Supplementary Material for SimDA: Simple Diffusion Adapter for Efficient Video Generation

This supplementary Appendix contains the following.

- Section A: The Hyperparameter settings of our SimDA.
- Section B: The introduction of the datasets and evaluation metrics.
- Section C: Additional experiments details for our methods.
- Section D: More Text-to-Video and Text-guided video editing visualization results of our SimDA.

A. Hyperparameter Settings

For video data, we evenly sample 16 frames from a two-second clip. Subsequently, we perform image resizing and center cropping to obtain dimensions of 256 × 256. The latent space is configured as 32 × 32 × 4. Optimization is carried out using Adam [5], with a learning rate set to $1e-4$, and the number of diffusion steps (T) set to 1000. During the inference phase, the number of sampling steps (T) is set to 100, and the guidance scale (s) is set to 15. Table 1 details the hyperparameter settings for our models.

Table 1. The hyper-parameter setting of our models. AE denotes the auto-encoder to encode and decode videos.

Hyper-parameter (common)	Value	Hyper-parameter (common)	Value
Image Size	256	Num Frame	16
Guidance Scale	15	Text Seq Length	77
Text Encoder	CLIP ViT-L/14	First Stage Model	AutoencoderKL
AE Double z	True	AE z channel	4
AE Resolution	256	AE In Channel	3
AE Out Channel	3	AE Channel	128
AE Channel Multiplier	$[1, 2, 4, 4]$	AE Num ResBlock	2
AE Atten Resolution	[]	AE Dropout	0.0
Store EMA	True	EMA FP32	True
EMA Decay	0.9999	Diffusion In Channel	4
Diffusion Out Channel	4	Diffusion Channel	320
Conditioning Key	crossattn	Noise Schedule	quad
Encoder Channel	1280	Atten Resolution	$[4, 2, 1]$
Num ResBlock	2	Channel Multiplier	$[1, 2, 4, 4]$
Transformer Depth	1	Batch Size	4
Learn Sigma	False	Diffusion Step	1000
Timestep Respacing	100	Sampling FP16	False
Learning Rate	$1e^{-4}$	Sample Scheduler	DDPM
Num Head	8		

B. Datasets and Metric Details

B.1. Datasets

When adapting Stable Diffusion [7] into a text-to-video generator, we leverage the WebVid-10M dataset [1]. WebVid-10M is an extensive dataset comprising short videos accompanied by textual descriptions sourced from stock footage sites. The videos exhibit diversity and richness in their content, with a total of 10.7 million video-caption pairs and a cumulative duration of 52,000 video hours.

B.2. Quantitative Evaluation

We conduct quantitative assessments across all datasets, employing the Fre'chet Video Distance (FVD) metrics [8]. Acknowledging the potential unreliability of FVD, as discussed in [3], we complement our evaluation with human evaluation. In the context of text-to-video evaluation, we also calculate CLIP Similarly scores (CLIPSIM).

FVD The FVD metric measures the similarity between real and generated videos [8]. Following the methodology outlined in [10], we generate 4,476 videos of the validation set of WebVid [1], each comprising 16 frames. Subsequently, we extract features using a pre-trained I3D [**?**] action classification model. To establish reference statistics, we extract features from random sequences of videos containing at least 16 frames from the dataset.

CLIPSIM In our text-to-video experiments on MSR-VTT, we also evaluate CLIP similarity (CLIPSIM) [6]. The MSR-VTT test set contains 2990 examples and 20 descriptions/prompts per example. We generate 2990 videos (16 frames) by using one random prompt per example. We then average the CLIPSIM score of the 47,840 frames. We use the ViT-B/32 [6] model to compute the CLIP score following VideoLDM [2].

Human evaluation We conduct a human evaluation (user study) to assess the realism of videos generated by our method compared to LVDM [4] and ModelScope [9]. In our user study, we create 150 videos, each comprising 16 frames. The study presents pairs of videos, with each pair containing one random video generated by our method and one from either LVDM or ModelScope. Participants are instructed to choose the more realistic video in a non-forced-choice response, allowing for the option to vote for "equally realistic." It's worth noting that the A-B order of the video pairs is randomly assigned. Each video pair is presented to twenty participants, resulting in 3,000 responses per dataset.

C. Additional Experimental Details

Details: Text-to-Video with Stable Diffusion We ran experiments with the publicly available Stable Diffusion v1.5 checkpoints as image LDM backbones. Most of the research project was conducted with the SD 1.5-based model.

