DreamVideo: Composing Your Dream Videos with Customized Subject and Motion —Supplementary Material—

Yujie Wei[1] Shiwei Zhang[2] Zhiwu Qing[2] Hangjie Yuan[2] Zhiheng Liu[2]
Yu Liu[2] Yingya Zhang[2] Jingren Zhou[2] Hongming Shan[1,3,4*]
[1] Institute of Science and Technology for Brain-inspired Intelligence, Fudan University
[2] Alibaba Group [3] MOE Frontiers Center for Brain Science, Fudan University
[4] Key Laboratory of Computational Neuroscience and Brain-inspired Intelligence, Fudan University
yjwei22@m.fudan.edu.cn, hmshan@fudan.edu.cn, {zhangjin.zsw, qingzhiwu.qzw}@alibaba-inc.com
{yuanhangjie.yhj, pingzhi.lzh, ly103369, yingya.zyy, jingren.zhou}@alibaba-inc.com

1. Experimental Details

In this section, we supplement the experimental details of each baseline method and our method. To improve the quality and remove the watermarks of generated videos, we further fine-tune ModelScopeT2V [17] for 30k iterations on a randomly selected subset from our internal data, which contains about 30,000 text-video pairs. For a fair comparison, we use the fine-tuned ModelScopeT2V model as the base video diffusion model for all methods except for AnimateDiff [5] and Tune-A-Video [18], both of which use the image diffusion model (Stable Diffusion [14]) in their official papers. Here, we use Stable Diffusion v1-5[1] as their base image diffusion model. During training, unless otherwise specified, we default to using AdamW [11] optimizer with the default betas set to 0.9 and 0.999. The epsilon is set to the default 1.0×10^{-8}, and the weight decay is set to 0. During inference, we use 50 steps of DDIM [16] sampler and classifier-free guidance [6] with a scale of 9.0 for all baselines. We generate 32-frame videos with 256 × 256 spatial resolution and 8 fps. All experiments are conducted using one NVIDIA A100 GPU. In the following, we introduce the implementation details of baselines from subject customization, motion customization, and arbitrary combinations of subjects and motions (referred to as video customization).

1.1. Subject Customization

For all methods, we set batch size as 4 to learn a subject.

DreamVideo (ours). In subject learning, we take ∼3000 iterations for optimizing the textual identity following [4, 10] with learning rate 1.0×10^{-4}, and ∼800 iterations for learning identity adapter with learning rate 1.0×10^{-5}. We set the hidden dimension of the identity adapter to be half the input dimension. Our method takes ∼12 minutes to train the identity adapter on one A100 GPU.

Textual Inversion [4]. According to their official code[2], we reproduce Textual Inversion to the video diffusion model. We optimize the text embedding of pseudo-word "S^*" with prompt "a S^*" for 3000 iterations, and set the learning rate to 1.0×10^{-4}. We also initialize the learnable token with the corresponding class token. These settings are the same as the first step of our subject-learning strategy.

Dreamix [12]. Since Dreamix is not open source, we reproduce its method based on the code[3] of ModelScopeT2V. According to the descriptions in the official paper, we only train the spatial parameters of the UNet while freezing the temporal parameters. Moreover, we refer to the third-party implementation[4] of DreamBooth [15] to bind a unique identifier with the specific subject. The text prompt used for target images is "a [V] [category]", where we initialize [V] with "sks", and [category] is a coarse class descriptor of the subject. The learning rate is set to 1.0×10^{-5}, and the training iterations are 100 ∼ 200.

Custom Diffusion [9]. We refer to the official code[5] of Custom Diffusion and reproduce it on the video diffusion model. We train Custom Diffusion with the learning rate of 4.0×10^{-5} and 250 iterations, as suggested in their paper. We also detach the start token embedding ahead of the class word with the text prompt "a S^* [category]". We simultaneously optimize the parameters of the key as well as value matrices in cross-attention layers and text embedding

[*] Corresponding author.

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

[2] https://github.com/rinongal/textual_inversion

[3] https://modelscope.cn/models/damo/text-to-video-synthesis

[4] https://github.com/XavierXiao/Dreambooth-Stable-Diffusion

[5] https://github.com/adobe-research/custom-diffusion

	A.D.	*M.S.*	LoRA	Ours
Dynamic Degree [%]	34.3	50.0	60.2	**68.9**

Table S1. **Dynamic Degree on video customization.** *A.D.* and *M.S.* are AnimateDiff and ModelScopeT2V.

of S^*. We initialize the token S^* with the token-id 42170 according to the paper.

1.2. Motion Customization

To model a motion, we set batch size to 2 for training from multiple videos while batch size to 1 for training from a single video.

DreamVideo (ours). In motion learning, we train motion adapter for ~1000 iterations with learning rate 1.0×10^{-5}. Similar to the identity adapter, the hidden dimension of the motion adapter is set to be half the input dimension. On one A100 GPU, our method takes ~15 and ~30 minutes to learn a motion pattern from a single video and multiple videos, respectively.

ModelScopeT2V [17]. We only fine-tune the temporal parameters of the UNet while freezing the spatial parameters. We set the learning rate to 1.0×10^{-5}, and also train 1000 iterations to learn a motion.

Tune-A-Video [18]. We use the official implementation[6] of Tune-A-Video for experiments. The learning rate is 3.0×10^{-5}, and training iterations are 500. Here, we adapt Tune-A-Video to train on both multiple videos and a single video.

Text2LIVE [1]. We use the official code[7] of Text2LIVE for experiments. We follow their default settings to set the learning rate to 0.0025 and the number of training epochs to 3000. Since Text2LIVE cannot edit the foreground and background simultaneously, we train it twice in some cases.

