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

Anonymous Authors

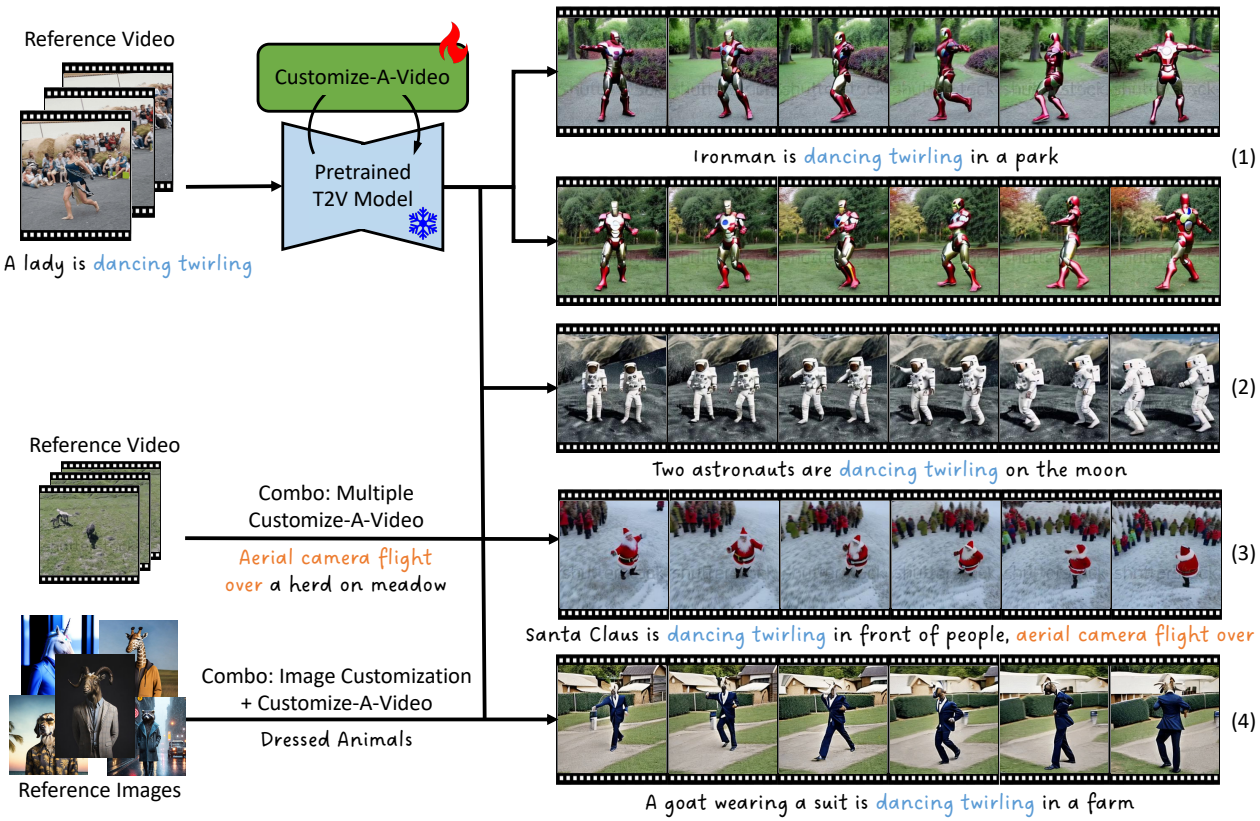

Figure 1. Customize-A-Video takes as input a single reference video (top left) and transfers its motion onto new generated videos with plausible variance. **(1-2)** Results of transferring the dancing twirling from the lady onto Ironman with two random output variants. **(3)** Results of transferring the motion onto multiple subjects. **(4)** Results of combining multiple motion customization together, i.e., both *dancing twirling* and with *aerial camera flight over*. **(5)** Results of combining proposed motion customization and existing image customization methods to support both appearance and motion customization.

Abstract

Image customization has been extensively studied in text-to-image (T2I) diffusion models, leading to impressive outcomes and applications. With the emergence of text-to-video (T2V) diffusion models, its temporal counterpart, motion customization, has not yet been well investigated. To address the challenge of one-shot motion customization, we propose Customize-A-Video that models the motion from a single reference video and adapting it to new subjects and scenes with both spatial and temporal varieties. It leverages low-rank adaptation (LoRA) on temporal attention layers to tailor the pre-trained T2V diffusion model for specific motion modeling from the reference videos. To disentangle the spatial and temporal information during the training pipeline, we introduce a novel concept of appearance absorbers that detach the original appearance from the single reference video prior to motion learning. The proposed modules are trained in a staged pipeline and inferred in a plug-and-play fashion, enabling easy extensions of our

method to various downstream tasks such as custom video generation and editing, video appearance customization and multiple motion combination. Our project page can be found at https://anonymous-314.github.io.

1. Introduction

Replicating an iconic motion in novel scenes is highly desirable for video creation. Recent large-scale diffusion-based text-to-video (T2V) generation models [5, 43] demonstrate impressive outcomes in generating imaginative videos based on text depictions. However, they struggle with precise motion control and often demand extensive prompt engineering. Another thread of work on video editing such as [4, 9, 45, 48] leverages large generative models' generalization capability on appearance alteration, and introduces frame-wise precise controls through techniques including DDIM inversion [31, 41] and ControlNet [52]. While achieving promising motion transfer results with variations in appearance and texture, these methods rigidly adhere to the reference frame structure and layout and fail to provide variability in the motion itself, such as adapting to new positions, intensities, camera views, or quantity of subjects.

