Customize-A-Video: One-Shot Motion Customization of Text-to-Video Diffusion Models Supplementary Materials

Yixuan Ren[1], Yang Zhou[2], Jimei Yang[2], Jing Shi[2], Difan Liu[2], Feng Liu[2], Mingi Kwon[3], and Abhinav Shrivastava[1]

[1] University of Maryland, College Park MD 20770, USA
[2] Adobe Research, San Jose CA 95110, USA
[3] Yonsei University, Seoul 03722, Republic of Korea
https://customize-a-video.github.io

A Implementation Details

A.1 Bypassing Temporal Layers in T2V Models

Many diffusion-based T2V models such as [1,2,4,8] have their denoising network structure adapted from T2I UNet with temporal convolution and attention layers injected. The new temporal layers are usually implemented as residual connections. The models are also usually trained on image and video datasets jointly to acquire both appearance and motion generative capability.

Based on this mechanism, we propose to train our appearance absorbers with the temporal layers bypassed and the model to perform image generation tasks on static frames. This shared design further enables us to load third-party image customization models pre-trained on external image data to serve as ready appearance absorbers or additional spatial customization modules in our video applications.

A.2 Patch Training of Appearance Absorbers

Some motions are intrinsically highly associated with postures, such as walking, running and sitting, and one image can primarily represent them. When the appearance absorbers have modeled the static postures to fit the appearance in the first training stage, T-LoRA might have little left to learn such as only the trivial perturbations across frames.

Therefore, we propose to crop the unordered frames into patches and encourage the appearance absorbers to mainly capture local shapes and textures in the first training stage. This prevent our appearance absorbers from overfitting on the global structures fundamentally. In practice, we find that setting the crop ratio randomly between 0.33 to 0.67 yields the best effect to retain the desired motion evidently in the second training stage.

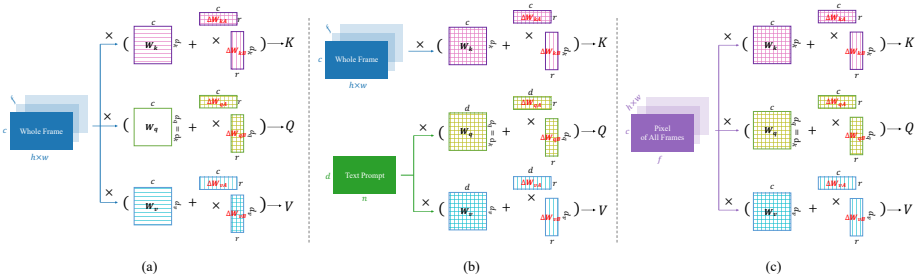

Fig. 1: (a) Spatial self-attention between each frame and itself; (b) Spatial cross-attention between each frame and the text prompt; (c) Temporal cross-frame attention among pixels of all frames in a video. Batch size is omitted for simplicity.

A.3 Attentions and LoRAs in T2V Diffusion Models

A base T2V diffusion model involves spatial self-attention (SSA) between a frame and itself, spatial cross-attention (SCA) between each frame and the text prompt, and temporal cross-frame attention (TCFA) among a pixel across all time in each 3D UNet block. We display their computations in Fig. 1. Three types of input and their corresponding K, Q and V are marked in respective colors. SSA is calculated between each frame and itself (K, Q, V). SCA is between a frame and the text prompt (K, Q, V). TCFA is among pixels of all frames (K, Q, V). Our LoRAs are applied to all attention weights W_* (ΔW_k, ΔW_q, ΔW_v).

A.4 Model Hyperparameters and Training Time

LoRA [5] typically features very few additional parameters attached to the base model. Its rank r controls the shape of the residual matrix, and α represents its scale when added to the pre-trained model weights. In experiments we discovered that setting the rank of T-LoRA $r_T = 4$ and the rank of S-LoRA in the appearance absorber $r_S = 1$ yields satisfactory results. Meanwhile, we empirically determined the alpha values $\alpha_T = 1$ for T-LoRAs and $\alpha_S = 0.5$ for S-LoRAs. For textual inversion [3] as the appearance absorber, we set the length of new learnable tokens to 2-6 depending on the content complexity.

We run experiments on a single NVIDIA RTX A5000 GPU with half-precision floats. Our T-LoRA takes approximately 7 minutes to converge. S-LoRA takes around 0.5 minute and the textual inversion takes 1 minute to converge in the first training stage. This is comparable to Tune-A-Video [9] (6 minutes), Video-P2P [6] (8 minutes in fast mode; 14 minutes in full mode on A6000 for bigger VRAM) and concurrent work MotionDirector [10] (8 minutes) on the same device for the same frame resolution and clip length. The learning rate is set to 5×10^{-4} for T-LoRA and 5×10^{-5} for appearance absorbers to prevent overfitting.