1.3. Video Customization

DreamVideo (ours). We combine the trained identity adapter and motion adapter for video customization during inference. No additional training is required. We also randomly select an image provided during training as the appearance guidance. We find that choosing different images has a marginal impact on generated videos.

AnimateDiff [5]. We use the official implementation[8] of AnimateDiff for experiments. AnimateDiff trains the motion module from scratch, but we find that this training strategy may cause the generated videos to be unstable and temporal inconsistent. For a fair comparison, we further fine-tune the pre-trained weights of the motion module provided by AnimateDiff and carefully adjust the hyperparameters. The learning rate is set to 1.0×10^{-5}, and training iterations are 50. For the personalized image diffusion model, we use the third-party implementation[4] code to train a DreamBooth model. During inference, we combine the DreamBooth model and motion module to generate videos.

ModelScopeT2V [17]. We train spatial/temporal parameters of the UNet while freezing other parameters to learn a subject/motion. Settings of training subject and motion are the same as Dreamix in Sec. 1.1 and ModelScopeT2V in Sec. 1.2, respectively. During inference, we combine spatial and temporal parameters into a UNet to generate videos.

LoRA [7]. In addition to fully fine-tuning, we also attempt the combinations of LoRAs. According to the conclusions in Sec. 3.3 of our main paper and the method of Custom Diffusion [9], we only add LoRA to the key and value matrices in cross-attention layers to learn a subject. For motion learning, we add LoRA to the key and value matrices in all attention layers. The LoRA rank is set to 32. Other settings are consistent with DreamVideo. During inference, we merge spatial and temporal LoRAs into corresponding layers.

2. More Results

In this section, we conduct further experiments and showcase more results to illustrate the superiority of our DreamVideo.

2.1. Video Customization

We provide more results compared with the baselines, as shown in Fig. S1. The videos generated by AnimateDiff suffer from little motion, while other methods still struggle with the fusion conflict problem of subject identity and motion. In contrast, our method can generate videos that preserve both subject identity and motion pattern.

We also use an additional quantitative metric, Dynamic Degree [8], to evaluate motion intensity. Since Temporal Consistency cannot comprehensively evaluate the generated motions (1 for static videos), it is important to also evaluate the degree of dynamics (*i.e.*, whether it contains large motions) generated by the model. Tab. S1 shows that high Temporal Consistency often results in a lower Dynamic Degree, and our DreamVideo achieves a higher Dynamic Degree score, which exceeds AnimateDiff about 34.6%.

2.2. Subject Customization

In addition to the baselines in the main paper, we also compare our DreamVideo with another state-of-the-art method, Custom Diffusion [9]. Both the qualitative comparison in Fig. S2 and the quantitative comparison in Tab. S2 illustrate that our method outperforms Custom Diffusion and can generate videos that accurately retain subject identity and conform to diverse contextual descriptions with fewer parameters.

[6] https://github.com/showlab/Tune-A-Video

[7] https://github.com/omerbt/Text2LIVE

[8] https://github.com/guoyww/AnimateDiff

Method	CLIP-T	CLIP-I	DINO-I	T. Cons.	Para.
Custom Diffusion [9]	0.284	0.699	0.471	0.962	24M
DreamVideo (ours)	**0.295**	**0.701**	**0.475**	**0.964**	11M

Table S2. **Quantitative comparison of subject customization between our method and Custom Diffusion [9].** "T. Cons." and "Para." denote Temporal Consistency and parameter number, respectively.

Method	Text Alignment	Subject Fidelity	Temporal Consistency
ours *vs*. Textual Inversion [4]	58.4 / 41.6	79.8 / 20.2	69.7 / 30.3
ours *vs*. Dreamix [12]	63.7 / 36.3	56.0 / 44.0	50.8 / 49.2
ours *vs*. Custom Diffusion [9]	70.4 / 29.6	69.1 / 30.9	63.0 / 37.0

Table S3. **Human evaluations** on customizing subjects between our method and alternatives.

Method	Text Alignment	Motion Fidelity	Temporal Consistency
ours *vs*. ModelScopeT2V [17]	64.1 / 35.9	52.6 / 47.4	62.8 / 37.2
ours *vs*. Tune-A-Video [18]	73.8 / 26.2	52.4 / 47.6	74.1 / 25.9

Table S4. **Human evaluations** on customizing motions between our method and alternatives.

As shown in Fig. S3, we provide the customization results for more subjects, further demonstrating the favorable generalization of our method.

2.3. Motion Customization

To further evaluate the motion customization capabilities of our method, we show more qualitative comparison results with baselines on multiple training videos and a single training video, as shown in Fig. S4. Our method exhibits superior performance than baselines and ignores the appearance information from training videos when modeling motion patterns.

We showcase more results of motion customization in Fig. S5, providing further evidence of the robustness of our method.

2.4. User Study

For subject customization, we generate 120 videos from 15 subjects, where each subject includes 8 text prompts. We present three sets of questions to participants with 3$\sim$5 reference images of each subject to evaluate Text Alignment, Subject Fidelity, and Temporal Consistency. Given the generated videos of two anonymous methods, we ask each participant the following questions: (1) Text Alignment: "Which video better matches the text description?"; (2) Subject Fidelity: "Which video's subject is more similar to the target subject?"; (3) Temporal Consistency: "Which video is smoother and has less flicker?". For motion customization, we generate 120 videos from 20 motion patterns with 6 text prompts. We evaluate each pair of compared methods through Text Alignment, Motion Fidelity, and Temporal Consistency. The questions of Text Alignment and Temporal Consistency are similar to those in subject customization above, and the question of Motion Fidelity is like: "Which video's motion is more similar to the motion of target videos?" The human evaluation results are shown in Tab. S3 and Tab. S4. Our DreamVideo consistently outperforms other methods on all metrics.