Image customization of T2I models has been explored [10, 38] where a specific unique appearance is modeled and composed into novel roles and scenes. These models are trained on a small set of images that share the same concept. They are then able to produce specific output needless of complex prompt engineering, while also allowing for diversity in poses, views, lighting, etc. compared to direct stitching and editing approaches. Inspired by this idea, we present a novel video motion customization method named *Customize-A-Video* based on T2V diffusion model. It customizes the model using the motion learned from the reference video, enabling it to be easily adjusted to new subjects and scenes. This includes not only precise transfer but also variations in motion intensities, positions, quantity of subjects, and camera views. These variations make the output videos more dynamic and engaging, as opposed to the robotic or unnatural appearance of per-frame replication.

Specifically, we utilize a common image customization technique, Low-Rank Adaptation (LoRA) [19], applied on the pre-trained T2V diffusion model [43] to capture the motion signature in the reference video. However, applying LoRA directly to the entire T2V models proves less effective in motion preservation, as spatial and temporal characteristics are intricately entangled and both spatial and temporal information are learned simultaneously. Therefore, we apply LoRA only on temporal 3D cross-frame attention layers, creating *Temporal LoRA (T-LoRA)*, which is more effective in capturing temporal motion dynamics from the reference video. In comparison to other popular customization techniques, LoRA also offers a portable model size and requires minimal training data, as well as the simplicity of plug-and-play for easy extensibility to collaborate with additional customization modules.

While LoRA works well on few-shot customization tasks through the residual module weights, a portion of spatial features still leak into it during the training on a single reference video without diverse appearances accompanying a consistent motion concept. Concurrent efforts attempt to address this challenging yet significant issue by either concentrating on video customization demanding a small dataset with the same motion [30, 46] or stopping training early and supplementing the underfitted temporal modules with direct control signals from the reference video [20]. To tackle this issue and facilitate one-shot video customization, we introduce an innovative training approach based on a new *Appearance Absorber* module to further decompose spatial information from motions. The key idea of this module is that it can be trained to *absorb* the appearance of the reference video, leaving only the motion information for the proposed Temporal LoRA to learn.

We introduce a three-stage training and inference pipeline as illustrated in Fig. 2 to connect all the components we have proposed. In the first stage, we load and train the appearance absorber on unordered reference video frames to capture frame-wise spatial information, such as the subject's appearance and the background scene. In the second stage, we load the trained appearance absorber with frozen parameters, and train the proposed Temporal LoRA on the temporal layers of the T2V model. The appearance absorber has learned to reconstruct the static frames and therefore helps the Temporal LoRA focus primarily on temporal signals, minimizing the spatial information leaked into motion customization modules. During the final inference stage, we remove the appearance absorber that encodes the appearances from the reference video, and load solely the trained Temporal LoRA. Given a text prompt containing novel subjects and scenes, our model not only accurately transfers the learned motion signature to the new appearance, but also produces diverse motions in terms of their intensities, positions, and camera views, enhancing the dynamism and engagement of the generated videos for the motion customization task.

To summarize, our contributions involve:

- We present a novel one-shot motion customization method for single reference video based on pre-trained text-to-video diffusion models;
- We introduce *Temporal LoRA* module to learn the specific motion from a reference video, facilitating motion transfer with not only accuracy but also variety;
- We propose *Appearance Absorber* module to dedicatedly decompose the spatial information, effectively excluding it from the motion customization process;
- Our modules feature the plug-and-play and staged nature

