Explorative Inbetweening of Time and Space

Haiwen Feng[†1], Zheng Ding[†3], Zhihao Xia[2], Simon Niklaus[2],
Victoria Abrevaya[1], Michael J. Black[1], and Xuaner Zhang[2]

[1] Max Planck Institute for Intelligent Systems
[2] Adobe
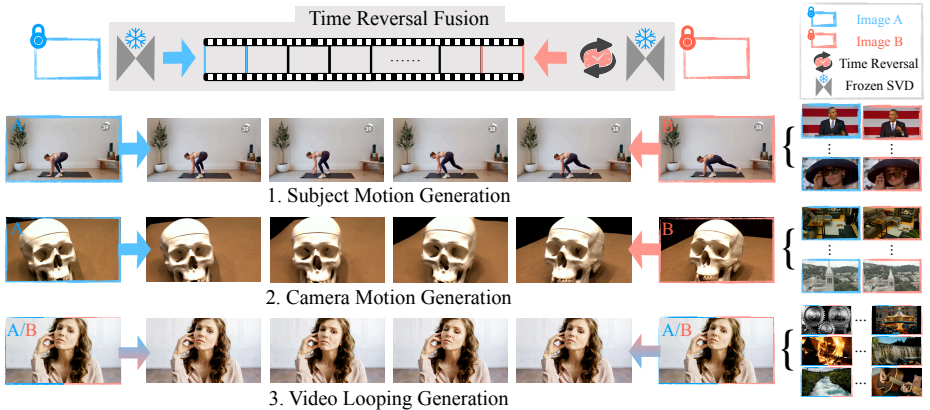[3] University of California San Diego

Fig. 1: Bounded generation in three scenarios: 1) Generating subject motion with the two bound images capturing a moving subject. 2) Synthesizing camera motion using two images captured from different viewpoints of a static scene. 3) Achieving video looping by using the same image for both bounds. We propose a new sampling strategy, called Time Reversal Fusion, to preserve the inherent generalization of an image-to-video model while steering the video generation towards an exact ending frame.

Abstract. We introduce bounded generation as a generalized task to control video generation to synthesize arbitrary camera and subject motion based only on a given start and end frame. Our objective is to fully leverage the inherent generalization capability of an image-to-video model without additional training or fine-tuning of the original model. This is achieved through the proposed new sampling strategy, which we call Time Reversal Fusion, that fuses the temporally forward and backward denoising paths conditioned on the start and end frame, respectively. The fused path results in a video that smoothly connects the two frames, generating inbetweening of faithful subject motion, novel views of static scenes, and seamless video looping when the two bounding frames

† H. Feng partially performed and Z. Ding performed the work while interning at Adobe.

are identical. We curate a diverse evaluation dataset of image pairs and compare against the closest existing methods. We find that Time Reversal Fusion outperforms related work on all subtasks, exhibiting the ability to generate complex motions and 3D-consistent views guided by bounded frames. See project page at `https://time-reversal.github.io`.

1   Introduction

The recent success of large image-to-video (I2V) models [2,5,8] suggests that they have remarkable generalization capacity, and can hallucinate complex dynamic scenes from just a single image after exposure to millions of videos. However, they lack an important form of user control: generation between two endpoints. We often want to be able to generate the frames in between two image frames, which may be captured quite far apart in space or time. We call this general task of inbetweening from sparse endpoint constraints *bounded generation*. Existing I2V models are incapable of doing bounded generation, since they lack the ability to control the motion towards an exact end state. What we seek is a form of generalized control for video generation, that can synthesize both camera and object motion without making underlying assumptions on the type of dynamics. For example, when the initial and final frames capture a dynamic subject, the task is to generate in-between object motion (e.g. Fig. 1, row 1 shows complex articulated human motion). In instances where the snapshots capture a static scene from different viewpoints, the task is to fill in the camera trajectory (e.g. Fig. 1, row 2 illustrates a rigid scene). If the start and end frames are identical, the task is to produce a looping video that starts and ends at the same frame (e.g. Fig. 1, row 3 shows natural head, hand, and eye movements). We define a general method to address all these problems in a unified manner.

On the surface, bounded generation is akin to several classic topics in the field, but with important differences: 1) Frame interpolation [41, 44] aims to fill in content between frames by taking the shortest path along the arrow of time, whereas bounded generation exploits different plausible trajectories and can handle much larger frame gaps. 2) Novel view synthesis from wide-baseline stereo [12,13] relies on generating new camera viewpoints through triangulation, necessitating the visibility of 3D points in both frames and the knowledge of camera poses, while bounded generation can generate novel views for any points present in either frame without any pose information. 3) Single-image video looping [23, 29, 35] hallucinates a flow field using specific motion models and requires scene segmentation, while bounded generation applies to arbitrary object motion without localizing any region. These previous methods cannot solve the general bounded generation problem because they are constrained by inductive biases originating from either the domain-specific training data, or the embedded physical model that addresses only certain types of motion. We seek instead a method that can do "explorative inbetweening", providing a set of possible solutions that connect one endpoint to the other.

