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

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 [2]Adobe Research [3]Yonsei University

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 video motion customization, we propose Customize-A-Video that models the motion from a single reference video and adapts 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. To disentangle the spatial and temporal information during training, we introduce a novel concept of appearance absorbers that detach the original appearance from the 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 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://customize-a-video.github.io.

Keywords: Video Motion Customization · Text-to-Video Diffusion Models · Low-Rank Adaptation

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 [5,45] 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 [4,9,47,50] leverages large image generative models for appearance alteration, and introduces frame-wise precise controls via DDIM inversion [32,42] or ControlNet [54]. 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 new positions, intensities, camera views, or quantity of subjects.

Image customization of T2I models has been widely explored [10,39] where a specific unique appearance is modeled and composed into novel roles and

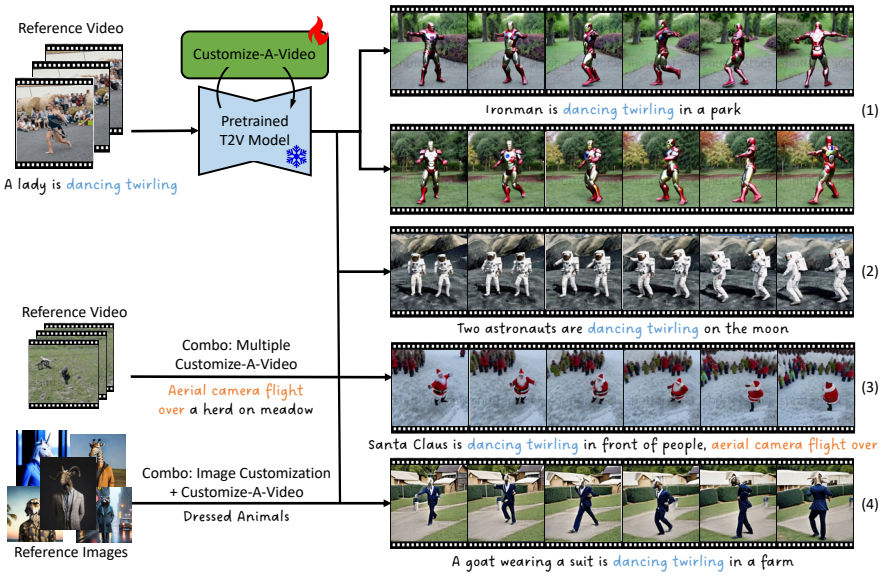

Fig. 1: Customize-A-Video takes as input a single reference video (top left) and transfers its motion onto new generated videos with plausible variance. **(1)** Transferring the dancing twirling from the lady onto Ironman with two random output variants. **(2)** Transferring the motion onto multiple subjects. **(3)** Combining multiple motion customization together, i.e., both *dancing twirling* and with *aerial camera flight over*. **(4)** Combining proposed motion customization and existing image customization methods ([21] in the example) to support both appearance and motion customization.

scenes. These modules are trained on a small set of images that share the same concept. They are then able to reproduce the desired concept 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, we introduce a new task of video motion customization and present a novel one-shot method named *Customize-A-Video* (Fig. 1) built upon T2V diffusion models. It customizes the pre-trained model with the motion learned from the reference video, enabling it to be easily adapted 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 rhythm or unnatural appearance of frame-wise tampering.

Specifically, we start from utilizing a common customization technique, Low-Rank Adaptation (LoRA) [19], applied on a pre-trained T2V diffusion model [45] to capture the motion signature in the reference video. Applying LoRA directly to the entire T2V models proves less effective in motion preservation, as spatial and temporal characteristics are intricately entangled and both will be learned simultaneously. Therefore, we apply LoRA only on temporal cross-frame atten-

tion layers, creating Temporal LoRA (T-LoRA), which is more concentrated on capturing motion dynamics from the video. In comparison to other popular customization algorithms, 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 when trained on a single reference video. Concurrent efforts attempt to address this challenging yet significant issue by either demanding a small dataset with diverse appearances and the consistent motion [31,48], or stopping training early and supplementing the underfit 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 Appearance Absorber module to further decompose static signals from dynamics. The key idea of this module is to *absorb* the appearance out of the reference video, leaving only the desired motion information for the Temporal LoRA to model.

We introduce a staged training and inference pipeline as illustrated in Fig. 2 to connect all the components we have proposed while keeping them independent. In the first stage, we build and train the appearance absorber on unordered reference video frames to capture frame-wise spatial information, such as the subject appearance and the background scene. In the second stage, we load the trained appearance absorber in frozen state, and construct the Temporal LoRA on the temporal layers of the T2V model to train. The appearance absorber has encoded the static frames and therefore helps the Temporal LoRA focus primarily on temporal signals, minimizing the spatial information leakage into motion customization modules. During the inference stage, we remove the appearance absorber 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.