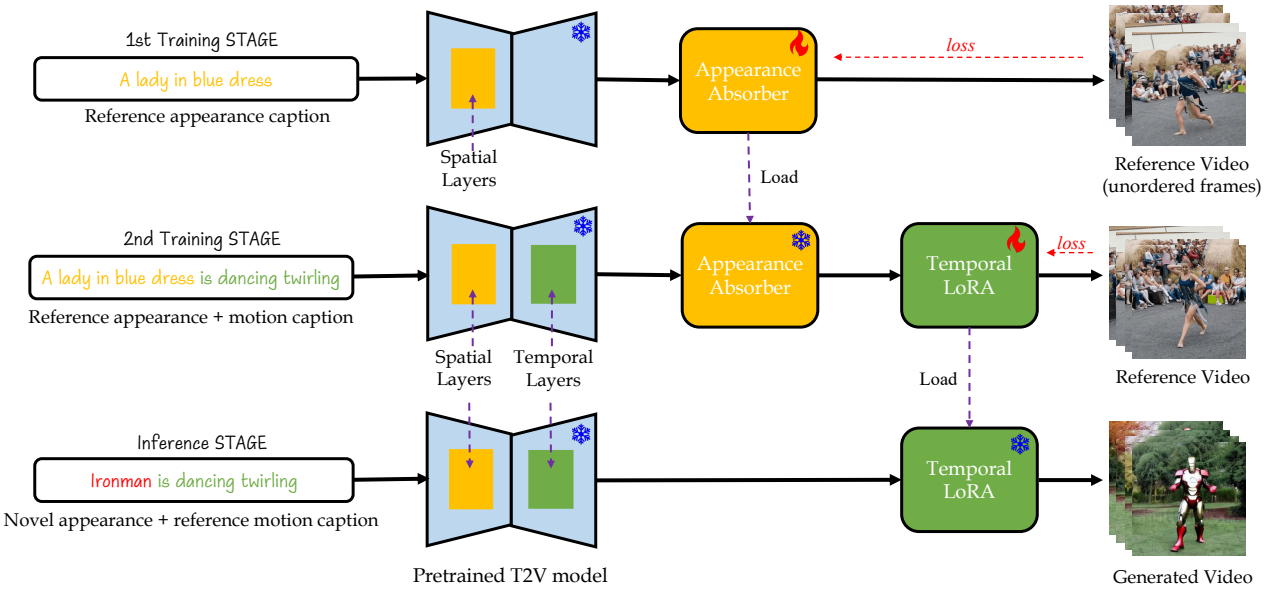

Figure 2. Our proposed Temporal LoRA and its training and inference processes. All noise and denoising schedules are omitted for simplicity. **(1)** We bypass all temporal layers in a base T2V diffusion model and apply appearance absorber such as S-LoRA or Textual Inversion on its spatial attention layers. The module is trained on unordered video frames. **(2)** We apply T-LoRA on all temporal attentions in the full base T2V model. The trained appearance absorber is also loaded and frozen. The module is trained on the target video data. **(3)** During inference, only the trained T-LoRA is loaded. A new video with the customized motion is generated by a prompt describing the new appearance and the desired motion.

and can be smoothly extended to various downstream applications.

2. Related Work

Text-to-Video Generation Models Text-to-video (T2V) generation task generates videos from given text prompts specifying the expected appearances and motions. It has been widely explored previously using GANs [22, 24, 34] and transformers [11, 18, 42, 47, 50]. With the emergency of T2I diffusion models, T2V diffusion models become subsequently under fast development. [21] reprogram the 2D spatial attentions into 3D temporal attentions to handle the new temporal dimension. [3, 6, 14, 16, 17, 26, 40, 43, 59] insert spatio-temporal 3D convolutions and/or cross-frame attentions to regulate the output temporal consistency from the random input noise. [12, 21, 26, 28] design explicitly disentangled noise prior between key frames and residues to enforce temporal coherency. T2V models designate the generated content through text prompts, demanding significant engineering effort to prompt it to produce desired motions in details.

T2I-based Video Editing Leveraging the control signal directly from a reference video by editing it into new appearances is a popular practice to transfer the motion and has been studied by various methods. [13, 27, 36, 39, 44, 48, 55] leverage the inverse denoising process or degraded images of the reference video frames to maintain the desired motion while altering its appearance through T2I generation. [1, 8, 9, 25, 29, 45, 54] adopt controllable image generation approaches [33, 52] and extract the low-level reference signals such as their depth or edge maps to guide the generation process. [4, 7, 51, 56, 58] make use of the combination of above techniques. However, such methods fall short as they focus more on adopting novel appearances, and merely duplicate the original motion exactly but with no temporal diversity to vary in the frame structure, motion intensity, subject position and quantity etc.

Video Motion Customization Customizing a diffusion model is the task defined as adapting the original output to a new specific domain by adjusting the pre-trained model weights. It was first introduced for T2I models to personalize in spatial aspects such as identity, art style and pose etc [10, 23, 38]. Recently, the idea of customizing the motion given reference videos has also been emerging and evolving rapidly. [27, 39, 48, 55] add temporal attentions from scratch on pre-trained T2I models and finetune them on a single video. Concurrent work [20, 30] finetunes the temporal layers in place in a pre-trained T2V model. [30] trains with special tokens following [38] on multiple videos with an extra regularization set. [20] replaces the vanilla re-

construction objective with a frame residual vector loss. Instead, our method appends residual weights to the original model based on LoRA [19] and enables advanced training strategy and inference utility.

[32] represents the first attempt to finetune the spatial and temporal attentions respectively for the appearance and motion of a reference video, which are however not inferred with independently. Concurrent work [46] adds specially designed adapters over the pre-trained temporal attentions conditioned on one frame to decompose pure motion from its appearance. It requires this additional image input while ours has the minimal input requirement of text prompt only. Another concurrent work [57] applies dual-path LoRAs on spatial and temporal attentions and trains them jointly with an appearance-debiased temporal debiased loss. In contrast, our approach adopts a staged training pipeline, where our appearance absorbers can be easily extended to more candidate categories than LoRA, or third-party modules pre-trained on external images or other videos.

3. Method

We present a novel motion customization method based on base T2V diffusion models for a single input video. We suggest learning the motion concept from the input video through a LoRA module designed for temporal layers of the T2V model. Given the challenging nature of working with a single input video, we have developed an innovative training strategy based on an appearance absorber module to disentangle spatial information from motion. An illustration of each proposed module and its connection to the existing T2V model is shown in Fig. 2.

3.1. Preliminary

Text-to-Video Diffusion Models. A text-to-video (T2V) diffusion model trains a 3D UNet ϵ_θ to generate videos in a series denoising steps conditioned on a input text prompt. The 3D UNet usually consists of spatial self- and cross-attentions, 2D and 3D convolutions, and temporal cross-frame attentions. Given the F frames $x^{1\cdots F}$ of a video, the 3D UNet is trained by

$$L_\theta = \mathbb{E}_{x^{1\cdots F},\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t^{1\cdots F}, t, \tau(y))\|] \quad (1)$$

at every denoising step $t = T, ..., 0$, where $\epsilon \sim \mathcal{N}(0, 1)$ is Gaussian noise, τ is the text encoder and y is the input text prompt.

Low-Rank Adaptation. Low-Rank Adaptation (LoRA) [19] was proposed for adapting pre-trained large language models to downstream tasks. It has also been widely developed for image customization models. LoRA applies a residue path of two low-rank matrices $B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}$ in attention layers, whose original weight is $W_0 \in \mathbb{R}^{d\times k}$, $r \ll \min(d, k)$. The new forward path is

$$W = W_0 + \alpha\Delta W = W_0 + \alpha BA, \quad (2)$$

where α is a coefficient adjusting the strength of the added LoRA.

