Supplementary Material for Structure and Content-Guided Video Synthesis with Diffusion Models

Patrick Esser Johnathan Chiu Parmida Atighehchian
Jonathan Granskog Anastasis Germanidis
Runway
https://research.runwayml.com/gen1

Training time and cost After initialization from the pretrained image weights, our model was trained for an additional 125k steps taking approximately 32,640 GPU hours on A100. This corresponds to roughly 20% of the compute costs for the image pretraining, showing the benefits of reusing weights. See section 3 of the main paper for details about the different stages of training.

Training data creation As described in section 3, we train on a mixed dataset consisting of both images and videos. Our video dataset was created from a set of videos that have an aspect ratio within 10% of $16:9$ and a smaller side larger than 700 pixels. We split the videos into clips by detecting cuts using both content-aware and threshold-based detection as implemented in PySceneDetect. We further split long clips until each of them is less than 20 seconds long and discard clips that have less than 18 frames. For our internal collection of videos, this process resulted in 6.4M clips that we use for training. For training, we additionally crop and downsample the videos to a lower resolution.

Evaluation data creation For evaluation, we test on stock footage and videos from DAVIS [4]. We create edited captions to test with by first producing an initial caption of the input video with BLIP [2] and then asking GPT-3 [1] to create edit prompts and matching captions given the initial caption. The initial caption is generated from the first frame of the video. The edited caption is achieved by requesting GPT-3 to modify the original caption such that it describes an edited version of the video instead of the original video. For example, if the BLIP-predicted caption is *there is a bear that is walking through the forest* then GPT-3 might suggest an edit such as *edit it so that the bear is in space* and a final caption such as *a space bear walking through the stars*. See Tab. 3 for a few examples of captions, edit prompts and the edited captions for some videos from the DAVIS dataset.

The full input prompt to GPT-3 to produce edit prompts and modified captions was:

> You are the most creative VFX artist in the world, extremely skilled at coming up with original ideas for editing videos. You are tasked with coming up with triplets of "caption, edit, modified caption" for possible video editing tasks. Cover a wide variety of creative edits, including editing the content, background, or style of the video. Try to be creative and come up with edits that are not obvious. For example, avoid changing the color of objects.
> Examples:
>
> - a man looking at the camera, in the the style of pinscreen animation, pinscreen animation of a man looking at the camera
> - kite-surfer in the ocean, same scene during sunset, kite-surfer in the ocean at sunset,
> - car on a road in the countryside, make the car cyberpunk, cyberpunk neon car on a road in the countryside
>
> New Edit:
> { BLIP caption },

where { BLIP caption } represents the initial caption predicted for the first frame of the video.

Parameterization In early experiments on moving MNIST, we observed fewer color artifacts with a v-parameterization compared to an ϵ-parameterization. Thus, we conducted all further experiments with v-parameterizations. Further investigation points to a problem with the diffusion schedule, where the terminal signal-noise ratio (SNR) of the diffusion process remains too large [3]. The ϵ-parameterization weighs the loss by the SNR [5], giving less importance to steps with a low SNR. In contrast, the v-parameterization results in a "SNR+1" weighting, giving higher weights at low SNR steps and counteracting issues caused by schedules with non-zero terminal SNR.

method	frame consistency	prompt consistency	w/ depth preferred
w/ depth	0.9648 $\pm$0.0031	0.2805 $\pm$0.0065	–
w/ hed	0.9813 $\pm$0.0031	0.2652 $\pm$0.0072	68.57%
w/o struct.	0.9482 $\pm$0.0034	0.2769 $\pm$0.0062	64.71%
w/o temp.	0.9396 $\pm$0.0035	0.2838 $\pm$0.0060	57.14%

Table 1. Ablations on structure conditioning and temporal UNet.

Figure 1. Top-left to bottom-right: Three frames of input, w/ depth, w/o temp and w/ hed for "a space bear walking through the stars".

Format of structure representation We also compare our depth-conditioned model against a model receiving edge maps extracted with holistically-nested edge detection (HED) [6] as input instead of depth maps. We denote this model *w/ hed* in Tab. 1. We observe a significantly lower prompt consistency ($\times 2.23\sigma$) for the edge-conditioned model. This suggests that edge maps mix content and structure more strongly than depth maps. We show a qualitative example of this in Fig. 1, where the edge map reveals the background to be a wall and the edge-conditioned model fails to edit the background to stars. However, the frame consistency of this model is higher, suggesting that increased structure information can improve temporal stability. Still, through a user study, we show that depth-conditioning is preferred (68.57%), demonstrating a better trade-off between frame- and prompt consistency. We note that the optimal structure representation will depend on the application.

Additionally, a model not provided a representation of structure (*w/o struct*) achieves lower frame consistency and slightly worse prompt consistency. Users also report higher preference for our model that receives depth maps.

Temporal connections In Tab. 1, we also show metrics for a model trained without temporal connections (*i.e.* only spatial convolutions and attention), which is labeled with *w/o temp*. It achieves similar prompt consistency to our spatio-temporal model but much lower frame consistency. Users also show a slight preference for our model with additional processing along the time axis. Note also that temporal connections have a larger impact on the temporal consistency compared to providing a structure representation. However, this is likely partly due to our structure representation being predicted independently for each frame. A depth map predictor built for video input could increase temporal stability.

