Dreamix: Video Diffusion Models are General Video Editors

Anonymous authors
Paper under double-blind review

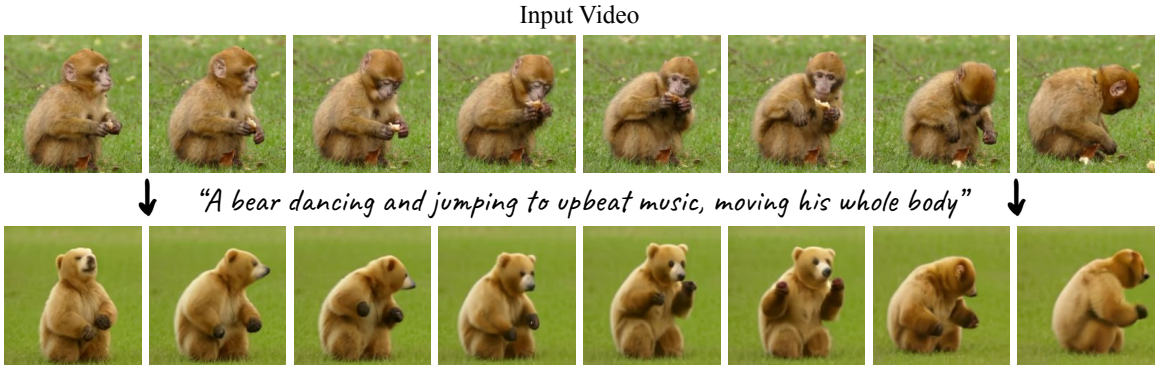

Figure 1: ***Video Editing with Dreamix***: By conditioning on the text prompt "A bear dancing and jumping to upbeat music, moving his whole body", Dreamix transforms the eating monkey (top row) into a dancing bear (bottom row), affecting motion and appearance. It maintains fidelity to color and pose, and results in a temporally consistent video. *We strongly encourage the reviewer to view the supplementary videos*

Abstract

Text-driven image and video diffusion models have recently achieved unprecedented generation realism. While diffusion models have been successfully applied for image editing, few can edit **motion** in video. We present a diffusion-based method that is able to perform text-based motion and appearance editing of general, real-world videos. Our approach uses a video diffusion model to combine, at inference time, the low-resolution spatio-temporal information from the original video with new, high resolution information that it synthesized to align with the guiding text prompt. As maintaining high-fidelity to the original video requires retaining some of its high-resolution information, we add a preliminary stage of finetuning the model on the original video, significantly boosting fidelity. We propose to improve motion editability by using a mixed objective that jointly finetunes with full temporal attention and with temporal attention masking. We extend our method for animating images, bringing them to life by adding motion to existing or new objects, and camera movements. Extensive experiments showcase our method's remarkable ability to edit motion in videos.

1 Introduction

Recent advancements in generative models (Ho et al., 2020; Chang et al., 2022; Yu et al., 2022; Chang et al., 2023) and multimodal vision-language models (Radford et al., 2021), paved the way to large-scale text-to-image models capable of unprecedented generation realism and diversity (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022b; Nichol et al., 2021; Avrahami et al., 2023). These models have ushered in a new era of creativity, applications, and research. Although these models offer new creative processes, they are limited to synthesizing *new* images rather than editing *existing* ones. To bridge this gap, intuitive image editing methods offer text-based editing of generated and real images while maintaining some of their original attributes (Hertz et al., 2022; Tumanyan et al., 2023; Brooks et al., 2023; Kawar et al., 2023; Valevski et al.,

Figure 2: ***Video Motion Editing:*** Dreamix can significantly change the actions and motions of subjects in a video (e.g. making a puppy leap) while maintaining temporal consistency

2023). Similarly to images, text-to-video models have recently been proposed (Ho et al., 2022c;a; Singer et al., 2022; Yu et al., 2023), however, editing the **motion** of real videos remains a challenging task.

In text-guided video editing, the user provides an input video and a text prompt describing the desired attributes of the resulting video (Fig. 1). The objectives are three-fold: i) alignment: the edited video should conform with the input text prompt ii) fidelity: the edited video should preserve the content of the original input iii) quality: the edited video should be of high-quality. Video editing is more challenging than image editing, as it requires synthesizing new motion, not merely modifying appearance. It also requires temporal consistency. As a result, applying image-level editing methods e.g. SDEdit (Meng et al., 2021) or Prompt-to-Prompt (Hertz et al., 2022) sequentially on the video frames is insufficient.

We present a new method, Dreamix, to adapt a text-conditioned video diffusion model (VDM) for video editing, in a manner inspired by UniTune (Valevski et al., 2023). The core of our method is enabling a text-conditioned VDM to maintain high fidelity to an input video via two main ideas. First, instead of using pure-noise as initialization for the model, we use a degraded version of the original video, keeping only low spatio-temporal information by downscaling it and adding noise. This is similar to SDEdit but the degradation includes not merely noise, but also downscaling. Second, we further improve the fidelity to the original video by finetuning the model on the original video. Finetuning ensures the model has knowledge of the high-resolution attributes of the original video. Naively finetuning on the input video results in relatively low motion editability as the model learns to prefer the original motion instead of following the text prompt. We propose a novel use for the mixed finetuning approach, suggested in VDM (Ho et al., 2022c) and Imagen-Video (Ho et al., 2022a), in which the VDMs are trained on both images and video. In our approach, we finetune the model on both the original video but also on its (unordered) frames individually. This allows us to perform significantly larger motion edits with high fidelity to the original video.

As a further contribution, we leverage our video editing model to add motion to still images (see Fig. 3) e.g., animating the objects and background in an image or creating dynamic camera motion. To do so, we first create a coarse video by simple image processing operations, e.g., frame replication or geometric image transformation. We then edit it with our Dreamix video editor. Our framework can also perform subject-driven video generation (see Fig. 3), extending the scope of current image-based methods e.g., Dreambooth(Ruiz et al., 2023) to video and motion editing. We evaluate our method extensively, demonstrating its remarkable capabilities unmatched by the baseline methods.

To summarize, our main contributions are:

1. Proposing a method for text-based *motion* editing of real-world videos.
2. Repurposing mixed training as a finetuning objective that significantly improves motion editing.
3. Presenting a new framework for text-guided image animation, by applying our video editor method on top of simple image preprocessing operations.

2 Related Work

2.1 Diffusion Models for Synthesis

Deep diffusion models recently emerged as a powerful new paradigm for image generation (Ho et al., 2020; Song et al., 2020), and have their roots in score-matching (Hyvärinen & Dayan, 2005; Vincent, 2011; Sohl-Dickstein et al., 2015). They outperform (Dhariwal & Nichol, 2021) the previous state-of-the-art approach, generative adversarial networks (GANs) (Goodfellow et al., 2020). While they have multiple formulations, EDM (Karras et al., 2022) showed they are equivalent. Outstanding progress was made in text-to-image generation (Saharia et al., 2022b; Ramesh et al., 2022; Rombach et al., 2022; Avrahami et al., 2023), where new images are sampled conditioned on an input text prompt. Extending diffusion models to video generation is a challenging computational and algorithmic task. Early work include (Ho et al., 2022c) and text-to-video extensions by (Ho et al., 2022a; Singer et al., 2022; Blattmann et al., 2023; Voleti et al., 2022; Wang et al., 2023a; Ge et al., 2023; Esser et al., 2023). Another line of work extends synthesis to various image reconstruction tasks (Saharia et al., 2022c;a; Ho et al., 2022b; Lugmayr et al., 2022; Chung et al., 2022), (Horwitz & Hoshen, 2022) extracts confidence intervals for reconstruction tasks.

2.2 Diffusion Models for Editing

Image editing with generative models has been studied extensively, in past years many of the models were based on GANs (Vinker et al., 2021; Patashnik et al., 2021; Gal et al., 2021; Roich et al., 2022; Wang et al., 2018b; Park et al., 2019; Bau et al., 2020; Skorokhodov et al., 2022; Jamriška et al., 2019; Wang et al., 2018a; Tzaban et al., 2022; Xu et al., 2022; Liu et al., 2022). Another recent line of works demonstrated preliminary generation and editing capabilities using masked image models (Yu et al., 2023; Villegas et al., 2022; Yao et al., 2021; Nash et al., 2022). However, most of the recent editing methods adopt diffusion models (Avrahami et al., 2022b;a; Voynov et al., 2023). SDEdit (Meng et al., 2021) proposed to add targeted noise to an input image, and then use diffusion models for reversing the process. Prompt-to-Prompt (Hertz et al., 2022; Tumanyan et al., 2023; Mokady et al., 2023) perform semantic edits by mixing activations extracted with the original and target prompts. For InstructPix2Pix (Brooks et al., 2023) this is only needed for constructing the training dataset. Other works (e.g. (Gal et al., 2022; Ruiz et al., 2023)) use finetuning and optimization to allow for personalization of the model, learning a special token describing the content. UniTune (Valevski et al., 2023) and Imagic (Kawar et al., 2023) finetune on a single image, allowing better editability while maintaining good fidelity. However, the methods are image-centric and do not use temporal information. Neural Atlases (Kasten et al., 2021) and Text2Live (Bar-Tal et al., 2022) allow some texture-based video editing, however, unlike our method they cannot edit the *motion* of a video. Recently, many methods proposed to adapt an image diffusion model for video editing (Wu et al., 2023; Geyer et al., 2023; Chai et al., 2023; Yang et al., 2023; Wang et al., 2023b; Qi et al., 2023; Kim et al., 2023; Khachatryan et al., 2023; Liu et al., 2023; Wang et al., 2024; Guo et al., 2023). Despite their promising results, they use a text-to-image backbone that can edit video appearance *but not motion*. Their results are also not fully temporally consistent. In contrast, our method uses a text-to-video backbone, enabling *motion* editing while maintaining smoothness and temporal consistency.