3. More Ablation Studies

3.1. More Qualitative Results

We provide more qualitative results in Fig. S6 to further verify the effects of each component in our method. The conclusions are consistent with the descriptions in the main paper. Remarkably, we observe that without appearance guidance, the generated videos may learn some noise, artifacts, background, and other subject-unrelated information from training videos.

3.2. Effects of Parameters in Adapter and LoRA

To measure the impact of the number of parameters on performance, we reduce the hidden dimension of the adapter to make it have a comparable number of parameters as LoRA. For a fair comparison, we set the hidden dimension of the adapter to 32 without using textual identity and appearance guidance. We adopt the DreamBooth [15] paradigm for subject learning, which is the same as LoRA. Other settings are the same as our DreamVideo.

As shown in Fig. S7, we observe that LoRA fails to generate videos that preserve both subject identity and motion. The reason may be that LoRA modifies the original param-

Method	CLIP-T	CLIP-I	DINO-I	T. Cons.	Para.
LoRA	0.286	0.644	0.409	0.964	6M
Adapter	**0.298**	**0.648**	**0.412**	**0.967**	3M

Table S5. **Quantitative comparison of video customization between Adapter and LoRA.** "T. Cons." denotes Temporal Consistency. "Para." means parameter number.

Self-Attn	FFN	CLIP-T	T. Cons.
Serial	Serial	0.303	0.969
Serial	Parallel	0.306	0.971
Parallel	Serial	0.308	**0.975**
Parallel	Parallel	**0.309**	**0.975**

Table S6. **Quantitative comparison of adapter type on motion customization.** "Serial" and "Parallel" mean using serial and parallel adapters in the corresponding layer, respectively.

Method	CLIP-T	T. Cons.	Human
ModelScopeT2V [17]	0.290	0.950	22.4%
Tune-A-Video [18]	0.287	0.943	13.2%
Text2LIVE [1]	0.275	0.952	4.3%
DreamVideo (**ours**)	**0.318**	**0.964**	**60.1%**

Table S7. **Extra quantitative comparison on motion customization from DAVIS dataset.** Human denotes human evaluation.

eters of the model during inference, causing conflicts and sacrificing performance when merging spatial and temporal LoRAs. In contrast, the adapter can alleviate fusion conflicts and achieve a more harmonious combination.

The quantitative comparison results in Tab. S5 also illustrate the superiority of the adapter compared to LoRA in video customization tasks.

3.3. Effects of Adapter Type

To evaluate which adapter is more suitable for customization tasks, we design four combinations of adapters and parameter layers for motion customization, as shown in Tab. S6. We consider the serial adapter and parallel adapter along with self-attention layers and feed-forward layers. The results demonstrate that using parallel adapters on all layers achieves better performance. Therefore, we uniformly employ parallel adapters in our approach.

4. Extra Results of Motion Customization

We conduct extra motion customization experiments on the DAVIS dataset [13] to further verify the effectiveness of our method. We also design 116 text prompts used for experimental validation.

Method	CLIP-T	T. Cons.
w/o appearance guidance	0.292	0.956
DreamVideo (**ours**)	**0.318**	**0.964**

Table S8. **Extra ablation study of appearance guidance on motion customization from DAVIS dataset.**

Architecture	CLIP-T	T. Cons.
a linear layer	0.315	0.952
a linear layer with nonlinearity	0.308	0.949
bottleneck w/o nonlinearity	0.313	0.960
DreamVideo (**ours**)	**0.318**	**0.964**

Table S9. **Extra ablation study of adapter architecture on motion customization from DAVIS dataset.**

4.1. Comparison with More Baselines

Besides ModelScopeT2V [17] and Tune-A-Video [18], we also compare our method with Text2LIVE [1], as shown in Fig. S8 and Tab. S7. Qualitative and quantitative results demonstrate that our DreamVideo consistently outperforms baselines. For human evaluation in Tab. S7, we ask 5 people to rate the overall quality of 205 videos by four methods, and our method is most preferred by users.

4.2. Extra Ablation Studies

Effects of appearance guidance. Fig. S9 and Tab. S8 further verify the effects of appearance guidance on motion customization. When appearance guidance is removed, we observe that the appearance of subjects in the generated videos is corrupted.

Extra multiple seeds results. We provide multi-seed results in Fig. S9 to reduce randomness and verify the consistency of model performance.

Effects of adapter architecture. Tab. S9 shows the effects of different adapter architectures on motion customization. We find that our designed bottleneck adapter with nonlinearity achieves better performance.

5. Social Impact and Discussions

Social impact. While training large-scale video diffusion models is extremely expensive and unaffordable for most individuals, video customization by fine-tuning only a few images or videos provides users with the possibility to use video diffusion models flexibly. Our approach allows users to generate customized videos by arbitrarily combining subject and motion while also supporting individual subject customization or motion customization, all with a small computational cost. However, our method still suffers from the risks that many generative models face, such as fake data generation. Reliable video forgery detection techniques may be a solution to these problems.

Discussions. We provide some failure examples in Fig. S10. For subject customization, our approach is limited by the inherent capabilities of the base model. For example, in Fig. S10(a), the basic model fails to generate a video like “a wolf riding a bicycle”, causing our method to inherit this limitation. The possible reason is that the correlation between “wolf” and “bicycle” in the training set during pre-training is too weak. For motion customization, especially fine single video motion, our method may only learn the similar (rough) motion pattern and fails to achieve frame-by-frame correspondence, as shown in Fig. S10(b). Some video editing methods may be able to provide some solutions [2, 3, 18]. For video customization, some difficult combinations that contain multiple objects, such as “cat” and “horse”, still remain challenges. As shown in Fig. S10(c), our approach confuses “cat” and “horse” so that both exhibit “cat” characteristics. This phenomenon also exists in multi-subject image customization [9]. One possible solution is to further decouple the attention map of each subject.