Raw data and additional comparisons We include the raw data of Fig. 6 and Fig. 7 in Tab. 2. Fig. 2-8 contain additional results for text based edits, Fig. 9-13 for image based edits. Fig. 14 shows a qualitative comparison.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 1, 17

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 17

[3] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed, 2023. 1

[4] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 1, 17

[5] Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*, 2021. 1

[6] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 2

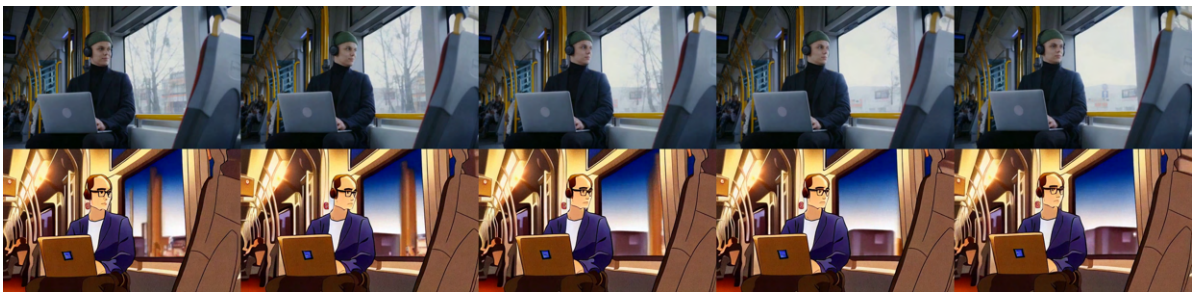
Prompt	Driving Video (top) and Result (bottom)
pencil sketch of a man looking at the camera, black and white	
a man using a laptop inside a train, anime style	
a woman and man take selfies while walking down the street, claymation	
oil painting of a man driving	
low-poly render of a man texting on the street	

Figure 2. Additional results for text-to-video-editing.

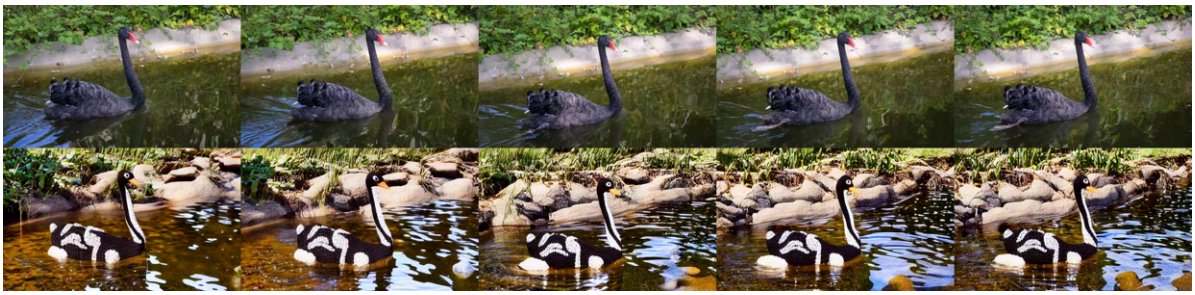
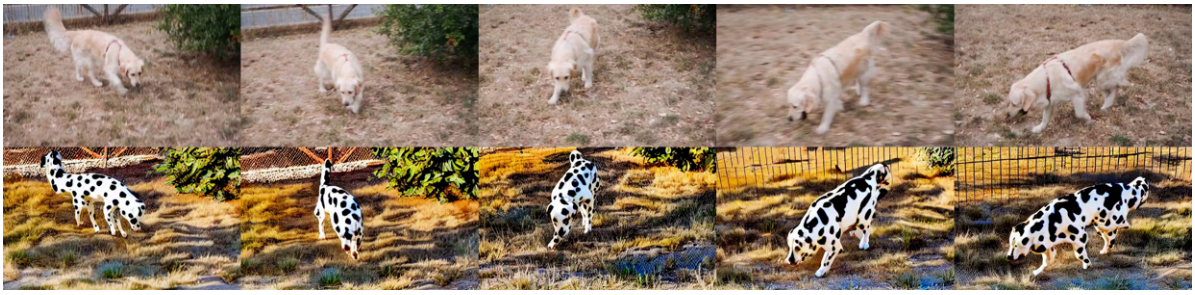
Prompt	Driving Video (top) and Result (bottom)
2D vector animation of a group of flamingos standing near some rocks and water	
cartoon animation of an elephant walks through dirt surrounded by boulders	
cyberpunk neon car on a road in the countryside	
a crochet black swan swims in a pond with rocks and vegetation	
a dalmatian dog is walking away from a fence	

Figure 3. Additional results for text-to-video-editing.