3 Background: Video Diffusion Models

Denoising Model Training. Diffusion models rely on a deep denoising neural network denoted by D_θ. Let us denote the ground truth video as v, an i.i.d Gaussian noise tensor of the same dimensions as the video as $\epsilon \sim N(0, \mathbf{I})$, and the noise level at time s as σ_s. The noisy video is given by: $z_s = \gamma_s v + \sigma_s \epsilon$, where $\gamma_s = \sqrt{1 - \sigma_s^2}$. Furthermore, let us denote a conditioning text prompt as t and a conditioning video c (for super-resolution, c is a low-resolution version of v). The objective of the denoising network D_θ is to recover the ground truth video v given the noisy input video z_s, the time s, prompt t and conditioning video c. The model is trained on a (large) training corpus $\mathcal{V}$ consisting of pairs of video v and text prompts t.

Sampling from Diffusion Models. The key challenge in diffusion models is to use the denoiser network D_θ to sample from the distribution of videos conditioned on the text prompt t and conditioning video c, $P(v|t, c)$.

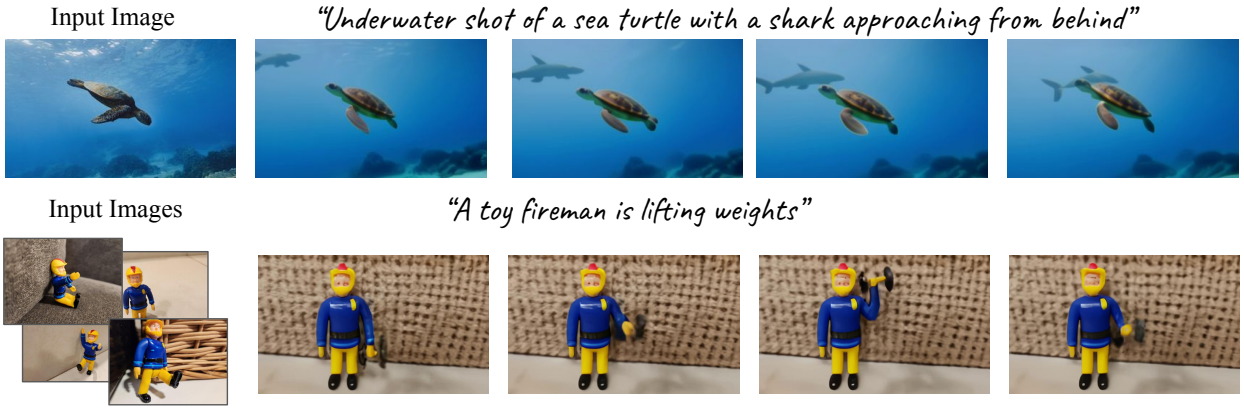

Figure 3: ***Image-to-Video editing with Dreamix:*** Dreamix instills complex motion in a static image (first row), adding a moving shark and making the turtle swim. In this case, visual fidelity to object location and background was preserved but the turtle direction was flipped. In the subject-driven case (second row), Dreamix extracts the visual features of a subject given multiple images and animates it in different scenarios such as weightlifting

While the derivation of such sampling rule is non-trivial (see e.g. Karras et al. (2022)), the implementation of such sampling is relatively simple in practice. We follow Ho et al. (2022a) in using stochastic DDIM sampling. At a heuristic level, at each step, we first use the denoiser network to estimate the noise. We then remove a fraction of the estimated noise and finally add randomly generated Gaussian noise, with magnitude corresponding to half of the removed noise.

Cascaded Video Diffusion Models. Training high-resolution text-to-video models is very challenging due to the high computational complexity. Several diffusion models overcome this by using cascaded architectures. We use a model that follows the architecture of Ho et al. (2022a), which consists of a cascade of 7 models. The base model maps the input text prompt into a 5-second video of $24 \times 40 \times 16$ frames. It is then followed by 3 spatial super-resolution models and 3 temporal super-resolution models. For implementation details, see Appendix C.

4 Editing by Video Diffusion Models

We propose a new method for video editing using text-guided video diffusion models. We extended it to image animation in Sec. 5.

4.1 Video Editing by Inverting Corruptions

We wish to edit an input video using the guidance of a text prompt t describing the video **after** the edit. In order to do so we leverage the power of a cascade of VDMs. The key idea is to first corrupt the video by downsampling followed by adding noise. We then apply the sampling process of the cascaded diffusion models from the time step corresponding to the noise level, conditioned on t, which upscales the video to the final spatio-temporal resolution. The effect is that the VDM will use the low-resolution details provided by the degraded input video, but synthesize new high spatio-temporal resolution information using the text prompt guidance. While this procedure is essentially a text-guided version of SDEdit (Meng et al., 2021), for complex edits e.g., *motion editing* this by itself does not result in sufficiently high-fidelity videos. To mitigate this issue, we use a mixed-finetuning objective described in Sec. 4.2.

Input Video Degradation. We downsample the input video to the resolution of the base model (16 frames of 24×40). We then add i.i.d Gaussian noise with variance σ_s^2 to further corrupt the input video. The noise strength is equivalent to time s in the diffusion process of the base model. For $s = 0$, no noise is added,

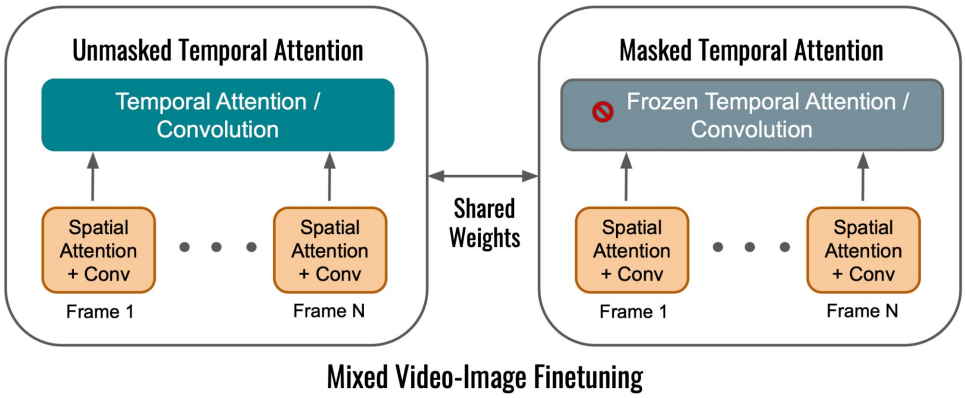

Figure 4: ***Mixed Video-Image Finetuning:*** Finetuning the VDM on the input video alone limits the extent of motion change. Instead, we use a mixed objective that beside the original objective (bottom left) also finetunes on the unordered set of frames. We use "masked temporal attention" to prevent the temporal attention and convolution from changing (bottom right). This allows adding motion to a static video

while for $s = 1$, the video is replaced by pure Gaussian noise. Note, that even when no noise is added, the input video is highly corrupted due to the extreme downsampling ratio.

Text-Guided Corruption Inversion. We can now use the cascaded VDMs to map the corrupted, low-resolution video into a high-resolution video that aligns with the text. The core idea here is that given a noisy, very low spatio-temporal resolution video, there are many perfectly feasible, high-resolution videos that correspond to it. We use the target text prompt t to select the feasible outputs that not only correspond to the low-resolution of the original video but are also aligned to edits desired by the user. The base model starts with the corrupted video, which has the same noise as the diffusion process at time s. We use the model to reverse the diffusion process up to time 0. We then upscale the video through the entire cascade of super-resolution models (see Appendix C for additional information). All models are conditioned on the prompt t.

4.2 Mixed Video-Image Finetuning

The naive method presented in Sec. 4.1 relies on a corrupted version of the input video which does not include enough information to preserve high-resolution details such as fine textures or object identity. We tackle this by adding a preliminary stage of finetuning the model on the input video v. Note that this only needs to be done once for the video, which can then be edited by many prompts without further finetuning. We would like the model to separately update its prior both on the appearance and the motion of the input video. Our approach therefore treats the input video, both as a single video clip and as an unordered set of M frames, denoted by $u = \{x_1, x_2, .., x_M\}$. We use a rare string t^* as the text prompt, following Ruiz et al. (2023). We finetune the denoising models by a combination of two objectives. The first objective updates the model prior on both motion and appearance by requiring it to reconstruct the input video v given its noisy versions z_s.

$$\mathcal{L}_\theta^{vid}(v) = \mathbb{E}_{\epsilon\sim N(0,\mathbf{I}),s\in\mathcal{U}(0,1)}\|D_{\theta'}(z_s, s, t^*, c) - v\|^2 \tag{1}$$

Additionally, we train the model to reconstruct each of the frames individually given their noisy version. This enhances the appearance prior of the model, separately from the motion. Technically, the model is trained on a sequence of frames u by replacing the temporal attention layers by trivial fixed masks ensuring the model only pays attention within each frame, and also by masking the residual temporal convolution blocks. We denote the attention masked denoising model as D_θ^a. The masked attention objective is:

$$\mathcal{L}_\theta^{frame}(u) = \mathbb{E}_{\epsilon\sim N(0,\mathbf{I}),s\in\mathcal{U}(0,1)}\|D_{\theta'}^a(z_s, s, t^*, c) - u\|^2 \tag{2}$$

