

Dreamix: Video Diffusion Models are General Video Editors

Eyal Molad^{* 1}, Eliahu Horwitz^{* ‡ 1,2}, Dani Valevski^{* 1}, Alex Rav Acha¹, Yossi Matias¹, Yael Pritch¹, Yaniv Leviathan^{† 1}, Yedid Hoshen^{† ‡ 1,2}

¹Google Research, ²The Hebrew University of Jerusalem

<https://dreamix-video-editing.github.io/>

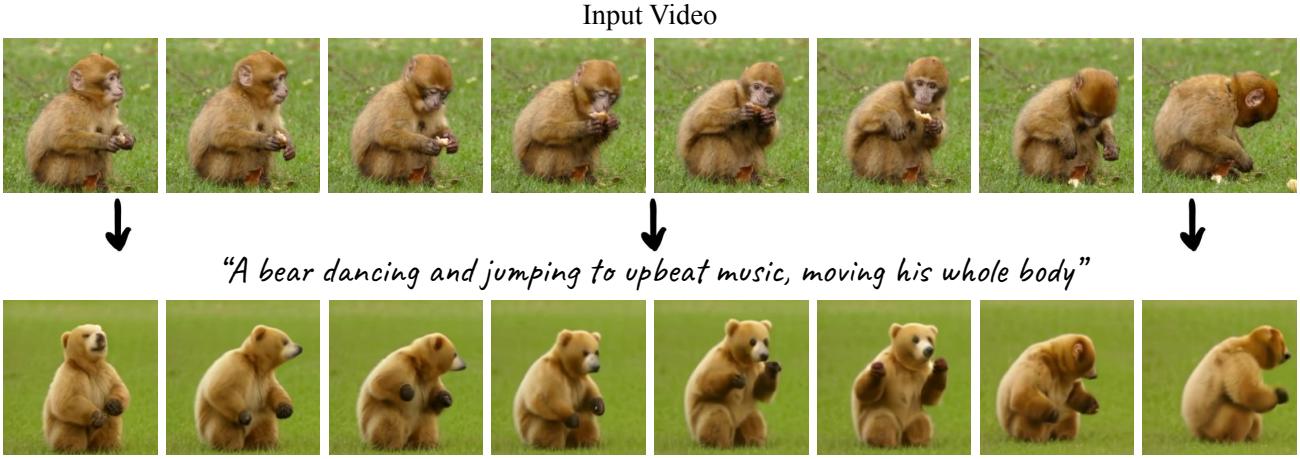


Figure 1. *Video Editing with Dreamix*: Frames from a video conditioned on the text prompt “A bear dancing and jumping to upbeat music, moving his whole body”. Dreamix transforms the eating monkey (top row) into a dancing bear, affecting appearance and motion (bottom row). It maintains fidelity to color, posture, object size and camera pose, resulting in a temporally consistent video

Abstract

Text-driven image and video diffusion models have recently achieved unprecedented generation realism. While diffusion models have been successfully applied for image editing, very few works have done so for video editing. We present the first diffusion-based method that is able to perform text-based motion and appearance editing of general 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 obtaining 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 a new, mixed objective that jointly finetunes with full temporal attention and with temporal attention masking. We further introduce a new framework for image animation. We first transform the image into a coarse video by simple image processing operations such as replication and perspective geometric projections, and then use our general video editor to animate it. As a fur-

ther application, we can use our method for subject-driven video generation. Extensive qualitative and numerical experiments showcase the remarkable editing ability of our method and establish its superior performance compared to baseline methods.

1. Introduction

Recent advancements in generative models [7, 16, 47] and multimodal vision-language models [29] have paved the way to large-scale text-to-image models capable of unprecedented generation realism and diversity [3, 26, 31, 32, 35]. These models have ushered in a new era of creativity, applications, and research efforts. Although these models offer new creative processes, they are limited to synthesizing *new* images rather than editing *existing* ones. To bridge this gap, intuitive text-based image editing methods offer text-based editing of generated and real images while maintaining some of their original attributes [6, 13, 22, 40, 41]. Similarly to images, text-to-video models have recently

^{*} Equal contribution.

[†] Equal advising.

[‡] Performed this work while working at Google.

been proposed [15, 18, 37, 48], but there are currently very few methods using them for video editing.

In text-guided video editing, the user provides an input video and a text prompt which describes 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 standard image editing, as it requires synthesizing new motion, not merely modifying visual appearance. It also requires temporal consistency. As a result, applying image-level editing methods e.g. SDEdit [24] or Prompt-to-Prompt [13] 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 [41]. 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. Second, we further improve the fidelity to the original video by finetuning the generation model on the original video. Finetuning ensures the model has knowledge of the high-resolution attributes of the original video. A naive 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, mixed finetuning approach, in which the VDMs are also finetuned on the collection of individual frames of the input video while discarding their temporal order. Technically, this is achieved by masking the temporal attention. Mixed finetuning significantly improves the quality of motion edits.

As a further contribution, we leverage our video editing model to propose a new framework for image animation (see Fig. 2). This has several applications including: animating the objects and background in an image, creating dynamic camera motion, etc. We do this by simple image processing operations, e.g. frame replication or geometric image transformation, to create a coarse video. We then edit it with our Dreamix video editor. We also use our novel finetuning approach for subject-driven video generation, i.e. a video version of Dreambooth [33]. We perform an extensive qualitative study and a human evaluation, showcasing the remarkable abilities of our method. We compare our method against the state-of-the-art baselines, demonstrating superior results. To summarize, our main contributions are:

1. Proposing the first method for general text-based appearance and motion editing of real-world videos.

2. Proposing a novel mixed finetuning model that significantly improves the quality of motion edits.
3. Presenting a new framework for text-guided image animation, by applying our video editor method on top of simple image preprocessing operations.
4. Demonstrating subject-driven video generation from a collection of images, leveraging our novel finetuning method.