In this paper we make two contributions towards this end. First, we introduce the task of bounded generation, that unifies and extends three computer vision

tasks: frame interpolation, novel-view synthesis, and cinemagraph. Second, we bring bounded generation to I2V models by proposing a new sampling strategy: *Time Reversal Fusion* (TRF). TRF is training- and tuning-free, thus it can harness the inherent generation capability of an I2V model. We are motivated by empirical findings that existing I2V models are trained to generate content along the arrow of time, thereby lacking the ability to propagate image conditions backwards in time to preceding frames. TRF simultaneously denoises the temporally forward path conditioned on a given start frame and the backward path conditioned on an end frame, followed by the fusion of these two paths into a unified trajectory. We show that fusing the forward and backward paths can be achieved through a joint optimization towards both paths, resulting in a simple weighted average form of the denoised latents. TRF combines bidirectional trajectories without relying on pixel correspondence or motion assumptions, resulting in video generation that predictably ends with the bounding frame. While simple in design, the solution is not naive, and we demonstrate that alternative baselines are not sufficient to achieve bounded generation. Constraining both ends of the generated video make the problem challenging and can quickly become stuck in local minima, resulting in abrupt frame transitions. To mitigate this, we introduce stochasticity through noise re-injection to ensure smooth frame transitions.

To evaluate videos created with bounded generation, we curate a dataset of 426 image pairs as the start and end bounds. These images contain snapshots ranging from multi-view imagery of complex static scenes to kinematic motions of humans and animals, as well as stochastic movement like fire and water. As our experiments show, bounded generation, when combined with large I2V models, not only opens up the possibility of numerous downstream tasks that were previously deemed hard, but also enables probing into the generated motion to understand the 'mental dynamics' of I2V models. In summary, we propose:

1. the task of *bounded generation* for large image-to-video (I2V) models, where the goal is to synthesize the in-between frames given an arbitrary context by leveraging the generalization ability of these models.
2. a novel sampling method that enables pretrained I2V models to perform bounded generation without fine-tuning or training.
3. a dataset for bounded generation and a systematic evaluation of both our method and the closest existing work. The empirical results indicates substantial improvements of our method over the state of the art. We will release the code and data to the research community for academic purposes.

2   Related Works

2.1   Control-based Video Generation

Diffusion-based video generation methods have recently achieved impressive results, with a focus on controllability—providing user-friendly ways to generate

videos under controlled conditions. Initial efforts, inspired by the success of text-to-image models, concentrate on text-to-video generation [6,15,22,33,35,54,59]. Recognizing the limitations of text prompts in capturing complex scenes, later research [5, 17, 28, 29] leverages image-conditioned video generation for a more direct approach. Notably, cinemagraph generation techniques [23, 29, 34, 35] focus on transforming still images into animated looping videos but are typically restricted to Eulerian motion, limiting their applicability to scenes with fluid continuous dynamics. Further innovations have introduced additional control mechanisms for video generation, such as structural guides [14,61], edge maps [27,61], and dynamic controls like motion trajectories [53,55,57], camera poses [55], and sequences of human poses [24]. Our work introduces a unique concept, *bounded generation*, as a novel control mechanism for video generation, leveraging both start and end frames to guide the generation process. Using the same frame as the start and end guidance, our approach also enables the creation of looping videos, without relying on predefined motion models.

2.2   Bounded Frame Generation

Several existing sub-fields can be viewed as special cases of bounded frame generation. Our formulation unifies these, addressing them with a cohesive framework that leverages a large video diffusion model.

Frame Interpolation. There is an extensive history of research on frame interpolation, with early work focusing on finding heuristics for block-level motion compensation [9,18], while current research leverages machine learning instead [31,41,41]. Regardless of the underlying approach, video frame interpolation aims to find the most probable arrow of time that occurred between two frames. Looking at it differently, given two input frames it postulates that all motion follows the shortest path, which implies a single solution. This holds true even for techniques that aim to perform extreme versions of frame interpolation [44,48], or ones that take more than two input frames and then assume a quadratic path [30,56]. In contrast, we target inbetweening of distant inputs to increase the diversity of solutions. On this note, such dissimilar inputs go beyond the typical scenario that current frame interpolation techniques can handle.

Sparse Novel View Synthesis. Recent advancements in novel view synthesis, sparked by the introduction of neural radiance fields (NeRF) [37], have made significant strides [4,26,39,49,50]. The core idea is to utilize correspondences across multiple images with small baseline separations to reconstruct 3D geometry and appearance for generating new views of the observed 3D points. Efforts have been made to achieve synthesis from very sparse observations [12,16,40,58,62], often leveraging priors learned from large datasets, including image priors in diffusion models [16,62]. Du et al. [13] introduce a method for rendering new views from wide-baseline stereo pairs by employing cross-attention to match epipolar features between two frames. However, this approach requires known camera

intrinsics and extrinsics and struggles with occluded points not visible in both views. In contrast, our method diverges fundamentally from existing approaches to novel view synthesis. We do not rely on explicit 3D geometry modeling or a rendering pipeline. Instead, we generate novel views, even for 3D points visible in only one view, by harnessing the capabilities of a video diffusion model to perform bounded generation from two views of a static scene, without needing any information on camera poses.

2.3   Sampling-based Guided Image Generation

Adopting new sampling techniques for manipulating the generative process of pretrained diffusion models is effective in a range of controlled image generation tasks [3, 10, 20, 32, 36, 38, 51]. For instance, RePaint [32] integrates observed regions into the sampled area during denoising for inpainting. SDEdit [36] applies noise to a user's stroke-guided image, then denoises it using a pre-trained diffusion model. To create large content images like panoramas, DiffCollage [60] and MultiDiffusion [3] use a pre-trained diffusion model to generate segments of the content in parallel, merging the outputs at each denoising step for a cohesive large-scale image. Our approach aligns with these concepts but targets video generation. By running two parallel I2V generations guided by start and end frames, we merge the outputs of each denoising step. This produces a coherent video bounded by the initial and final frames, marking a unique application of manipulating the generative process for video creation.

