VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models (Supplementary Material)

Hyeonho Jeong* Geon Yeong Park* Jong Chul Ye
Korea Advanced Institute of Science and Technology (KAIST)
{hyeonho.jeong, pky3436, jong.ye}@kaist.ac.kr

This supplementary material is organized as follows. Section 1 introduces the pseudo training algorithm behind our Video Motion Customization (VMC) framework. In Section 2, we provide a discussion on related works in the field of generative model customization. Following this, we delve into the details on our training and inference configurations in Section 3. Concluding the document, Section 4 features a showcase of additional results obtained from our VMC framework.

1. Pseudo Training Algorithm

Algorithm 1 Temporal Attention Adaption

Input: N-frame input video sequence $(\boldsymbol{v}_0^n)_{n\in\{1,\dots,N\}}$, appearance-invariant training prompt $\mathcal{P}_{\text{inv}}$, textual encoder ψ, Training iterations M, key-frame generator parameterized by θ and its temporal attention parameters θ_{TA}.
Output: Fine-tuned temporal attention layers θ^*_{TA}.

for $step = 1$ **to** M **do**
  Sample timestep $t \in [0, T]$ and Gaussian noise $\boldsymbol{\epsilon}_t^{1:N}$, where $\boldsymbol{\epsilon}_t^n \in \mathbb{R}^d \sim \mathcal{N}(0, I)$
  Prepare text embeddings $c_{\text{inv}} = \psi(\mathcal{P}_{\text{inv}})$
  $\boldsymbol{v}_t^n = \sqrt{\bar{\alpha}_t}\boldsymbol{v}_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t^n, \forall n.$
  $\delta\boldsymbol{\epsilon}_{\theta,t}^n = \boldsymbol{\epsilon}_\theta^{n+1}(\boldsymbol{v}_t^{1:N}, t, c_{\text{inv}}) - \boldsymbol{\epsilon}_\theta^n(\boldsymbol{v}_t^{1:N}, t, c_{\text{inv}}), \forall n \le N-1.$
  $\delta\boldsymbol{\epsilon}_t^n = \boldsymbol{\epsilon}_t^{n+1} - \boldsymbol{\epsilon}_t^n, \forall n \le N-1$
  Update θ_{TA} with $\frac{1}{N-1}\sum_n \ell_{\cos}(\delta\boldsymbol{\epsilon}_t^n, \delta\boldsymbol{\epsilon}_{\theta,t}^n)$
end for

We express Gaussian noises as $\boldsymbol{\epsilon}_t^{1:N}$ to avoid confusion. In our observation, aligning the image-space residuals ($\delta\boldsymbol{v}_0^n$ and $\delta\hat{\boldsymbol{v}}_0^n(t)$) corresponds to aligning the latent-space epsilon residuals ($\delta\boldsymbol{\epsilon}_t^n$ and $\delta\boldsymbol{\epsilon}_{\theta,t}^n$) across varying time steps $t \in [0, T]$. This relationship stems from expressing the motion vector $\delta\boldsymbol{v}_0^n$ and its estimation $\delta\hat{\boldsymbol{v}}_0^n(t)$ in terms of $\delta\boldsymbol{v}_t^n$, $\delta\boldsymbol{\epsilon}_t^n$, and $\delta\boldsymbol{\epsilon}_{\theta,t}^n$. Consequently, the proposed optimization framework fine-tunes temporal attention layers by leveraging diverse diffusion latent spaces at time t which potentially contains multi-scale rich descriptions of video frames. Therefore, this optimization approach seamlessly applies to video diffusion models trained using epsilon-matching, thanks to the equivalence between $\delta\boldsymbol{\epsilon}_t^n$-matching and $\delta\boldsymbol{v}_0^n$-matching.

*indicates co-first authors

2. Related Works

Image Customization. Prior methodologies in text-to-image customization, termed personalization [2–4, 7–10, 12], aimed at capturing specific subject appearances while maintaining the model's ability to generate varied contents. However, this pursuit of personalization poses challenges in time and memory demands [8]. Fine-tuning each personalized model requires substantial time costs while storing multiple personalized models may strain storage capacity. To address these hurdles, some approaches prioritize efficient parameter customization, leveraging techniques like LoRA [3, 5] or HyperNetwork [9] rather than training the entire model.

