffer from off-manifold issues, where samples deviate from the learned data manifold. As a result, their linearly interpolated results often lead to oscillations and

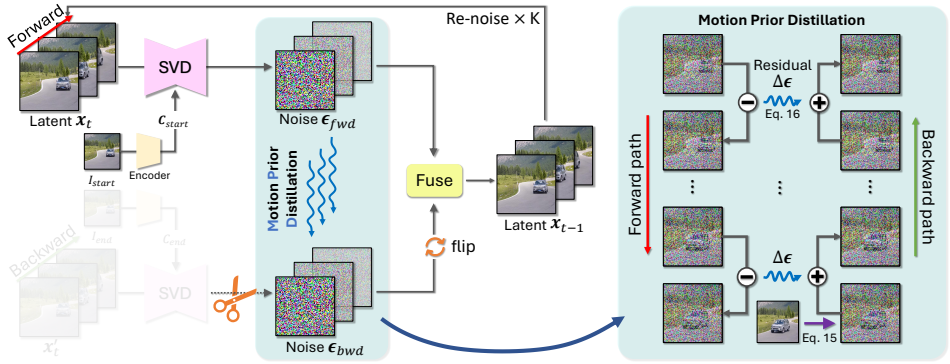

Figure 2: **Denoising process of our MPD.** Our algorithm is employed on time reversal sampling framework to distill forward motion prior into the backward path.

undesirable artifacts. Furthermore, they do not resolve the conflicting motion priors induced by the two conditions, so motion fidelity can still degrade.

Sequential Method. An alternative approach adopts the sequential time reversal sampling strategy (Yang et al., 2025a). Instead of fusing two temporal paths in parallel, this method sequentially denoises the forward and backward paths. On-manifold generation can be achieved by inserting a single re-noising step before switching from the forward to the backward path:

$$\boldsymbol{x}_{t-1,\boldsymbol{c}_{\text{start}}} = \hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}} + \frac{\sigma_{t-1}}{\sigma_t}\left(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}\right), \tag{8}$$

$$\boldsymbol{x}_{t,\boldsymbol{c}_{\text{start}}} = \boldsymbol{x}_{t-1,\boldsymbol{c}_{\text{start}}} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\,\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \boldsymbol{I}), \tag{9}$$

$$\boldsymbol{x}_{t-1} = (\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{end}}} + \frac{\sigma_{t-1}}{\sigma_t}\left(\boldsymbol{x}_{t,\boldsymbol{c}_{\text{start}}} - \hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{end}}}\right))'. \tag{10}$$

Unlike the parallel approach, this sequential structure maintains a more consistent and manifold-aligned path. Nevertheless, alternating two denoised paths results in conflicting motion priors, as each path relies on its own conditioning frame. This highlights the need to align two temporal paths without the motion prior conflicts.

4 Method

Given a pair of two frames $\{I_{\text{start}}, I_{\text{end}}\}$, our goal is to align two temporal paths with both temporal coherence and visual fidelity. Fig. 2 provides an overview of our method. To begin with, we recast the time reversal sampling process as an optimization problem to solve a bidirectional path misalignment problem. In this work, we present a simple yet effective approach, called Motion Prior Distillation (MPD) which propagates a motion residual from a forward path into a backward path.

4.1 Motivation

The existing time reversal sampling methods can be interpreted as a sampling procedure in which each denoising path approximately minimizes the following loss function $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}; \theta, \boldsymbol{c}_{\text{start}}, \boldsymbol{c}_{\text{end}}, \sigma) &= \left\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}; \sigma, \boldsymbol{c}_{\text{start}}) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}'; \sigma, \boldsymbol{c}_{\text{end}})'\right\|_2^2 \\ &= \left\|\frac{\boldsymbol{x} - \hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}}{\sigma_t} - \frac{(\boldsymbol{x}')' - (\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}_{\text{end}}})'}{\sigma_t}\right\|_2^2 \\ &= \frac{1}{\sigma_t^2}\left\|\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}} - (\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}_{\text{end}}})'\right\|_2^2.\end{aligned} \tag{11}$$