2. Related Work

2.1. Diffusion Models for Synthesis

Deep diffusion models recently emerged as a powerful new paradigm for image generation [16, 39], and have their roots in score-matching [20, 38, 42]. They outperform [9] the previous state-of-the-art approach, generative adversarial networks (GANs) [12]. While they have multiple formulations, EDM [21] showed they are equivalent. Outstanding progress was made in text-to-image generation [3, 31, 32, 35], 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 [18] and text-to-video extensions by [15, 37]. Another line of work extends synthesis to various image reconstruction tasks [8, 17, 23, 34, 36], [19] 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 [11, 27, 28, 43, 45]. Editing methods have recently adopted diffusion models [2, 4, 44]. Several works proposed to use text-to-image diffusion models for editing rather than text-conditioned synthesis. SDEdit [24] proposed to add targeted noise and other corruptions to an input image, and then use diffusion models for reversing the process. It can perform significant image edits, while losing some fidelity to the original image. Prompt-to-Prompt [13] (and later Plug-and-Play [40] and [25]) perform semantic edits by mixing activations extracted with the original and target prompts. For InstructPix2Pix [6] this is only needed at test time. Other works (e.g. [10, 33]) use finetuning and optimization to allow for personalization of the model, learning a special token describing the content. UniTune [41] and Imagic [22] finetune on a single image, allowing better editability while maintaining good fidelity. However, the above methods are image-centric and do not take temporal information into account. Text2Live [5] allows some texture-based video editing but are not diffusion-based and cannot edit motion. A concurrent paper, Tune-a-Video [46] preform video editing by inflating a text-to-image model to

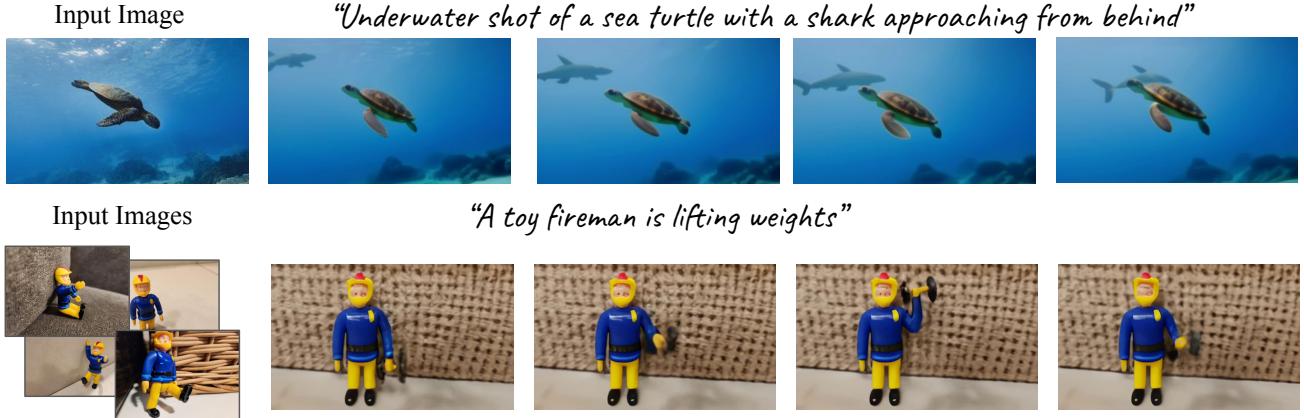


Figure 2. *Image-to-Video editing with Dreamix*: Dreamix can create videos based on image and text inputs. In the single image case (first row) it is able to instill complex motion in a static image, adding a moving shark and making the turtle swim. In this case, visual fidelity to object location and background was preserved but the turtle direction was flipped. In the subject-driven case (second row) Dreamix is able to extract the visual features of a subject given multiple images and then animate it in different scenarios such as weightlifting

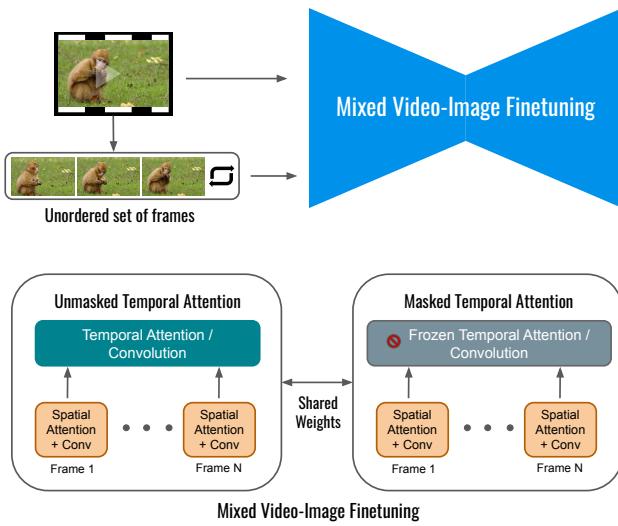


Figure 3. *Mixed Video-Image Finetuning*: Finetuning the VDM on the input video alone limits the extent of motion change. Instead, we use a mixed objective that beside the original objective (bottom left) also finetunes on the unordered set of frames. This is done by using “masked temporal attention”, preventing the temporal attention and convolution from being finetuned (bottom right). This allows adding motion to a static video

learn temporal consistency. 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 video smoothness.