Video Customization. Building on the success of text-to-image customization, recent efforts have adopted text-to-image or text-to-video diffusion models for customizing videos in terms of appearance or motion. These endeavors, such as frameworks proposed by [1, 13], focus on creating videos faithful to given subjects or motions. Moreover, works by [16] or [14] delve into motion-centric video customization, employing various fine-tuning approaches ranging from temporal-spatial motion learning layers to newly introduced LoRAs. In this paper, the proposed VMC framework emphasizes efficient motion customization with explicit motion distillation objectives, specifically targeting temporal attention layers. This approach, facilitated by cascaded video diffusion models, efficiently distills motion from a single video clip while minimizing computational

burdens in terms of both time and memory.

3. Training & Inference Details

For our work, we utilize the cascaded video diffusion models from Show-1 [15], employing its publicly accessible pre-trained weights [1]. Our approach maintains the temporal interpolation and spatial super-resolution modules in their original state while focusing our temporal optimization solely on the keyframe generator. In specific, we finetune Query, Key, Value projection matrices W^Q, W^K, W^V of temporal attention layers of the keyframe UNet. We use AdamW [6] optimizer, with weight decay of 0.01 and learning rate 0.0001. By default, we employ 400 training steps. During the inference phase, we perform DDIM inversion [11] for 75 steps. For the temporal interpolation and spatial super resolution stages, we follow the default settings of Show-1.

4. Additional Results

This section is dedicated to presenting further results in motion customization. We display keyframes (7 out of the total 8 frames) from input videos in Figures S1, S2, S3, and S4, accompanied by various visual variations that maintain the essential motion patterns. Specifically, Figure S1 showcases input videos featuring car movements. In Figure S2, we exhibit input videos capturing the dynamics of airplanes in flight and the blooming of a flower. Figure S3 focuses on bird movements, including walking, taking off, floating, and flying. Lastly, Figure S4-top highlights input videos of mammals, while S4-bottom illustrates the motion of pills falling. Moreover, for a comprehensive comparison between the motion in the input and generated videos, complete frames from these videos are presented in Figures S5, S6, S7, S8, and S9. In each of these figures, the left columns show the 8-frame input video, while the adjacent three columns on the right exhibit 29 frames from the generated videos, replicating the same motion pattern.