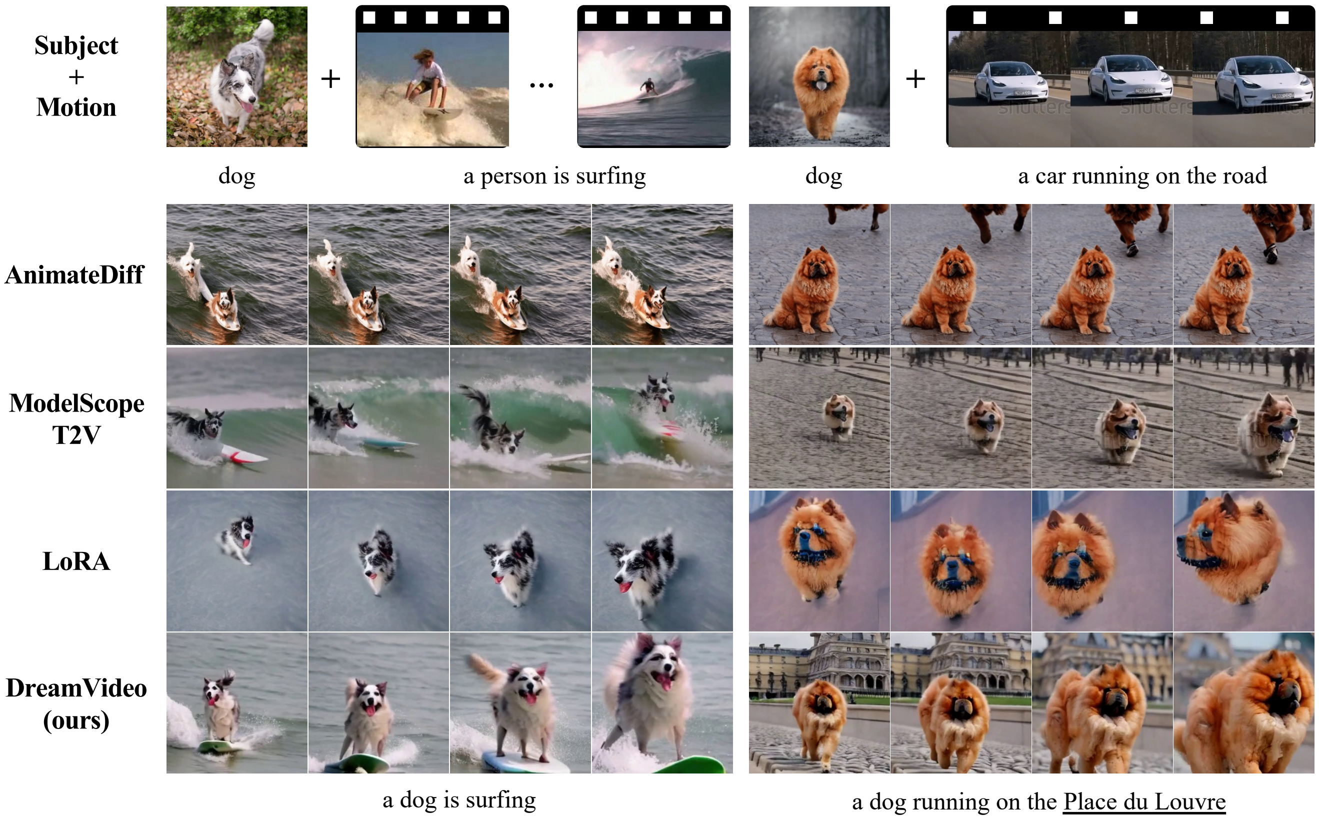


Figure S1. **Qualitative comparison of customized video generation with both subjects and motions.**

