Grid Diffusion Models for Text-to-Video Generation

Supplementary Material

Note: We provide the implementation details, figures of more generated samples, CGcaption prompts, UCF-101 prompts and human evaluation details that were not included in the main paper due to space constraints.

1. Implementation Details

1.1. Experiments setup

We fine-tune the key grid image generation and interpolation models using Stable Diffusion 1.5 [7]. To train our models, we use two NVIDIA A100 80GB GPUs. We train the key grid image generation model with a batch size of 28 for 82k steps, using AdamW as the optimizer and setting the learning rate to 0.00001. For both the 1-step and 2-step interpolation models, we train with a batch size of 20 for 54k steps, using AdamW as the optimizer and a learning rate of 0.00005. For inference, we employ the DDIM sampler [9] for key grid image generation model and Euler Ancestral Discrete Scheduler [4] for both interpolation models. We set the inference steps at 80 for key grid image generation and at 20 for both interpolation models.

1.2. Quantitative evaluation

MSR-VTT experiment. To evaluate the MSR-VTT [12] test set in a zero-shot manner, following prior work [3], we generate 2,990 videos. Our generated videos have 16 frames which are randomly selected from among 28 frames.

UCF-101 experiment. For the IS score [8], we generate 20 videos for each prompt and to calculate FVD [11], we sample 2,048 videos for evaluation in a zero-shot manner, following prior work [3]. We use the text prompts for each class, as provided by previous work [3]. The prompts are listed in Section 6.

CGcaption experiment. We generate 500 videos using generated prompts from GPT-4. As a baseline, we utilize VideoFusion (Modelscope) [5], for which the source code is publicly available. The prompts are listed in Section 5.

1.3. Human evaluation

We conduct human evaluation on Amazon Mechanical Turk (AMT) to evaluate our videos in the following criteria: text matching, video quality, temporal consistency, and motion quality. We compare our model to VideoFusion [5] and VideoCrafter [2] which is publicly available. For human evaluation, we randomly sample 100 generated videos from each of MSR-VTT [12], UCF-101 [10], and CGcaption datasets, in total 300 samples. We conduct the surveys with 30 participants. For text matching, we provide videos and text pairs and ask participants to respond to the question, “Choose a video that matches the text better.”. For video quality, we ask participants to respond to the question, “Choose a video with better video quality.”. For temporal consistency, we ask participants to respond to the question, “Choose a video with better temporal consistency.”. For motion quality, we ask participants to respond to the question, “Choose a video with better motion quality.”. The screenshot of the user study including the instructions is shown in Figure 1.

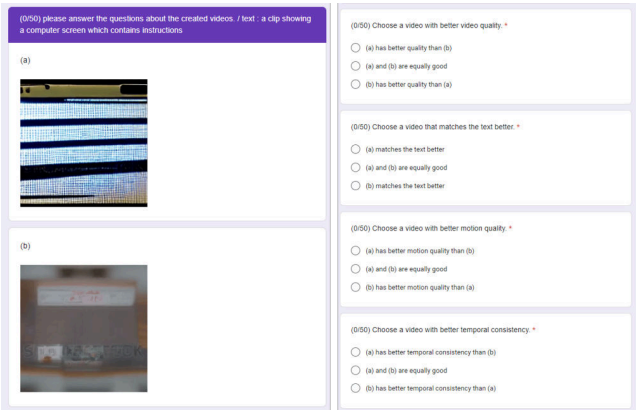


Figure 1. **Screenshot of instructions provided to participants during the human evaluation.**

2. Samples : Text to Video Generation

We provide the following generated samples in the figures. As shown in Figures 2 and 3, our model demonstrates better text alignment than VideoFusion [5] and exhibits more dynamic motion and better video quality. In the case of the ablation study, while there is a slight difference when viewed as Figures 4 and 5, our model appears smoother than the ablation models and more effectively maintains consistency with the previous frames.

- Comparison on MSR-VTT [12] and CGcaption with our model and VideoFusion : Figure 2, Figure 3
- Comparison on MSR-VTT [12] with ablation models : Figure 4, Figure 5

3. Samples : Text to Video Generation with More Frames

We provide the following generated samples in the figures. To evaluate text-to-video generation with more frames, we qualitatively compare our model with VideoFusion [5] and FreeNoise [6] on MSR-VTT [12] for videos with 64 frames. To illustrate in a figure, we extract 8 frames from the entire video at one-second intervals. As shown in Figures 6 and 7, our model generates videos with richer motion and more dynamic camera movements compared to VideoFusion [5] and FreeNoise [6], which produces more static motion. Also, as shown in Figures 8 and 9, we qualitatively compare our model with VideoFusion [5] and FreeNoise [6] for videos with 128 frames.