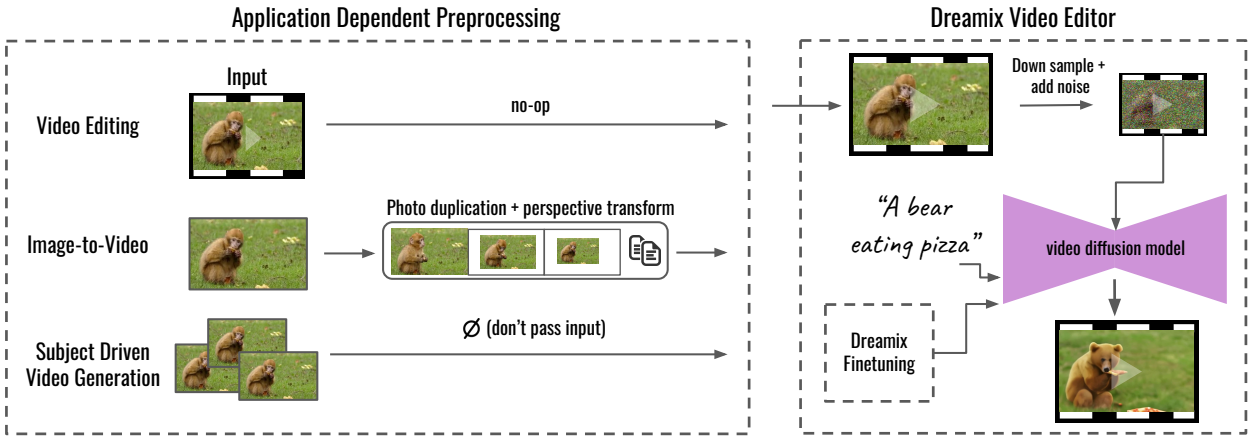

Figure 5: ***Inference Overview:*** Our method supports multiple applications by converting the input into a uniform video format (left). For image-to-video, the input image is duplicated and transformed using perspective transformations, synthesizing a coarse video with some camera motion. For subject-driven video generation, the input is omitted - finetuning alone takes care of the fidelity. This coarse video is then edited using our general "Dreamix Video Editor" (right): we first corrupt the video by downsampling followed by adding noise. We then apply the finetuned text-guided VDM, which upscales the video to the final spatio-temporal resolution

We train the joint objective:

$$\theta = arg\min_{\theta'} \alpha \mathcal{L}^{vid}_{\theta'}(v) + (1-\alpha)\mathcal{L}^{frame}_{\theta'}(u) \tag{3}$$

Where α is a constant factor, see Fig. 4. Training on a single video or a handful of frames can easily lead to overfitting, reducing the editing ability of the original model. To mitigate overfitting, we use a small number of finetuning iterations and a low learning rate (see Appendix C). Note that while such a training objective was used by Imagen-Video (Ho et al., 2022a) and VDM (Ho et al., 2022c), its purpose was different. There, the aim was to increase dataset size and diversity by training on large image datasets. Here, the aim is to enforce the style of the video in the model, while allowing motion editing.

5 Applications of Dreamix

The method proposed in Sec. 4, can edit *motion* and appearance in real-world videos. In this section, we propose a framework for using our Dreamix video editor for general, text-conditioned *image-to-video* editing, see Fig. 5 for an overview.

Dreamix for Single Images. Provided our general video editing method, Dreamix, we now propose a framework for image animation conditioned on a text prompt. The idea is to transform the image or a set of images into a coarse, corrupted video and edit it using Dreamix. For example, given a single image x as input, we can transform it to a video by replicating it 16 times to form a static video $v = [x, x, x...x]$. We can then edit its appearance and motion using Dreamix conditioned on a text prompt. Here, we do not wish to incorporate the motion of the input video (as it is static and meaningless) and therefore use only the masked temporal attention finetuning ($\alpha = 0$). To create "cinematic" effects, we can further control the output video by simulating camera motion, such as panning and zoom. We perform this by sampling a smooth sequence of 16 perspective transformations $T_1, T_2..T_{16}$ and apply each on the original image. When the perspective requires pixels outside the input image, we simply outpaint them using reflection padding. We concatenate the sequence of transformed images into a low quality input video $v = [T_1(x), T_2(x)..T_{16}(x)]$. While this does not result in realistic video, Dreamix can transform it into a high-quality edited video. See Appendix D for details on the applied transformations.

Dreamix for subject-driven video generation. We propose to use Dreamix for text-conditioned video generation given an image collection. Differently from existing methods, e.g., Dreambooth (Ruiz et al.,

Figure 6: ***Additional Image-to-Video Results:*** Dreamix can perform animated camera motion based on a single image, as in the first row where the image is zoomed out to reveal a bathing buffalo. Dreamix can also instill motion in a static image as in the second row where the glass is gradually filled with coffee. In addition, Dreamix can add animated subjects into an image, in the third row, multiple penguins are animated into the image

2023), it can add motion and not only change appearance. The input to our method is a set of images, each containing the subject of interest. This can also use different frames from the same video, as long as they show the same subject. Higher diversity of viewing angles and backgrounds is beneficial for the performance of the method. We then use the finetuning method from Sec. 4.2, where we only use the masked attention finetuning ($\alpha = 0$). After finetuning, we use the text-to-video model *without* a conditioning video, but rather only using a text prompt (which includes the special token t^*).

6 Experiments

In this section, we establish that Dreamix is able to edit motion in real-world videos and images, a major improvement over the existing methods. To fully experience our results, please see the supplementary videos.

6.1 Qualitative Results

Video Editing. In Fig. 1, we change the motion to dancing and the appearance from monkey to bear while keeping the coarse attributes of the video fixed. Dreamix can also generate new motion that does not necessarily align with the input video (puppy in Fig. 2, orangutan Fig. 15), and can control camera movements (zoom-out example in Fig. 16). Dreamix can generate smooth visual modifications that align with the temporal information in the input video. This includes adding effects (field in Fig. 13 and saxophone in Fig. 16), adding or replacing objects (cake in Fig. 7, hat and robot in Fig. 13, skateboard in Fig. 14), and changing the background (truck in Fig. 16).

Image-driven Videos. When the input is a single image, Dreamix can use its video prior to add new moving objects (penguins in Fig. 6 and camel in Fig. 17), inject motion into the input (turtle in Fig. 3 and coffee in Fig. 6), or new camera movements (buffalo in Fig. 6). Our method is unique in adding large motions and moving objects into general, real-world images.

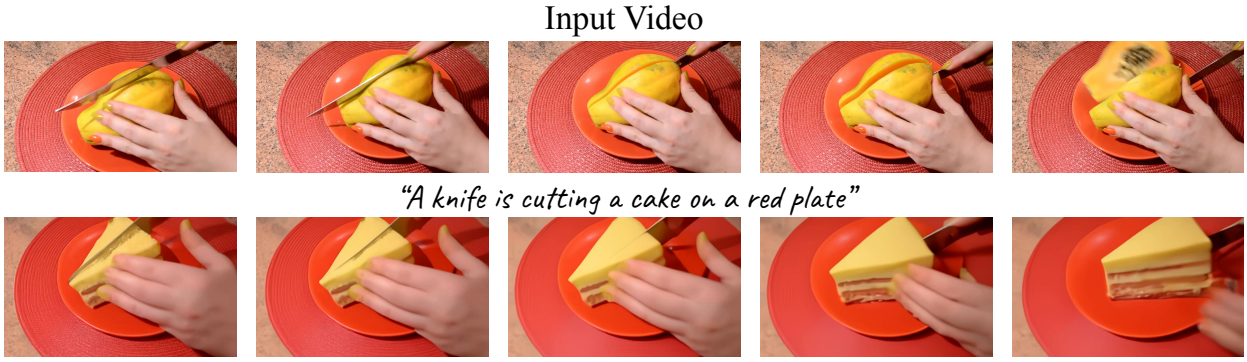

Figure 7: ***Additional Video Editing Results:*** Dreamix can replace objects and appearance in real videos, for example, replacing the papaya with a cake

Figure 8: ***Additional Subject-driven Video Generation Results:*** Dreamix can use a small number of independent *images* to learn the visual appearance of a subject and animate it in a realistic manner. In the first row, the bear animated to walk while maintaining fidelity to the original bear. Dreamix can also compose a real subject with objects present in a prompt. In the second row, the magnifying glass is composed and animated into a scene with the caterpillar

Subject-driven Video Generation. Dreamix can take an image collection showing the same subject and generate new videos with this subject in motion. We demonstrate this on a range of subjects and actions including: the weight-lifting toy fireman in Fig. 3, walking and drinking bear in Fig. 8 and Fig. 18. It can also place the subjects in new surroundings, e.g., moving caterpillar on a leaf or even under a magnifying glass (Fig. 18 and Fig. 8).

6.2 Baseline Comparisons

Baselines. We compare our method against three baselines: *Unconditional.* Directly mapping the text prompt to a video, without conditioning on the input video using a model similar to Imagen-Video. *Plug-and-Play (PnP).* Applying PnP(Tumanyan et al., 2023) on each video frame independently. *Tune-a-Video (TaVid).* Finetuning Tune-a-Video(Wu et al., 2023) on the input video.