Here, the objective of Eq. (11) is to enforce the consistency between one path and a temporally reversed path in both directions by optimizing the noisy samples $\boldsymbol{x}$ as follows:

$$\bar{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}; \theta, \boldsymbol{c}_{\text{start}}, \boldsymbol{c}_{\text{end}}, \sigma), \tag{12}$$

where $\bar{\boldsymbol{x}}$ denotes the latent that minimizes the discrepancy between the two temporal paths. However, incompatible motion priors induced by two frame conditions introduce the ambiguity between the two denoising paths, especially in early denoising steps. Without resolving this problem, the loss $\mathcal{L}$ is optimized to make the misaligned path worse, causing implausible motions in generated videos.

For a clearer understanding of this problem, we show the denoised estimates at the midpoint of the time reversal sampling process in Fig. 1 (d). We take the forward denoised estimate $\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}$ and the backward denoised estimate $\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}_{\text{end}}}$, align the latter to the temporal order, and inspect their difference. Under the existing time reversal sampling, the two paths are optimized with the conflicting motion priors, which exhibits the implausible motion of the intermediate frames. When there is a significant gap between two motion priors, we could observe unrealistic motions like reverse play. Note that various types of visual artifacts come from incompatible motion priors, which will be discussed in Sec. 5.2.

4.2 Bidirectional Path Alignment with Motion Prior Distillation

Since subsequent denoising steps primarily focus on restoring high-frequency details, the previous works (Feng et al., 2024; Yang et al., 2025a) often fail to correct this misaligned trajectory. To resolve this issue, we introduce a single path sampling scheme that distills the motion prior induced by the start conditioning frame $\boldsymbol{c}_{\text{start}}$ into the backward path.

Here, our key intuition is that the forward motion residual Δ of the denoised estimates $\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}$ contain useful motion information, which can be written as:

$$\Delta\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}{}^{(i)} := \hat{\boldsymbol{x}}^{(i)}_{0,\boldsymbol{c}_{\text{start}}} - \hat{\boldsymbol{x}}^{(i-1)}_{0,\boldsymbol{c}_{\text{start}}}, \tag{13}$$

where $i \in \{2, ...N\}$ denotes the frame index, given N frames. Then, using the relation between $\boldsymbol{D}_\theta$ and $\boldsymbol{\epsilon}_\theta$ in Eq. (2), the residual of noise from the forward path $\Delta\boldsymbol{\epsilon}_{\text{fwd}}$ is given as:

$$\Delta\boldsymbol{\epsilon}_{\text{fwd}} = \frac{\Delta\boldsymbol{x}_t - \Delta\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}}{\sigma_t}, \tag{14}$$

where $\Delta\boldsymbol{x}_t = \boldsymbol{x}_t^{(i)} - \boldsymbol{x}_t^{(i-1)}$ represents the residual of the noisy sample $\boldsymbol{x}_t$. Now, given the encoded latent $\boldsymbol{z}_{\text{end}}$ of the end frame I_{end}, we initialize the first index of the backward denoised estimate $\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}_{\text{end}}}{}^{(1)}$ with $\boldsymbol{z}_{\text{end}}$ as:

$$\boldsymbol{\epsilon}^{(1)}_{\text{bwd}} = \frac{(\boldsymbol{x}'_t)^{(1)} - \boldsymbol{z}_{\text{end}}}{\sigma_t}. \tag{15}$$

Next, we reconstruct the backward noise residual $\boldsymbol{\epsilon}_{\text{bwd}}$ by cumulatively subtracting the forward noise residual from the initial backward noise $\boldsymbol{\epsilon}^{(1)}_{\text{bwd}}$:

$$\boldsymbol{\epsilon}^{(i)}_{\text{bwd}} = \boldsymbol{\epsilon}^{(1)}_{\text{bwd}} - \sum_{k=2}^{i} \Delta\boldsymbol{\epsilon}^{(k)}_{\text{fwd}}. \tag{16}$$