- Comparison on MSR-VTT [12] with our model, VideoFusion [5] and FreeNoise for 64 frames : Figure 6, Figure 7
- Comparison with our model, VideoFusion [5] and FreeNoise [6] for 128 frames : Figure 8, Figure 9

4. Text-Guided Video Manipulation

As mentioned in our main paper, we can easily manipulate videos with text by using our approach. Figure 12 depicts the process. The input image is a grid image created by selecting four frames from the Webvid-10M video. The video caption corresponds to the video description of the Webvid-10M. The prompt is the set of conditions for manipulating the grid image. The inter prompt is the prompt condition for the interpolation model. We also provide more examples of video manipulation in Figures 10 and 11. As shown in Figure 10 and 11, we can edit the content by adding glasses or changing the shape of a hat, and also transform the style of the video.

5. CGcaption prompts

We generate a total of 500 prompts using GPT-4 and conduct zero-shot evaluation. The samples for CGcaption are as follows. Please refer to **cgcaption.txt** for all prompts.

A majestic dragon flying over a medieval castle.
A group of children playing soccer in the rain.
A curious squirrel exploring a bustling city.
A grandmother knitting a scarf under a cherry blossom tree.
A group of aliens playing jazz music on Mars.
A group of friends hosting a backyard film festival.
A woman crafting homemade candles.
A group of people participating in a community bike ride.
A snail racing against a turtle in a garden obstacle course.

6. UCF-101 prompts

We simply use all the prompts available in PYoCo [3].

applying eye makeup, applying lipstick, archery, baby crawling, gymnast performing on a balance beam, band marching, baseball pitcher throwing baseball, a basketball player shooting basketball, dunking basketball in a basketball match, bench press, biking, billiards, blow dry hair, blowing candles, body weight squats, a person bowling on bowling alley, boxing punching bag, boxing speed bag, swimmer doing breast stroke, brushing teeth, clean and jerk, cliff diving, bowling in cricket gameplay, batting in cricket gameplay, cutting in kitchen, diver diving into a swimming pool from a springboard, drumming, two fencers have fencing match indoors, field hockey match, gymnast performing on the floor, group of people playing frisbee on the playground, swimmer doing front crawl, golfer swings and strikes the ball, haircuting, a person hammering a nail, an athlete performing the hammer throw, an athlete doing handstand push up, an athlete doing handstand walking, massagist doing head massage to man, an athlete doing high jump, group of people racing horse, person riding a horse, a woman doing hula hoop, ice dancing man and woman dancing on the ice, athlete practicing javelin throw, a person juggling with balls, a young person doing jumping jacks, a person skipping with jump rope, a person kayaking in rapid water, knitting, an athlete doing long jump, a person doing lunges with barbell, military parade, mixing in the kitchen, mopping floor, a person practicing nunchuck, gymnast performing on parallel bars, a person tossing pizza dough, a musician playing the cello in a room, a musician playing the daf, a musician playing the indian dhol, a musician playing the flute, a musician playing the guitar, a musician playing the piano, a musician playing the sitar, a musician playing the tabla, a musician playing the violin, an athlete jumps over the bar, gymnast performing pommel horse exercise, a person doing pull ups on bar, boxing match, push ups, group of people rafting on fast moving river, rock climbing indoor, rope climbing, several people rowing a boat on the river, couple salsa dancing, young man shaving beard with razor, an athlete practicing shot put throw, a teenager skateboarding, skier skiing down, jet ski on the water, sky diving, soccer player juggling football, soccer player doing penalty kick in a soccer match, gymnast performing on still rings, sumo wrestling, surfing, kids swing at the park, a person playing table tennis, a person doing TaiChi, a person playing tennis, an athlete practicing discus throw, trampoline jumping, typing on computer keyboard, a gymnast performing on the uneven bars, people playing volleyball, walking with dog, a person standing doing pushups on the wall, a person writing on the blackboard, a kid playing Yo-Yo.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 13

[2] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 1

[3] 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*, pages 22930–22941, 2023. 1, 2

[4] 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. 1

[5] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 1, 2, 4, 5, 10, 11

[6] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 2, 8, 9, 10, 11

[7] 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

[8] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[11] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1