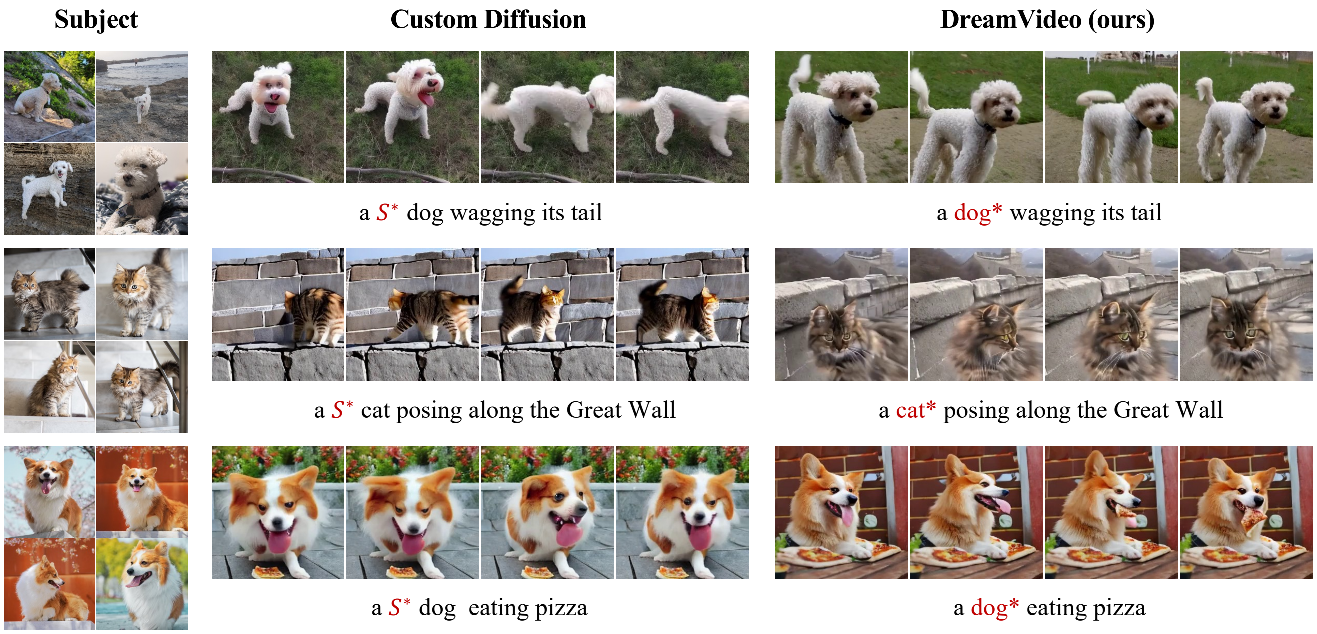
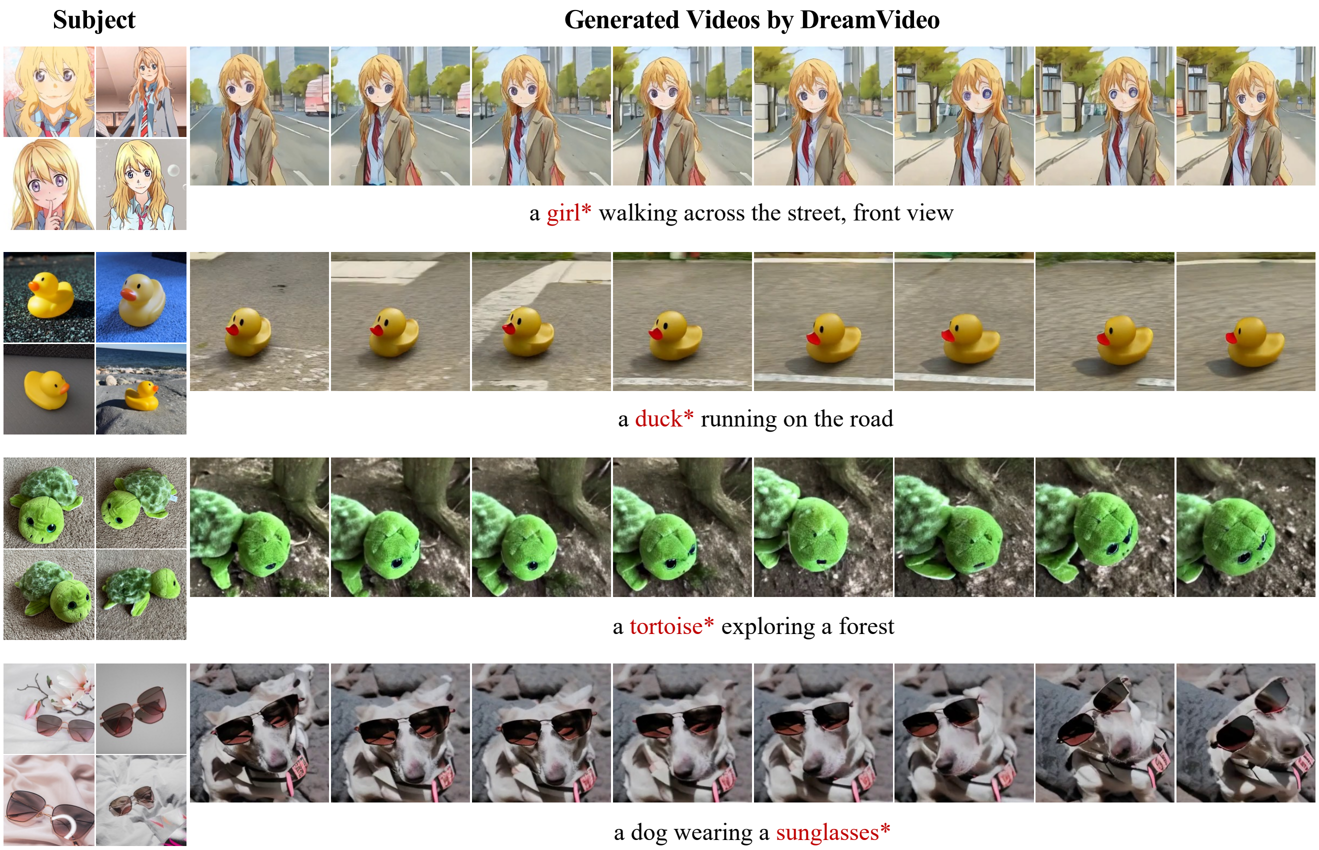


Figure S2. **Qualitative comparison of subject customization between our method and Custom Diffusion [9].**