3.2. Customize-A-Video

3.2.1 Temporal LoRA

Inspired by [19], we introduce Temporal LoRA (T-LoRA), a technique for capturing motion characteristics from input videos and enabling motion customization for new appearance via text prompts. We apply LoRAs on all temporal cross-frame attention layers of the base T2V model [43] to maximize modeling motion signals. Our ablation studies reveal that T-LoRA outperforms applying LoRA to other non-temporal attention layers, as T-LoRA targets at preserving motion while discarding unnecessary input appearance (see Sec. 4.2).

3.2.2 Appearance Absorbers

We enhance the appearance absorber to separate spatial signals from temporal signals within a single video. Its objective is to absorb the spatial information, including the identity, texture, scene, etc., out of the training video, in order that the reference motion can be exclusively modeled by our T-LoRA. To achieve this, we construct the absorber using a set of image customization modules including:

Spatial LoRA: We apply LoRA on only the spatial attention layers in a T2V model to adopt solely the spatial information out of the unordered video frames. LoRA modules are injected in all self-attention layers of the frames and cross-attention layers between frames and the text prompt. We call it spatial LoRA or S-LoRA to distinguish from our T-LoRA for temporal modeling.

Textual Inversion: We utilize textual inversion [10] as an alternative approach to gather spatial features from the training video. It creates learnable placeholder tokens, initialized with briefly depicting words of the video appearance, to assimilate relevant spatial information via the pre-trained text tokenizer.

These image customization modules are adept at modeling appearance signals from limited number of frames of a single video in a few-shot manner, and thus we prefer less finetuning-based customization methods such as [38] since they require a considerable amount of training and regularization data. In addition, all types of appearance absorbers can be employed individually or jointly.

3.2.3 Training and Inference Pipelines

Our motion customization pipeline consists of two training stages for appearance absorbers and T-LoRA respectively,

Figure 3. Results of one-shot motion customization. **Middle-left:** ModelScope is the base T2V model which controls motions via text and fails to transfer the reference motion faithfully. **Top-right:** Tune-A-Video relies on DDIM inverted latent as the input and thus mostly duplicates the original frame image structure. **Middle-right:** Video-P2P upgrades Tune-A-Video with decouple-guidance cross-attention control but still yields deterministic output. **Bottom:** Our methods generates motion with both accuracy as well as variety in details such as view perspective and layout. We show two variants of our generated results on bottom left and right.

and one final inference stage to generate output videos with novel text prompts.

First training stage We train appearance absorber modules first. Since they are originated from T2I models, we propose to specially train them by bypassing all temporal layers in the T2V model, including temporal attention layers and 3D convolution layers in the denoising UNet. We train them with the appearance description of the ground truth caption so that they focuses on the spatial information to learn. The input images for training are composed by unordered frames from the reference video. We follow their native loss as in [10, 19] to train each type of appearance absorber. Formally, for S-LoRA

$$L_{\Delta\theta_s} = \mathbb{E}_{x,\epsilon,t}[\|\epsilon - \epsilon_{\theta_0+\Delta\theta_s}(x_t^f, t, \tau(y))\|], \tag{3}$$

and for textual inversion

$$L_\tau = \mathbb{E}_{x,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t^f, t, \tau(y))\|]. \tag{4}$$

Second training stage We inject previously trained appearance absorbers into the T2V model and maintain their frozen state. Our T-LoRA is meanwhile injected into the temporal attention layers of the T2V model. It is trained with the given reference video and full ground truth caption consisting of both motion verbs and appearance nouns, by which the appearance absorber is also triggered to yield spatially customized content in static frames. We use the standard reconstruction loss as employed in the diffusion model [37]:

$$L_{\Delta\theta_t} = \mathbb{E}_{x^{1\cdots F},\epsilon,t}[\|\epsilon - (\epsilon_{\theta_0+\Delta\theta_t}(x_t^{1\cdots F}, t, \tau(y)))\|]. \quad (5)$$

Inference stage During the final inference stage, solely the trained T-LoRA is loaded onto the base T2V model. Given a new text prompt depicting the learned motion with new appearance, the customized model generates videos animated by the desired motion following the standard denoising process. As a result of the customized residual weights in T-LoRA, our output video transfers the reference motion faithfully as well as with diversity in motion intensities, positions, and camera views etc.

4. Experiments

Base T2V models Our methods are applicable to general T2V diffusion models. In the following experiments, we employ the ModelScope T2V model [43] as the pre-trained base model. All videos are pre-processed and generated for 2 seconds, 8 FPS and 256×256 resolution. Other training hyperparameters and model size statistics can be found in the appendix.

Datasets Since there is no dataset exactly following our single video motion customization setup, we select videos from multiple sources, including LOVEU-TGVE-2023 [49], WebVid-10M [2] and DAVIS [35] datasets to evaluate our method. We also apply our method on in-the-wild videos and demonstrate its generalization ability.

Comparison methods To our best knowledge there is no previous approach that performs the identical task of one-shot video motion customization with motion variety as we proposed. We pick the most relevant work, Tune-A-Video [48] and Video-P2P [27], which appends raw temporal layers to pre-trained T2I models and finetune both the spatial and temporal attentions on a single reference video prior to subsequent editing. It is worth noting that they additionally rely on DDIM inverted reference video latent as the input during inference and thus only produce temporally deterministic videos with frame layout and fixed view angle. When we remove this additional input they recover some motion generalization while sacrificing temporal coherency.