[12] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 2, 4, 6, 7, 8, 9

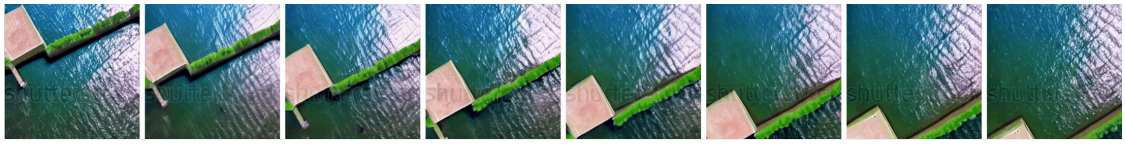
Prompt : *A boy plays a guitar*

(a) VideoFusion

(b) Ours

Prompt : *A camera flies over waterscape*

(a) VideoFusion



(b) Ours

Prompt : *2 kids are fun talking with food on the snow covered mountains*

(a) VideoFusion

(b) Ours

Prompt : *A girl is talking on her phone*

(a) VideoFusion

(b) Ours

Figure 2. **Text-to-video generation comparison with VideoFusion [5] on MSR-VTT [12].**

Prompt : *A street musician playing a violin in a city center*

(a) VideoFusion

(b) Ours

Prompt : *A snowman sunbathing on a beach*

(a) VideoFusion

(b) Ours

Prompt : *A businessman reading a newspaper on the morning commute*

(a) VideoFusion

(b) Ours

Prompt : *A grandmother knitting a scarf under a cherry blossom tree*

(a) VideoFusion

(b) Ours

Figure 3. **Text-to-video generation comparison with VideoFusion [5] on CGcaption**.

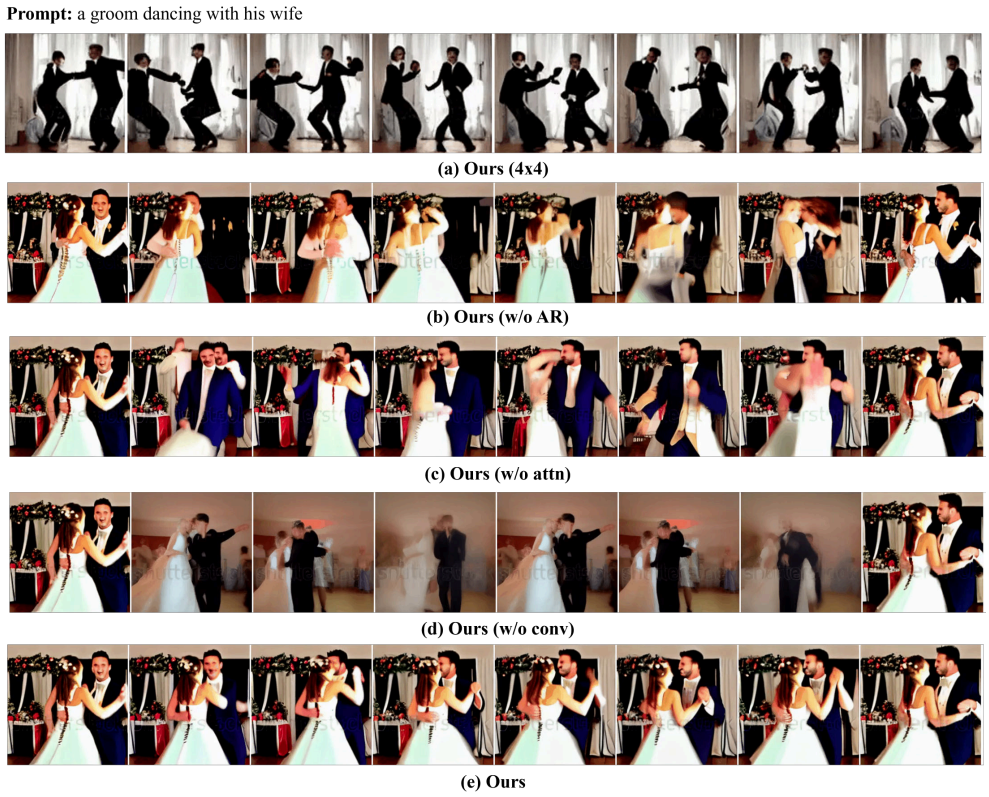


Figure 4. **Text-to-Video generation comparison for ablation study on MSR-VTT [12]**. For the 4×4 model, we generate videos with a resolution of 126×126, while the other models, we generate videos with a resolution of 254×254.

Figure 5. **Text-to-Video generation comparison for ablation study on MSR-VTT [12]**. For the 4×4 model, we generate videos with a resolution of 126×126, while the other models, we generate videos with a resolution of 254×254.