It is noteworthy that we should ignore the end frame condition $\boldsymbol{c}_{\text{end}}$. Therefore, we reformulate Eq. (2) as follows:

$$\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}^*_{\text{start}}} = \boldsymbol{x}_t - \sigma_t\,\boldsymbol{\epsilon}_{\text{bwd}}. \tag{17}$$

Here, the reconstructed $\boldsymbol{\epsilon}_{\text{bwd}}$ from the residual of the forward noise $\Delta\boldsymbol{\epsilon}_{\text{fwd}}$ provides us with a denoised estimate $\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}^*_{\text{start}}}$ from $\boldsymbol{c}^*_{\text{start}}$, which implies the flipped motion prior of $\boldsymbol{c}_{\text{start}}$. To curb off-manifold behaviors, we adopt CFG++ (Chung et al., 2025). Consequently, the Euler step of SVD in Eq. (4) denoises the sample $\boldsymbol{x}_t$:

$$\tilde{\boldsymbol{x}}_{0,c_{\text{start}}} = (1-\lambda)\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}} + \lambda(\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}^*_{\text{start}}})', \tag{18}$$

$$\boldsymbol{x}_{t-1} = \tilde{\boldsymbol{x}}_{0,c_{\text{start}}} + \frac{\sigma_{t-1}}{\sigma_t}(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_{0,\varnothing}), \tag{19}$$

where $\lambda \in [0, 1]$ serves as the interpolation scale. Note that during this process, we do not denoise the temporally backward path with the end frame condition $\boldsymbol{c}_{\text{end}}$. This enables the direct transfer of the forward motion prior toward the end-frame constraint without introducing additional sources of misalignment. In addition, the proposed update in Eq. (19) can be seen as satisfying the proposed objective in Eq. (11) in a relaxed form. In our single-path update, the end-conditioned estimate is

effectively replaced with the reconstructed estimate $\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}^*_{\text{start}}}$ distilled from the forward motion prior. Thus, the loss that we define in Eq. (11) is simplified as:

$$\mathcal{L}(\boldsymbol{x};\theta,\boldsymbol{c}_{\text{start}},\boldsymbol{c}^*_{\text{start}},\sigma_t)=\frac{1}{\sigma_t^2}\left\|\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}-(\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}^*_{\text{start}}})'\right\|_2^2. \tag{20}$$

By replacing the end frame condition $\boldsymbol{c}_{\text{end}}$ with $\boldsymbol{c}^*_{\text{start}}$, we can reduce the gap between the two temporal paths only with the start frame condition $\boldsymbol{c}_{\text{start}}$:

$$\bar{\boldsymbol{x}}=\arg\min_{\boldsymbol{x}}\ \mathcal{L}(\boldsymbol{x};\theta,\boldsymbol{c}_{\text{start}},\boldsymbol{c}^*_{\text{start}},\sigma), \tag{21}$$

This reformulated loss shows that the backward path no longer introduces independent motion prior; instead, it is aligned through the forward motion prior. This gives the denoiser $\boldsymbol{D}_\theta$ the opportunity that reconciles the original denoised path with its reconstructed counterpart within a timestep, producing a more stable trajectory.

Specifically, we disable our proposed process in later denoising steps. This choice follows the coarse-to-fine property of diffusion sampling as analyzed in the previous work (Kim et al., 2023): early denoising steps with large σ primarily represent global and low-frequency structure, whereas later denoising steps refine high frequency details. Hence, we apply our method during the earlier denoising steps with additional re-noising steps $k > 0$ to steer the trajectory onto the correct direction. Then, we switch to existing time reversal sampling to enhance endpoint consistencies, which will be discussed in Sec. 5.3. The detail of this distillation process is provided in Algorithm 1.