Table 1. Quantitative comparisons on [49] dataset. *Text. Align.* represents the text alignment, *Temp. Consist.* represents the temporal consistency and *Div.* represents the comprehensive diversity of appearance and motion. $\sim$ *w/o DDIM Inv* represents the above method not inputting DDIM inverted latent of the reference video at inference time. Note that Video-P2P outputs video clips of 4 FPS with 512×512 resolution.

Method	Text Align.↑	Temp. Consist.↓	Div.↑
ModelScope [43]	<u>31.484</u>	0.175	**0.647**
Tune-A-Video [48]	31.141	0.180	-
$\sim$ w/o DDIM Inv	30.304	0.206	0.348
Video-P2P [27]	31.001	**0.162**	-
$\sim$ w/o DDIM Inv	30.876	0.251	0.469
Ours No AA	31.348	0.178	0.604
Ours S-LoRA AA	30.816	<u>0.164</u>	0.612
Ours TextInv AA	**32.194**	0.167	0.623
Ours Both AA	31.449	0.171	<u>0.631</u>

Besides, we compare our method against the pre-trained T2V model, i.e. ModelScope, to prove that our method enhances the base model to produce faithful motions following the reference video that are not trivial to depict via simply prompt engineering.

Quantitative metrics We measure the performance quantitatively over a subset of videos from [49] which has standard 2-second clips with both ground truth and modification captions. We consider comparisons in terms of three metrics: **Text alignment** between the generated video frames and the inference prompt, in the form of CLIPScore [15]; **Temporal consistency** between consecutive frames of the generated video, in the form of LPIPS [53]; **Diversity** among multiple generated videos with the same prompt and different random noise, in the form of LPIPS. It involves both the appearance and motion diversity by collating aligned frames at the same timestamp.

4.1. Motion Customization from Single Video

Qualitative Results Fig. 3 illustrates the comparative visual results of one-shot motion customization. The base ModelScope T2V model, although proficient in inferring general motion concepts from its extensive pre-training on large-scale datasets, fails to accurately replicate specific motions guided by reference videos. Contrastingly, Tune-A-Video [48] and Video-P2P [27] leverage DDIM inverted latents extracted from reference videos, resulting in temporally deterministic output with structural constraints by the reference frame layouts. In contrast to both of them, our approach demonstrates the capability to transfer the modeled

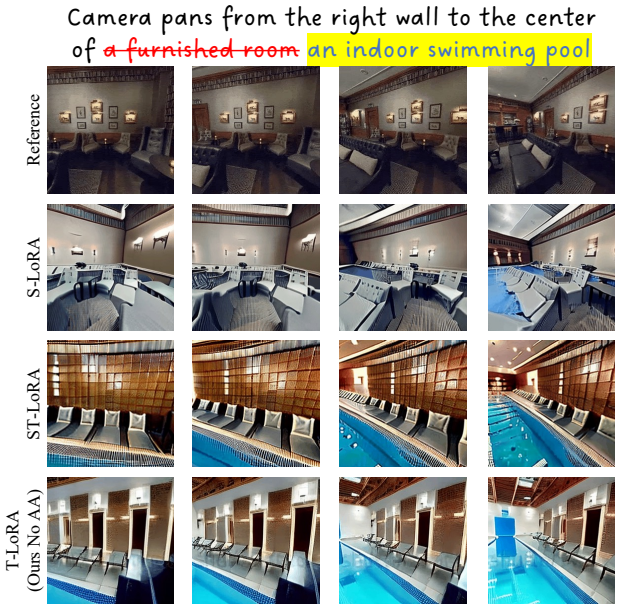

Figure 4. Ablation results on applying LoRAs on different attention layers. When transferring the camera movement, S-LoRA memorizes the indoor furniture and wall decorations and T-LoRA converts paintings to entrances and sofas to pool benches.

Figure 5. Ablation results on training T-LoRAs with different types of appearance absorbers. No AA remains the bare hands and white sleeves from the original appearance. Both AA reaches the best spatial clearance and displays tidy wall and desk in addition to the evident new dress-up.

reference motion to new scenarios and subjects while introducing temporal variations through random noise input. The last row of each example presents two output samples generated by our method, prompted by the same textual cue. Our outcomes not only exhibit diverse subject appearances and background scenes but also showcase variability in motion attributes such as action range, intensity, velocity, and camera perspective.

Quantitative Results. The quantitative results and comparisons are listed in Tab. 1. In need of universally applicable motion similarity metric encompassing not only human actions but also non-human object and camera movements, text alignment serves to gauge motion accuracy relative to the provided text prompt and inadequately reflects the nature of customized iconic reference motion transfer. Video-P2P leads on temporal consistency with DDIM inverted latent input, but deteriorates considerably in its absence. Ours exhibits comparable coherency against these deterministic methods and the foundation T2V model.

Due to the lack of generic motion representations irrespective of visual appearance, we calculate the per-frame diversity among generated videos that registers both spatial and temporal variety. Comparative evaluation unveils that base ModelScope provides the highest diversity while our methods sacrifice it subtly to gain significant improvements in accurately customizing the exemplar motion, as well as retaining rich varieties in motion details. Tune-A-Video and Video-P2P have deterministic results with DDIM inverted input and yield more yet not as much diverse output without it.