3   Method

The goal of this work is to enable training-free *bounded generation* within a diffusion image-to-video (I2V) framework – that is, the use of contextual information in the form of a start and an end frame. We focus in particular on Stable Video Diffusion [5] (SVD) which has shown impressive realism and generalization capacity for *unbounded* video generation. While bounded generation can in principle be addressed by fine-tuning the model with paired data, this would inevitably lead to a compromise in model generalization [42]. Therefore, our study focuses on a training-free approach.

We begin by reviewing SVD in Sec 3.1, and discuss two alternative and straightforward strategies for training-free bounded generation: condition manipulation and inpainting. We then analyze the reasons why these approaches are insufficient for our setting. Motivated by this, we present our proposed approach, Time Reversal Fusion, in Section 3.2.

3.1   Preliminaries

Stable Video Diffusion (SVD) has achieved state-of-the-art performance in image-to-video generation, producing high-fidelity video sequences. Given an initial input frame, SVD generates a sequence of N video frames, denoted by $\boldsymbol{x} =$

$\{x^0, x^1, ..., x^{N-1}\}$. This sequence is constructed through a denoising diffusion process where, at each denoising step t, a conditional 3D-UNet Φ is used to iteratively denoise the sequence:

$$\boldsymbol{x}_{t-1} = \Phi(\boldsymbol{x}_t, t, c). \tag{1}$$

Here, c represents the condition of the initial input frame, which includes its CLIP [43] embedding as well as its VAE latent, and ensures a consistent reference to the original frame throughout the video generation process. Note that SVD operates within a latent diffusion framework, meaning that, at the conclusion of the denoising steps, each frame within $\boldsymbol{x_0}$ is subsequently decoded back to pixel space using a VAE decoder.

There are two straigthforward solutions for incorporating *bounded* generation within SVD: (1) condition manipulation, and (2) temporal inpainting. We discuss each of these in the following, and elaborate on the reasons why these simple approaches do not work in our setting.

Condition manipulation. As mentioned, SVD conditions each frame-wise latent noise on the initial input frame. A straightforward solution to incorporate end-frame control is to condition the beginning of the sequence on the start frame, while the later part is conditioned on the end frame. This can be achieved by conditioning on a linear interpolation between the first and last frame, with the weight set as a function of time. We implemented this and observed that the video produced did not align with the condition set on the later frames; in other words, the condition specified during the later frames was largely ignored by the model. An example of this is shown in Fig. 2, where the top row is generated with the aforementioned strategy, while the middle row is generated by setting random noise as the end frame. In both cases we can observe a similar generation, suggesting that *only the initial frames are responsible for the structure and dynamics of the output*. We hypothesize that this is due to the nature of the training data, which was constructed to ensure significant disparity between the start and last frame. Due to this, and as observed in our own experiments, the network is trained to ignore the conditioning image on the latter frames, and focus instead on following the right dynamics based on the previous frames. Therefore, the intuitive idea of modulating towards the end frame by altering the condition is not a viable solution.

Temporal Inpainting. A second alternative for end-frame control is to apply diffusion image inpainting techniques [32] to video data along the temporal axis. However, there are fundamental differences between videos and images that make these methods not applicable. First, images are static, and hence do not exhibit a preference for a direction, whereas videos are embedded with sequential influences that dictate a flow of time. The architecture of SVD incorporates positional encoding of time stamps, imposing a temporal order to the generated video content. The learning process is designed to maintain temporal consistency, starting with a condition image (the first frame). As the sequence progresses, the later
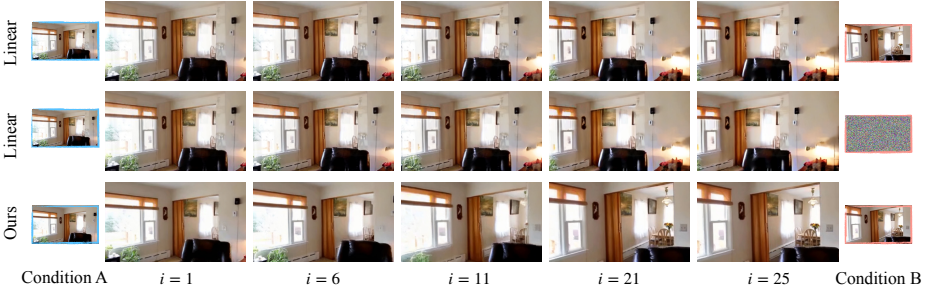
Fig. 2: The impact of conditioning on video generation. We experiment with different conditioning strategies and show their effects on the generated video. (Row 1) Using a linear interpolation of A and B as the image condition, the generated video does not end at B. (Row 2) Swapping B with random noise yields similar results, indicating B imposes minimal influence on the generated contents. (Row 3) With the proposed time reversal fusion, our generated video starts with A and ends at B.
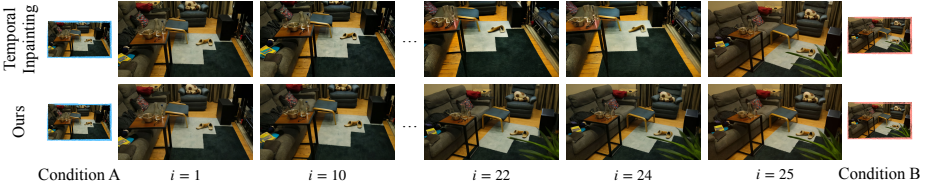