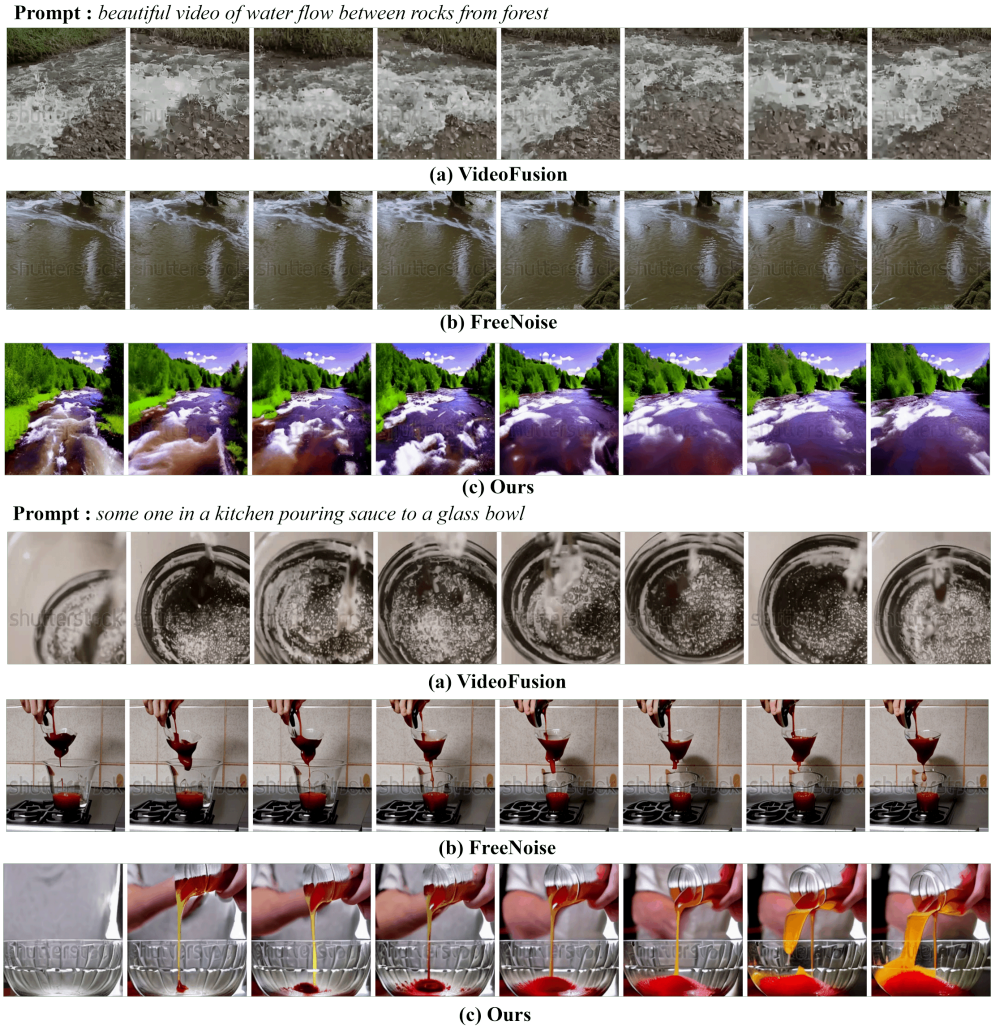


Figure 6. **Text-to-video generation with more frames comparison with FreeNoise [6] on MSR-VTT [12]**. Each video has 64 frames at 8 fps.