Given that Stable Diffusion (SD) is trained on images at a resolution of 512×512, directly applying it to the smaller-sized videos from the WebVid-10M dataset would result in significant degradation in image quality. To address this, we initially conduct fine-tuning on the spatial layers of the Stable Diffusion image backbone using WebVid-10M data. Specifically, we resize and center-crop the WebVid-10M videos to a resolution of 256×256 and subsequently fine-tune the SD latent space diffusion model on independently encoded frames extracted from the videos. Standard SD training hyperparameters are employed, with a learning rate set to $1e-4$.

Upsampler Training We also conducted video fine-tuning on the publicly available text-guided Stable Diffusion 4× upscaler, which itself is a latent diffusion model. We trained the upsampler for temporal alignment in a patch-wise manner on 320×320 cropped videos (WebVid-10M [1]), embedded into an 80×80 latent space. The 80×80 low-resolution conditioning videos are concatenated with the 80×80 latents. The learned temporal alignment layers are text-conditioned, similar to our base text-to-video Latent Diffusion Models (LDMs). During training, we randomly sampled $t \in \{0, ..., 250\}$ and perturbed the low-resolution conditioning following our variance-preserving diffusion process, utilizing the same linear noise schedule as the main upsampling diffusion model. At inference time, we applied the model at an extended resolution, providing 256×256 resolution videos as low-resolution input, predicting 256×256 resolution latents, and decoding to 1024×1024 resolution videos.

Text-guided Video Editing Our approach can be extended to text-guided video editing. To ensure a fair comparison with Tune-A-Video [11], we replicated their methodology, also utilizing Stable Diffusion v1.5 as the backbone. During the training phase, we sampled 32 uniform frames at a resolution of 512×512 and fine-tuned the model for 200 steps with a learning rate of $5e-5$ and a batch size of 1. In the inference phase, we employed the DDIM sampler with classifier-free guidance in our experiments. For a single video, the fine-tuning process takes approximately 2.5 minutes, and sampling takes about 1 minute on an NVIDIA A100 GPU.

D. More Visualizations

In this section, we present more visualization of SimDA, the text-to-video results are shown in Fig. 1 and Fig. 2. In addition, we present the comparison results of SimDA and Tune-A-Video [11] in Fig. 3, it is evident that our approach maintains better temporal consistency. For fully rendered videos, we primarily refer the reader to our anonymous project page (https://simda-v1.github.io/).

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 2

[2] 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 *CVPR*, 2023. 2

[3] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *NeurIPS*, 2022. 2

[4] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[6] 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 *ICML*, 2021. 2

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[8] 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. 2

[9] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[10] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 2

[11] 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. In *ICCV*, 2023. 3, 6

Monkey learning to play the piano.

Nature Through a front Car Window, van Gogh's style, 4k, high resolution.

Beer pouring into glass, low angle video shot.

Time lapse at the snow land with aurora in the sky, 4k, high resolution

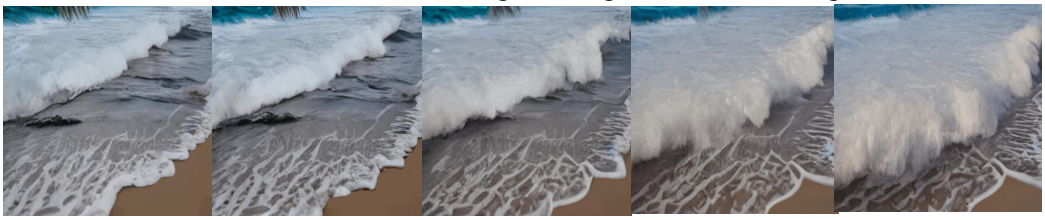
Coffee pouring into a cup, 4k, high resolution.

A cat wearing sunglasses and working as a lifeguard at a pool.

Figure 1. Results of extending our SimDA to text-to-video generation.

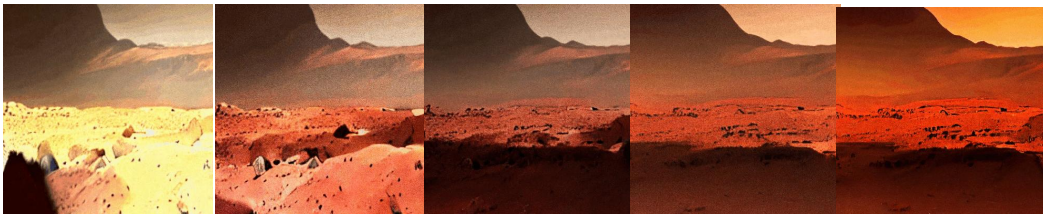
A panda is taking a selfie, 4k, high resolution.