[1] https://huggingface.co/showlab/show-1-base

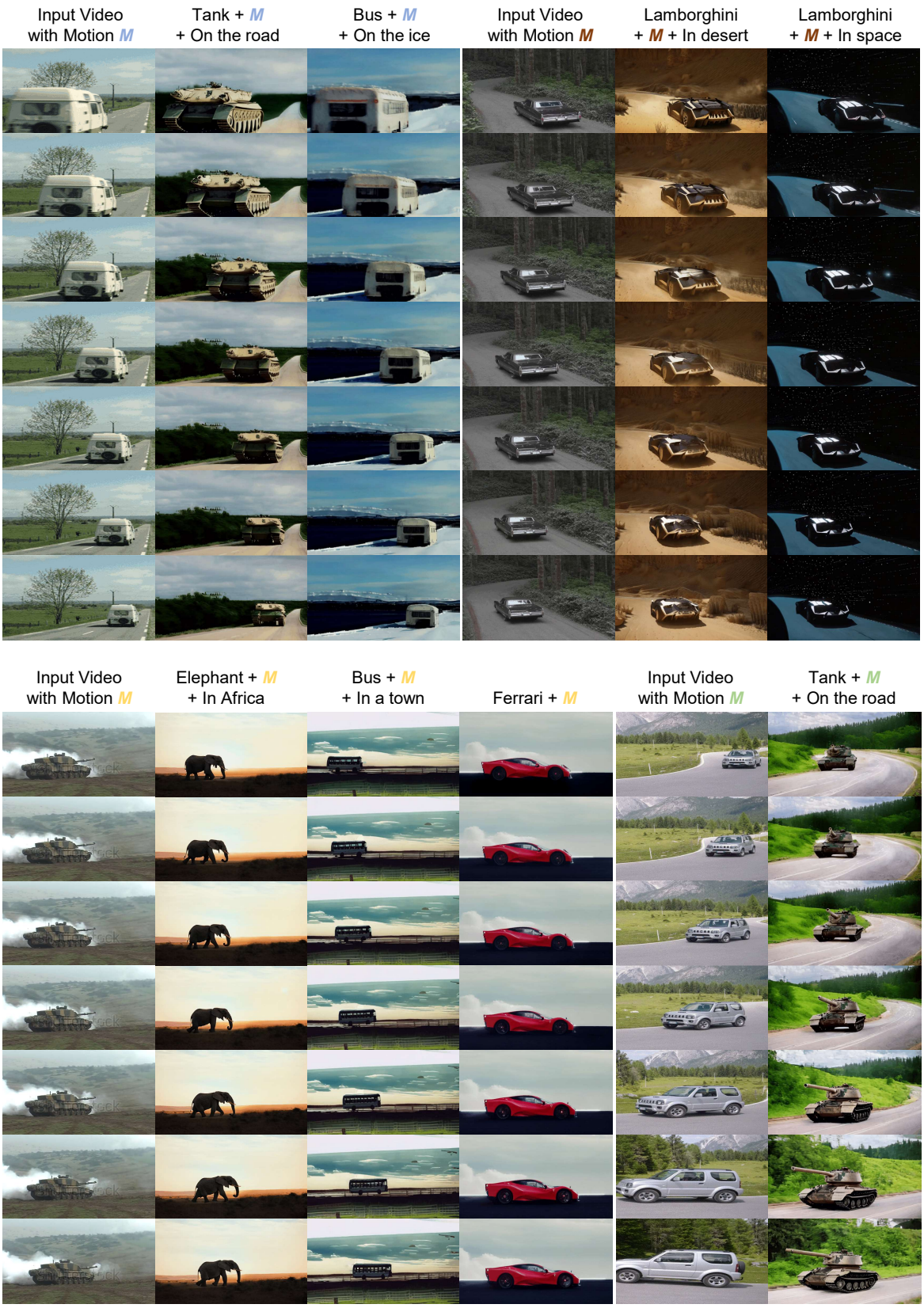

Figure S1. Video Motion Customization results: Keyframes visualized.