Algorithm 1 MOTION PRIOR DISTILLATION

Input: $\boldsymbol{x}_T\sim\mathcal{N}(0,\sigma_T^2\boldsymbol{I})$, $\boldsymbol{z}_{\text{start}},\boldsymbol{z}_{\text{end}}$, $\{\sigma_t\}_{t=1}^T$, Guidance scale λ, Renoising steps k, Distillation ratio γ
Output: Improved inbetweening results $\boldsymbol{x}_0$

$\boldsymbol{c}_{\text{start}},\boldsymbol{c}_{\text{end}}\leftarrow\text{encode}(\boldsymbol{z}_{\text{start}},\boldsymbol{z}_{\text{end}})$
for $t=T$ **to** $(1-\gamma)T$ **do**
  for $j=0$ **to** $k-1$ **do**
    $\hat{\boldsymbol{x}}_{0,\varnothing},\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}\leftarrow D_\theta(\boldsymbol{x}_t;\sigma,\boldsymbol{c}_{\text{start}})$ ▷ denoise forward path with $\boldsymbol{c}_{\text{start}}$ (Eq. (1))
    $\Delta\boldsymbol{x}_t^{(i)}\leftarrow\boldsymbol{x}_t^{(i)}-\boldsymbol{x}_t^{(i-1)}$ ▷ forward path residuals
    $\Delta\boldsymbol{x}_{0,\boldsymbol{c}_{\text{start}}}^{(i)}\leftarrow\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}^{(i)}-\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}^{(i-1)}$ ▷ forward denoised estimate residuals (Eq. (13))
    $\Delta\epsilon_{\text{fwd}}\leftarrow(\Delta\boldsymbol{x}_t-\Delta\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}})/\sigma_t$ ▷ forward noise residuals (Eq. (14))
    $\boldsymbol{x}'_t\leftarrow\text{flip}(\boldsymbol{x}_t)$ ▷ temporal flip
    $\epsilon_{\text{bwd}}^{(1)}\leftarrow((\boldsymbol{x}'_t)^{(1)}-\boldsymbol{z}_{\text{end}})/\sigma_t$ ▷ initialize first index of ϵ_{bwd} (Eq. (15))
    $\epsilon_{\text{bwd}}^{(i)}\leftarrow\epsilon_{\text{bwd}}^{(1)}-\sum_{k=2}^{i}\Delta\epsilon_{\text{fwd}}^{(k)}$ ▷ reconstruct ϵ_{bwd} (Eq. (16))
    $\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}^*_{\text{start}}}\leftarrow\boldsymbol{x}_t-\sigma_t\epsilon_{\text{bwd}}$ ▷ reconstruct $\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}^*_{\text{start}}}$ (Eq. (17))
    $\tilde{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}\leftarrow(1-\lambda)\hat{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}+\lambda(\hat{\boldsymbol{x}}'_{0,\boldsymbol{c}^*_{\text{start}}})'$ ▷ fuse two estimates (Eq. (18))
    $\boldsymbol{x}_{t-1}\leftarrow\tilde{\boldsymbol{x}}_{0,\boldsymbol{c}_{\text{start}}}+\frac{\sigma_{t-1}}{\sigma_t}(\boldsymbol{x}_t-\hat{\boldsymbol{x}}_{0,\varnothing})$ ▷ update with Euler step (Eq. (19))
    $\boldsymbol{x}_t=\boldsymbol{x}_{t-1}+\sqrt{\sigma_t^2-\sigma_{t-1}^2}\varepsilon$ ▷ re-noise
  end for
end for

5 EXPERIMENTAL RESULTS

5.1 EXPERIMENTAL SETTINGS

Evaluation Dataset. Following the previous works (Yang et al., 2025a; Wang et al., 2025b), we compare our method with relevant SOTA methods on two representative datasets. Specifically, we utilize 100 video-keyframe pairs from DAVIS dataset (Pont-Tuset et al., 2017), and 45 from Pexels [1]. To simulate typical inbetweening conditions where long-range temporal reasoning is required between sparsely spaced keyframes, those videos exhibit diverse and large motions such as driving, dancing, and so on.

Implementation Details. We plug our method into both TRF (parallel) and ViBiD (sequential) building on SVD-XT model of SVD (Blattmann et al., 2023a) on a single NVIDIA RTX 4090 GPU.