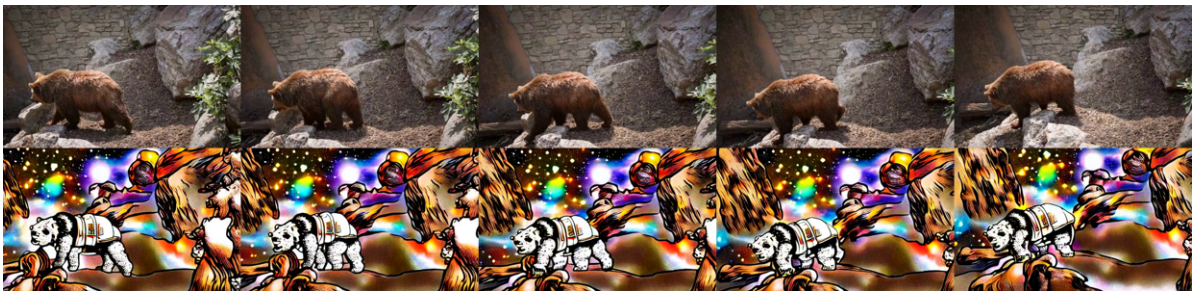
Prompt	Driving Video (top) and Result (bottom)
kite-surfer in the ocean at sunset	
car on a snow-covered road in the countryside	
small grey suv driving in front of apartment buildings at night	
a space bear walking through the stars	
white swan swimming in the water	

Figure 4. Additional results for text-to-video-editing.

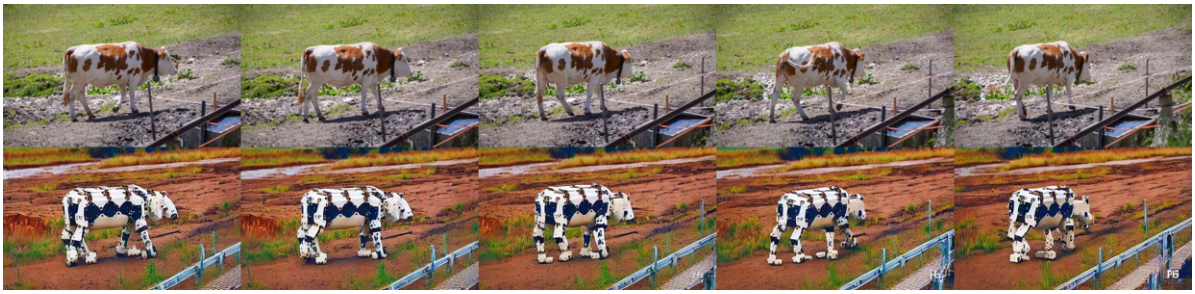
Prompt	Driving Video (top) and Result (bottom)
man riding a bicycle up the side of a dirt slope in a graphic novel style	
blue and white bus driving down a city street with a backdrop of snow-capped mountains	
toy camel standing on dirt near a fence	
8-bit pixelated car driving down the road	
a robotic cow walking along a muddy road	

Figure 5. Additional results for text-to-video-editing.

Prompt	Driving Video (top) and Result (bottom)
oil painting of four pink flamingos wading in water	
paper cut-out mountains with a hiker	
alien explorer hiking in the mountains	
man hiking in the starry mountains	
magical flying horse jumping over an obstacle	

Figure 6. Additional results for text-to-video-editing.

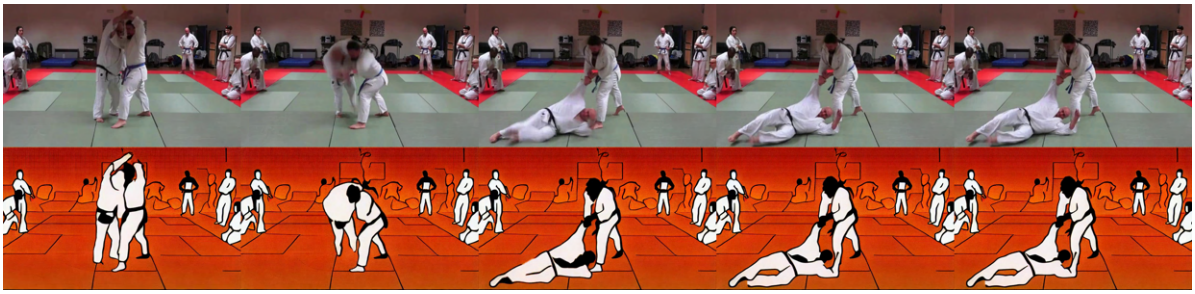
Prompt	Driving Video (top) and Result (bottom)
person rides on a horse while jumping over an obstacle with an aurora borealis in the background.	
martial artists practicing on grassy mats while others watch	
silhouetted martial artists practicing while others watch	
3D animation of a small dog running through grass	
hyper-realistic painting of a person paragliding on a mountain	

Figure 7. Additional results for text-to-video-editing.

Prompt	Driving Video (top) and Result (bottom)
paraglider soaring on a mountain under a starry sky	
cartoon-style animation of a man riding a skateboard down a road	
robot skateboarder riding down a road	
a man riding a skateboard down a magical river	
man playing tennis on the surface of the moon	

Figure 8. Additional results for text-to-video-editing.