It worth noting that due to the difference in LoRA applications between our method and concurrent work MotionDirector, the quantities of parameters

Table 1: Quantitative and model size comparison with concurrent work.

Method	Text Align.↑	Temp. Consist.↓	Div.↑	LoRA Rank		#Params	
				Temp.	Spat.	Temp.	Spat.
Ours No AA	31.687	0.166	0.613	4	-	831.5K	-
Ours S-LoRA AA	31.913	0.163	0.618		1		207.5K
Ours TextInv AA	**32.632**	**0.160**	<u>0.621</u>		-		4K
Ours Both AA	32.193	0.164	**0.631**		1		211.5K
MotionDirector [10]	31.842	0.166	0.595	2	1	779.5K	274.5K
MotionDirector [10]	<u>32.500</u>	<u>0.163</u>	0.606	4	1	1559K	274.5K

and module sizes are not aligned with the same LoRA rank. We apply LoRAs on temporal cross-frame attentions (TCFAs), while MotionDirector moreoever add them to the following feed-forward networks (FFNs). This lead to approximately twice the quantity of parameters to tune. We apply LoRAs on all spatial attentions including the self-attentions (SSAs) and the cross-attentions (SCAs). MotionDirector excludes the SCAs and additionally involves the following FFNs. Thus our spatial LoRAs have comparable amounts of parameters. Tab. 1 expands the quantitative comparison with these model size differences.

B More Visualizations

B.1 Video Generation Results

More video results generated by our models are displayed in Fig. 2. We present two random output samples for each reference video.

B.2 Appearance Absorber Results

We exhibit the output of our appearance absorbers trained on unorder reference frames with the spatial text prompt in Fig. 3. The 2nd and 4th rows show the generation results with the appearance absorbers (S-LoRA and textual inversion respectively, same below) loaded and all temporal layers bypassed in the base T2V model, and the spatial part of the text prompt is used. It yields individual static frame replicating the reference appearances with random postures. The dynamic information is successfully left for our temporal customization module to learn in the next stage. The 3rd and 5th rows show the output videos with the appearance absorbers loaded on the full base T2V model, and the full text prompt is used. With the temporal description, the model can still only produce generic motions upon the learned appearances, indicating the necessity and effectiveness of our temporal customization module training. It can be further noticed that S-LoRA and textual inversion have different flavors of spatial modeling due to their different mechanisms, and thus loading both of them achieves the best performance with comprehensive and thorough appearance absorbing.

Fig. 2: Additional generation results of our method.

Fig. 3: Appearance absorbers' generation output. The 2nd and 3rd rows have S-LoRA loaded. The 4th and 5th rows have textual inversion loaded. The training prompts and special tokens are noted above for each sample.

B.3 Training Schedule Variances

Our modules fit on each reference video individually to model its unique motion signal. Fig. 4 displays cases whose optimal iterations vary across different reference videos. In general we observed that the convergence steps increase along with the complexity the specific reference motion and that of the original appearance.

We also observed that different types of appearance absorbers may exhibit different characteristics that affect the optimal checkpoint step and the output details. In Fig. 5 we present some cases where appearance absorbers vary in their effect of assisting following stages to capture the accurate motion or to generate novel scenes in certain iterations.

C User Study Questionnaire Design

We present example questions in our user studies in Fig. 6. Every reference video is presented with the output videos by random 4 out of 5 algorithms to be evaluated. For motion fidelity, 1-star represents the most dissimilar and 5-star represents the most faithful transferred motions w.r.t. the reference video. For motion diversity, 1-star indicates the most identical and 5-star indicates the most diverse generated motions among the two output videos.