4.2. Ablations Studies

LoRAs on Non-temporal Attentions While it is intuitive to apply LoRA on only temporal attention layers to learn video motions without original appearance, we also validate the effects of applying LoRA on the spatial attentions only (*S-LoRA*), or on both spatial and temporal attentions (*ST-LoRA*) in the base T2V model. Fig. 4 displays the visualization that adding LoRA to spatial attentions significantly impairs the modeling of motions. The models with spatial customization modules primarily memorize the video based on spatial layout, resulting in a substantial degradation of both video appearance and motion customization compared to our T-LoRA approach.

Comparisons across Appearance Absorbers Our method explores four different configurations of Appearance Absorbers (AA). **No AA**: no appearance absorber

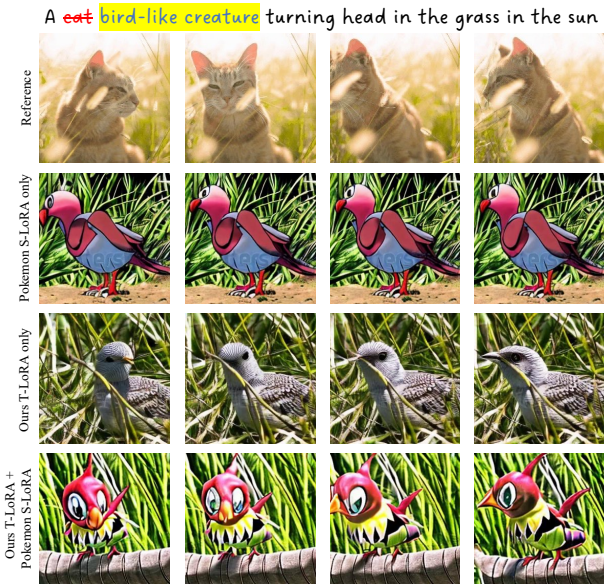

Figure 6. Video Appearance Customization. (1) Reference video. (2) Result of image customization S-LoRA only fails to transfer the motion. (3) Result of motion customization. (4) Result combined both motion and appearance customizations.

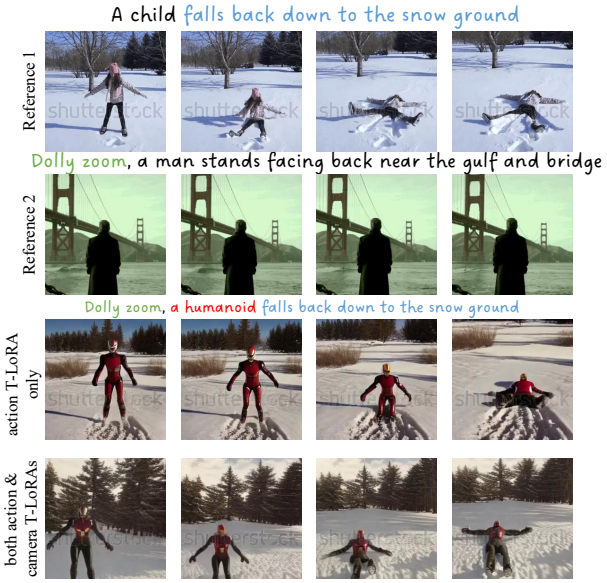

Figure 7. Multiple Motion Combination. (1-2) two reference videos. (3) Result only customized from the first reference video, only displaying a normal zoom in instead of the *dolly zoom*, where the foreground remain fixed in scale and the background zoom in or out. (4) Result with two T-LoRAs trained from both reference videos, yielding videos with merged motions. The scale of the humanoid barely changes through the frames while the trees fast zooms out.

is used, **S-LoRA AA**: a spatial LoRA based appearance absorber is used, **TextInv AA**: a textual inversion based appearance absorber is used, and **Both AA**: two appearance absorbers of both above types are used. The comparison results are shown in Fig. 5. No AA remains some original appearance in addition to modeling the motion. S-LoRA AA and TextInv AA are both able to capture the pure action with minimal appearance leakage. We notice that S-LoRA AA is easier to overfit and sometimes causes spatial artifacts while TextInv AA might tend to underfit and leave spatial residues on the other hand. We attribute these properties to the spatial structure of S-LoRA weights inside U-Net blocks while textual inversion works via a 1D learnable embedding as new tokens. Both AA unites their advantages and leads to a comprehensive result with both the reference motion and new appearance clearly reflected.

5. Applications

With the plug-and-play nature of LoRA and our staged training pipeline, we present three downstream applications that demonstrate the collaborative potential of our proposed modules.

5.1. Video Appearance Customization

Our motion customization modules work on the temporal layers and thus can cooperate with image customization methods to manipulate both the temporal and spatial layers in the base T2V model at the same time. In Fig. 6, we inject a Temporal LoRA to present the reference action as well as an image spatial LoRA to reflect the comic style in one comprehensive output. Intermediate results with one LoRA loaded has merely either the spatial or temporal aspect customized on the other hand.

5.2. Multiple Motion Combination

T-LoRA is applied to the original layers with residual connection. Therefore we can customize the base model with multiple T-LoRA modules trained on different reference videos to integrate various movements into one outcome. Fig. 7 demonstrates that our approach can merge the human action of ”falling back” and camera movement of ”dolly zoom” into one target scenario using two T-LoRA modules, achieving combined motions in the generated video.

5.3. Third-Party Appearance Absorbers

Our staged training pipeline enables loading third-party image customization modules pre-trained on in-the-wild image data as a ready appearance absorber when they share the similar appearance. This skips the first training stage and facilitates reusing modules across videos. In Fig. 8, the reference video focuses on a certain fictional character, and we load a public image LoRA of this character shared in

Figure 8. Third-Party Appearance Absorbers. (1) Reference video, whose subject is a popular fictional character. (2) Result trained with no AA. (3) Result trained with a third-party S-LoRA that is pre-trained on in-the-wild images of the character rather than the frames of the reference video.

open-source communities without tuning on the unordered video frames. Trained with its enhancement our T-LoRA avoids the leakage of the original headgear to the replaced subject.