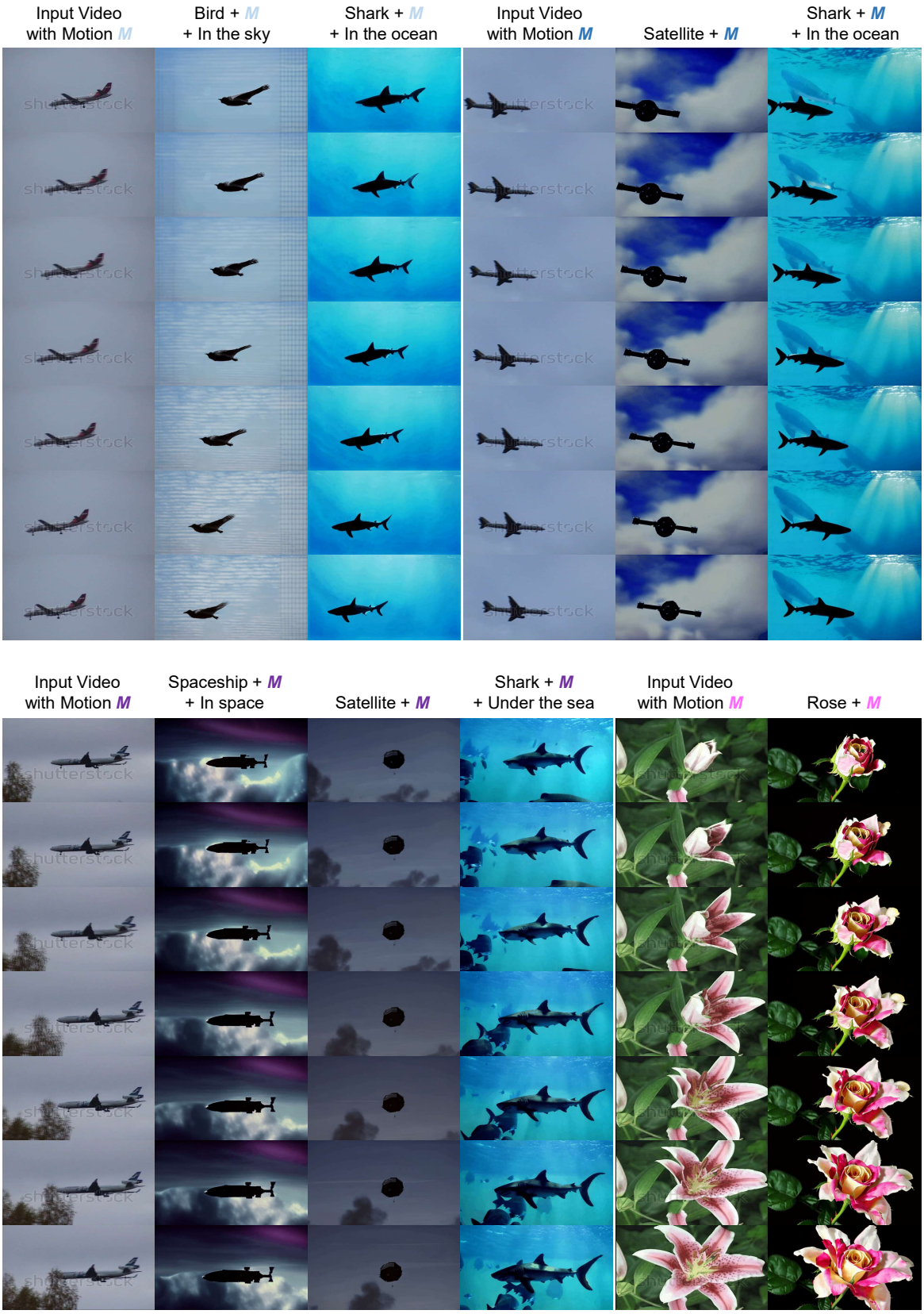

Figure S2. Video Motion Customization results: Keyframes visualized.