Fig. 3: Image inpainting strategies do not apply to videos. We follow the standard diffusion inpainting method by replacing the last frame with the target frame at each denoising step. However, this results in a video that satisfies the end frame condition but with abrupt content changes, as indicated in the last frames in row 1. Our method, on the other hand, generates a smooth video (row 2) that ends at the given condition.

frames are trained to align more closely with the preceding frames rather than adapting to subsequent frames (as also analyzed in the paragraph above). In other words, each frame is temporally consistent primarily with its preceding frames, establishing an influence that follows the arrow of time. The reason why this hinders the application of an inpainting method is illustrated in the top row of Fig. 3. Here, the last frame is replaced with the target end-frame (with corresponding noise) at every denoising step, as done in e.g. [32, 36]. This standard inpainting strategy leads however to a video where the end frame is correctly satisfied, but the rest do not naturally lead to it, resulting in abrupt changes.

To summarize, the unique characteristics of video data, coupled with the model architecture and learning patterns, highlight why techniques that are effective for images cannot be applied to videos. The directional bias, time-encoded architecture, and forward-leaning temporal consistency are integral to how the model processes and generates video content.

3.2   End-Frame Guidance using Time Reversal Fusion

Based on our analysis, we observe that SVD follows a *forward* arrow of time, where the conditioning image initializes the video but its influence decreases over time. The challenge then lies in introducing a *backward* influence to video generation, without fine-tuning the model.

Algorithm 1 Method

Input:
$\boldsymbol{x}_T$: Random initialized noise;
c_s: Start frame conditions;
c_e: End frame conditions;
M: Number of noise injection steps;
t_0: Cutoff timestep for noise injection;
σ_t: Std at timestep t.
for $t = T..1$ **do**
    $\boldsymbol{x}_{t-1,s} = \Phi(\boldsymbol{x}_t, c_s, t)$
    $\boldsymbol{x}_{t-1,e} = \Phi(\boldsymbol{x}_t, c_e, t)$
    $\boldsymbol{x}_{t-1} = \text{Fuse}(\boldsymbol{x}_{t-1,s}, \boldsymbol{x}_{t-1,e})$ ▷ Eq.3
    if $t > t_0$ **then**
        for $m = 0..M - 1$ **do**
            $\epsilon \sim \mathcal{N}(0, \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\boldsymbol{I})$
            $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \epsilon$
            $\boldsymbol{x}_{t-1,s} = \Phi(\boldsymbol{x}_t, c_s, t)$
            $\boldsymbol{x}_{t-1,e} = \Phi(\boldsymbol{x}_t, c_e, t)$
            $\boldsymbol{x}_{t-1} = \text{Fuse}(\boldsymbol{x}_{t-1,s}, \boldsymbol{x}_{t-1,e})$ ▷ Eq.3
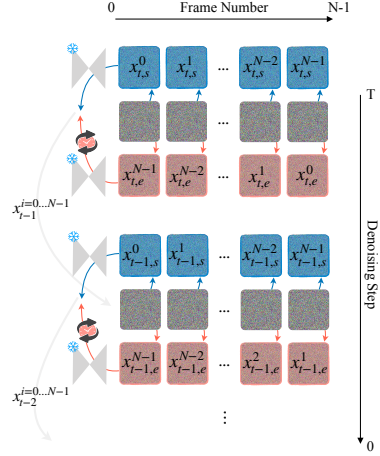   **end for**
    end if
end for

Fig. 4: Pseudo code and illustration of Time Reversal Fusion. Initiated with identical noise and conditioned on the start and end frame, the two paths undergo the SVD (frozen) denoiser. The forward path is fused with a time reversed backward path to produce the output for the subsequent step. Noise is re-injected to the fused output to add stochasticity in the sampling process.

Our key idea is to generate two *reference trajectories*: one conditioned on the starting frame c_s, which we call forward generation, and another one conditioned on the ending frame c_e, called the backward generation. We initiate both forward and backward denoising paths from the same noise, conditioned on the corresponding frame. At each denoising step, the denoised output from the backward path is reversed such that its dynamics integrate better with the forward one. Drawing inspiration from large content generation works [3,60], we then composite both models into a single coherent video using a single sampling strategy Ψ, designed to align each denoising path closely with SVD's reference trajectories through the following optimization objective:

$$\Psi(\boldsymbol{x}_t, c_s, c_e) = \arg\min_{\boldsymbol{x}} \|\boldsymbol{\alpha} \odot (\boldsymbol{x} - \Phi(\boldsymbol{x}_{t+1}, c_s))\|^2 + \|(\boldsymbol{1} - \boldsymbol{\alpha}) \odot (R(\boldsymbol{x}) - \Phi(R(\boldsymbol{x}_{t+1}), c_e))\|^2 .$$
(2)

Here, $R(\boldsymbol{x})$ represents the reverse of the sequence $\boldsymbol{x}$, and $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, ..., \alpha_{N-1}\}$ denotes a per-frame weighting factor, which is adjusted based on the proximity to the start or end guidance frame, either linearly or exponentially.