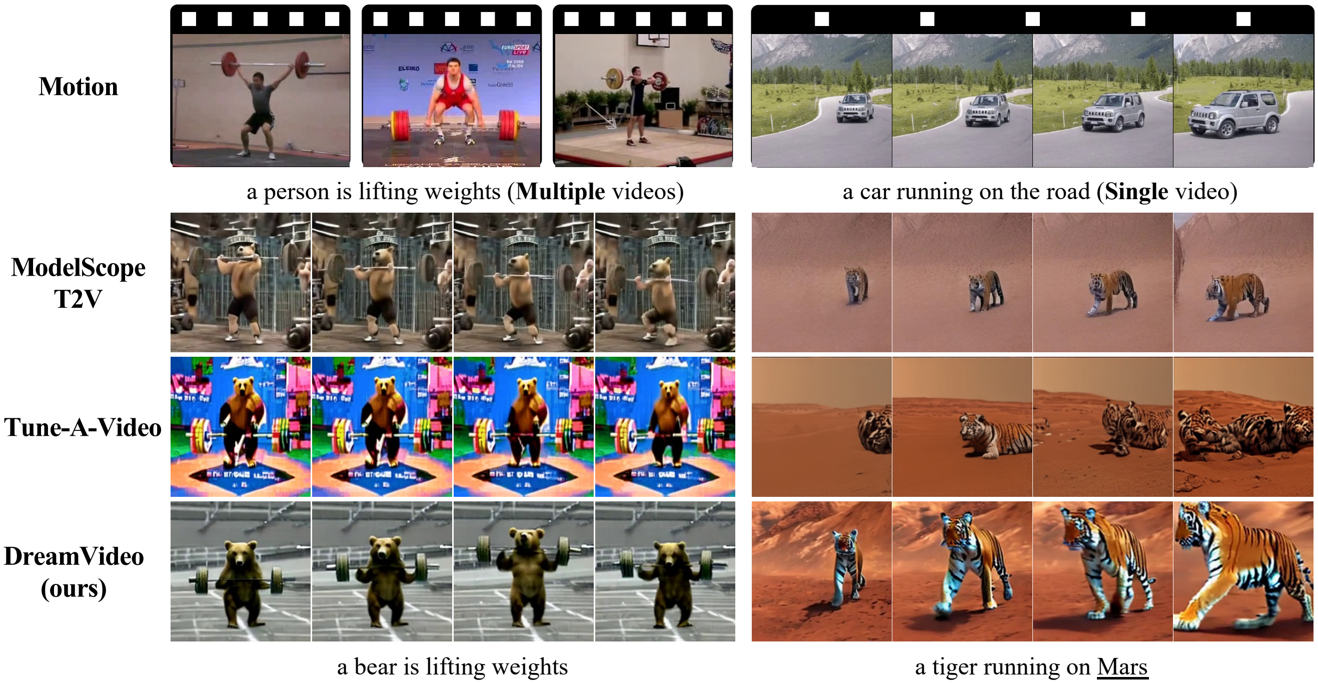
a girl* walking across the street, front view

a duck* running on the road

a tortoise* exploring a forest

a dog wearing a sunglasses*

Figure S3. **More results of subject customization for Our DreamVideo**.



a person is lifting weights (**Multiple** videos)

a car running on the road (**Single** video)

a bear is lifting weights

a tiger running on Mars

Figure S4. **Qualitative comparison of motion customization.**

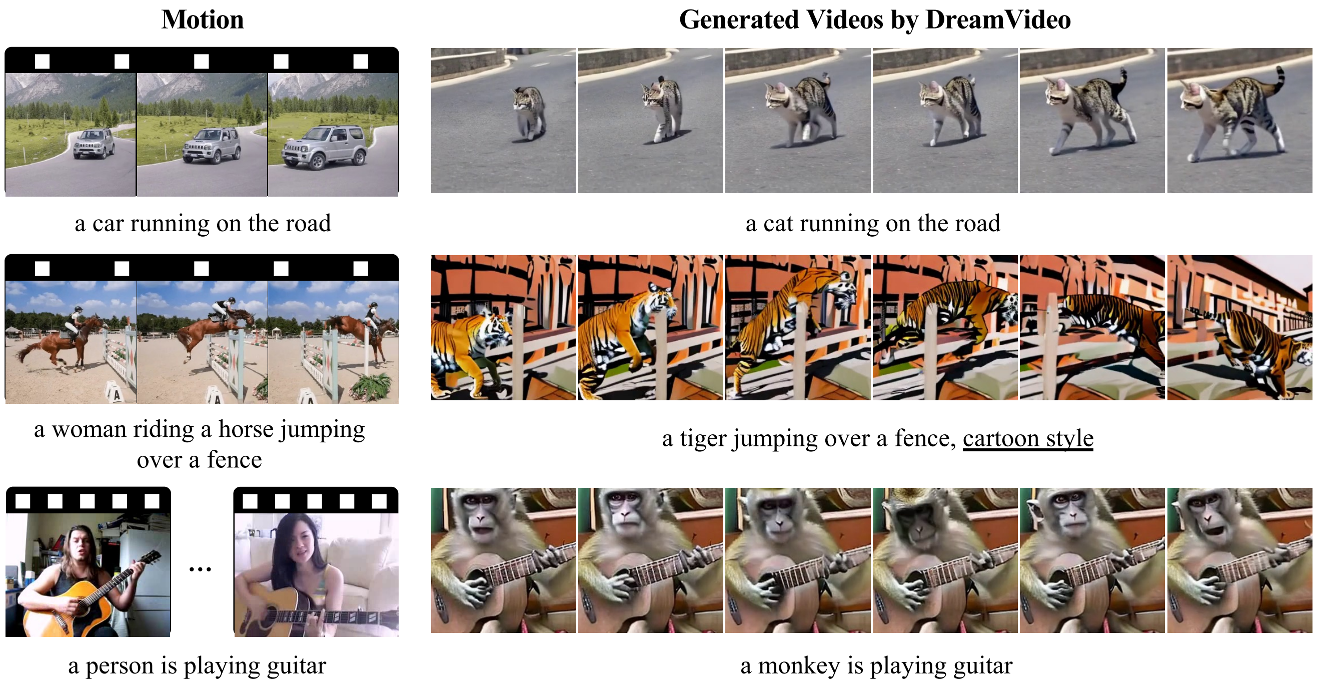

Motion
Generated Videos by DreamVideo
a car running on the road
a cat running on the road
a woman riding a horse jumping over a fence
a tiger jumping over a fence, cartoon style
...
a person is playing guitar
a monkey is playing guitar

Figure S5. **More results of motion customization for Our DreamVideo**.