Data. We created a dataset of 29 videos taken from YouTube-8M (Abu-El-Haija et al., 2016) and 127 text prompts, spanning different categories (see Appendix E).

Quantitative Comparison. We measure alignment by the frame-level CLIP Score (Hessel et al., 2021) and quality (stability) with LPIPS (Zhang et al., 2018) between consecutive frames. As automatic metrics

Table 1: ***User Study:*** Users rated editing results by quality, fidelity to the base video, and alignment with the text prompt. Based on visual inspection, we require an edit to score greater than 2.5 in all dimensions to be successful and observe that Dreamix is the only method to achieve the desired trade off

Method	Quality	Fidelity	Alignment	Success
PnP	2.16 ±1.13	3.78 ±0.99	3.39 ±1.38	20%
TaVid	1.99 ±0.92	3.29 ±1.21	2.69 ±1.55	13%
Ours	3.58 ±1.04	3.55 ±1.09	3.79 ±1.33	76%
Uncond.	3.43 ±1.09	2.49 ±1.12	4.28 ±1.02	45%

Table 2: ***Baseline Comparisons:*** Our method achieves better temporal consistency than PnP and Tune-a-Video (TaVid). Moreover, Dreamix is successful at *motion* editing while other methods cannot. This is reflected in the better quality (low LPIPS) and alignment (high CLIP Score). While the unconditional method seems to outperform Dreamix, it has poor fidelity as it is not conditioned on the input video

Metric	PnP	TaVid	Ours	Uncond.
LPIPS ↓	0.209	0.145	0.112	0.101
CLIP Score ↑	0.304	0.303	0.317	0.320
Fidelity	See user study (Tab. 1)			

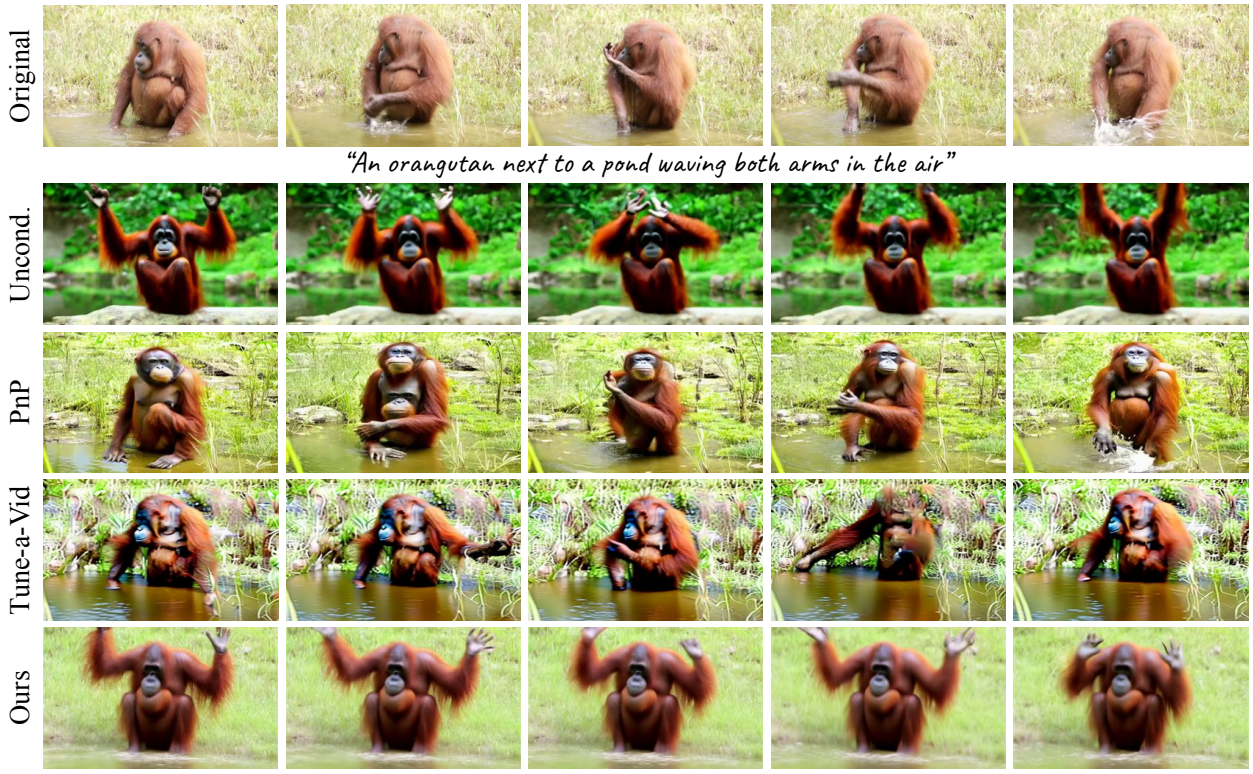

Figure 9: ***Comparison to Baseline Methods for Motion Edits:*** Although the quality and alignment of unconditional generation are high, there is no resemblance to the original video (low fidelity). While PnP and Tune-a-Video preserve the scene, they fail to edit the motion according to the prompt (no waving) and suffer from poor temporal consistency (flickering). Our method is able to edit the *motion* according to the prompt while preserving the fidelity and generating a high quality video. Moreover, video-based methods (Uncond. and ours) exhibit motion blur, also present in real videos

do not measure fidelity and are imperfectly aligned with human judgement, we also conduct a user study. A panel of 20 evaluators rated each video/prompt pair on a scale of $1-5$ to evaluate its quality, fidelity and alignment. When visually inspecting the results we discover that videos that received a score lower than 2.5 in any of the dimensions are usually clear failure cases. Therefore we also report the percentage of items where all dimensions are larger than 2.5 (i.e. "Success"). See Appendices F and G for additional details on the evaluation protocol.

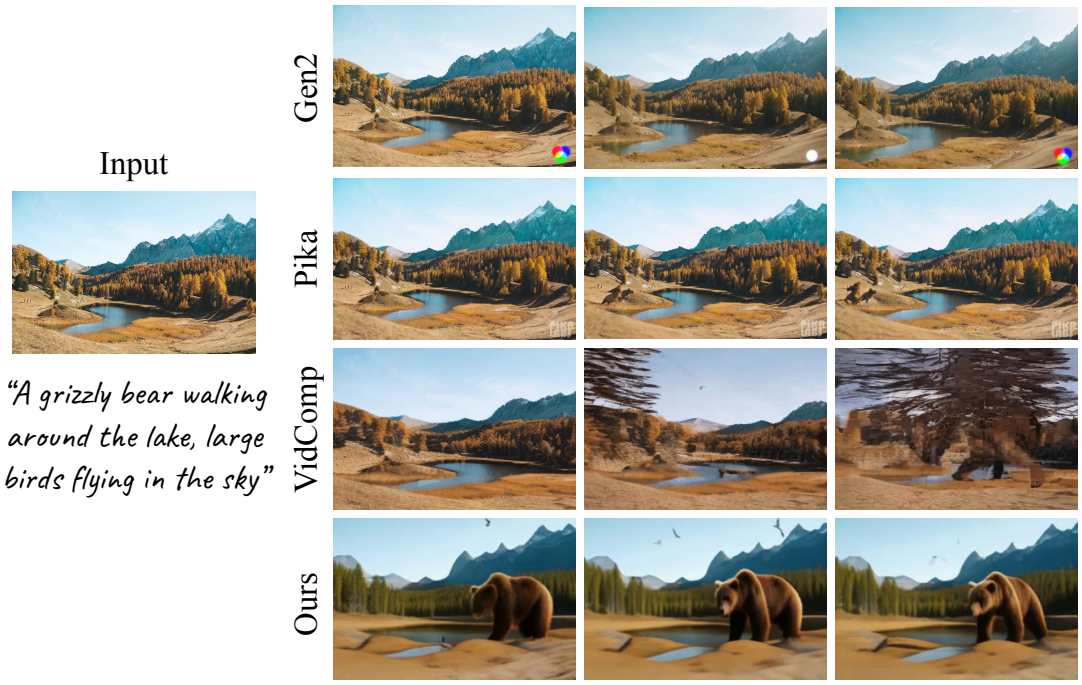

Figure 10: ***Image-driven Videos Comparison:*** We compare Dreamix against different SoTA image-driven video baselines. Gen2 and Pika are commercial products, although they maintain great visual fidelity to the input image, they did not add motion and animated objects that align with our input prompt. VideoComposer fails to maintain fidelity to the original image and underperforms in terms of temporal consistency and alignment. Dreamix is able to add large motion and animated objects to the image

The evaluation and user study are presented in Tab. 2 and Tab. 1. Image-based methods (PnP, Tune-a-Video) exhibit impaired temporal consistency, resulting in low quality. Moreover, they are unable to perform *motion* edits, resulting in poor alignment and high fidelity. Video-based methods maintain temporal consistency while allowing motion editing. Although unconditional generation outperforms our method in the automatic evaluations (Tab. 2), it has poor fidelity (Tab. 1) as it is not conditioned on the input video. Overall, our method has the highest success rate.

Qualitative Comparison. Figure 9 presents an example of motion editing by Dreamix compared to the baselines. The text-to-video model achieves low fidelity edits as it is not conditioned on the original video. PnP preserves the scene but fails to perform the edit and lacks consistency between different frames. Tune-a-Video exhibits better temporal consistency but still fails to perform the motion edit. Dreamix performs well on all three objectives, adding the desired motion while preserving fidelity and high-quality.