This optimization approach, a form of least squares approximation, naturally leads to a closed-form solution representing a weighted average of the forward and backward generations:

$$x_t^n = \alpha_n x_{t,s}^n + (1 - \alpha_n) x_{t,e}^{N-n-1}. \tag{3}$$

In this formula, $x_{t,s}^n$ is the n-th frame from the SVD denoising UNet conditioned on the start frame, while $x_{t,e}^n$ corresponds to the n-th frame conditioned on the end frame.

Enhancing Fusion with Noise Re-Injection While time reversal fusion at each step effectively facilitates bounded generation, we occasionally observe blending cuts or undesirable artifacts, as highlighted in Fig. 5, row 1. These issues often stem from significant disparities in the dynamics between the forward and backward generation processes. When such discrepancies are pronounced, the solution proposed in Eq. 3 may result in poor quality, attributable to the lack of harmony between the two processes. Although subsequent denoising steps with SVD have the potential to enhance the quality, this improvement is typically short-lived, succumbing once again due to the same integration issue. Whereas the original denoising diffusion process ensures incremental quality enhancement with each step, the introduction of information and constraints from an alternate process can inadvertently alter the sampling trajectory.

To mitigate these discrepancies, we advocate for the introduction of additional stochasticity into the sampling process, thereby affording the network more opportunities to reconcile the two generative pathways. The EDM sampling strategy [25] employed during SVD inference incorporates a 'churn' term to introduce noise at each step. We empirically find that it is not sufficient (Row 2 in Fig. 5), as the small amount of noise introduced at each step is not strong enough to influence the generation at earlier denoising stages. To address this, we propose to augment each denoising step by injecting supplementary noise, followed by a denoising phase, iterating this several times before advancing to the subsequent denoising step. This noise injection approach allows the sampling method to realign the fused generation at each step closer to the sampling trajectory defined by the pre-trained SVD, resulting in bounded generations that have similar visual fidelity. Our algorithmic approach is depicted in Fig. 4.

4 Experiments

We evaluate here our approach for bounded generation using the proposed Time Reversal Fusion (TRF). We consider three different scenarios for bounded generation: using dynamic bounds (different start and end frame with object/scene motion), view bounds (different start and end frame with camera motion), and identical bounds (using a single image as start and end frame). These scenarios are akin to the classic frame interpolation, novel view synthesis and loop video generation tasks but with more challenging constraints, and are detailed in
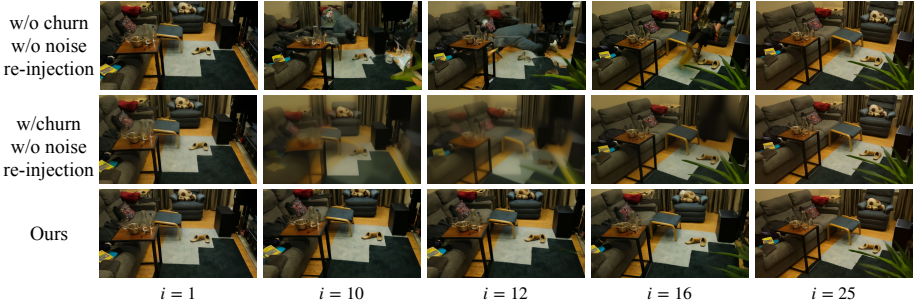
Fig. 5: The impact of noise re-injection on fusion. (Row 1) Without any stochasticity, the video suffers from random dynamics and unsmooth transitions. (Row 2) Tuning the churn term in SVD leads to blurry and low-quality frames. (Row 3) Using noise re-injection leads to smooth and natural frame transitions.

Sec. 4.1. In addition, we curate a new evaluation dataset for the aforementioned tasks containing difficult dynamics, which we present in Sec. 4.2. We compare against the closest state-of-the-art method for each of the tasks in Sec. 4.3, using standard metrics as well as a perceptual study, and show that our method significantly outperforms competitors. For video results please see our project page.

4.1   Evaluation Setting

We categorize the test scenarios into the following three distinct settings, which cover diverse types of bounded generation:
Dynamic Bound: With two frames capturing snapshots of a moving subject or object, the model should generate motion that seamlessly connects the frames.
View Bound: When the two frames capture the same static object from different views, the model should synthesize plausible in-between camera trajectories, which also allows us to gauge the 3D consistency of the I2V model.
Identical Bound: When the two frames are identical, the model should generate looping videos that involve stochastic or periodic motion.

4.2   The Bounded Generation Dataset

To evaluate bounded generation on the three settings above, we curated a high resolution image / video dataset for each of the tasks, consisting of:
– 125 image pairs sampled from YouTube videos, including kinematic motions of humans and animals, camera motion of complex scenes (e.g. landscape, cityscape, drone shots, etc.) and human-object interaction from movies, offering a broad spectrum of dynamic contents, paired with ground-truth clips.
– 25 wide-baseline image pairs sampled from 6 indoor / yard scenes, plus 15 out-of-distribution image pairs ranging from underwater reef to crowded table, which

Table 1: Quantitative results on different downstream tasks: 1) dynamics bound (Dyn.Bnd.) in terms of FVD [52], 2) identical bound (Id.Bnd.) in terms of FVD, and 3) view bound (View Bnd.) in terms of (a) FID with different feature dimensions, and (b) number of correspondences matched using COLMAP [45, 46].