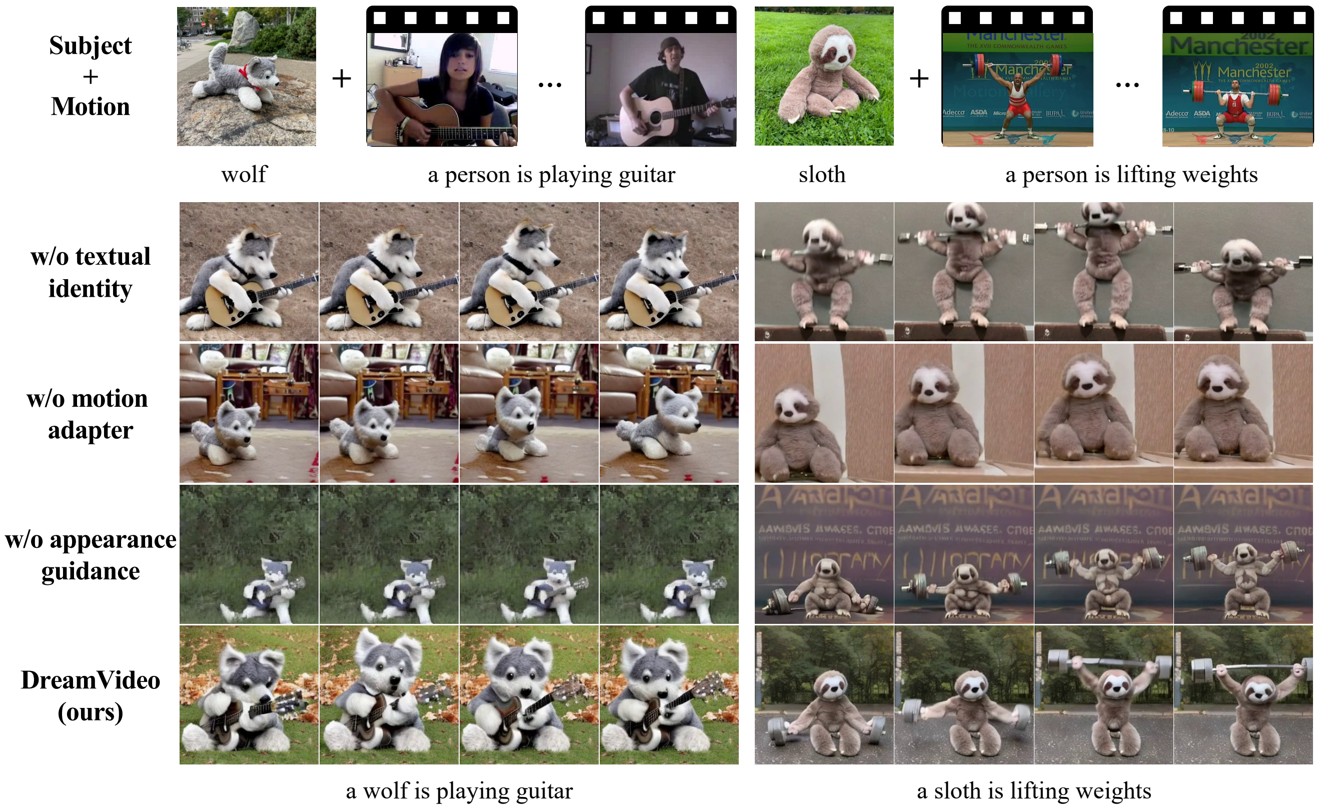


Figure S6. **Qualitative ablation studies** on each component.

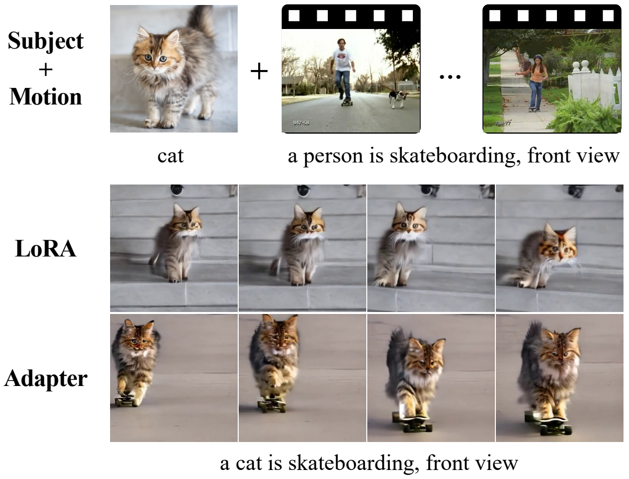

Subject
+
Motion
+
...
cat
a person is skateboarding, front view
LoRA
Adapter
a cat is skateboarding, front view

Figure S7. **Qualitative comparison of video customization between Adapter and LoRA**. The Adapter and LoRA here have the same hidden dimension (rank) and a comparable number of parameters.