3. Background: Video Diffusion Models

Denoising Model Training. Diffusion models rely on deep denoising neural network D_θ . Let us denote the groundtruth 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 groundtruth video v given the noisy input video z_s , the time s , prompt t and conditioning video c . The model is trained on a (typically very large) training corpus \mathcal{V} consisting of pairs of video v and text prompts t . The optimization objective is:

$$\mathcal{L}_\theta(v) = \mathbb{E}_{\epsilon \sim N(0, \mathbf{I}), s \in \mathcal{U}(0, 1)} \|D_\theta(z_s, s, t, c) - v\|^2 \quad (1)$$

Sampling from Diffusion Models. The key challenge in diffusion models is to use the denoiser network D to sample from the distribution of videos conditioned on the text prompt t and conditioning video c , $P(v|t, c)$. While the derivation of such sampling rule is non-trivial (see e.g. [21]), the implementation of such sampling is relatively simple in practice. We follow [15] 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

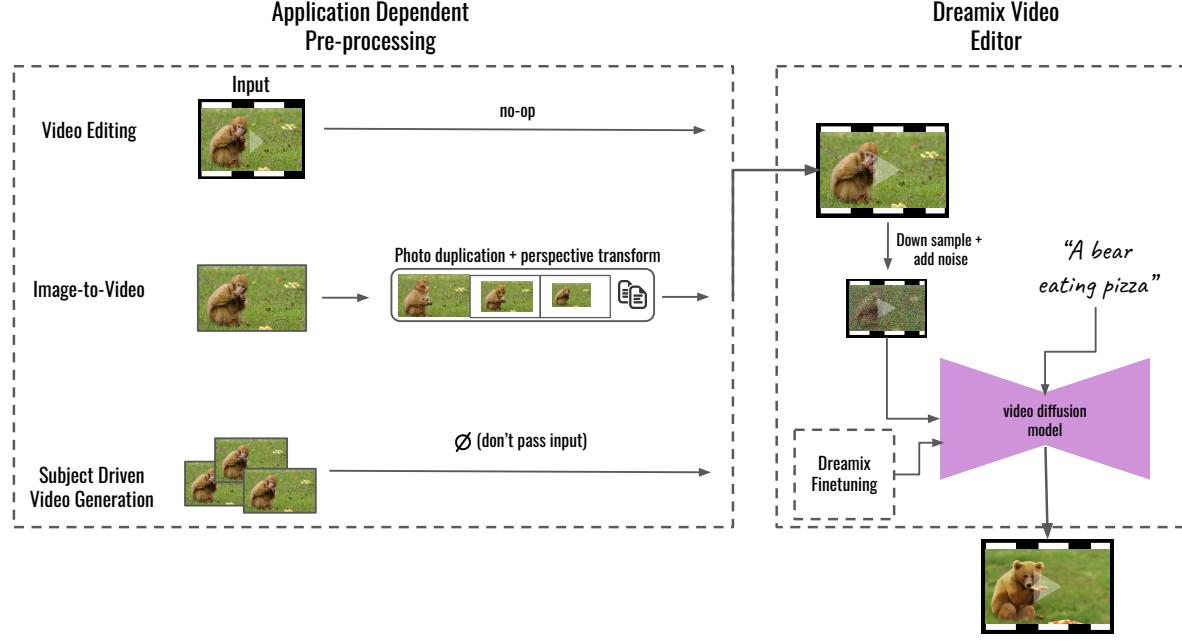


Figure 4. **Inference Overview:** Our method supports multiple applications by application dependent pre-processing (left), converting the input content into a uniform video format. For image-to-video, the input image is duplicated and transformed using perspective transformations, synthesizing a coarse 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

the high computational complexity. Several diffusion models overcome this using cascaded architectures. We use Imagen-Video [15], 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 A.

4. General 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. Text-Guided 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 [24], there are some video specific technical challenges that we will describe below. Note, that this by itself does not result in sufficiently high-fidelity video editing. We present a novel finetuning objective for mitigating this issue 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 text-to-video model. For $s = 0$, no noise is added, 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. For the non-finetuned base model, values of $s \in [0.4, 0.85]$ typically worked best.

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-



Figure 5. **Video Motion Editing:** Dreamix can significantly change the actions and motions of subjects in a video, making a puppy leap in this example. The resulting video maintains temporal consistency while preserving the unedited details

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 A). 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 issue 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 $u = \{x_1, x_2, \dots, x_M\}$. We use a rare string t^* as the text prompt, following [33]. 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 exactly reconstruct the input video v given its noisy versions z_s .

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

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 given by:

$$\mathcal{L}_{\theta}^{frame}(u) = \mathbb{E}_{\epsilon \sim N(0, I), s \in \mathcal{U}(0, 1)} \|D_{\theta'}^a(z_s, s, t^*, c) - u\|^2 \quad (3)$$

We train the objectives jointly and denote this *mixed finetuning*:

$$\theta = \arg \min_{\theta'} \alpha \mathcal{L}_{\theta'}^{vid}(v) + (1 - \alpha) \mathcal{L}_{\theta'}^{frame}(u) \quad (4)$$

Where α is a hyperparameter weighting between the two objectives, (see Fig. 3). 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 A).

4.3. Hyperparameters

Our method has several hyperparameters. For inference time, we have the noise scale $s \in [0, 1]$ where $s = 1$ corresponds to standard sampling without using the degraded input video. For finetuning, we have the number of finetuning steps FT_{steps} , learning rate lr , and mixing weight α between the video and frames finetuning objectives (see Sec. 4.2). See Fig. 7 for a qualitative analysis of hyperparameter impact, and Sec. 6.3 for a quantitative analysis. Additional implementation details may be found in Appendix A.

5. Applications of Dreamix

The method proposed in Sec. 4, can naturally be used to edit motion and appearance in real-world videos. In this section, we propose a framework for using our Dreamix