Application Comparison. We compare Dreamix to recent methods designed specifically for Image-driven video generation (e.g. VideoComposer (Wang et al., 2024), Gen2 (gen), PikaLabs (pik)). We present the results in Fig. 10, Dreamix is the only method that is able to add large motion and animated objects to the image. For more comparisons see the supplementary videos, for more details see Appendix C.4.

6.3 Ablation Study

We ablate the use of finetuning and the mixed video-image finetuning by performing a user study using the dataset described above. The ablation indeed supports the idea of using finetuning in cases where high-editability is required. We can see that *Motion* changes require high-editability and are thus improved by finetuning. Moreover, as the noising corrupts the video, preserving fine-details in *background, color or texture* edits requires finetuning. In contrast, denoising without finetuning worked well for *style* edits, where finetuning was often detrimental. This is expected as *style* edits are often conflicted with high fidelity

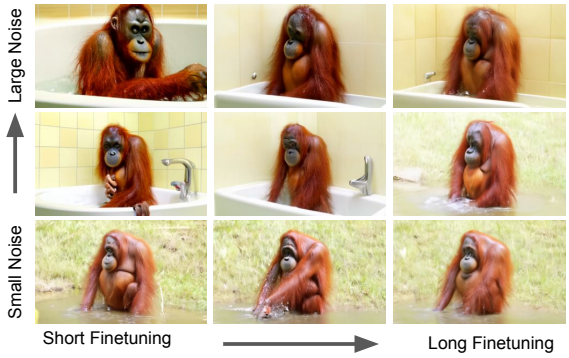

Figure 11: ***Noise-Finetuning Tradeoffs:*** We compare the effect of noise magnitude and number of finetuning iterations on edited videos. Higher noise allows for larger edits but reduces fidelity, more iterations improve fidelity at higher noises. The best results are obtained for high noise and a large number of finetuning iterations (bottom left is input)

Figure 12: ***Ablation Study:*** **Left:** Users were asked to compare text-guided video edits of with (w/ Ft) and without (w/o Ft) finetuning. "None" indicates failure of both methods according to user. Apart from style-based edits, where high fidelity is not needed, finetuning significantly improves the results. **Right:** Users were asked to compare video finetuning (Vid) with mixed video-image finetuning (Mix). Mixed finetuning significantly improves the results for most cases

Edit Type	# of Edits	w/o Ft.	w/ Ft.	None	Vid	Mix
Motion	36	17%	72%	11%	35%	65%
Object	44	36%	48%	16%	62%	38%
Background	32	19%	77%	9%	36%	64%
Style	15	67%	27%	6%	26%	74%

preservation (e.g. changing the texture of an object means reducing fidelity). The ablation shows that in most cases mixed finetuning improves the results by a wide margin. We present a visual ablation in Fig. 11 and user study results in Fig. 12.

7 Limitations

While Dreamix is the first diffusion-based video method that can edit motion, it has limitations.

Computational Cost. VDMs are computationally expensive. Finetuning our model using 4 TPU v4 accelerators requires around 30 minutes per video. Once finetuned, sampling takes roughly 2 minutes on similar hardware. Speeding it up will allow Dreamix to be used for more applications.

Comparison to Image-based Methods. Dreamix uses VDMs while previous approaches used image-level methods. As VDMs are nascent and have lower resolution than image DMs, this presents an interesting trade-off. Dreamix has the ability to edit motion and has high temporal consistency, while previous methods e.g., PnP, Tune-a-Video, Gen-2, can have higher spatial resolution. Although Tune-a-Video can achieve high alignment for appearance editing on videos with limited motion, it suffers from poor temporal consistency (see supplementary videos). This highlights the importance of using a VDM backbone that provides temporal consistency and enables *motion* editing.

Model Availability. Since our VDM is not currently publicly available, we conducted a preliminary investigation on the publicly available "ModelScope" (Wang et al., 2023a). It is important to note that "ModelScope" modified an image diffusion model (StableDiffusion) into a video diffusion model. It exhibits some temporal modeling capabilities but not as much as a full VDM model which we use. We were nonetheless able to reproduce our main findings by applying Dreamix on "ModelScope". For more details see Appendix C.5 and the supplementary videos. We note that when using the "ModelScope" backbone, the computational cost of our method drops significantly and is on par with other editing methods (< 15 minutes for training, < 2 minutes for inference, and roughly $24GB$ of GPU memory).

8 Conclusion

We presented a diffusion-based method that can edit *motion* in real-world videos. Our method can be applied to image animation and subject-driven video generation. Extensive experiments demonstrated the unprecedented capabilities of our method.

References

Gen-2, `https://research.runwayml.com/gen2`.

Pikalabs, `https://pikalabs.org`.

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022a.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022b.

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18370–18380, 2023.

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pp. 707–723. Springer, 2022.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.

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*, pp. 22563–22575, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23040–23050, 2023.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12413–12422, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7346–7356, 2023.

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22930–22941, October 2023.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022b.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022c.

Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. Stylizing video by example. *ACM Transactions on Graphics*, 38(4), 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6007–6017, 2023.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

Gyeongman Kim, Hajin Shim, Hyunsu Kim, Yunjey Choi, Junho Kim, and Eunho Yang. Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6091–6100, 2023.

Feng-Lin Liu, Shu-Yu Chen, Yu-Kun Lai, Chunpeng Li, Yue-Ren Jiang, Hongbo Fu, and Lin Gao. DeepFaceVideoEditing: Sketch-based deep editing of face videos. *ACM Transactions on Graphics*, 41(4):167:1–167:16, 2022.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 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.

Charlie Nash, João Carreira, Jacob Walker, Iain Barr, Andrew Jaegle, Mateusz Malinowski, and Peter Battaglia. Transframer: Arbitrary frame prediction with generative models. *arXiv preprint arXiv:2203.09494*, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2337–2346, 2019.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022b.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022c.

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

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3626–3636, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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.

Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Trans. Graph.*, 42(4), jul 2023. doi: 10.1145/3592451.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7): 1661–1674, 2011.

Yael Vinker, Eliahu Horwitz, Nir Zabari, and Yedid Hoshen. Image shape manipulation from a single augmented training sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13769–13778, October 2021.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018a.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018b.

Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023b.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.

Yiran Xu, Badour AlBahar, and Jia-Bin Huang. Temporally consistent semantic video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pp. 357–374. Springer, 2022.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13789–13798, 2021.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

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 conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Attached Videos

In addition to this file, we include a number of videos, **we highly encourage the reviewer to view them**. The included videos are:

1. "1_dreamix_overview_video.mp4" - An overview video of our method with audio narration.
2. "2_dreamix_video_editing_examples.mp4" - A number of video editing examples generated by our method (Dreamix).
3. "3_dreamix_image2video_examples.mp4" - A number of image-to-video examples generated by our method (Dreamix).
4. "4_dreamix_subject_driven_video_generation_examples.mp4" - A number of subject-driven video generation examples generated by our method (Dreamix).
5. "5_dreamix_baseline_comparisons.mp4" - A number of videos comparing our method (Dreamix) to the other baselines.
6. "6_dreamix_image2video_baseline_comparison.mp4" - A number of videos comparing our method (Dreamix) to the other baselines for the task of image to video generation.
7. "7_dreamix_modelscope_comparison.mp4" - A number of videos of our method (Dreamix) running on a ModelScope VDM backbone. We also ablate our method as opposed to regular finetuning with the ModeScope backbone.

Note: to match the requirement of maximum 100MB for the supplementary, all the videos are compressed, the uncompressed versions will be released in the final revision.

B Social Impact

Our primary aim in this work is to advance research on tools to enable users to animate their personal content. While the development of end-user applications is out of the scope of this work, we recognize both the opportunities and risks that may follow from our contributions. As discussed above, we anticipate multiple possible applications for this work that have the potential to augment and extend creative practices. The personalized component of our approach brings particular promise as it will enable users to better align content with their intent, despite potential biases present in general VDMs. On the other hand, our method carries similar risks as other highly capable media generation approaches. Malicious parties may try to use edited videos to mis-lead viewers or to engage in targeted harassment. Future research must continue investigating these concerns.

C Implementation Details

C.1 Architecture

All of our experiments were preformed on a VDM that is similar to Imagen-Video (Ho et al., 2022a), a pertrained cascaded text-to-video diffusion model, with the following components:

1. A T5-XXL(Raffel et al., 2020) text encoder, that computes embeddings from the textual prompt. This embeddings are then used as conditioning by all other models.
2. A base video diffusion model, conditioned on text. It generates videos at $16 \times 24 \times 40 \times 3$ resolution (frames X height X width X channels) at 3 fps.

3. 6 super-resolution video diffusion models, each conditioned on the text and the output video of the previous model. Each model is either spatial (SSR), i.e. upscales resolution, or temporal (TSR), i.e. fills in intermediate frames between the input frames. The order of super resolution models is TSR (2x), SSR (2x), SSR(4x), TSR(2x), TSR(2x), and SSR(4x). The multiplier in the parenthesis for output frames (for TSR), and for output pixels in height and width (for SSR). The final output video is in $128 \times 768 \times 1280 \times 3$ at 24 fps.

Note that the diffusion models are pretrained on both videos and images, with frozen temporal attention and convolution for the latter. Our mixed finetuning approach treats video frames as if they were images.