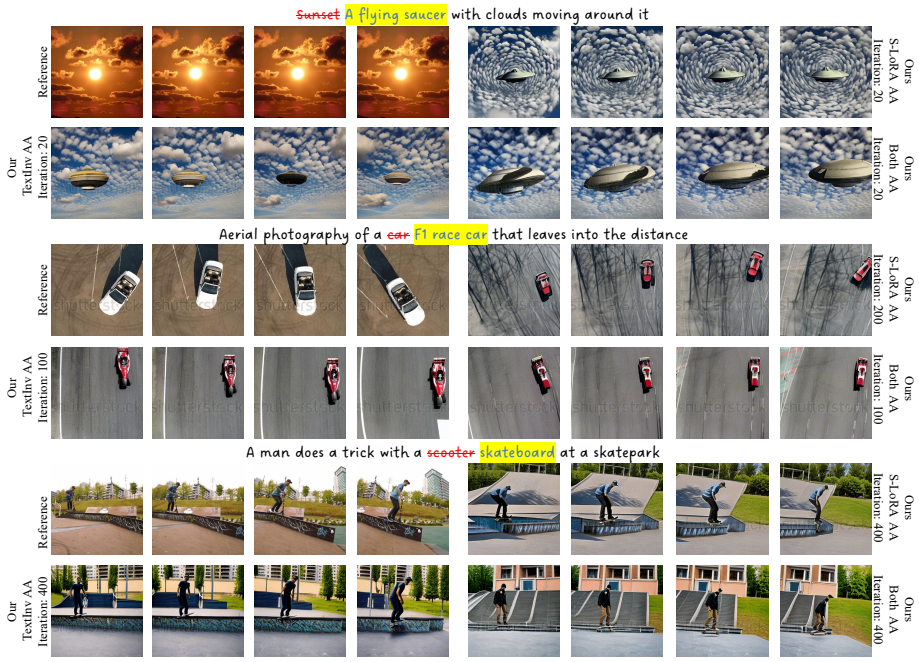

Fig. 4: Examples where different reference videos require different tuning iterations. (1-2) Simpler motions such as camera movements usually converge faster. (3) More complicated motions such as animal or human actions would demand more tuning steps.

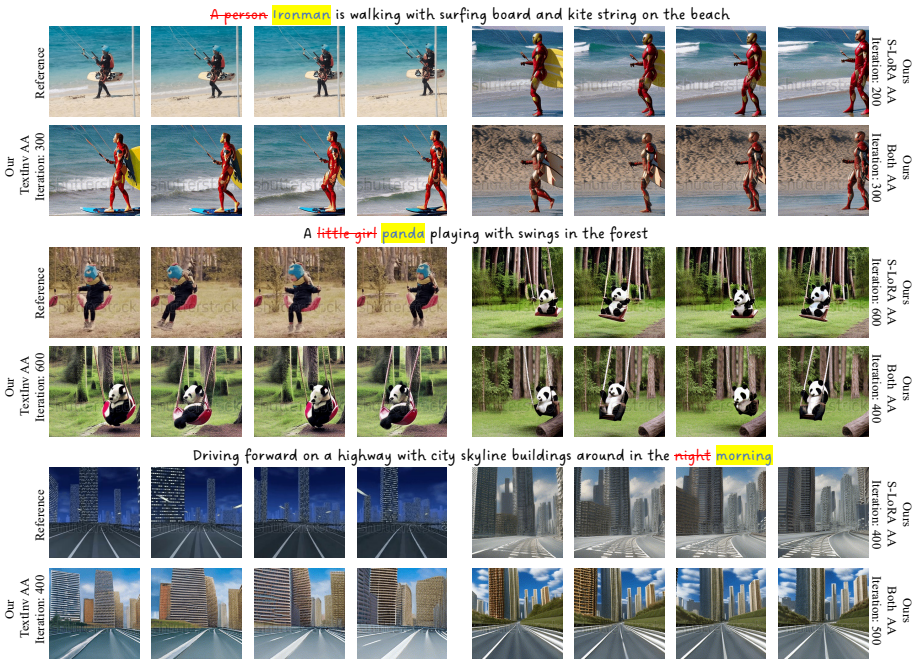

Fig. 5: Examples where different appearance absorbers exhibit different characteristics. (1) Our Both AA absorbs the original appearance more thoroughly, leading to more diverse new background generated. (2) Our Both AA may reduce the necessary convergence step compared to a single appearance absorber. (3) Our Both AA may be more stable and enable more tuning iterations without collapse to thoroughly clean up the original art style and generate a new one.