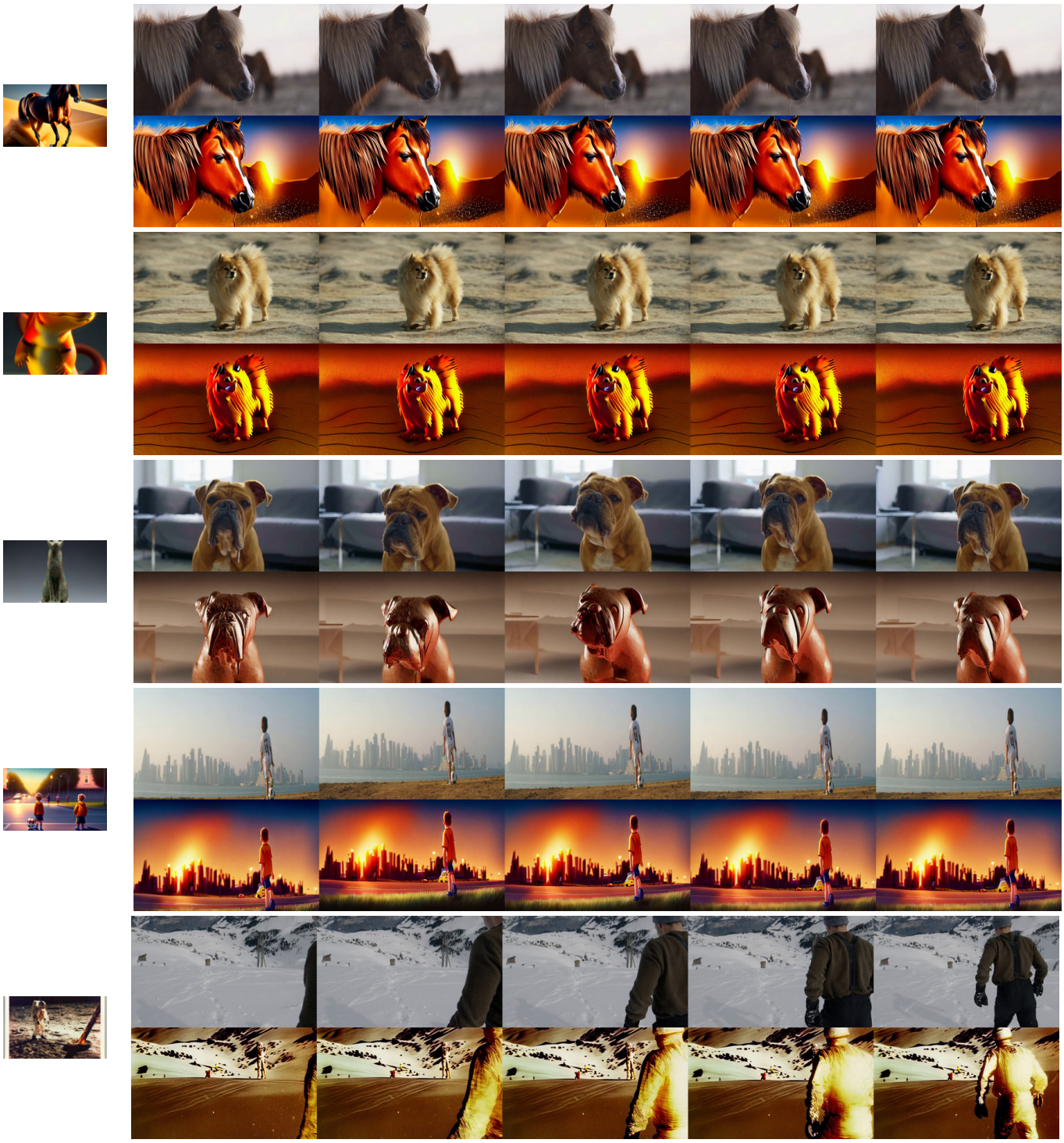


Figure 9. Additional results for image-to-video-editing.