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

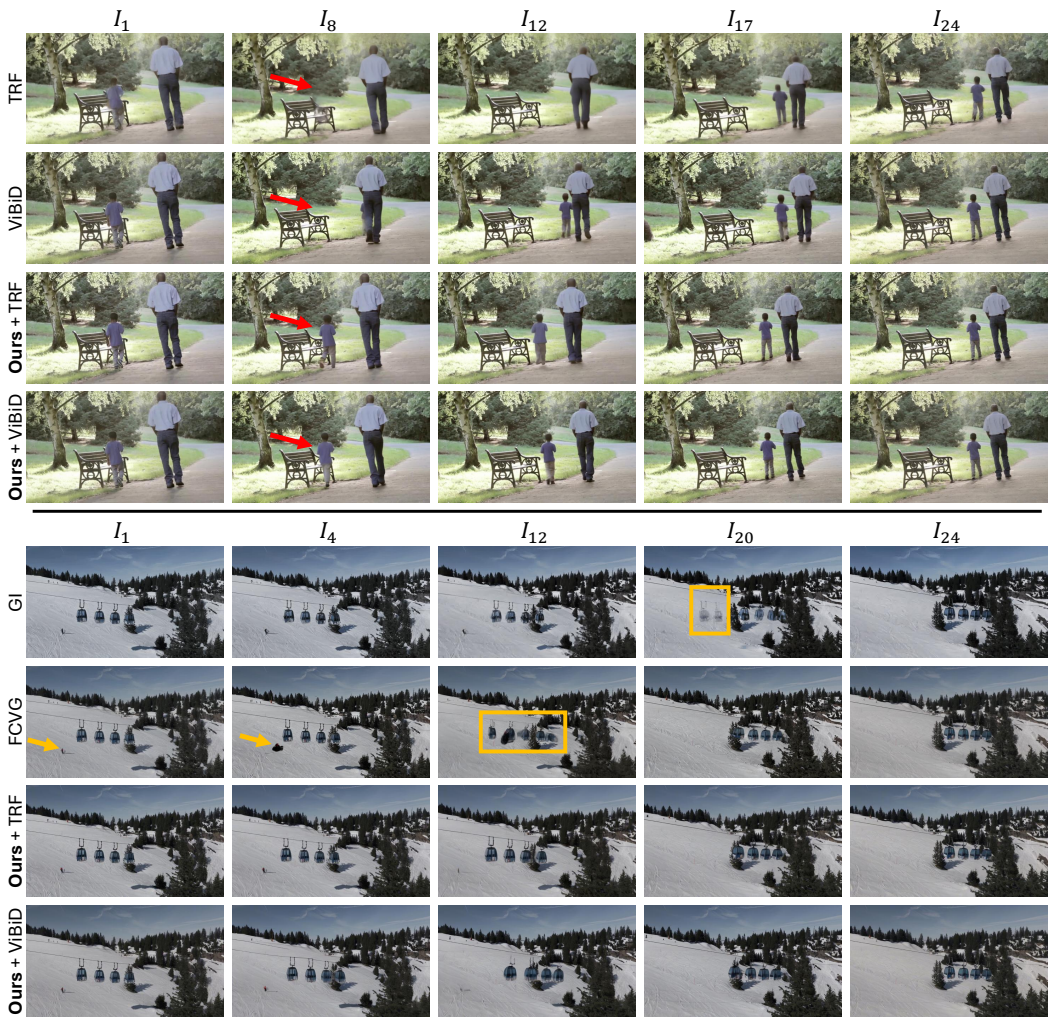

Figure 3: **Qualitative baseline comparisons.** TRF and ViBiD suffer from back-and-forth motion and intermittent disappearance, while GI and FCVG exhibit noticeable artifacts and ghosting effects. Our method yields more temporally consistent motion than the comparison methods. Additional examples are provided in the supplementary videos.

For the sampling process, we use the Euler scheduler with 25 timesteps with the default settings of SVD. Additionally, we configure each TRF and ViBiD with our settings: interpolation scale λ as 1.0 and 0.5, the number of re-noising steps k as 2 and 3, and the distillation step ratio γ as 0.2 and 0.3.