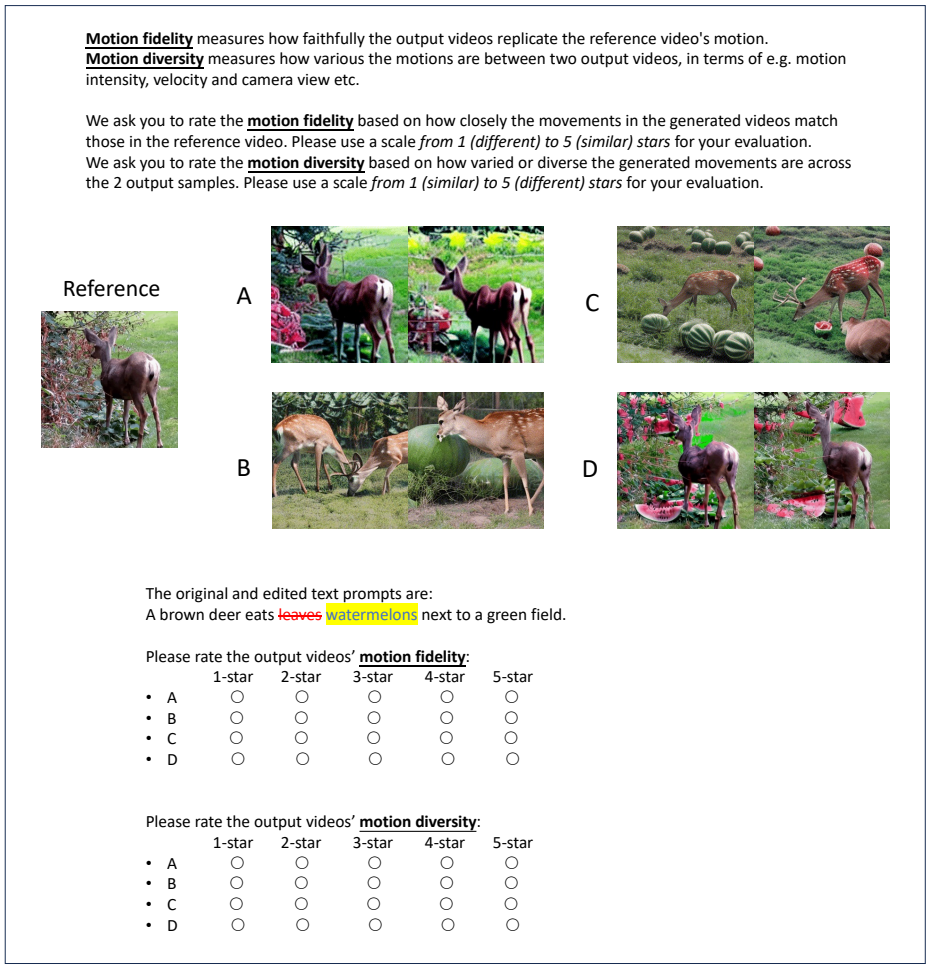

Fig. 6: An example question in the human user study. Participants are asked to rate each algorithm's output videos from 1 to 5 stars.

D Limitation Discussions

Per Instance Finetuning. Our method tunes on each reference video individually. Similar to comparing approaches [6, 9, 10], our method needs specialized recipes for different videos of diverse appearances and motions. The training configurations and iterations depend on the target video and can vary a lot as analysed in Sec. B.3. The trade-off balance between the object motion fidelity and its diversity also relies on dedicated hyperparameters and adjustments between underfitting and overfitting, like its image customization counterparts [3,7] have described. Though, our staged training pipeline and plug-and-play designs enable reusing both the appearance absorbers and T-LoRAs for future training and compositional inference, which improve their usability.

Spatial Domain Shift. The standalone finetuning of partial layers might have the risk of breaking the consistency among the pre-trained weights if the appearance absorbers overfit on static content reconstruction. If the reference frames are out of the T2V model's pre-trained generalization capacity, the spatial customization might shift its output domains during training and the subsequent temporal layers will be unable to parse the altered feature maps properly in the next stage. We suggest smaller learning rate and LoRA scale to pick the checkpoint when the reference video has complex appearances such as uncommon contents or extraordinary styles. Applying our methods on advanced base T2V models with leading capabilities also helps.

Text Encoding Conflict. While extensive spatial customization modules can be alternatively utilized as our appearance absorbers, some of them might encounter text mapping conflict when collaborating with the temporal customization modules. For example, we choose not to apply LoRA on the text encoder in T2V diffusion models although it can enhance the spatial modelling and appearance decomposition. Modifications on the pre-trained text encoder could tamper the original mapping from text to its embedding, and then T-LoRA will learn the motion associated with the altered text tokens. Finally it might not be triggered properly by the vanilla text encoder without the appearance absorbers during inference. The null-text prompt training trick for LoRA without triggering words might help to handle this issue.

E Future Work

Abundant image customization approaches with various tuning techniques have been developed for T2I diffusion models. We leverage some of them to serve as our appearance absorbers for their training stability on few-shot learning and inference simplicity in the staged scheme. In the next step we plan to investigate more options to discover their characteristics and further enhance our method's performance and usability. Besides, generative video foundation models are also rapidly evolving and our modules are inherently compatible with various types

of temporal attentions, regardless of the specific generation process and input modalities.

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
2. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23040–23050 (2023)
3. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
4. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=Fx2SbBgcte
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
6. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv preprint arXiv:2303.04761 (2023)
7. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023)
8. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
9. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: 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. pp. 7623–7633 (2023)
10. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023)