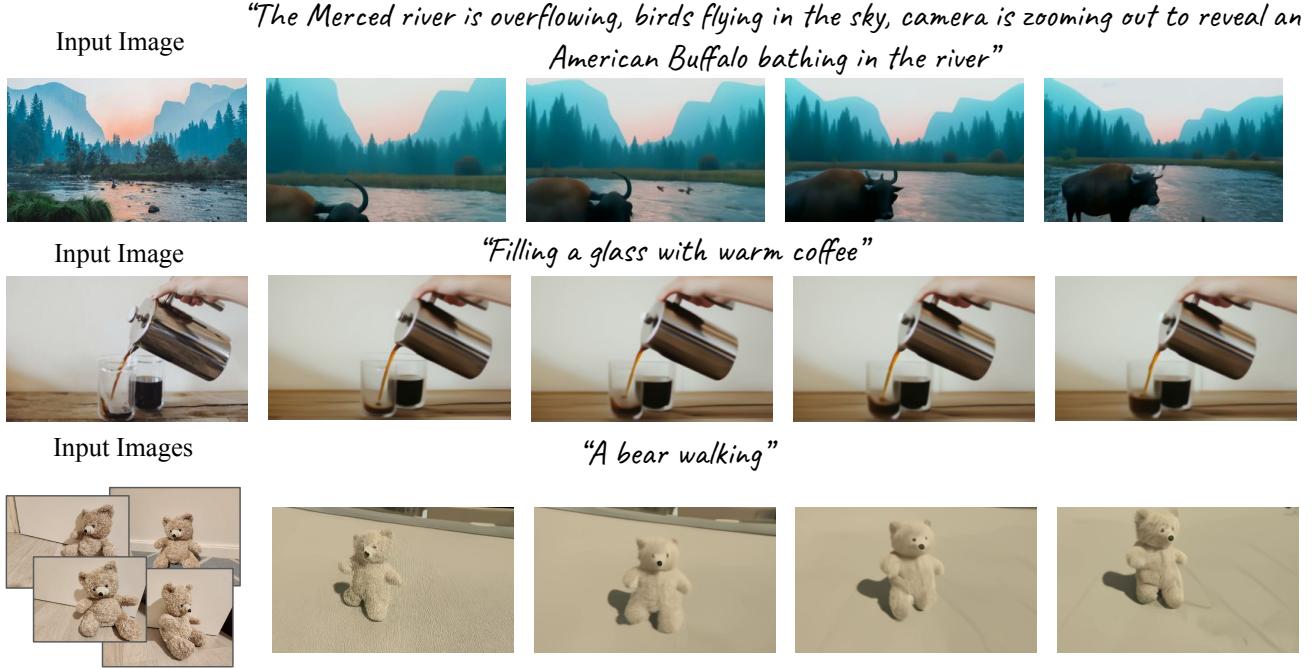


Figure 6. **Additional Image-to-Video Results:** Dreamix can generate camera effects such as zoom-out by combining a text prompt with a coarse video obtained by applying image-processing transformations on the input image. First row - the image is zoomed out to reveal a bathing buffalo. Dreamix can also instill motion in a static image as in the second row where the glass is gradually filled with coffee. Third row - we animate a provided subject based on a small number of independent images

video editor for general, text-conditioned image-to-video editing, see Fig. 4 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 ($\alpha = 0$). 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.

Dreamix for subject-driven video generation. We propose to use Dreamix for text-conditioned video generation

given an image collection. The input to our method is a set of images, each containing the subject of interest. This can potentially 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 our novel finetuning method from Sec. 4.2, where we only use the masked attention finetuning ($\alpha = 0$). After finetuning, we use the text-to-image model *without* a conditioning video, but rather only using a text prompt (which includes the special token t^*).

6. Experiments

6.1. Qualitative Results

We showcase the results of Dreamix, demonstrating unprecedented video editing and image animation abilities.

Video Editing. In Fig. 1, we change the motion to dancing and the appearance from monkey to bear, 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. 5, orangutan in Fig. 13), and can control camera movements (zoom-out example in Fig. 14). Dreamix can generate smooth visual modifications that align with the temporal information in the input video. This includes adding effects (field in Fig. 10, saxophone in the Fig. 14), adding objects (hat in Fig. 10 and skateboard

Table 1. **Ablation Study:** Users were asked to compare text-guided video edits of different variants of our method: no finetuning (no ft.), video-only finetuning (Video ft.), the proposed mixed finetuning (Mixed-ft). The object category includes adding/replacing objects. The background category includes background, color or texture changes. Mixed finetuning is important in motion editing and background change scenarios

Edit Type	# Edits	No ft.	Video-ft.	Mixed-ft.	None
Motion	36	17%	25%	47%	11%
Style	15	67%	7%	20%	6%
Object	44	36%	30%	18%	16%
Background	32	19%	28%	44%	9%

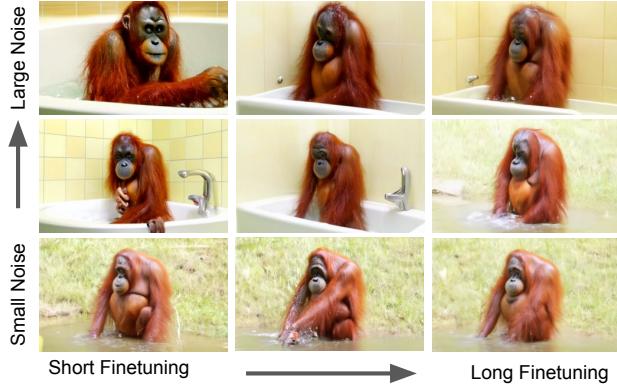


Figure 7. **Hyperparameter tradeoffs:** We compare the effect of noise magnitude and number of finetuning iterations on edited videos. The original frame is on the bottom left. The rest are frames generated by different parameters for the prompt "An orangutan with orange hair bathing in a bathroom". We can observe that higher noise allows for larger edits but reduces fidelity. More finetuning iterations improve fidelity at higher noises. The best results are obtained for high noise and a large number of finetuning iterations

in Fig. 11) or replacing them (robot in Fig. 10), changing the background (truck in the Fig. 14).

Image-driven Videos. When the input is a single image, Dreamix can use its video prior to add new moving objects (camel in Fig. 9), inject motion into the input (turtle in Fig. 2 and coffee in Fig. 6), or create new camera movements (buffalo in Fig. 6). Our method is unique in being able to do this for general, real-world images.