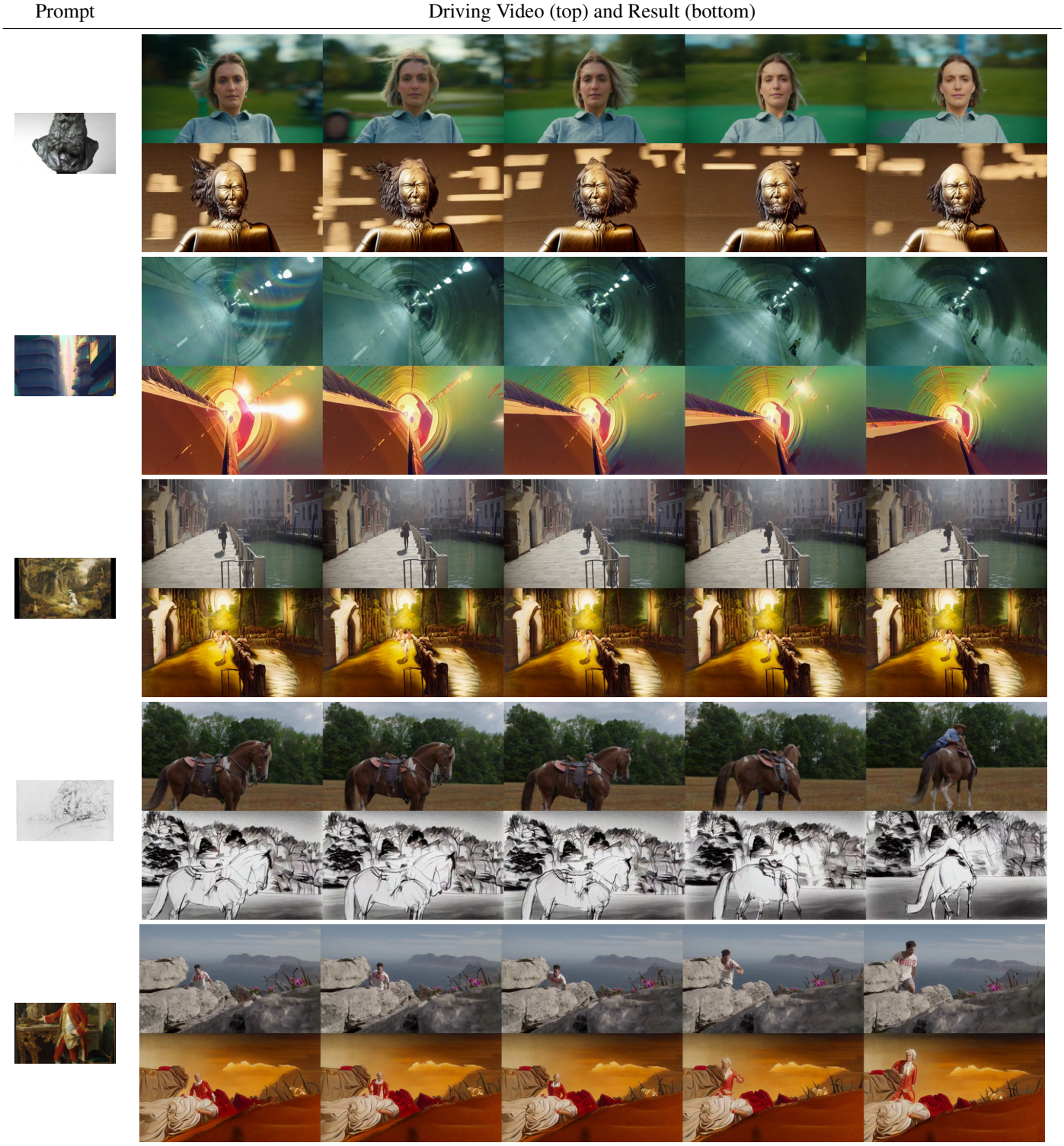


Figure 10. Additional results for image-to-video-editing.