Distillation. For some of these models, we use a distilled version to allow for faster sampling times. The base model is a distilled model with 64 sampling steps. The first two SSR models are non-distilled models with 128 sampling steps (due to finetuning considerations, see below). All other SR models use 8 sampling steps. All models use classifier-free-guidance weight of 1.0 (meaning that classifier free guidance is turned off).

C.2 Finetuning

To reduce finetuning time, we only finetune the base model and the first 2 SSR models. In our experiments, finetuning the first 2 SSR models using the distilled models (with 8 sampling steps) did not yield good quality. We therefore use the non-distilled versions of these models for all experiments (including non-finetuned experiments). When using "Mixed Video-Image Finetuning" we use $\alpha = 0.35$, and finetune for 300 steps. For all our experiments we use a learning rate of $6 \cdot 10^{-6}$.

C.3 Sampling

We use a DDIM sampler with stochastic noise correction, following (Ho et al., 2022a). For the last highest resolution SSR, for capacity reasons, we use the model to sample a sub-chunks of 32 frames of the input lower resolution videos, and then we concatenate all the outputs together back to 128 frame videos.

C.4 Image-driven Video Generation Baselines

We compare our method with three recent image-driven video generation baselines (see video "6_dreamix_image2video_baseline_comparison.mp4"). VideoComposer, Gen-2, and PikaLabs. For VideoComposer we use the official implementation, for each of the examples we generate 5 samples from different, random seeds. We display the best result out of the 5 samples. We can see the results fail to maintain fidelity to the original image and underperform in terms of temporal consistency and alignment. Since PikaLabs is a commercial product we only have access through the official discord channel. For each of the examples we generate 5 different samples, and we display the best result out of the 5 samples. We can see that although the video maintains great visual fidelity to the input image, it did not add motion and animated objects that align with our input prompt. As Gen-2 is a commercial product we only have access through the official website channel. Due to limited credits (that cost money), for each of the examples we generate 3 different samples, and we display the best result out of the 3 samples. In addition, Gen-2 allows for some "cinematic" effects, for the relevant examples we use the same transformations as the ones listed in Tab. 3. We can see that although the video maintains great visual fidelity to the input image, it did not add motion and animated objects that align with our input prompt. It is worth noting that all the methods also generate shorter videos than our method.

C.5 Dreamix with ModelScope Backbone

Since our VDM is not currently publicly available, we conducted a preliminary investigation on the publicly available "ModelScope" (Wang et al., 2023a) (see video "7_dreamix_modelscope_comparison.mp4"). For each of the videos, we show the results of our method when used with the ModelScope backbone (named "ModelScope + Ours" in the comparison video). In addition, we ablated the need for our method with the ModelScope backbone by running a simple finetuning of ModelScope on the input video (named "ModelScope

+ Finetuning" in the comparison video). We also include the results of Dreamix using the VDM backbone used for the rest of the paper. In both cases, we finetune ModelScope for 750 steps with a learning rate of 10^{-5}, for our method we use, $\alpha = 0.5$ with $s = 0.85$. For a fair comparison, we generate the results in the same spatial resolution as the results generated by our method (320×192), however, higher spatial generation resolution is possible when using ModelScope. We can see that our method can be applied to other, open source backbones and indeed improves the fidelity and temporal consistency of the generated results. It is important to note that "ModelScope" modified an image diffusion model (StableDiffusion) into a video diffusion model. It exhibits some temporal modeling capabilities but not as much as the full VDM model that we use.

D Image-to-Video Transformations

We only use perspective transformations to create "cinematic" effects, e.g., panning, zooming, and camera shake. In our supplementary videos, we included Image-to-Video examples with different perspective transformations applied to them. We detail these transformations in Tab. 3. Some of the examples did not use the perspective transformations at all. Also, ensuring the smoothness of the transformed sequence is unnecessary as this is fixed by the diffusion and super-resolution processes.

Table 3: ***Perspective Transformations***

Video	Timestamp	Transformation	Effect
Plant	00:00	Translate	Pan
Turtle	00:11	Rand. translate	Shake
Coffee	00:22	Translate	Pan
Camel	00:33	None	None
Volcano	00:43	Rand. translate	Shake
Bear	00:54	Perspective	Pan
Penguins	01:05	None	None
Unicorn	01:15	Scale	Zoom out
Buffalo	01:26	Scale	Zoom out
Bigfoot	01:37	Translate	Pan

E Evaluation dataset

In all evaluations described in the paper, we used a dataset of 29 videos with 127 edit prompts. The dataset videos were selected from YouTube-8M (Abu-El-Haija et al., 2016) and show animals, people performing actions, vehicles, and other objects. The edit prompt categories are motion, object, background, and style. In the motion category the prompts perform motion editing (e.g. adding motion with the prompt "An orangutan next to a pond waving both arms in the air"), the object category performs object level edits (e.g. adding a party hat with the prompt "A puppy walking with a party hat"), the background category performs edits of the background (e.g. adding a river with the prompt "A blue pickup truck crossing a deep river"), and the style category performs style-transfer like edits (e.g. changing the style to cartoon with the prompt "A cartoon of a man playing a saxophone").

F Human evaluation details

We performed human evaluations for the baseline comparison and the ablation analysis. The evaluations were conducted by a panel of 20 human raters using the dataset described in the main paper. The video resolution shown to raters was 350×200, except for tune-a-video where we used a resolution of 200×200 (because we observed it performs better with square outputs).

F.1 Ablation Study

In the ablation study the raters were asked to select the best edited video out of 12 hyperparameter combinations:

- no finetuning, with $s \in 0.4, 0.7, 0.8, 0.85$
- video finetuning for 64 steps, with $s \in 0.8, 0.9, 0.95, 0.98$
- mixed finetuning, with $(ft_{steps}, s) \in (150, 0.98), (200, 0.98), (200, 1.0), (300, 1.0)$

F.2 Baseline Comparison

In the baseline comparisons described in the paper, raters evaluated videos on quality, fidelity and alignment. The raters saw the original video alongside an edited video and answered the following questions on a scale of $1 - 5$:

1. "Rate the overall visual quality and smoothness of the edited video."
2. "How well does the edited video match the textual edit description provided?"
3. "How well does the edited video preserve unedited details of the original video?"

F.3 Direct comparison

We also conducted a direct comparisons between the different editing methods. In this comparison raters saw the videos simultaneously and selected the best edit. We conducted the comparison once with a fixed set of hyperparameters for Dreamix, and once more showing a single Dreamix video chosen among 12 hyperparameters sets. Results can be seen in Tab. 4.

Table 4: ***Direct comparison of editing methods:*** Users were shown editing results of different editing methods were asked to pick the best one. In the "Multiple HP" column, the Dreamix video was chosen from a set of 12 hyperparameters. We show the number of times each method got a majority vote (out of 5 ratings)

Method	Single HP	Multiple HP
Plug-and-Play	2%	1%
Tune-a-Video	6%	6%
Ours	34%	77%
No good edit / Uncond.	58%	16%

G Quantitative Evaluation

In the quantitative baseline comparison described in the paper we reported alignment and quality. We measure alignment by the frame-level CLIP Score (Hessel et al., 2021). That is, we compute the cosine similarity between the CLIP (Radford et al., 2021) embedding and the CLIP text embedding for each frame. For each video we take the average over all frames, finally we report the mean over all the videos. For quality (stability) we compute the LPIPS (Zhang et al., 2018) distance between all pairs of consecutive frames. For each video we take the average over all pairs of consecutive frames, finally we report the mean over all the videos. To perform a fair comparison with Tune-a-Video (which outputs videos of 24 frames at 5 fps) we subsampled the rest of the methods to match this framerate. Additionally, before passing through CLIP and LPIPS, all the frames are preprocessed to match the required format (i.e. resize to 224, center crop to 224, ImageNet normalization).

H Image Attribution

- Desert - https://unsplash.com/photos/PP8Escz15d8
- Fuji mountain https://unsplash.com/photos/9Qwbfa_RM94
- Tree in snow - https://unsplash.com/photos/aQNy0za7x0k
- Hut in snow - https://unsplash.com/photos/qV2p17GHKbs
- Lake with trees - https://unsplash.com/photos/dIQlgwq6V3Y
- Plant - https://unsplash.com/photos/LrPKL7jOldI
- Turtle - https://unsplash.com/photos/za9MCg787eI
- Yosemite - https://unsplash.com/photos/NRQV-hBF10M
- Foggy forest - https://unsplash.com/photos/pKNqyx_v62s
- Coffee - https://unsplash.com/photos/SMPe5xfbPT0
- Monkey - https://www.pexels.com/video/a-brown-monkey-eating-bread-2436088/

I Additional Results

Below we present additional results of our method, for the best experience see the included videos.