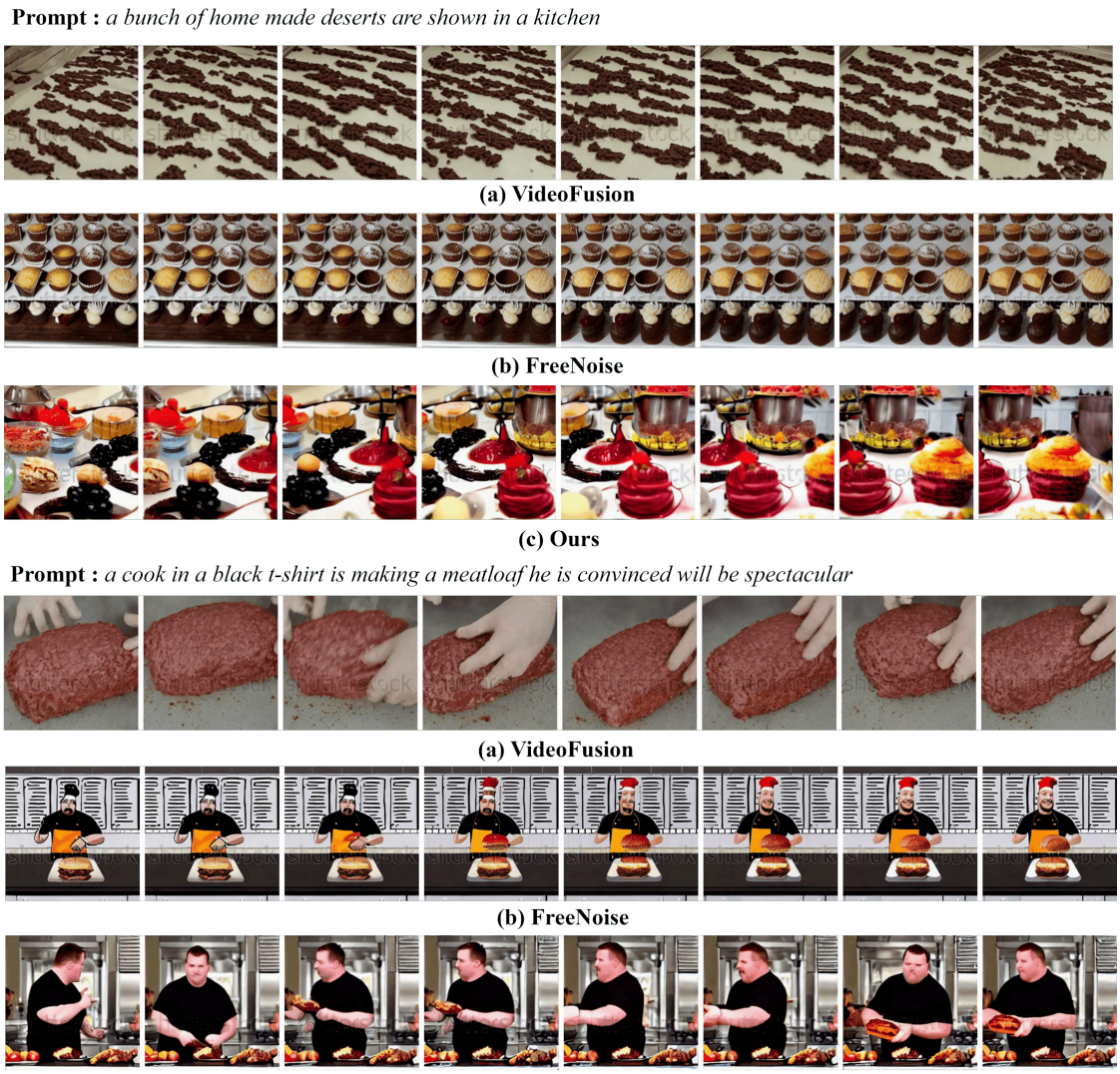


Figure 7. **Text-to-video generation with more frames comparison with FreeNoise [6] on MSR-VTT [12]**. Each video has 64 frames at 8 fps.