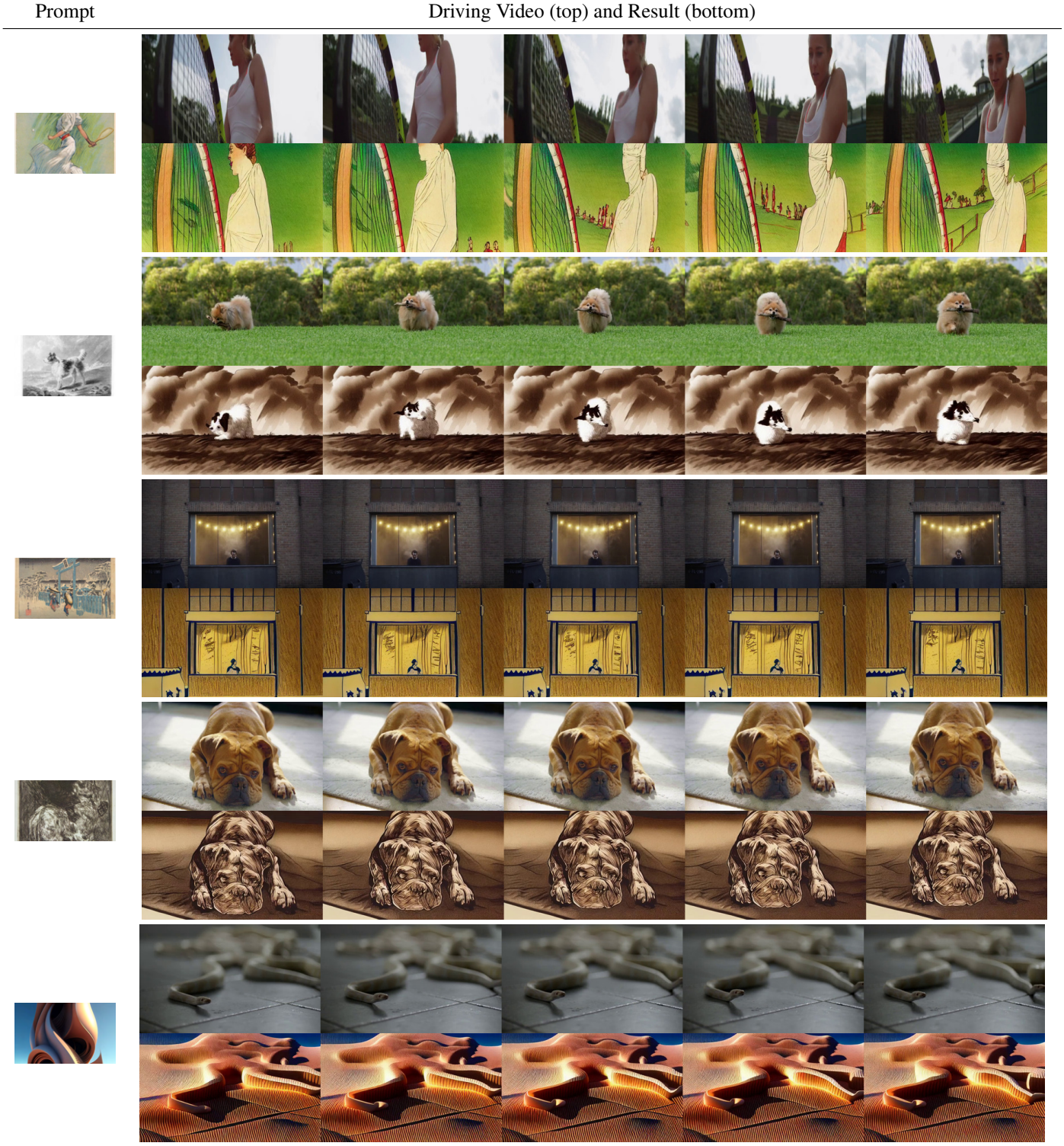


Figure 11. Additional results for image-to-video-editing.

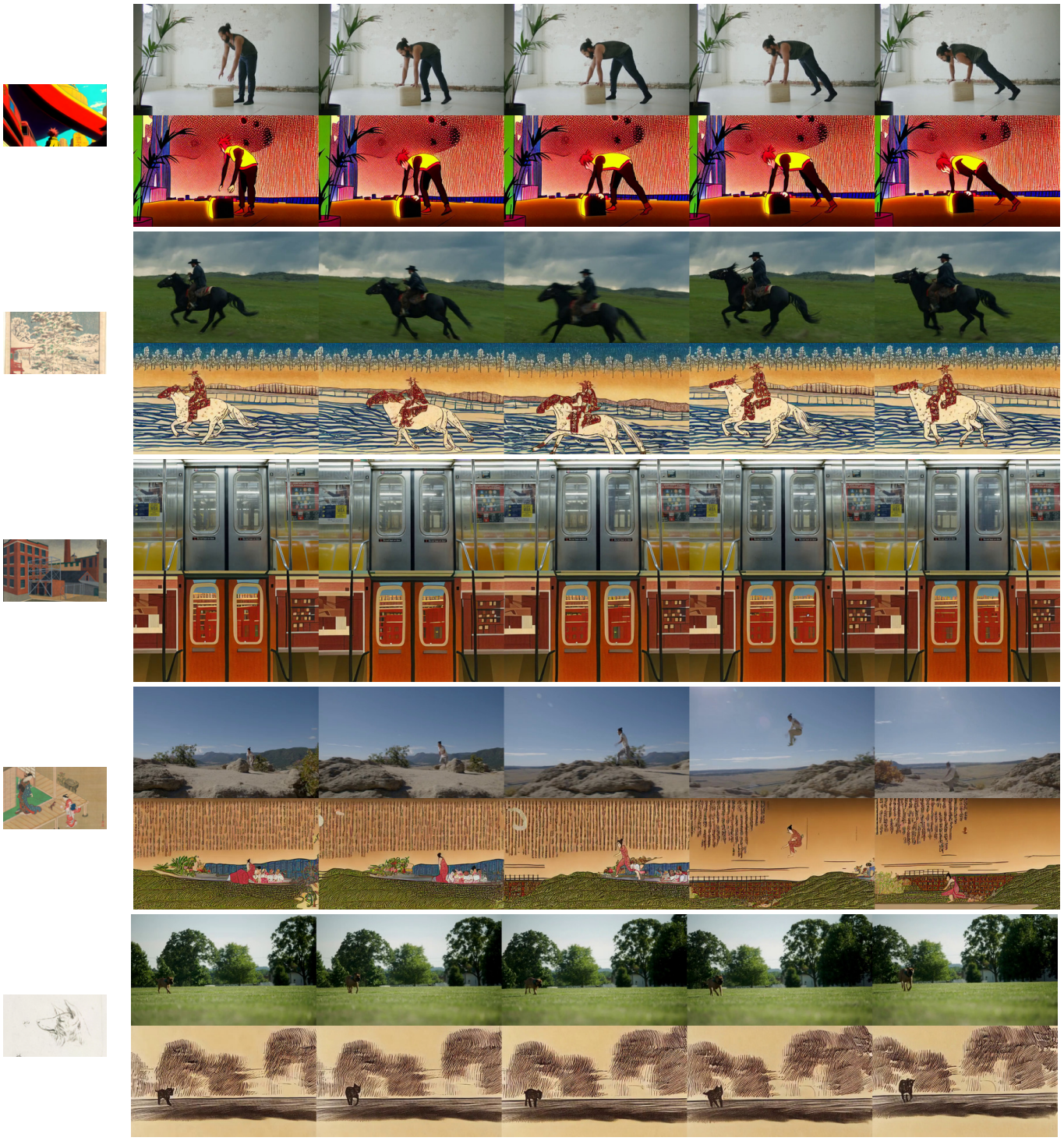


Figure 12. Additional results for image-to-video-editing.