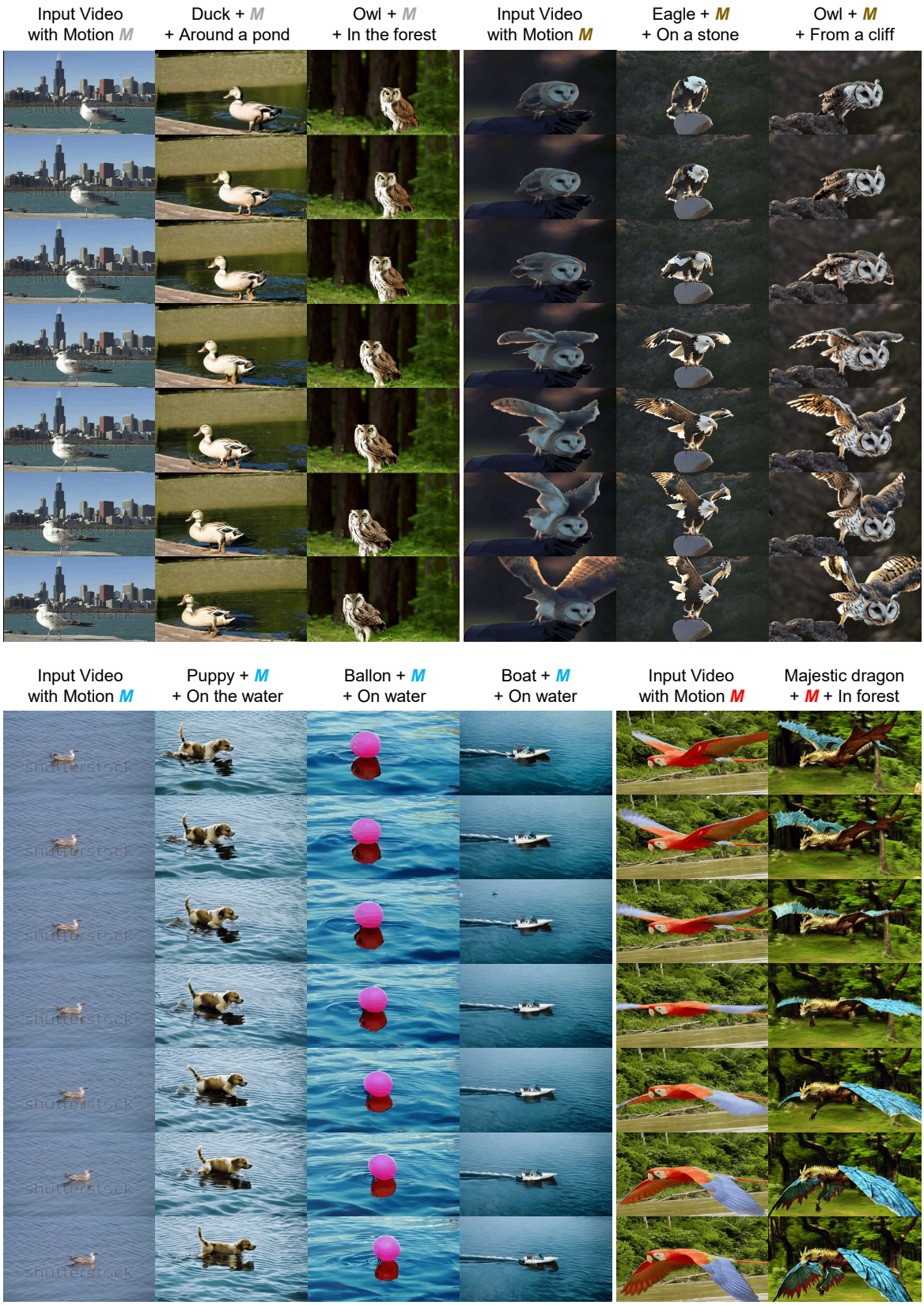

Figure S3. Video Motion Customization results: Keyframes visualized.

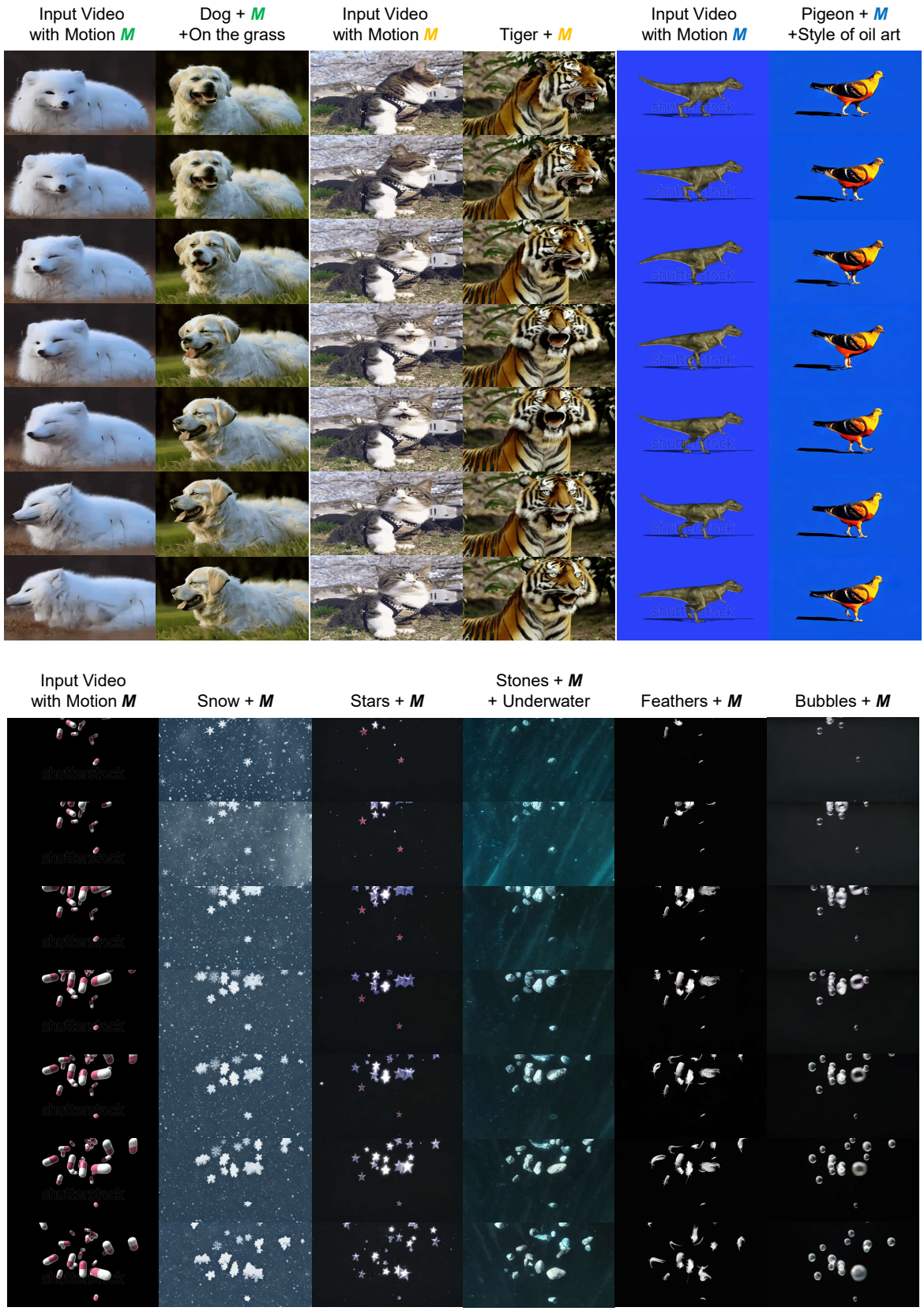

Figure S4. Video Motion Customization results: Keyframes visualized.