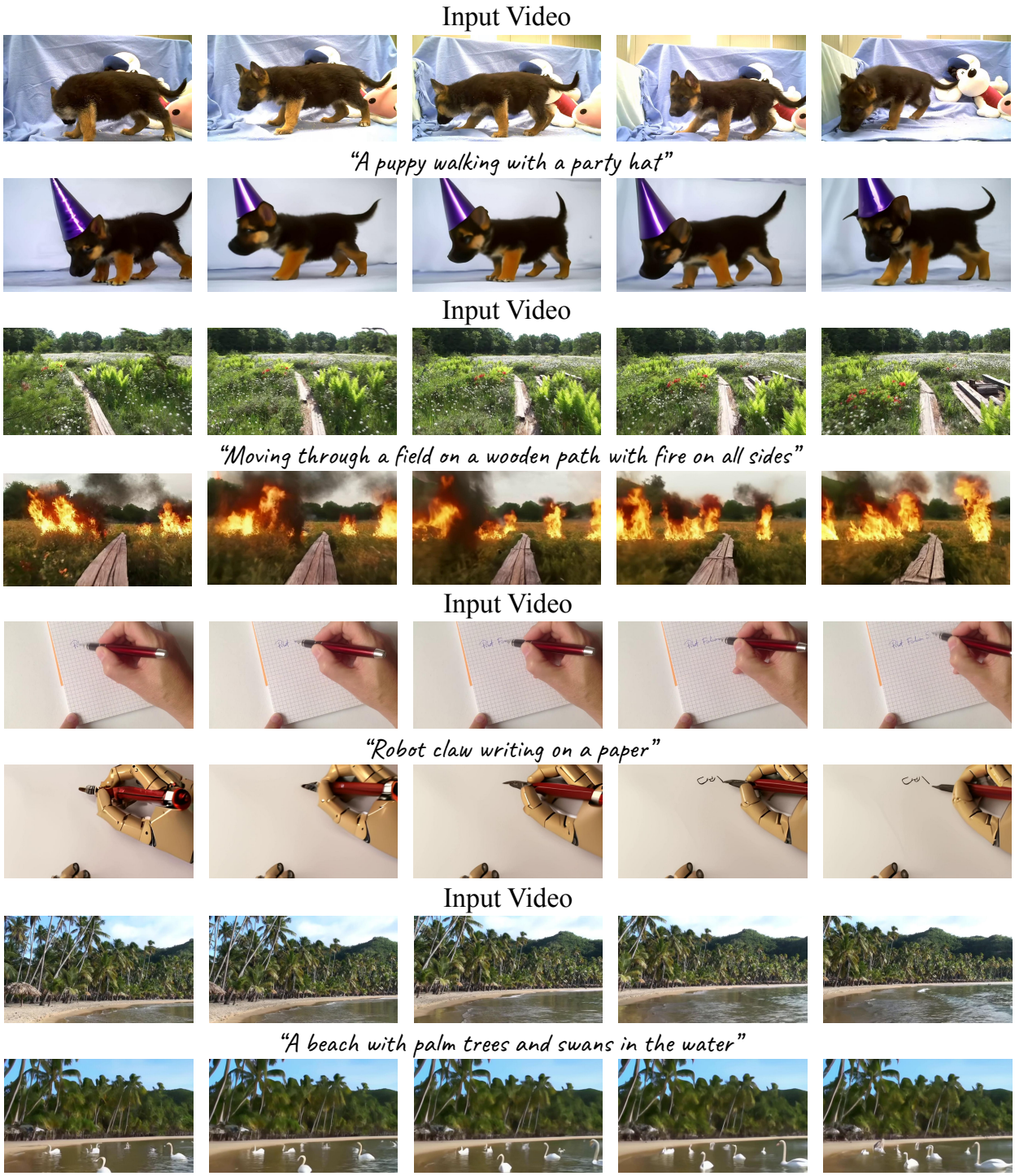

Figure 13: ***Additional Video Editing Results (1/4)***

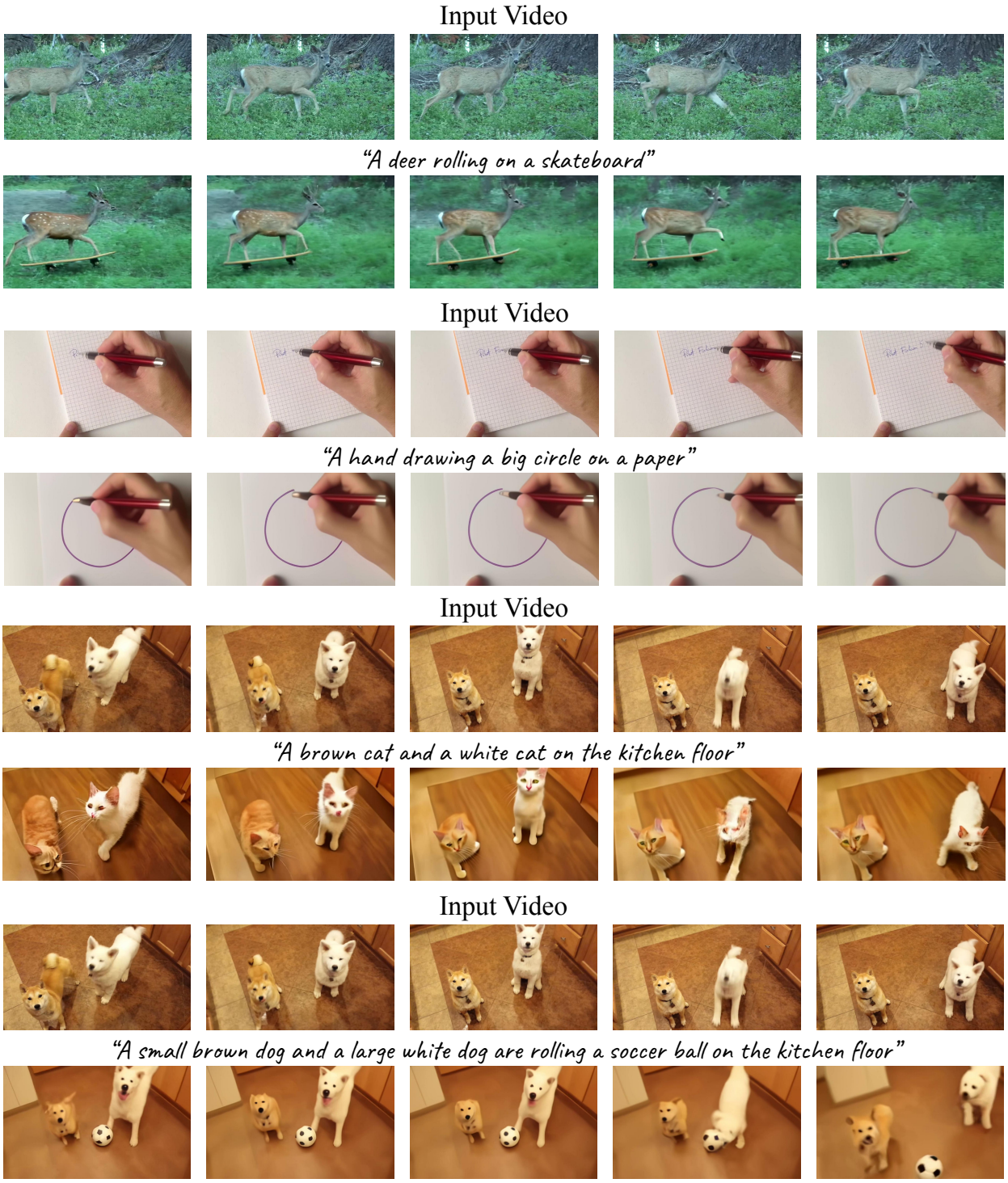

Figure 14: ***Additional Video Editing Examples (2/4)***

Figure 15: ***Additional Video Editing Examples (3/4)***