Subject-driven Video Generation. Dreamix can take an image collection showing the same subject and generate new videos with this subject in motion. This is unique, as previous approaches could only do this for images. We demonstrate this on a range of subjects and actions including: the weight-lifting toy fireman in Fig. 2, walking and drinking bear in Fig. 6 and Fig. 9. It can place the subjects

Table 2. **Baseline Comparison:** Users were asked to rate videos edited by different methods by visual quality, fidelity to the base video and alignment with the text prompt. We define an edit as successful when it receives a mean score larger than 2 in all dimensions. We observe that our method significantly outperforms the others in producing successful edits

Method	Quality	Fidelity	Alignment	Success
ImagenVid	2.99 ± 0.95	2.21 ± 0.97	4.04 ± 1.12	40%
PnP	1.76 ± 0.78	3.61 ± 0.96	3.09 ± 1.35	15%
Ours	3.09 ± 0.83	3.29 ± 0.99	3.50 ± 1.34	73%

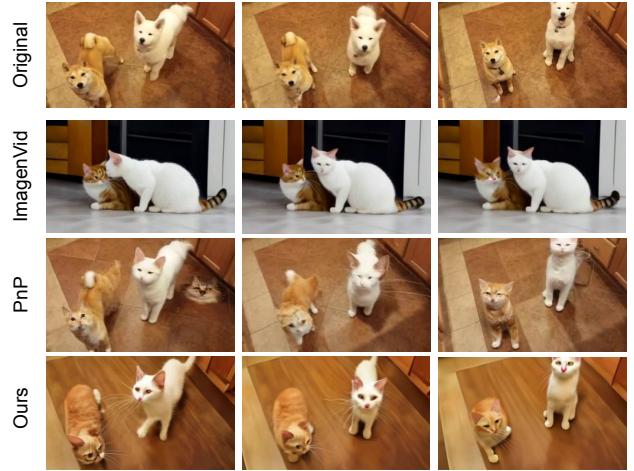


Figure 8. **Comparison to Baseline Methods:** Top row - original video. Second row - Imagen-Video. Although the quality and text alignment are high, there is no resemblance to our original video. Third row - PnP, an image-based text-guided editing method. Although the scene is well preserved (e.g. tiles), the temporal consistency is low. Bottom - ours. The edit aligns well with the text prompt while preserving many original details and generating a high quality video

in new surroundings, e.g., moving caterpillar to a leaf in Fig. 9 and even under a magnifying glass in Fig. 9.

6.2. Baseline Comparisons

Baselines. We compare our method against two baselines:

Text-to-Video. Directly mapping the text prompt to a video, without conditioning on the input video using Imagen-Video.

Plug-and-Play (PnP). A common approach for video editing is to apply text-to-image editing on each frame individually. We apply PnP [40] (a SoTA method) on each frame independently and concatenate the frames into a video.

Quantitative Comparison. We performed a human-

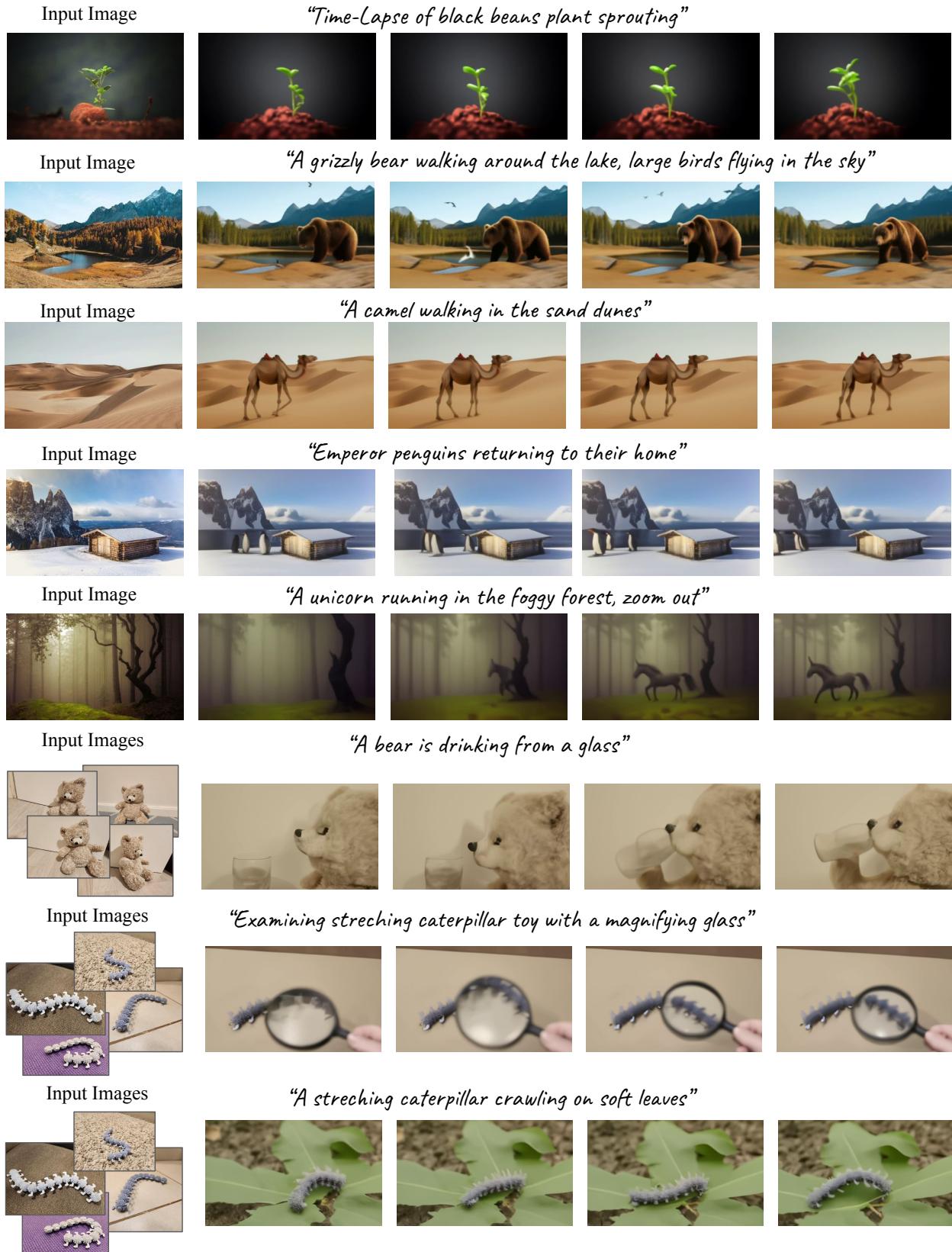


Figure 9. **Additional Results:** Image-to-Video, and subject-driven video generation