5.2 Comparative Results

As comparison methods, we choose representative time reversal sampling-based methods: TRF (Feng et al., 2024), GI (Wang et al., 2025b), FCVG (Zhu et al., 2025), and ViBiD (Yang et al., 2025a). We also include a flow-based VFI model, FILM (Reda et al., 2022), and a recent generative VFI model, DynamiCrafter (Xing et al., 2024b), for a broader comparison.

Quantitative Results. For quantitative evaluations, we use metrics for video frame interpolation, including FID (Heusel et al., 2017), FVD (Unterthiner et al., 2019), LPIPS (Zhang et al., 2018). FID and FVD measure the distances of generated frames/videos over ground-truth sequences. LPIPS assesses perceptual similarity at frame level. We also evaluate the overall quality of the videos using VBench (Huang et al., 2024a) and VBench++ (Huang et al., 2024b). VBench provides a comprehensive assessment across multiple dimensions such as subject consistency, background consistency, aesthetic quality, image quality, motion smoothness, and temporal flickering. VBench++ typically

Table 1: **Quantitative comparison results on DAVIS and Pexels dataset.** We compare against six baselines. *Ours + TRF* and *Ours + ViBiD* refer to our method applied to the parallel and sequential time reversal sampling schemes, respectively. Best is **bold**, and second best is underlined.

Method	DAVIS					Pexels				
	LPIPS ↓	FID ↓	FVD ↓	VB ↑	VB++ ↑	LPIPS ↓	FID ↓	FVD ↓	VB ↑	VB++ ↑
FILM	0.2946	55.160	1058.0	0.7978	**0.9740**	0.1157	43.935	761.60	0.8231	0.9734
DynamiCrafter	0.3158	46.739	678.92	0.7475	0.8735	0.2397	62.598	809.53	0.8211	0.9213
TRF	0.3127	56.894	674.31	0.7618	0.9352	0.2044	59.185	796.48	0.8008	0.9487
GI	0.2432	48.427	654.91	0.7747	0.9320	0.1114	47.990	476.93	0.8211	0.9566
FCVG	0.2347	38.997	621.82	0.7904	0.9353	0.1160	35.269	525.08	0.8245	0.9701
ViBiD	0.2492	39.883	559.49	0.7733	0.9387	0.1447	39.002	641.30	0.8130	0.9488
Ours + TRF	**0.2212**	**34.910**	612.17	**0.7992**	0.9330	0.1149	**34.470**	460.99	**0.8503**	**0.9862**
Ours + ViBiD	0.2220	37.241	**527.05**	0.7845	0.9474	**0.1028**	34.775	**412.66**	0.8235	0.9605

evaluates comprehensive performance of videos with a single reference frame. Because inbetweening must treat both endpoints symmetically, we compute VBench++ for start and end frame each, then average them.

As shown in Tab. 1, our method consistently outperforms the time reversal sampling-based methods, TRF and ViBiD, across all metrics. In particular, our method achieves significant improvements in terms of FID and FVD scores, highlighting its ability to produce temporally coherent sequences with smooth motion. For VBench++, FILM gets slightly better scores than our methods in DAVIS dataset. However, this can be attributed to flow-based warping that preserves local structures near each endpoint. This comes with blurry artifacts and weaker long-range temporal consistency, which yields the lower FVD score. Overall, our method effectively addresses the issue of conflicting motion priors and achieves both fidelity and perceptual quality over SOTA methods.

Qualitative Results. As shown in Fig. 3, our method produces a more temporally consistent motion than the comparison methods. In the first group, TRF and ViBiD fail to preserve the child's forward-walking trajectory. Near the end frames, the child appears to walk backward or partially vanish, indicating the misalignment issue between the two paths. In the second group, GI and FCVG exhibit oscillations and ghosting artifacts. In particular, FCVG, relying on line matching, results in ambiguous artifacts, which is observed with the the skier. In common, the comparison methods encounter difficulties when subjects' motion orientations is similar in both the start frame and end frame. Obviously, both GI and FCVG have tried to alleviate this issue, but are limited to linear motion in usual. In contrast, we validate that injecting forward motion residual into the backward path is enough to represent desirable object motions with fewer artifacts.