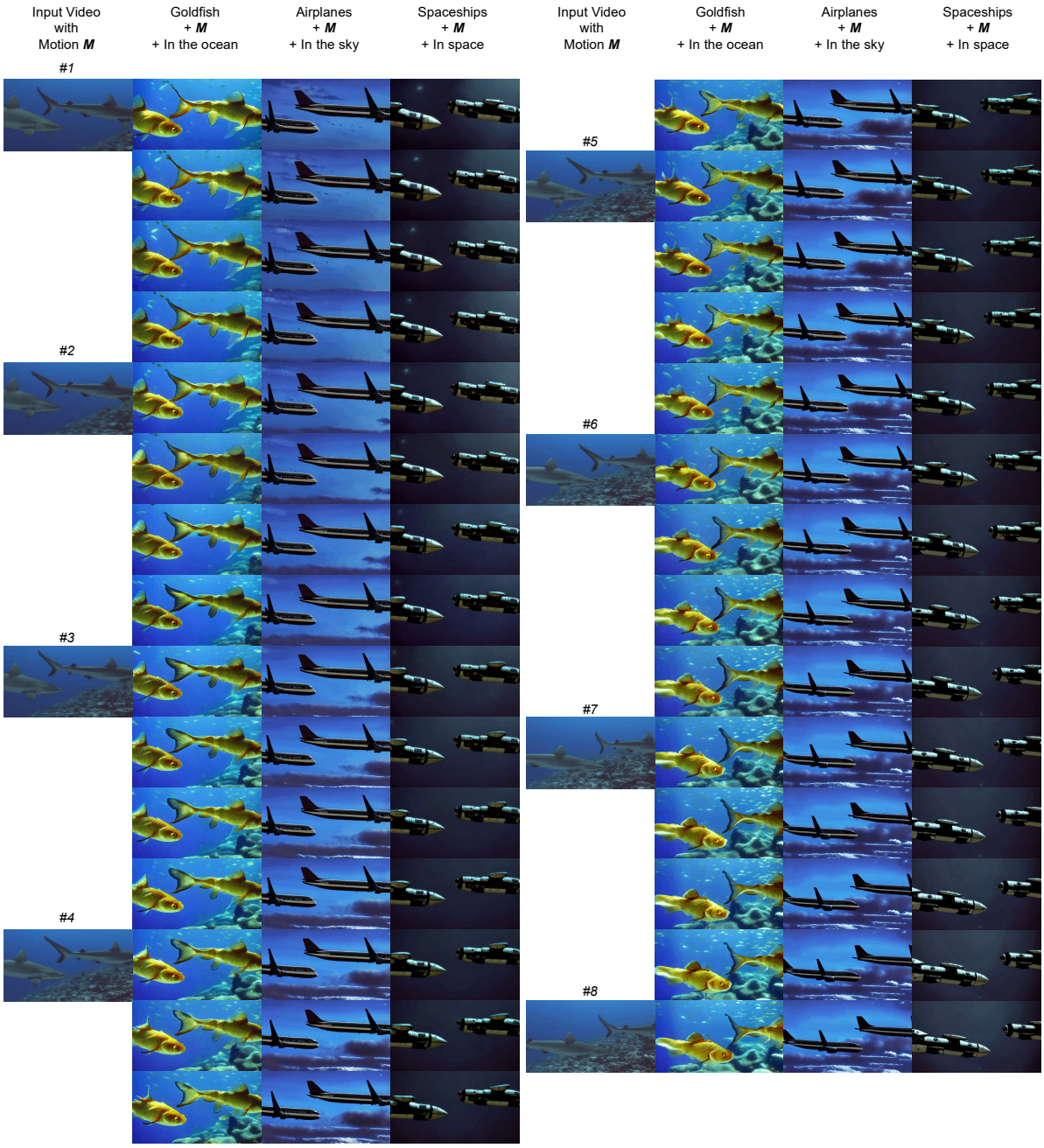

Figure S5. Full-frame results of Video Motion Customization: Text prompt “Sharks are moving” is used for training the keyframe generation UNet.