rated evaluation of Dreamix and the baselines on a dataset of 29 videos taken from YouTube-8M [1], and 127 text prompts, across different categories. We used a single hyperparameter set for all methods. Each edited video was rated on a scale of 1 – 5 to evaluate its visual quality, its fidelity to the unedited details of the base video and its alignment with the text prompt. We collected 4 – 6 ratings for each edited video. The results of the evaluation can be seen in Tab. 2. We also highlight the success rate of each method, where a successful edit is one that received a mean score larger than 2 in all dimensions. We observe that frame by frame methods like Plug-and-Play [40] perform poorly in terms of visual quality as they create flickering effects due to the lack of temporal input. Moreover, Plug-and-Play sometimes ignored the edit altogether, resulting in low alignment and high fidelity. The Text-to-Video baseline ignores the edited video, resulting in low fidelity. Our method balances between the three dimensions, resulting in a high success rate.

Qualitative Comparison. Fig. 8 presents an example of a video edited by Dreamix and the two baselines. The text-to-video model achieves low fidelity edits as it is not conditioned on the original video. PnP preserves the scene but lacks consistency between different frames. Dreamix performs well in all three objectives.

6.3. Ablation Study

We conducted a user study comparing our proposed mixed finetuning method (See Sec. 4.2) to two ablations: no finetuning and finetuning on the video only (but not the independent frames). Our dataset contained 29 videos (each of 5 seconds) taken from YouTube-8M [1], and a total of 127 text prompts. Additional details are provided in Appendix B. The results are presented in Tab. 1. Our main observations are: *Motion* changes require high-editability. Frame-based finetuning typically outperformed video-only finetuning. Denoising without finetuning worked well for *style transfer*, finetuning was often detrimental. Preserving fine-details in *background, color or texture* changes required finetuning.

7. Discussion

In this section, we analyse the limitations of our method, potential ways to address them and future applications.

Hyperparameter Selection. Optimal hyperparameter values e.g., noise strength, can change between prompts. Automating their selection will make our method more user friendly. It can be done by learning a regressor from (input video, prompt) to the optimal hyperparameters. Creating a training set with the optimal hyperparameters per-edit (e.g. as judged by users) is left for future work.

Automatic Evaluation Metrics. In our preliminary study, we found that automatic evaluation metrics (e.g.

CLIP Score [14] for alignment) are imperfectly correlated with human preference. Future work on automatic video text-editing metrics should address this limitation. Having effective metrics will also support labeling large datasets for the automatic hyperparameter selection suggested above.

Frequency of Objects in Dataset and Editability. Not all prompt-video pairs yield successful edits (as can be seen in Tab. 2). Being able to determine the successful pairs in advance, will speed up the creative editing process. In preliminary work, we found that edits containing objects and actions that frequently occurred in the training dataset resulted in better edits than rarer ones. This suggests that an automatic method for prompt engineering is a promising direction.

Computational Cost. VDMs are computationally expensive. Finetuning our model, containing billions of parameters, requires large hardware accelerators around 30 minutes per video. Speeding it up and lowering the computational cost, will allow our method to be used for a larger set of applications.

Future Applications. We expect Dreamix to have many future applications. Several promising ones are: motion interpolation between an image pair, text-guided inpainting and outpainting.

8. Conclusion

We presented a general approach for text-conditioned editing using video diffusion models. Beyond video editing, we introduced a new framework for image animation. We also applied our method to subject-driven video generation. Extensive experiments demonstrated the unprecedented results of our method.

9. 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.