6. Conclusion

We introduce the one-shot motion customization task that learns the motion signature from a single reference video and transfer it to new scenes and subjects with variety in both appearance and motion. We propose Temporal LoRA to model the target motion by adding LoRA residual weights on the temporal attention layers of a pre-trained text-to-video diffusion model. We further propose Appearance Absorbers to decouple the spatial information from the reference video so that Temporal LoRA can focus on motion modeling. Extensive experiments demonstrate that our methods yield faithful and diverse videos compared to both per-frame video editing approaches and the base T2V model. Our method is plug-and-play and supports various downstream tasks including precise video editing, video appearance customization and multiple motion combination.

References

[1] AILab-CVC. Ailab-cvc/videocrafter at 1f46314b6609712eea89b67f41d612557eec5b8e. 3

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 6

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3, 1

[4] 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. 2, 3

[5] 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. 2

[6] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3

[7] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 3

[8] Ernie Chu, Shuo-Yen Lin, and Jun-Cheng Chen. Video controlnet: Towards temporally consistent synthetic-to-real video translation using conditional image diffusion models. *arXiv preprint arXiv:2305.19193*, 2023. 3

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2, 3

[10] 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. 2, 3, 4, 5, 1

[11] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 3

[12] 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. 3

[13] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3

[14] 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. 3, 1

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[16] J Ho, T Salimans, A Gritsenko, W Chan, M Norouzi, and DJ Fleet. Video diffusion models. arxiv 2022. *arXiv preprint arXiv:2204.03458*. 3

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[18] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[19] 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, 4, 5, 1

[20] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. *arXiv preprint arXiv:2312.00845*, 2023. 2, 3

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 3

[22] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020. 3

[23] 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. 3

[24] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[25] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 3

[26] Binhui Liu, Xin Liu, Anbo Dai, Zhiyong Zeng, Zhen Cui, and Jian Yang. Dual-stream diffusion net for text-to-video generation. *arXiv preprint arXiv:2308.08316*, 2023. 3

[27] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 3, 6

[28] 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. 3

[29] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 3

[30] Joanna Materzyńska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Customizing motion in text-to-video diffusion models. *arXiv preprint arXiv:2312.04966*, 2023. 2, 3

[31] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2

[32] 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. 4

[33] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[34] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 3

[35] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 6

[36] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 3

[37] 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. 6

[38] 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. 2, 3, 4

[39] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023. 3

[40] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

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

[42] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3

[43] 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, 3, 4, 6, 1

[44] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3

[45] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2, 3

[46] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. *arXiv preprint arXiv:2312.04433*, 2023. 2, 4

[47] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022. 3

[48] 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. 2, 3, 6, 1

[49] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, et al. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*, 2023. 6

[50] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[51] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 3

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[54] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[55] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. Towards consistent video editing with text-to-image diffusion models. *arXiv preprint arXiv:2305.17431*, 2023. 3

[56] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023. 3

[57] 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. 4

[58] Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850*, 2023. 3

[59] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3

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

Supplementary Material

A. Implementation Details

A.1. Bypassing Temporal Layers in T2V Models

Many diffusion-based T2V models [3, 14, 43] have their denoising network structure adapted from T2I U-Net with injecting temporal convolution and attention layers. 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. Model Hyperparameters

LoRA [19] 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 $r_T = 1$ 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 [10] as the appearance absorber, we set the length of new learnable tokens to 3 and initialize them with 3 words depicting the static content in the frames. Under these settings, our T-LoRA model has merely 0.83M float parameters, compared to the full base T2V model's 552M dedicated temporal parameters [43].

B. More Results

B.1. Generation Results

More visual results generated by our models are displayed in Fig. 10. We present two random output samples for each reference video.

B.2. DDIM Inverted Latent Input

Our method can easily incorporate additional deterministic control signals to perform precise video editing. The comparison results to Tuna-A-Video[48] are shown in Fig. 9. Our models prove to be also able to benefit from the DDIM inverted latent of the reference video and yield output stuck to the exact original frame structures.

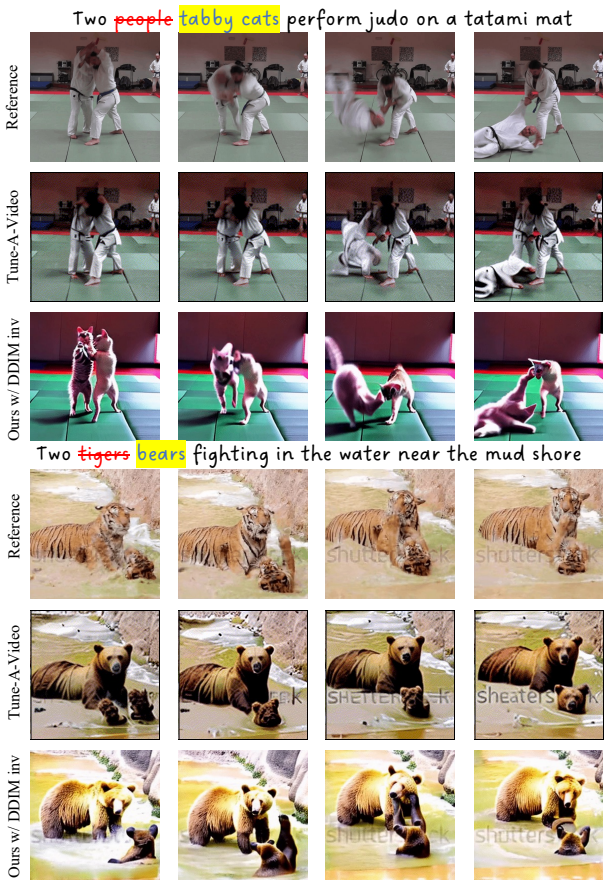

Figure 9. Comparison between our methods with DDIM inverted reference latent input and Tune-A-Video. Our models are also able to produce precise frame-wise editing output.