Table 2: **Comparison results of user study.** Best results are **bold**, second-best are underlined.

Method	Alignment ↑	Artifact ↓	Unrealistic ↓	Method	Alignment ↑	Artifact ↓	Unrealistic ↓
FILM	- 0.4060	62.74%	54.76%	FCVG	0.0988	20.36%	19.17%
DynamiCrafter	0.0190	34.64%	37.14%	ViBiD	- 0.0678	28.10%	25.24%
TRF	- 0.3119	28.09%	25.24%	**Ours** + TRF	**0.3060**	20.36%	22.62%
GI	0.1179	22.26%	13.57%	**Ours** + ViBiD	0.2440	**8.93%**	**9.88%**

User Study. To further evaluate human preference beyond the quantitative metrics, we conduct a comprehensive user study via Amazon Mechanical Turk (M-Turk) (Crowston, 2012). Each participant is presented with pairs of the start and end frames, followed by randomly 8 candidate videos generated by different methods. To avoid ordering bias, the display order of the videos is randomized for every sequence.

Our user study is designed with three types of questionnaires: (1) *Ranking:* participants are asked to rank videos in order of overall naturalness and temporal coherence, focusing on how plausibly the generated sequence links the start and end frames. These rankings are converted into scores in reciprocal order, ranging from 3.5 to -3.5. (2) *Artifact detection:* participants are asked to select all videos that exhibit noticeable visual artifacts, such as distortions, ghosting, or inconsistent textures. (3) *Unrealistic motion identification:* participants are asked to choose videos that contain unrealistic or physically implausible movements, which are closely related to perceptual plausibility.

Table 3: **Ablation results** for distillation steps ratio γ, re-noising steps k, and interpolation scale λ. *(a) - (c) Ours + ViBiD and (d) - (f) Ours + TRF*. Best results are **bold**.

(a) Distillation step ratio (γ)

γ	LPIPS ↓	FID ↓	FVD ↓
0.3	0.2421	39.855	574.05
0.2	**0.2220**	**37.241**	**527.05**
0.1	0.2374	39.949	545.25

(b) Re-noising steps (k)

k	LPIPS ↓	FID ↓	FVD ↓
1	0.2379	39.961	568.06
2	0.2341	39.368	545.25
3	**0.2220**	**37.241**	**527.05**

(c) Interpolation scale (λ)

λ	LPIPS ↓	FID ↓	FVD ↓
0.5	0.2242	37.837	539.99
1.0	**0.2220**	**37.241**	**527.05**
-	-	-	-

(d) Distillation step ratio (γ)

γ	LPIPS ↓	FID ↓	FVD ↓
0.3	**0.2212**	**34.910**	612.17
0.2	0.2238	35.408	576.01
0.1	0.2236	35.268	**573.13**

(e) Re-noising steps (k)

k	LPIPS ↓	FID ↓	FVD ↓
1	0.2246	35.149	**588.27**
2	**0.2212**	**34.910**	612.17
3	0.2248	35.765	662.75

(f) Interpolation scale (λ)

λ	LPIPS ↓	FID ↓	FVD ↓
0.5	**0.2212**	**34.910**	**612.17**
1.0	0.2264	35.612	654.72
-	-	-	-

We collected responses from a total of 30 participants across 28 randomly sampled video groups to ensure the statistical reliability of the study. As shown in Tab. 2, our method achieves the highest preference in the ranking task, while being selected least frequently in both the artifact and unrealistic motion categories. These results demonstrate that our approach outperforms competitive baselines in terms of perceptual plausibility and human preference, providing strong evidence of its effectiveness in practical scenarios.

5.3 Ablation Studies

We conduct ablation studies on DAVIS dataset to evaluate the impact of distillation step ratio γ, re-noising steps k, and interpolation scale λ. The results are as summarized in Table 3.

Within the parallel approach, LPIPS/FID are minimized at $\gamma = 0.3$ and $k = 2$, whereas FVD prefers weaker distillation and fewer re-noising steps. This is because TRF fuses the two conditional paths at every step, and the opposite motion prior is continually re-introduced after MPD, making the process more sensitive. Averaging the two paths partially cancels out the conflict and improves temporal coherence, while stronger MPD process can favor the frame-level fidelity at the cost of temporal smoothness.