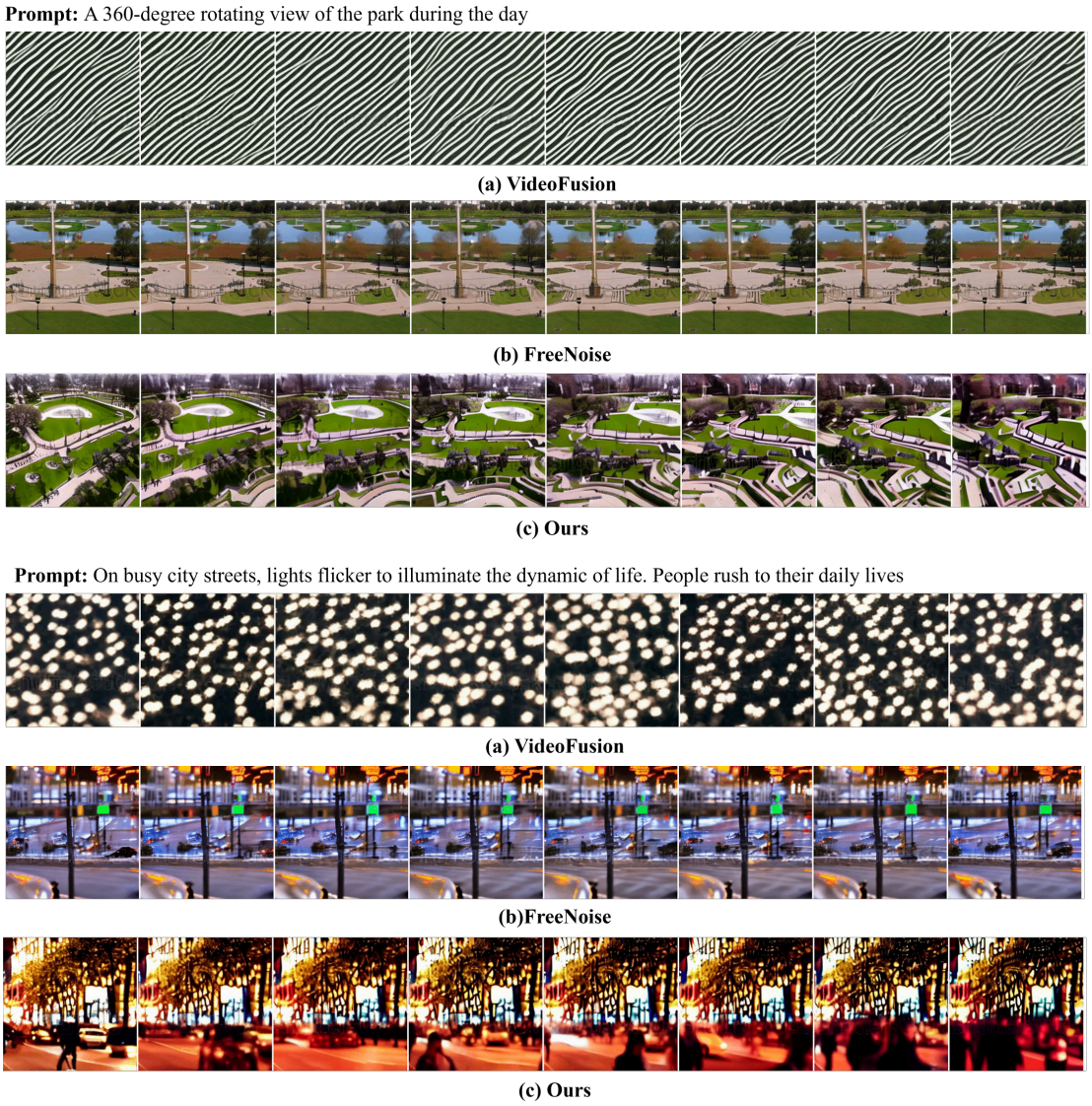


Figure 8. **Text-to-video generation with more frames comparison with VideoFusion [5] and FreeNoise [6]**. Each video has 128 frames at 8 fps.