Dyn.Bnd.	FVD_{25} ↓	Id.Bnd.	FVD_{25} ↓	View Bnd.	FID_{192} ↓	FID_{64} ↓	COLMAP ↑
FILM [44]	656.88	T2C [35]	911.67	Du et al. [13]	28.70	8.67	379.61
Ours	**431.16**	Ours	**458.91**	Ours	**10.31**	**3.43**	**884.08**

goes beyond the typical room tour distribution. The image pairs are selected from existing novel view synthesis datasets [4, 11, 19, 37].
– 276 static images from pexels.com [1] and [23], covering various dynamics such as natural phenomena (flaming, snowing), human activities (interaction with instruments, facial expressions), and larger scene dynamics (time-lapses, crowd movements) under 8 categories.

4.3 Comparative Analysis

We apply TRF to a frozen image-to-video model, Stable Video Diffusion [5], to generate videos that are conditioned on the image pairs from our curated dataset. While there is no existing method that accomplishes bounded generation across all three tasks, there are closely related works specific to each scenario. We compare against each state-of-the-art method, and demonstrate that they are not capable of fulfilling this new task. Given that both task and method inherently involve hallucination, the use of pixel-aligned metrics for quality evaluation is impractical. Instead, we use FVD [52] or FID [21] to compare the distribution between the generation and the ground truth. In the View Bound scenario, we further assess the 3D consistency of the generation based on the number of 3D corresponded points found by COLMAP, following the protocol in [7].

In the case of **Dynamics Bound**, the closest task is frame interpolation, which aims to smoothly interpolate between two given frames. We hence compare against FILM [44], a large motion frame interpolation method, using FVD. As per the results displayed in Table 1, TRF outperforms FILM by 30%. This large performance gap is due to the fact that FILM cannot handle interpolation of motion that is too further apart, or complicated motion that requires semantic understanding. As shown in Fig. 6 top, TRF is able to synthesize complex kinematics of human body and 3D consistent appearance in unseen regions, attributed to the underlying dynamic understanding and generalization ability of SVD, as well as the seamless dynamics trajectory fusing by TRF. In contrast, FILM primarily relies on flow-based correspondence and struggles to produce semantically meaningful motion when given sparse correspondences.

In the case of **View Bound**, the given frames are different viewpoints of the same static scene. We compare against the method of Du et al. [13] that addresses wide-baseline novel view synthesis using neural rendering. Note that their model is trained with known camera poses or correspondences, while our method has neither. We use FID [21, 47] to evaluate the quality of the rendered images in
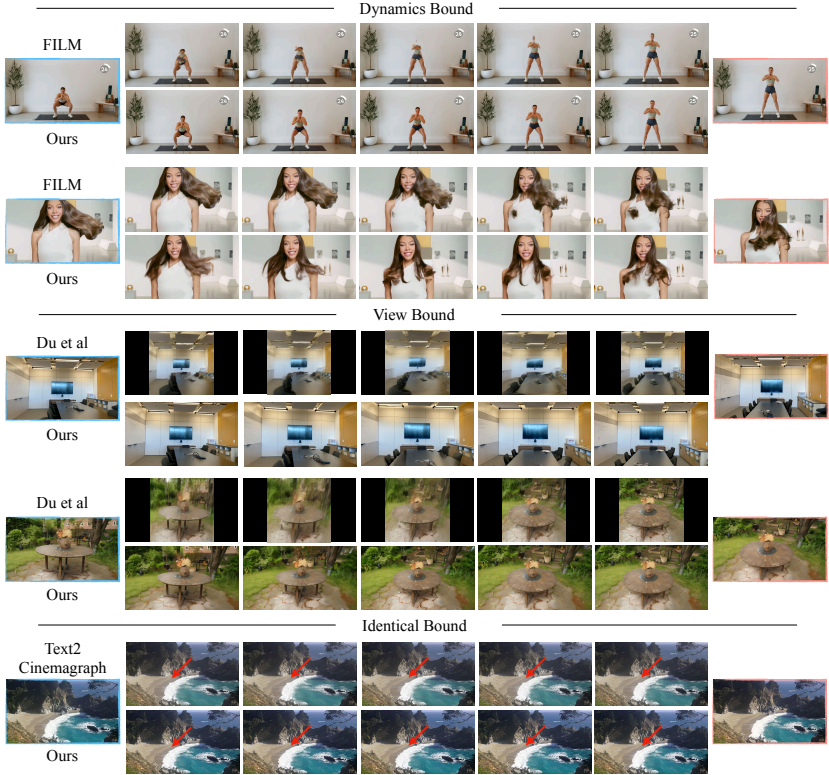
Fig. 6: Baseline comparisons. For Dynamics Bound, FILM fails to interpolate between large or complex motions such as kinematics. For View Bound, Du et al shows blurriness and stretching. On Identical Bound, our generated looping video depicts a more natural movement of the wave. We suggest viewing the videos in the project page.

terms of fidelity to the ground truth, and evaluate 3D consistency using sparse reconstruction (SFM) from the generated novel views with COLMAP [45, 46] following [7]. The number of extracted 3D correspondence points by COLMAP serves as an indicator of 3D consistency across all generated views. Given that Du et al. and SVD are trained under different resolution and aspect ratio, we resize and crop the ground-truth images to match their original size. COLMAP is applied on both methods in the same area. As shown in Table 1, TRF outperforms the baseline model by 60% on FID, indicating the superior quality of our method. This is also verified by the qualitative results as shown in Fig. 6 middle, where our rendered novel views contain less blurriness and stretching artifacts compared to the baseline, especially for 3D points that are only visible in one of the frames. The COLMAP reconstruction confirms that our rendered novel views are not only good in terms of visual quality, but also more 3D-consistent across the given wide-baseline paired views than the baseline method.