Time-lapse video of the passing and paths of the stars in the night sky over a house for an entire night using a wide wide angle.



Sea waves with foam on white tropical sandy beach.

Close up of Craftsman worker sawing a steel pipe, Technician concept, 4k.



A beautiful sunrise on mars, Curiosity rover, 4k, high resolution.

A beautiful sunrise on mars, Curiosity rover, 4k, high resolution.

Figure 2. Results of extending our SimDA to text-to-video generation.

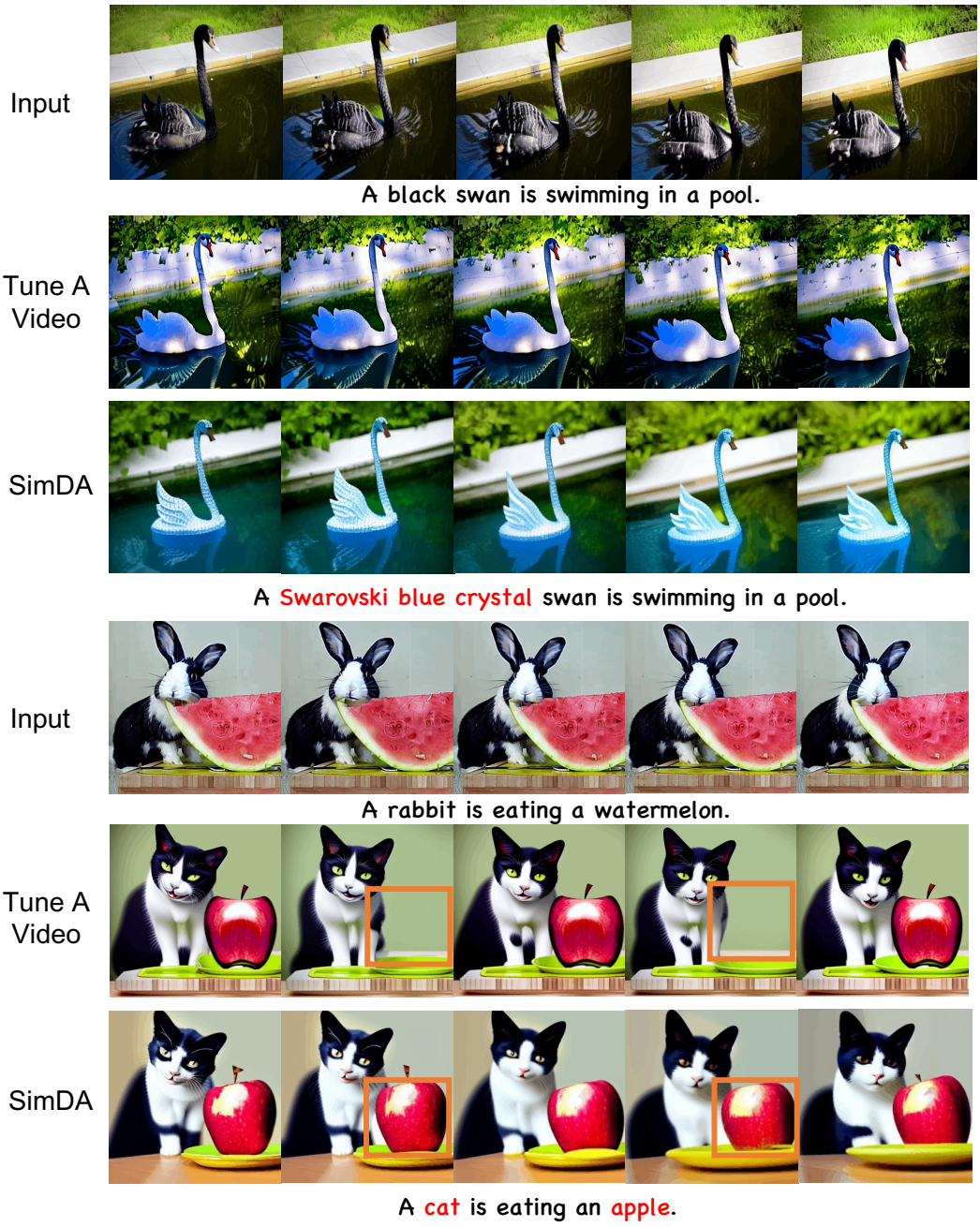


Figure 3. Results of comparisons of SimDA and Tune-A-Video [11] on text guided video editing task.