For the sequential time reversal sampling, we observe a clear and consistent optimum at $\gamma = 0.2$, $k = 3$, and $\lambda = 1.0$, achieving the best scores. This indicates that strong early single-prior distillation with no interpolation is beneficial for the sequential method. Once the backward path is aligned to the forward motion prior in the early phase, subsequent steps rarely introduce conflicting priors, so increasing k steadily helps to improve the temporal and perceptual quality of videos.

6 Conclusion

In this work, we analyze the bidirectional path misalignment problem in existing time reversal sampling through the lens of optimization. Based on the analysis, we introduce **Motion Prior Distillation (MPD)**, a training-free sampling method that resolves motion prior conflict in existing time reversal sampling to enhance generative inbetweening task. MPD replaces two conflicting temporal priors with a single motion prior from the start frame, and distills it through the backward denoising path, yielding a coherent trajectory that satisfies both endpoint constraints. By Integrating our MPD into existing time reversal sampling methods, we demonstrate that MPD synergistically reduces temporal discontinuities and visual artifacts, and achieves more appealing results both quantitatively and qualitatively than SOTA methods.

Limitations and Future Directions. Our method relies on the assumption that the start frame provides a reliable motion prior. However, when the motion is highly ambiguous or involves large non-rigid changes, this assumption may break down. To overcome this limitation, we plan to introduce a mechanism that adaptively adjusts the influence of the motion prior according to the difficulty and complexity of the motion.

7 REPRODUCIBILITY STATEMENT

For reproducibility, we have included the source code and sample key frames in the supplementary materials and have provided pseudocode for our method in Algorithms 1. Our code will be publicly released if the paper is accepted.

REFERENCES

Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):933–948, 2019.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

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*, 2023a.

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 (CVPR)*, 2023b.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13-14, 2012. Proceedings*, 2012.

Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

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. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Proceedings of the Neural Information Processing Systems Workshop (NeurIPSW)*, 2021.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2022b.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024b.

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 (CVPR)*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2022.

Yulhwa Kim, Dongwon Jo, Hyesung Jeon, Taesu Kim, Daehyun Ahn, Hyungjun Kim, et al. Leveraging early-stage robustness in diffusion models for efficient and high-quality image synthesis. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023.

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 (CVPR)*, 2022.

Dahyeon Kye, Changhyun Roh, Sukhun Ko, Chanho Eom, and Jihyong Oh. Acevfi: A comprehensive survey of advances in video frame interpolation. *arXiv preprint arXiv:2506.01061*, 2025.

Pengcheng Lei, Faming Fang, Tieyong Zeng, and Guixu Zhang. Flow guidance deformable compensation network for video frame interpolation. *IEEE Transactions on Multimedia*, 26:1801–1812, 2023.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Libo Long, Xiao Hu, and Jochen Lang. Learning to handle large obstructions in video frame interpolation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

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 (CVPR)*, 2022.

Zonglin Lyu, Ming Li, Jianbo Jiao, and Chen Chen. Frame interpolation with consecutive brownian bridge diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

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)*, 2018.

Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

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.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

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

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2022.

Wen Wang, Qiuyu Wang, Kecheng Zheng, Hao OUYANG, Zhekai Chen, Biao Gong, Hao Chen, Yujun Shen, and Chunhua Shen. Framer: Interactive frame interpolation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025a.

Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steve Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025b.

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

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024b.

Serin Yang, Taesung Kwon, and Jong Chul Ye. VibiDSampler: Enhancing video interpolation using bidirectional diffusion sampler. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025b.

Guozhen Zhang, Yuhan Zhu, Yutao Cui, Xiaotong Zhao, Kai Ma, and Limin Wang. Motion-aware generative frame interpolation. *arXiv preprint arXiv:2501.03699*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Tianyi Zhu, Dongwei Ren, Qilong Wang, Xiaohe Wu, and Wangmeng Zuo. Generative inbetweening through frame-wise conditions-driven video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.