The task of **Identical Bound** is directly related to single-image cinema-graph. We hence compare against the recent work Text2Cinemagraph (T2C) [35], that uses images and texts to generate cinemagraphs with a dedicated pipeline including segmentation and motion prior training. We follow their evaluation protocol to compare FVD score on the validation set of Holynski et al. [23]. Table 1 middle shows the substantial improvement of our method with around 50% lower FVD score. The qualitative evaluation in Fig. 6 bottom demonstrates that our results have more natural movement of wave, in contrast to the persistent wave of T2C, which often produces more subtle motions. We point the reader to our project page for the video version of the results

While T2C is specifically designed to animate the fluid motion of the segmented water region in the image, TRF can easily generalize to a larger range of motions, from stochastic dynamics of flame to non-rigid facial expressions of humans. This is thanks to the generative power of the pretrained I2V model, which enables generalization without any specific design choice or training data, as shown in Fig. 7 bottom. More video results can be found on the project page.

5   Discussion

Probing I2V models. The bounded generation task along with TRF can offer a unique lens to assess SVD's world dynamics understanding. Given two observations, we can assess how the I2V model connects the motion trajectory, allowing us to compare the generated and the observed real-world dynamics. For example, the results on Dynamics Bound on the top of Fig. 7 indicate the model's ability to understand and generate complex kinematics trajectories of articulated human bodies under different clothing, lighting, or with different image quality.

The importance of the motion bucket ID. While our Time Reversal Fusion (TRF) method successfully achieves bounded generation without additional training, it does require careful tuning of the temporal conditioning parameters, such as motion bucket ID and frames per second (fps), to produce visually coherent outputs for different inputs. A critical aspect to note is the necessity for a match between the image content and the motion ID.

We refer the reader to our supplemental material for a detailed discussion on probing I2V models, motion bucket ID, method limitation and performance.

6   Conclusion

We introduced bounded generation as a form of generalized control for pre-trained image-to-video models, and proposed Time Reversal Fusion, a new training-free sampling strategy to achieve bounded generation. We demonstrate results on three distinct settings that cover a diverse set of dynamics, and curate a bounded generation dataset to show the effectiveness of TRF. Our work opens up new opportunities for controlled video generationm and provides a valuable avenue for probing the underlying dynamics within existing I2V models.
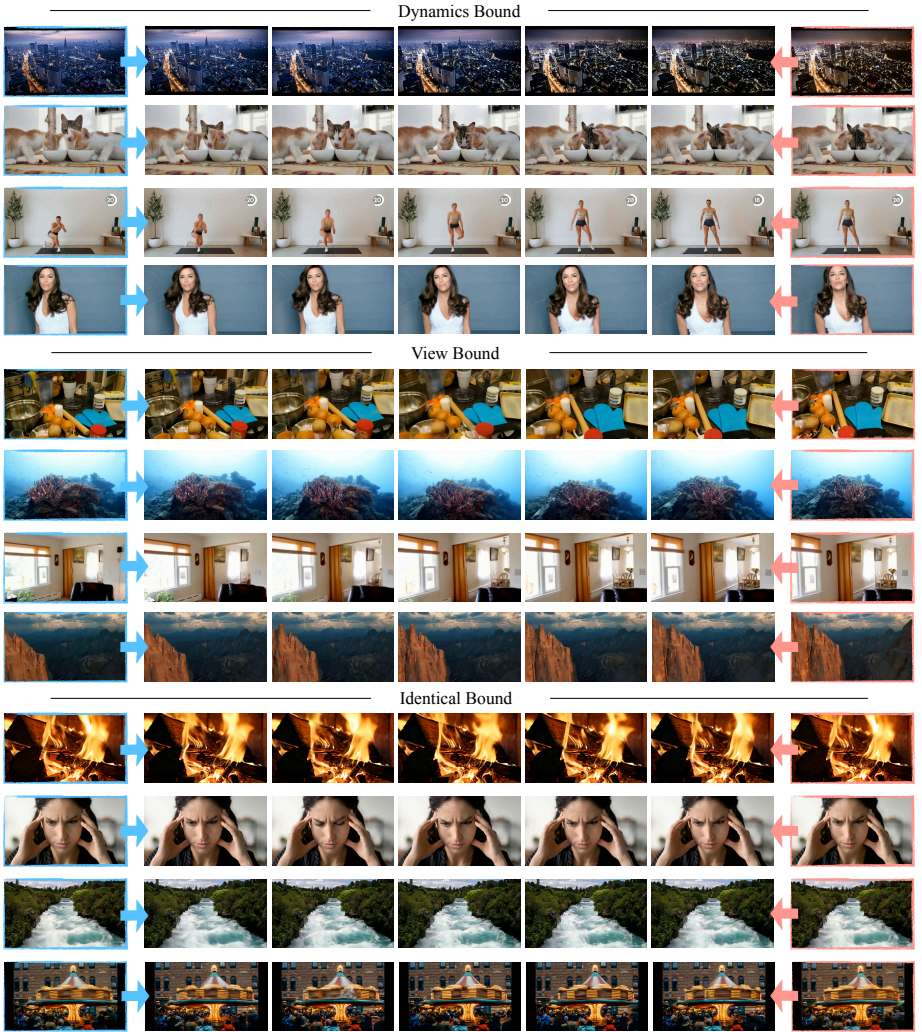
Fig. 7: Additional results on the Bounded Generation Dataset. We show that by enabling bounded generation on an I2V model with TRF, we can leverage its generalization capability to generate a diverse set of dynamics, including non-rigid body gestures, kinematics and nuanced hair movements. We can also synthesize camera trajectories across varied scenes and landscapes. The Identical Bound scenario produces looping videos capturing different movements in nature, which has not been shown before using a unified framework. Best viewed in video in the project page.