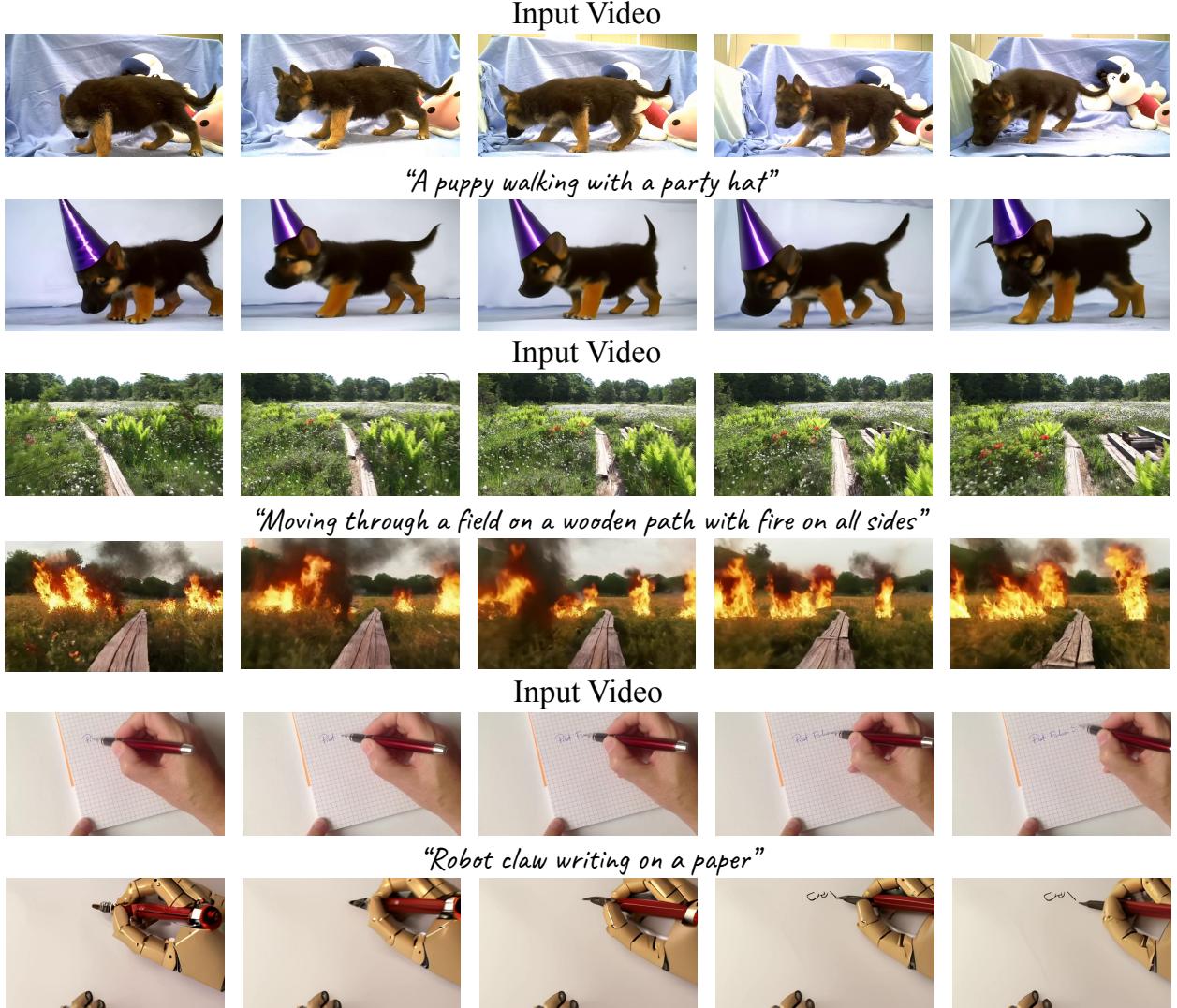


Figure 10. *Additional Video-to-Video results*

10. Acknowledgements

We thank Ely Sarig for creating the video, Jay Tenenbaum for the video narration, Amir Hertz for the implementation of our eval baseline, Daniel Cohen-Or, Assaf Zomet, Eyal Segalis, Matan Kalman and Emily Denton for their valuable inputs that helped improve this work.

References

- [1] 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. [9](#), [13](#)
- [2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. [2](#)
- [3] 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. *arXiv preprint arXiv:2211.14305*, 2022. [1](#), [2](#)
- [4] 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*, pages 18208–18218, 2022. [2](#)
- [5] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kassten, 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*, pages 707–723. Springer, 2022. [2](#)
- [6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. [1](#), [2](#)

- [7] 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*, pages 11315–11325, 2022. 1
- [8] 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*, pages 12413–12422, 2022. 2
- [9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2
- [10] 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. 2
- [11] 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. 2
- [12] 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. 2
- [13] 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. 1, 2
- [14] 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. 9
- [15] 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*, 2022. 2, 3, 4, 13
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2
- [17] 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, 2022. 2
- [18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2
- [19] Eliah Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022. 2
- [20] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 2
- [21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 2, 3
- [22] 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. *arXiv preprint arXiv:2210.09276*, 2022. 1, 2
- [23] 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*, pages 11461–11471, 2022. 2
- [24] 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. 2, 4
- [25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2
- [26] 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. 1
- [27] 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*, pages 2337–2346, 2019. 2
- [28] 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*, pages 2085–2094, 2021. 2
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1
- [30] 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. 13
- [31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2
- [33] 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. 2022. 2, 5
- [34] 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*, pages 1–10, 2022. 2
- [35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2
- [36] 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*, 2022. 2
- [37] 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. 2
- [38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2
- [39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [40] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 1, 2, 7, 9
- [41] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 1, 2
- [42] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 2
- [43] 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)*, pages 13769–13778, October 2021. 2
- [44] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*, 2022. 2
- [45] 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*, pages 8798–8807, 2018. 2
- [46] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022. 2
- [47] 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*, 2022. 1
- [48] 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. *arXiv preprint arXiv:2212.05199*, 2022. 2

Appendix

A. Implementation Details

A.1. Architecture

All of our experiments were preformed on Imagen-Video [15], a pretrained cascaded video diffusion model, with the following components:

1. a T5-XXL [30] text encoder, that computes embeddings from the textual prompt. This embeddings are then used as a condition 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 text and on 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).

A.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). Good combinations of finetuning hyperparameters are:

- $\alpha = 1.0$ (video only finetuning), $FT_{steps} = 64$
- $\alpha = 0.35$ (mixed video / video-frame finetuning), $FT_{steps} \in [200, 300]$
- $\alpha = 0$ (video-frame only finetuning), $FT_{steps} \in [50, 150]$

The learning rate (lr) we use in all experiments is $6 \cdot 10^{-6}$, much lower then the value used for pretraining the models.

A.3. Sampling

We use a DDIM sampler with stochastic noise correction, following [15]. 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.

Noise strength. We got the best results for the following values of noise strength s : for non-finetuned models, $s \in [0.4, 0.85]$ and for finetuned models, $s \in [0.95, 1.0]$.

B. Human evaluations details

We performed human evaluations for the baseline comparison and the ablation analysis. Both evaluations were conducted by a panel of 10 human raters, over a dataset of 29 videos with 127 edit prompts. The dataset videos were selected from YouTube-8M [1] and show animals, people performing actions, vehicles, and other objects. The edit prompt categories are detailed in Tab. I of the main paper. The video resolution shown to raters was 350×200 .

In the ablation analysis the raters selected the best edited video out of 12 hyperparameter combinations.

In the baseline comparison, the raters saw the original video alongside an edited video and answered the following questions:

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?