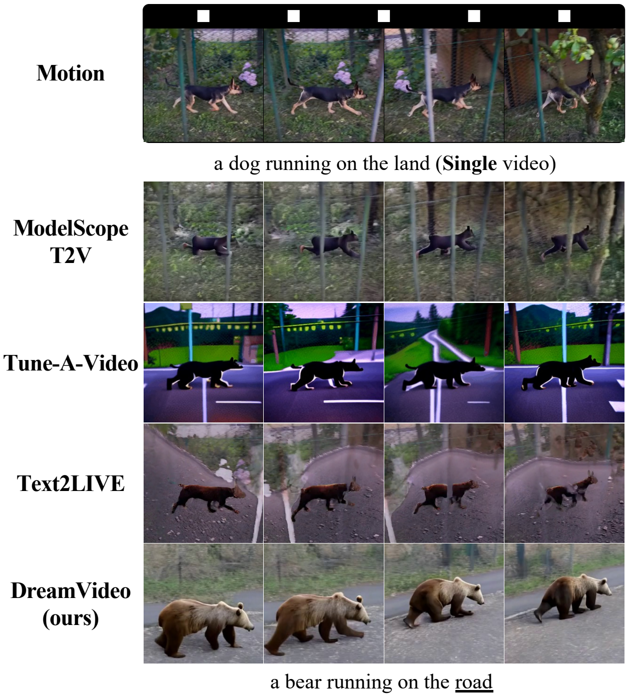

Motion
a dog running on the land (Single video)
ModelScope T2V
Tune-A-Video
Text2LIVE
DreamVideo (ours)
a bear running on the road

Figure S8. **Extra qualitative comparison of motion customization on DAVIS dataset.**

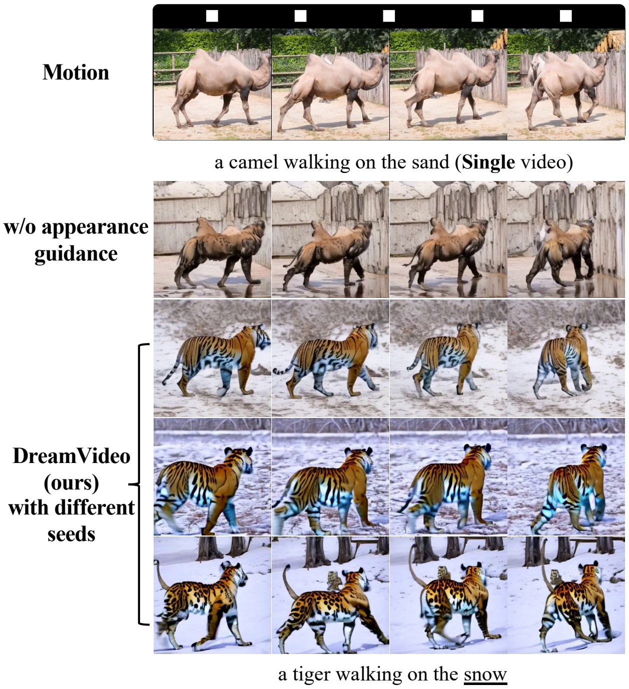
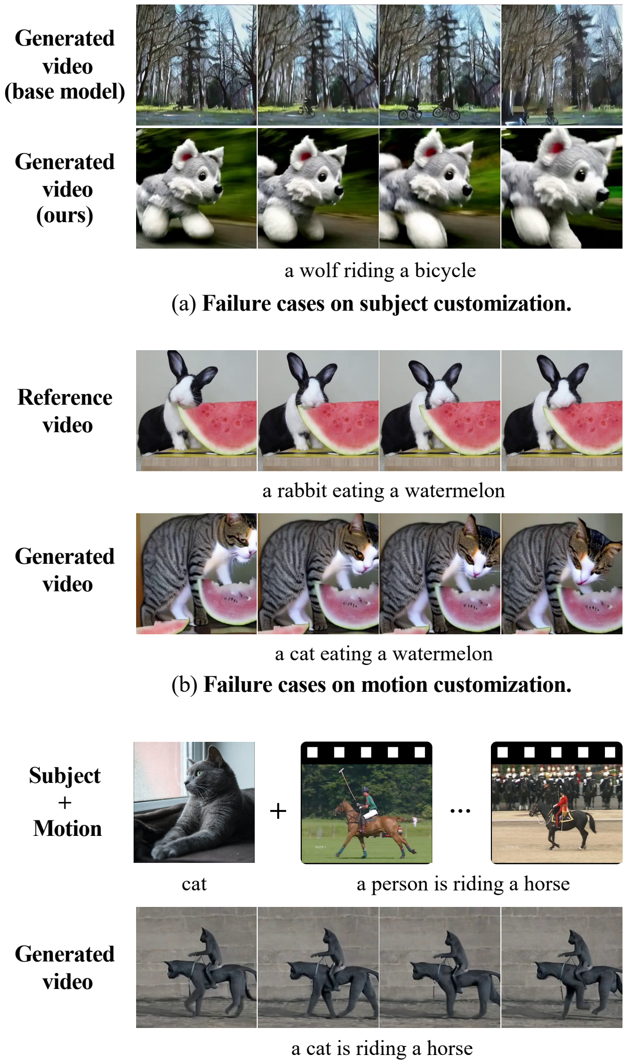

Motion
a camel walking on the sand (Single video)
w/o appearance guidance
DreamVideo (ours) with different seeds
a tiger walking on the snow

Figure S9. **Extra ablation study of appearance guidance and multi-seed results on motion customization from DAVIS dataset.**



(a) **Failure cases on subject customization.**

(b) **Failure cases on motion customization.**

(c) **Failure cases on video customization.**

Figure S10. **Failure cases**. (a) Our method is limited by the inherent capabilities of the base model. (b) Our method may only learn the similar motion pattern on a fine single video motion. (c) Some difficult combinations that contain multiple objects still remain challenges.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022. 2, 4

[2] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 5

[3] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 5

[4] 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. 1, 3

[5] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[8] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023. 2

[9] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2, 3, 5, 6

[10] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023. 1

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[12] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 1, 3

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

[14] 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

[15] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 3

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

[17] 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. 1, 2, 3, 4

[18] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 1, 2, 3, 4, 5