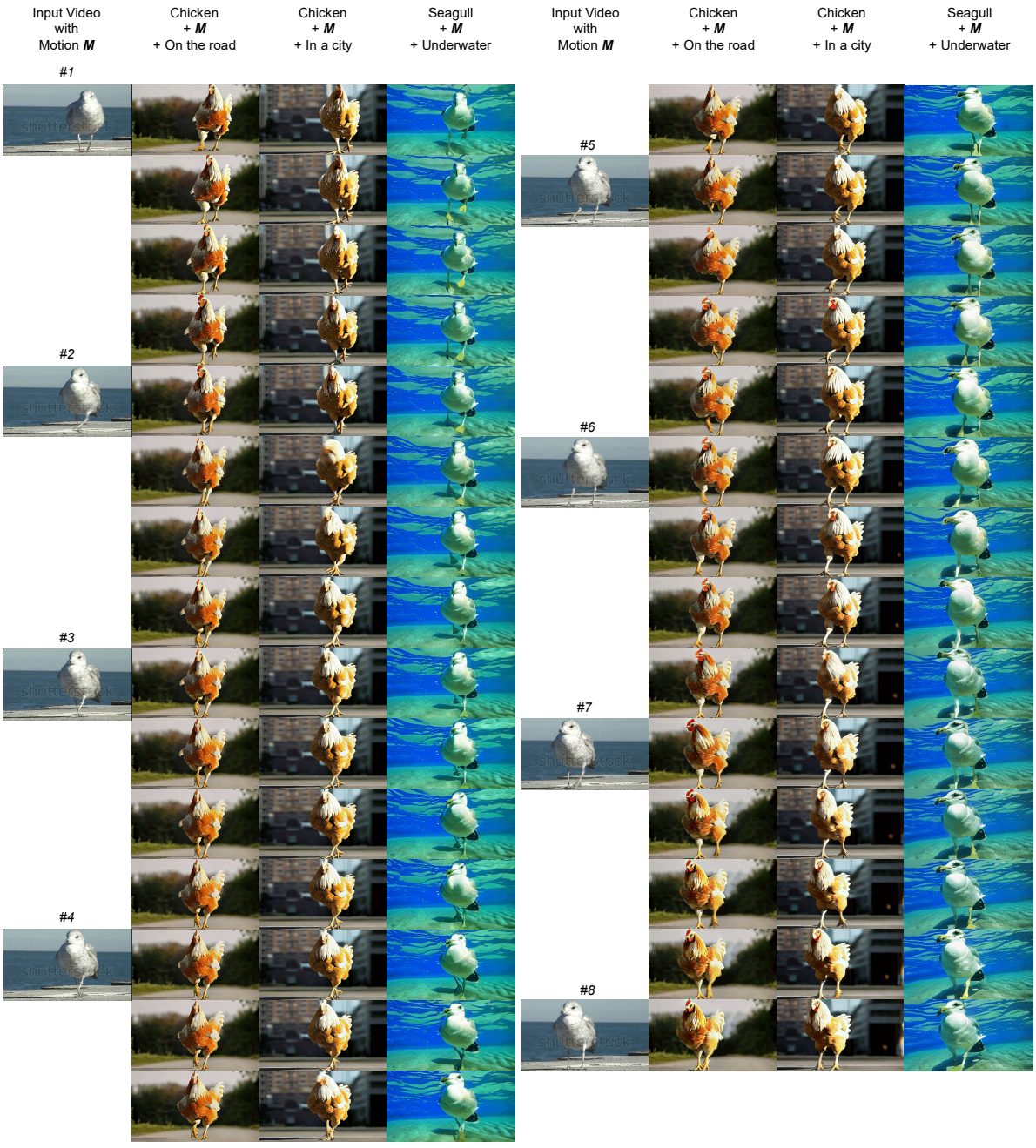

Figure S6. Full-frame results of Video Motion Customization: Text prompt "A seagull is walking" is used for training the keyframe generation UNet.