We used a single set of hyperparameters in the baseline eval: $\alpha = 0.35$; $FT_{steps} = 300$; $s = 1$.

C. 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/LrPKL7j0ldI>
- 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/>

Input Video



"A beach with palm trees and swans in the water"



Input Video



"A knife is cutting a cake on a red plate"



Input Video



"A deer rolling on a skateboard"



Input Video



"A hand drawing a big circle on a paper"



Figure 11. **Additional Video Editing Examples (1/4)**

Input Video



"A brown cat and a white cat on the kitchen floor"



Input Video



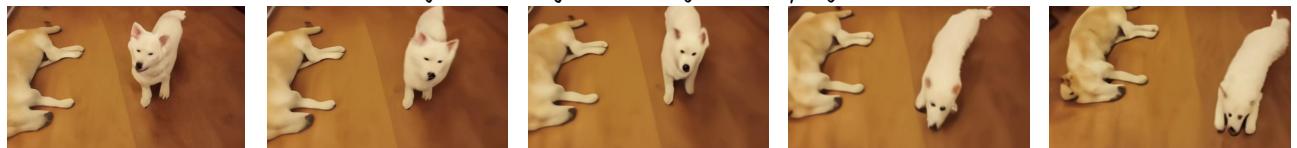
"A small brown dog and a large white dog are rolling a soccer ball on the kitchen floor"



Input Video



"A small brown dog and a large white dog are sleeping on the kitchen floor"



Input Video



"Stirring noodles in a pot"



Figure 12. *Additional Video Editing Examples (2/4)*

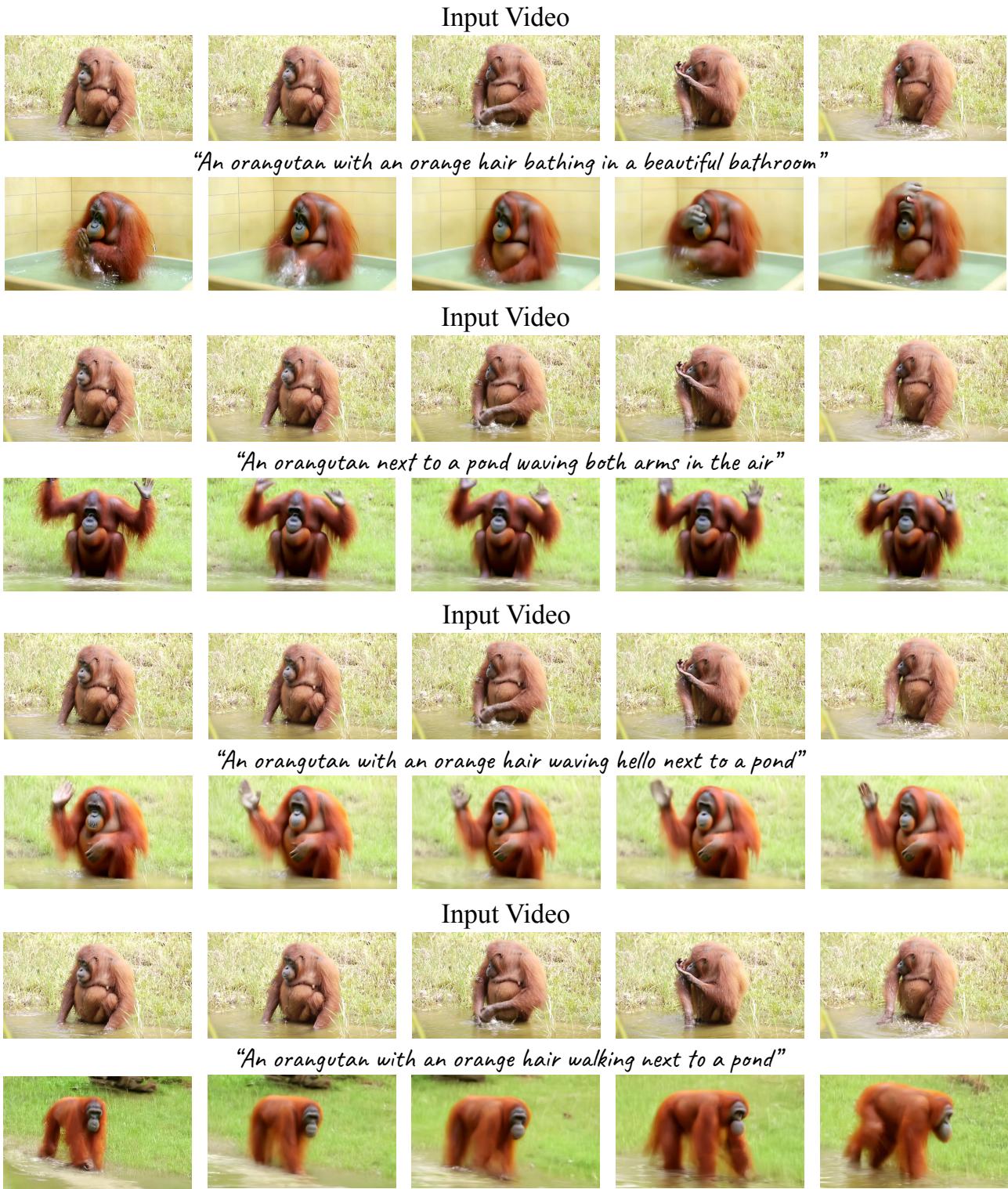


Figure 13. *Additional Video Editing Examples (3/4)*

Input Video



"An old pickup truck carrying wood logs"



Input Video



"A blue pickup truck crossing a deep river"



Input Video



"Zooming out from an old pickup truck"



Input Video



"A man playing a saxophone with musical notes flying out"



Figure 14. *Additional Video Editing Examples (4/4)*



Figure 15. *Additional Image-to-Video Examples*



Figure 16. *Additional Subject-Driven Video Generation*