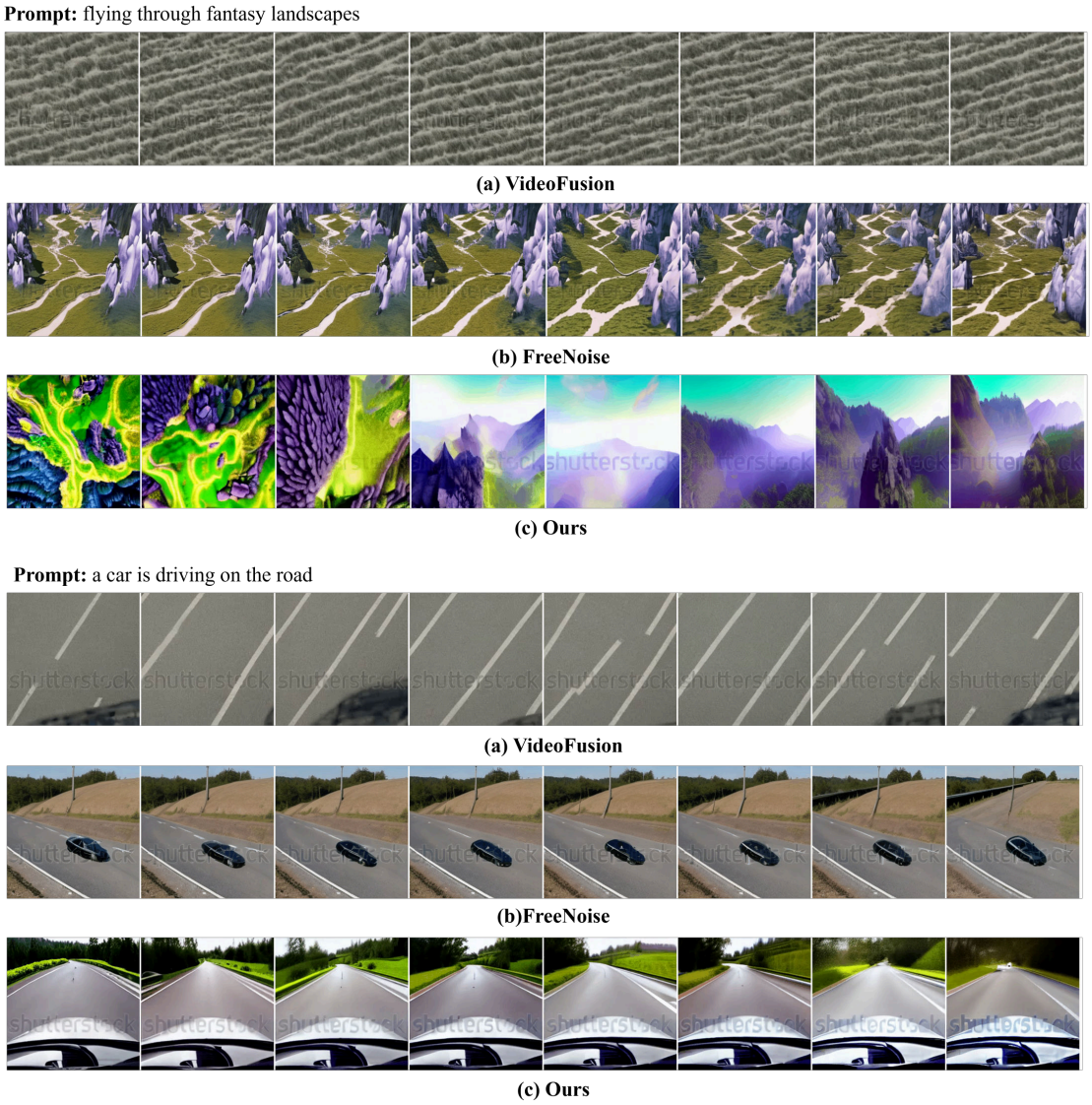


Figure 9. **Text-to-video generation with more frames comparison with VideoFusion [5] and FreeNoise [6]**. Each video has 128 frames at 8 fps.

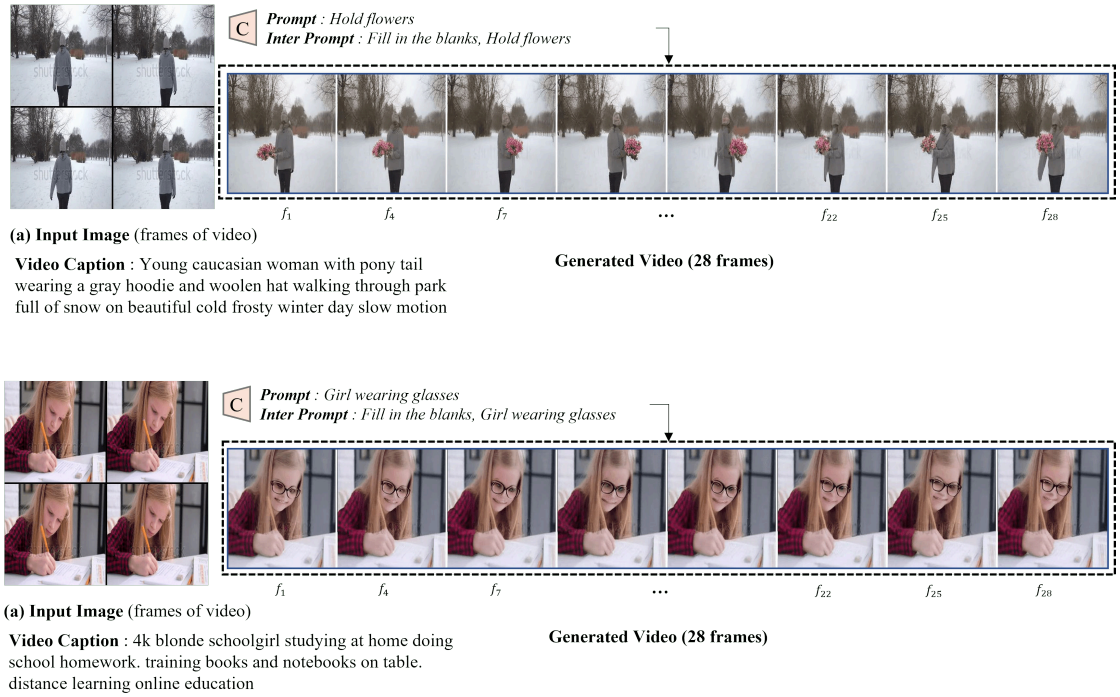


Figure 10. **The result of video manipulation.** The input image is a grid image created by selecting four frames from the Webvid-10M video.