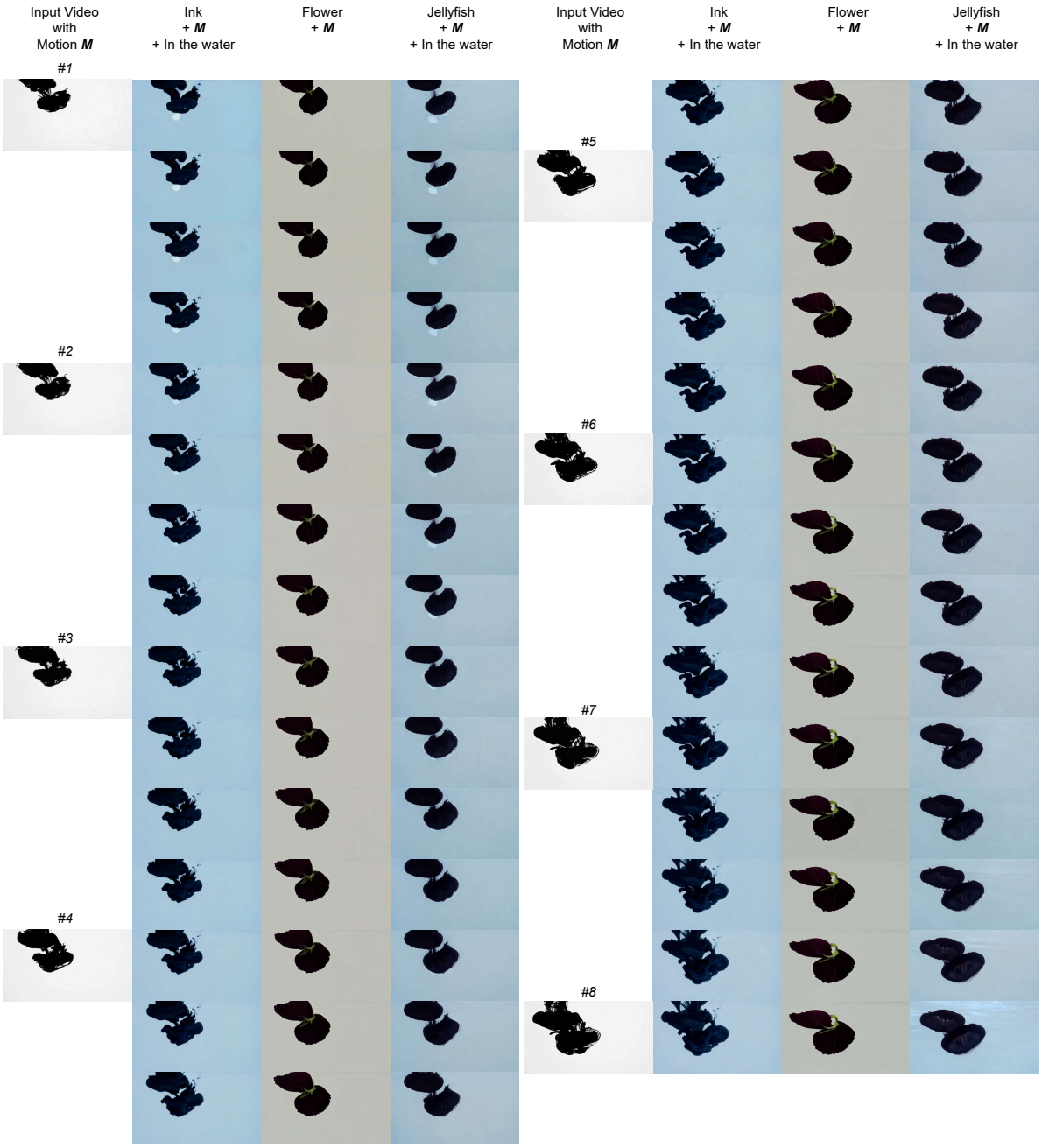

Figure S7. Full-frame results of Video Motion Customization: Text prompt “Ink is spreading” is used for training the keyframe generation UNet.

Figure S8. Full-frame results of Video Motion Customization: Text prompt "A man is snowboarding" is used for training the keyframe generation UNet.

Figure S9. Full-frame results of Video Motion Customization: Text prompt “A tiger is walking” is used for training the keyframe generation UNet.

References

[1] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*, 2023. 1

[2] 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] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023. 1

[4] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 1

[5] 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. 1

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[7] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14267–14276, 2023. 1

[8] 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

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 1

[10] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 1

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[12] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 1

[13] 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

[14] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*, 2023. 1

[15] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 2

[16] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 1