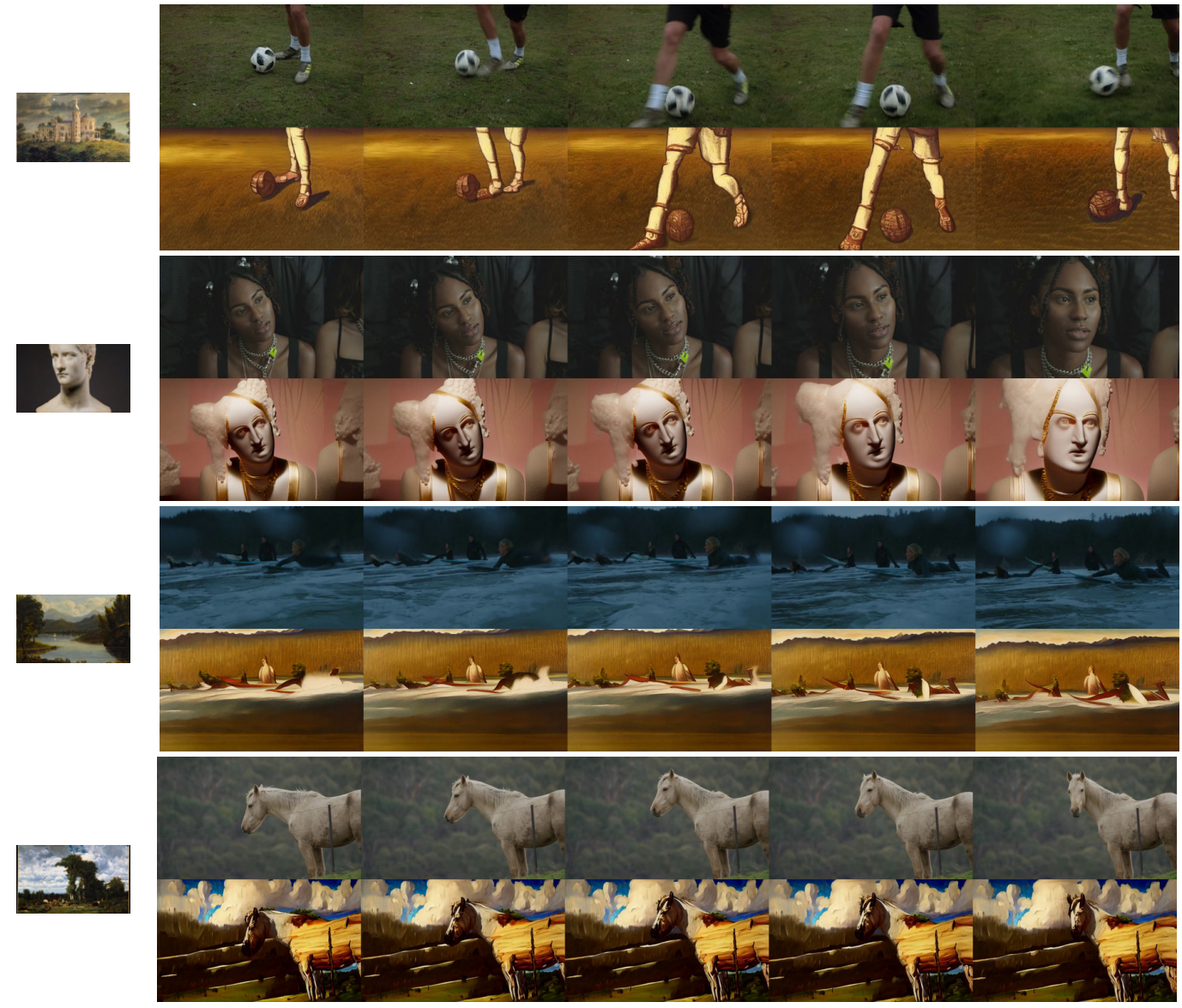


Figure 13. Additional results for image-to-video-editing.