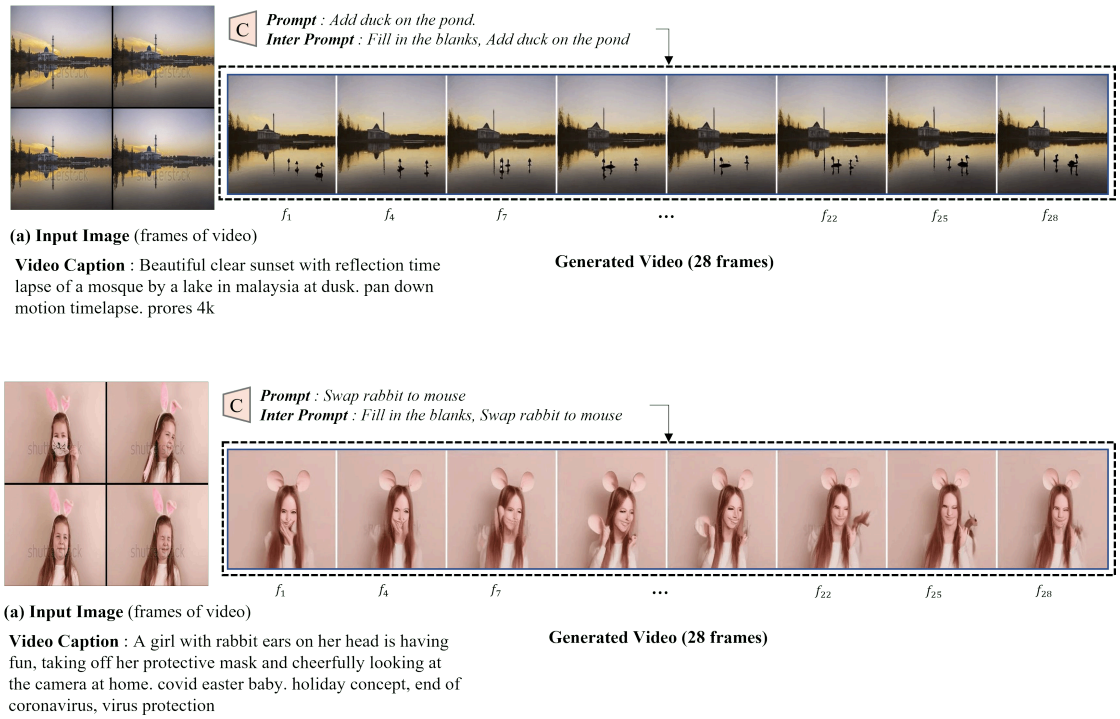


Figure 11. **The result of video manipulation.** The input image is a grid image created by selecting four frames from the Webvid-10M video.

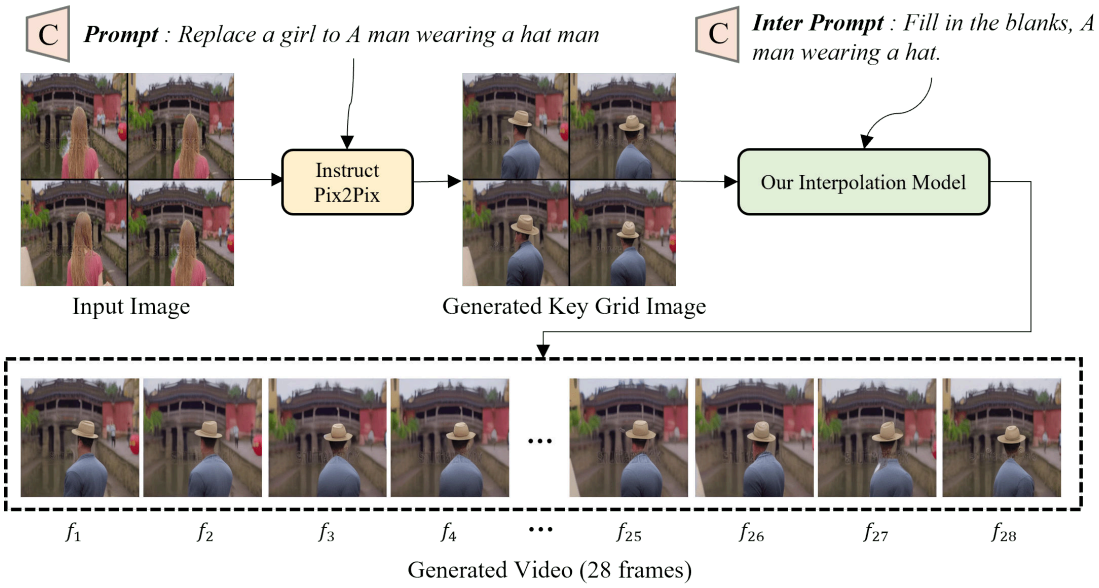


Figure 12. **Process of text-guided-video manipulation with our approach**. The input image is a grid image created by selecting four frames from original video. The Prompt is the conditions set for manipulating the grid image. The Inter prompt is the prompt conditions for the interpolation model. InstructPix2Pix [1] is pre-trained model for image manipulation.