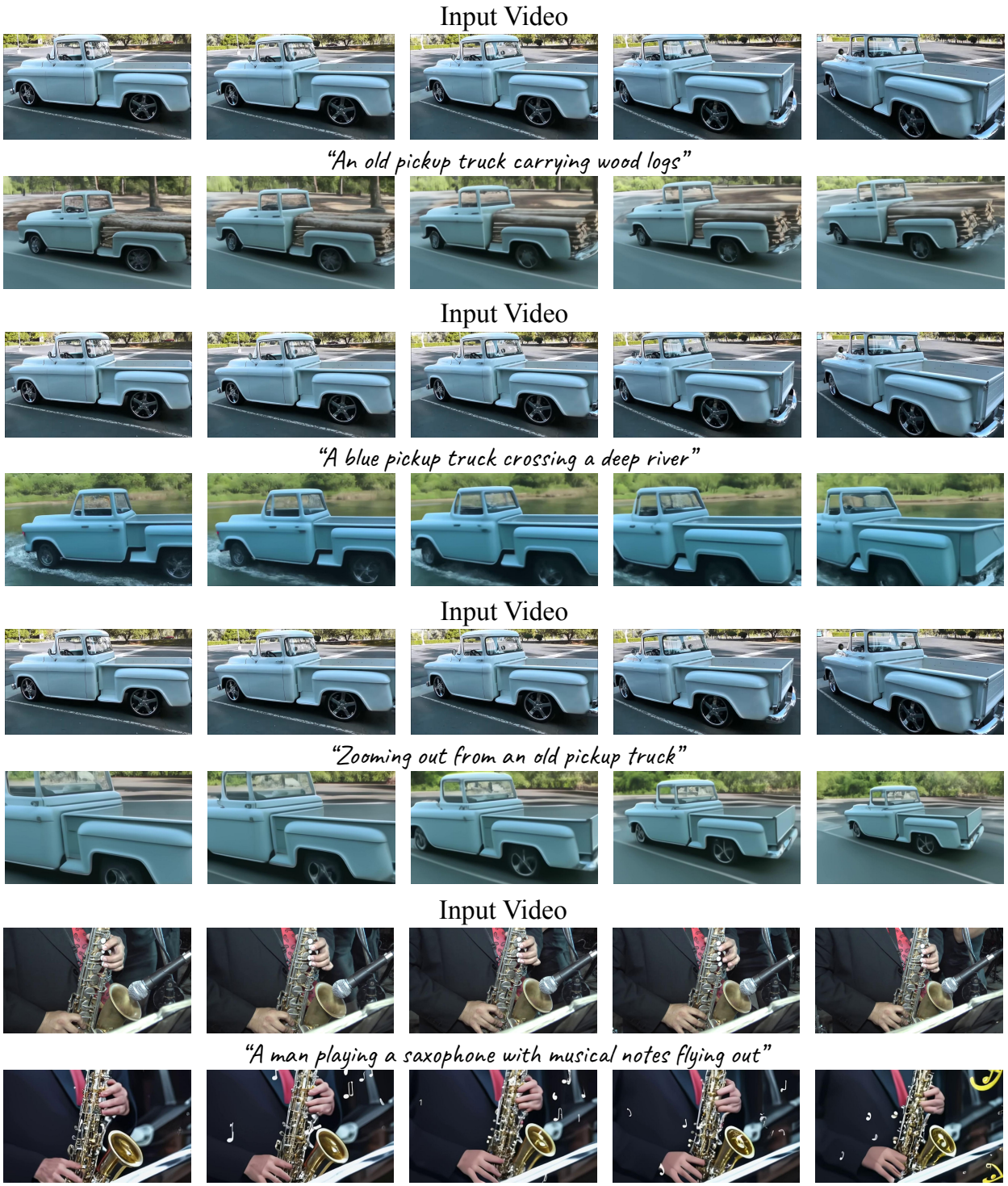

Figure 16: ***Additional Video Editing Examples (4/4)***

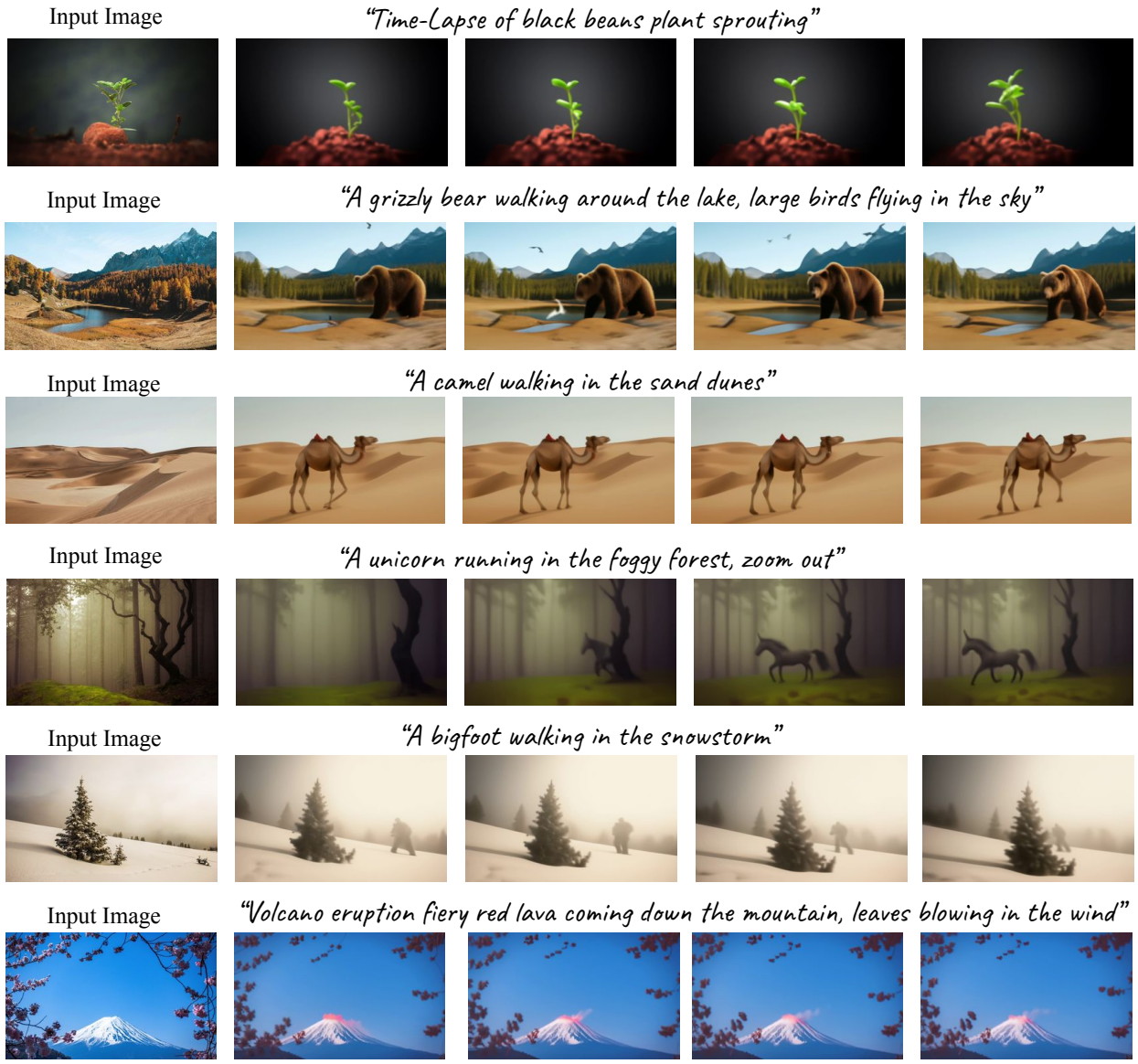

Figure 17: ***Additional Image-to-Video Examples***

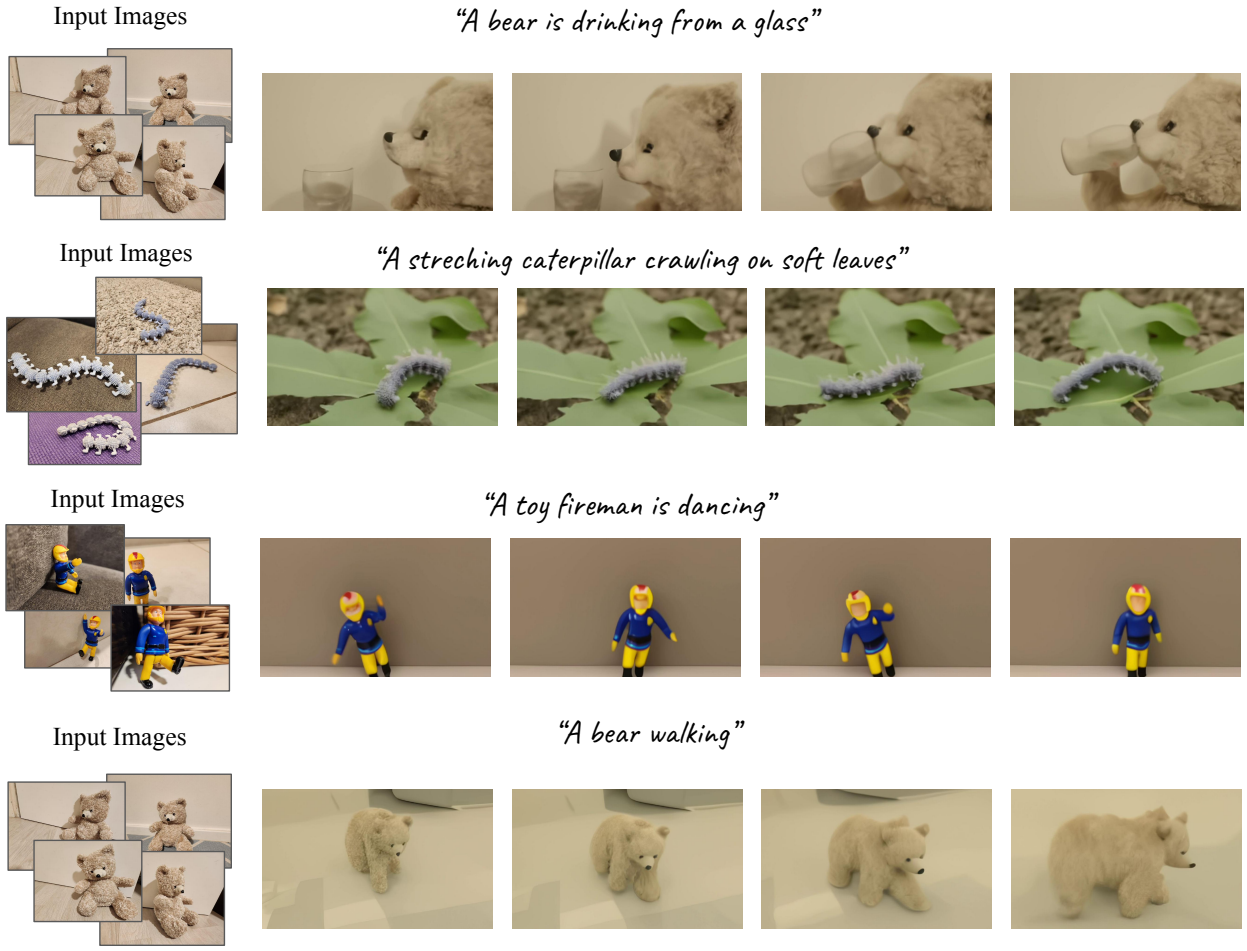

Figure 18: ***Additional Subject-Driven Video Generation***