Figure 14. Visual comparison between evaluated methods. From top to bottom: input, Deforum, ours, SDEdit, IVS, Depth-SD, Text2Live.

method	frame consistency	prompt consistency	ours preferred
Deforum	0.9087 ± 0.0079	0.2693 ± 0.0075	77.14%
SDEdit, strength=50%	0.9277 ± 0.0062	0.2454 ± 0.0073	85.29%
SDEdit, strength=75%	0.9189 ± 0.0078	0.2754 ± 0.0073	73.53%
IVS, strength=50%	0.9673 ± 0.0035	0.2401 ± 0.0076	79.41%
IVS, strength=75%	0.9668 ± 0.0030	0.2556 ± 0.0074	91.18%
Depth-SD	0.9126 ± 0.0064	0.2871 ± 0.0070	74.29%
Text2LIVE	0.9683 ± 0.0025	0.2732 ± 0.0078	88.24%
ours, $\sim s$, strength=50%	0.9541 ± 0.0039	0.2703 ± 0.0074	67.65%
ours, $\sim s$, strength=75%	0.9482 ± 0.0034	0.2769 ± 0.0062	64.71%
ours, $t_s = 0, \omega_t = 1.00, \omega = 7.50$	0.9648 ± 0.0031	0.2805 ± 0.0065	-
ours, $t_s = 0, \omega_t = 0.50, \omega = 7.50$	0.9238 ± 0.0039	0.2820 ± 0.0057	-
ours, $t_s = 0, \omega_t = 0.75, \omega = 7.50$	0.9521 ± 0.0030	0.2822 ± 0.0063	-
ours, $t_s = 0, \omega_t = 1.25, \omega = 7.50$	0.9702 ± 0.0026	0.2793 ± 0.0060	-
ours, $t_s = 0, \omega_t = 1.50, \omega = 7.50$	0.9722 ± 0.0024	0.2754 ± 0.0058	-
ours, $t_s = 4, \omega_t = 1.00, \omega = 7.50$	0.9678 ± 0.0025	0.2866 ± 0.0065	-
ours, $t_s = 6, \omega_t = 1.00, \omega = 7.50$	0.9717 ± 0.0023	0.2854 ± 0.0065	-
ours, $t_s = 7, \omega_t = 1.00, \omega = 7.50$	0.9790 ± 0.0025	0.2766 ± 0.0062	-

Table 2. Quantiative evaluations corresponding to Fig. 6 and Fig. 7. $\pm$ denotes standard error obtained with a sample size of 35.

BLIP caption	edit prompt from GPT-3	edited caption from GPT-3
there is a bear that is walking through the forest	make the bear seem like an illusion	the illusion of a bear walking through the forest
there is a black swan that is swimming in the water	morph the swan into a white swan	white swan swimming in the water
a man riding a bicycle up the side of a dirt slope	turn the man into a cartoon character	cartoon character riding a bicycle up the side of a dirt slope
a man doing a handstand on the pavement in front of a building	turn the man into a robot	robot doing a handstand on the pavement in front of a building
a blue and white bus driving down a city street	transform the bus into a colorful graffiti art piece	graffiti art piece bus driving down a city street
a camel is standing on dirt near a fence	turn the scene into a desert oasis	oasis in the desert with a camel near a fence
car with the passenger seat up on the road	replace car with a hot air balloon	hot air balloon with the passenger seat up on the road
there is a car that is driving down the road	make the car into a submarine	submarine driving down the road
the cow is walking along the muddy road	turn the cow into a robot	a robotic cow walking along a muddy road
four pink flamingos wading in water in a zoo	make the flamingos look like they are flying in the sky	four flamingos flying in the sky
several goldfish and other marine animals swimming inside an aquarium	make the aquarium bubble-filled and psychedelic	psychedelic aquarium with several goldfish and other marine animals swimming inside
a man hiking in the mountains	make the mountains out of paper cut-outs	paper cut-out mountains with a hiker
a person rides on a horse while jumping over an obstacle	person rides on a flying horse while jumping over an obstacle	magical flying horse jumping over an obstacle
men practicing martial on mats while others watch	turn the scene into a comic book	martial arts comic on mats with onlookers
the man is kiteboarding in the sea on his board	edit to make it look like he's riding a hoverboard	magical hoverboarder in the sea
a small dog that is running through some grass	make the grass appear to be made of bubble wrap	small dog running through bubble wrap grass
two ducks are standing on the grass next to a river	turn the ducks into robots	two robotic ducks standing on the grass next to a river
a motor bike rides on a roadway near a tree filled mountain	morph the bike into a spaceship	spaceship flying near a tree filled mountain.
a person in full gear is paragliding on a mountain	transform into a hyper-realistic painting	hyper-realistic painting of a person paragliding on a mountain
the young man is skateboarding through the gate	transform the gate into a giant whale	the young man skateboarding through the giant whale gate
a man riding a skateboard down a road	cartoon-style animation of the same scene	cartoon-style animation of a man riding a skateboard down a road
a person on a snowboard going down a hill	make the snowboarder go up the mountain	snowboarder ascending a mountain in a winter wonderland.
an old soccer ball is on the grass near some trees	make the soccer ball appear to be made of glass and filled with liquid	shimmering glass soccer ball filled with liquid on the grass near some trees
a person jumping a skateboard off of a ramp	make the ramp levitate in the air	person jumping a skateboard off of a levitating ramp
a man playing tennis is preparing to hit the ball	make it look like he's playing a game of pong	man playing tennis with a digital pong ball

Table 3. We generate captions for evaluation using BLIP [2] and GPT-3 [1]. An initial caption (left column) is predicted by BLIP from the first frame of the video and then GPT-3 predicts edit prompts and matching captions for images containing the suggested edit. All videos used here are from the DAVIS dataset [4].