References

1. Pexels. `pexels.com`, accessed: 2024-02-01
2. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Li, Y., Michaeli, T., et al.: Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945 (2024)
3. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
7. Cai, S., Chan, E.R., Peng, S., Shahbazi, M., Obukhov, A., Van Gool, L., Wetzstein, G.: Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In: ICCV (2023)
8. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: Videocrafter1: Open diffusion models for high-quality video generation (2023)
9. Choi, B.T., Lee, S.H., Ko, S.J.: New frame rate up-conversion using bi-directional motion estimation. IEEE Transactions on Consumer Electronics **46**(3), 603–609 (2000)
10. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
11. Debevec, P.E., Taylor, C.J., Malik, J., Levin, G., Borshukov, G., Yu, Y.: Image-based modeling and rendering of architecture with interactive photogrammetry and view-dependent texture mapping. In: 1998 IEEE International Symposium on Circuits and Systems (ISCAS). vol. 5, pp. 514–517. IEEE (1998)
12. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
13. Du, Y., Smith, C., Tewari, A., Sitzmann, V.: Learning to render novel views from wide-baseline stereo pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4970–4980 (2023)
14. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
15. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22930–22941 (2023)
16. Gu, J., Trevithick, A., Lin, K.E., Susskind, J.M., Theobalt, C., Liu, L., Ramamoorthi, R.: Nerfdiff: Single-image view synthesis with nerf-guided distillation from

3d-aware diffusion. In: International Conference on Machine Learning. pp. 11808–11826. PMLR (2023)

17. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)

18. Ha, T., Lee, S., Kim, J.: Motion compensated frame interpolation by new block-based motion estimation algorithm. IEEE Transactions on Consumer Electronics **50**(2), 752–759 (2004)

19. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G., Brostow, G.: Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics (ToG) **37**(6), 1–15 (2018)

20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)

21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

22. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)

23. Holynski, A., Curless, B.L., Seitz, S.M., Szeliski, R.: Animating pictures with eulerian motion fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5810–5819 (2021)

24. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023)

25. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)

26. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)

27. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)

28. Li, X., Cao, Z., Sun, H., Zhang, J., Xian, K., Lin, G.: 3d cinemagraphy from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4595–4605 (2023)

29. Li, Z., Tucker, R., Snavely, N., Holynski, A.: Generative image dynamics. arXiv preprint arXiv:2309.07906 (2023)

30. Liu, Y., Xie, L., Siyao, L., Sun, W., Qiao, Y., Dong, C.: Enhanced quadratic video interpolation. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 41–56. Springer (2020)

31. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: Proceedings of the IEEE international conference on computer vision. pp. 4463–4471 (2017)

32. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)

33. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10209–10218 (2023)

34. Mahapatra, A., Kulkarni, K.: Controllable animation of fluid elements in still images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3667–3676 (2022)

35. Mahapatra, A., Siarohin, A., Lee, H.Y., Tulyakov, S., Zhu, J.Y.: Text-guided synthesis of eulerian cinemagraphs. ACM Transactions on Graphics (TOG) **42**(6), 1–13 (2023)

36. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)

37. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)

38. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)

39. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)

40. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)

41. Niklaus, S., Mai, L., Liu, F.: Video frame interpolation via adaptive convolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 670–679 (2017)

42. Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., Schölkopf, B.: Controlling text-to-image diffusion by orthogonal finetuning. Advances in Neural Information Processing Systems **36** (2024)

43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

44. Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C., Curless, B.: FILM: Frame interpolation for large motion. In: European Conference on Computer Vision. pp. 250–266. Springer (2022)

45. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

46. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)

47. Seitzer, M.: pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid` (August 2020), version 0.3.0

48. Sim, H., Oh, J., Kim, M.: Xvfi: extreme video frame interpolation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14489–14498 (2021)

49. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32** (2019)
50. Trevithick, A., Yang, B.: Grf: Learning a general radiance field for 3d representation and rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15182–15192 (2021)
51. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
52. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: FVD: A new metric for video generation (2019)
53. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36** (2024)
54. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023)
55. Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. arXiv preprint arXiv:2312.03641 (2023)
56. Xu, X., Siyao, L., Sun, W., Yin, Q., Yang, M.H.: Quadratic video interpolation. Advances in Neural Information Processing Systems **32** (2019)
57. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
58. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
59. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint arXiv:2309.15818 (2023)
60. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. arXiv preprint arXiv:2303.17076 (2023)
61. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
62. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12588–12597 (2023)