C. Comparison with Concurrent Work

D. Limitations and Future Work

D.1. Spatial Domain Shift

In the first training stage the standalone finetuning of partial layers might have the risk of breaking the consistency among the pre-trained weights if the appearance absorbers are overfitted on static content reconstruction. If the reference frames are out of the pre-trained generalization capacity, the spatial customization might shift the output domain

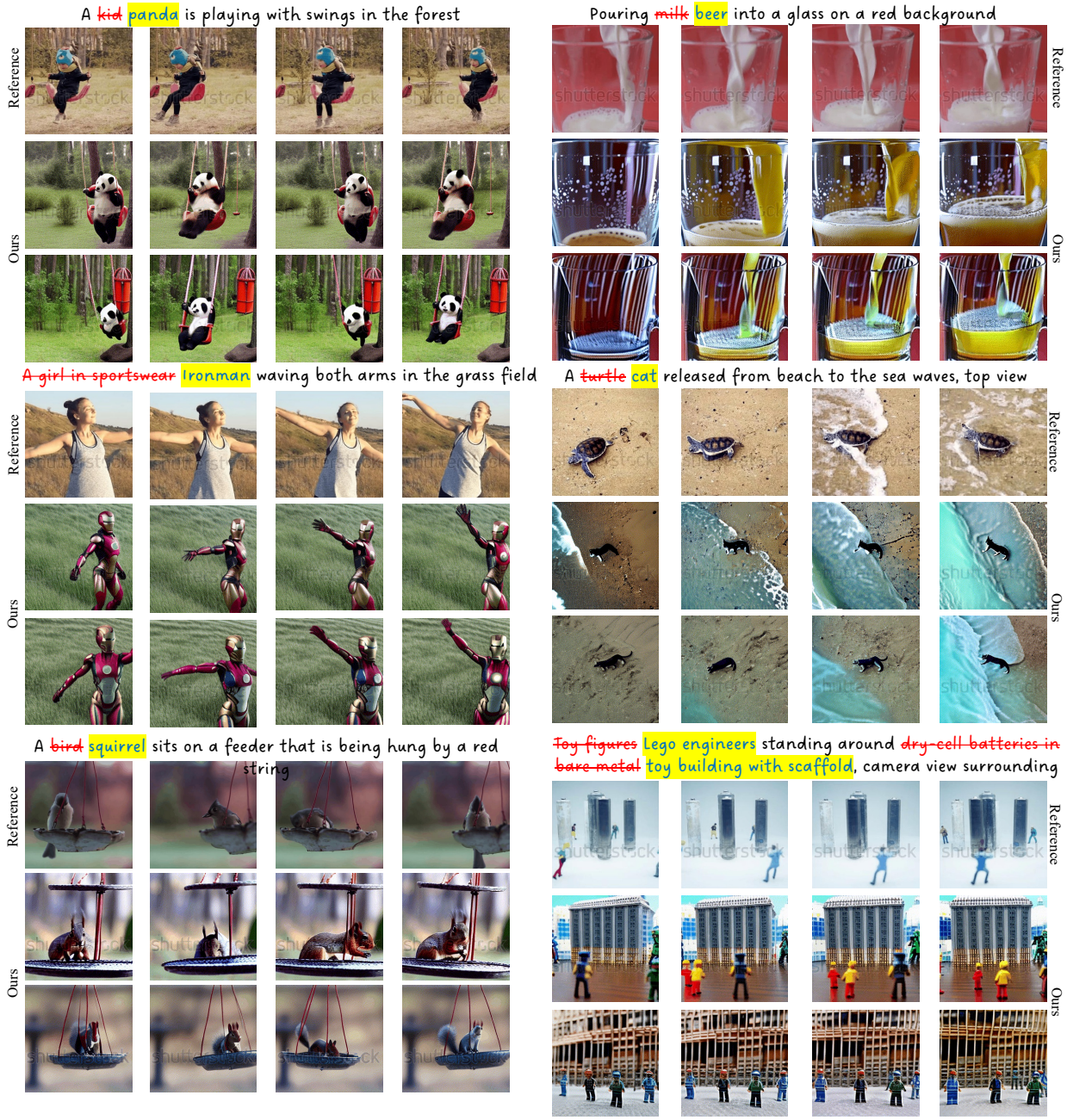

Figure 10. Additional generation results of our model.

and the subsequent temporal layers will not be able to parse the altered feature maps properly.

D.2. Small Actions and Poses

Some reference motions are intrinsically highly associated with poses, such as walking, running and sitting, and an image can primarily represent them. When the appearance absorbers have modeled the static postures to fit the appearance, in the second training stage the T-LoRA might have little left to learn such as the trivial perturbation across frames. Our model could fail to restore the desired motions evidently under such circumstance.

D.3. Future Plans

Abundant spatial customization approaches 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, video diffusion models are also rapidly evolving and our modules are inherently compatible with various types of temporal attentions regardless of the input modalities including image-to-video and video-to-video tasks.