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 to learn the motion from a single reference video, facilitating motion transfer with not only accuracy but also variety;
- We propose the general class of Appearance Absorbers to dedicatedly decompose the spatial information out of the reference video, effectively excluding it from the motion customization process;
- Our modules feature the plug-and-play and staged fashion 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 [23,25,35] and transformers [11,18,44,49,52]. With the boost of T2I diffusion models, T2V diffusion models become subsequently under fast development. [22] reprogram the 2D spatial attentions into 3D temporal attentions to handle the new temporal dimension. [3,6,14,16,17,27,41,45,62] insert spatio-temporal 3D convolutions and/or cross-frame attentions to regulate the output temporal consistency from the random input noise. [12,22,27,29] 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 an efficient practice to precisely transfer the motion and has been studied by various methods. [13,28,37,40,46,50,58] leverage the inverse denoising process or degradation of the reference video frames to maintain the desired motion while altering its appearance through T2I generation. [1,8,9,26,30,47,56] adopt controllable image generation approaches [34,54] and extract the low-level reference signals such as their depth or edge maps to guide the generation process. [4,7,53,59,61] 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 motion intensity and velocity, subject position and quantity etc.

Video Motion Customization. Model customization is the task of 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 or pose [10,24,39]. Recently, the idea of customizing the motion given reference videos has also been emerging and evolving rapidly. [28,40,50,58] add temporal attentions from scratch on pre-trained T2I models and finetune them on a single video. Concurrent work [20,31] tunes the temporal layers in place in a pre-trained T2V model with either a regularization set or a frame residual loss to reduce the impact of training videos' appearances. Instead, our method appends residual weights to the original model using LoRA [19] and enables ligthweight training strategy and flexible inference utility.

[33] represents the first attempt to finetune the spatial and temporal attentions independently for the appearance and motion of a reference video, which have however entangled inference. Concurrent work [57] employs two parallel UNets and tune one of them with an appearance normalization loss to disentangle the motion as well as maintain the generalization ability on new appearances. Another concurrent work [48] adds adapters conditioned on one frame to decompose pure motion from its appearance, requiring additional input of an image when inference while ours asks for the minimal input of text prompt only. Concurrent work [60] applies dual-path LoRAs and trains them jointly with an appearance-debiased loss. In contrast, our approach adopts a staged training pipeline with independent tuning configurations, where our appearance

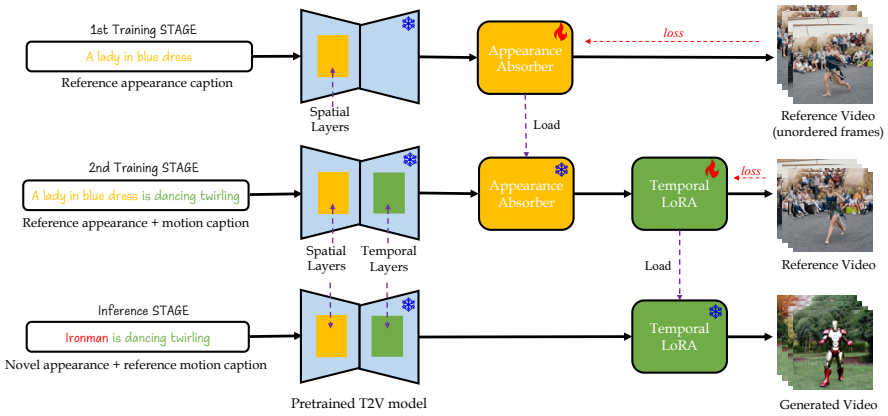

Fig. 2: Our Temporal LoRA, Appearance Absorbers and their training and inference processes. All noise and denoising schedules are omitted for simplicity. **(1)** We bypass all temporal layers in a base T2V diffusion model and apply appearance absorber such as S-LoRA or Textual Inversion on its spatial 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. **(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 target motion.

absorber class can be easily extended to more candidates than LoRA, or reusing third-party modules pre-trained on in-the-wild images or videos.

3 Method

We present a novel motion customization method based on pre-trained T2V diffusion models for a single reference video. We suggest learning the motion concept from the reference video through a LoRA module designed for temporal layers of the T2V model. Given the challenging nature of working with a single datum, we develop a staged training strategy with an appearance absorber module to disentangle spatial information from motion. Fig. 2 shows an illustration of each proposed module and its connection to the base T2V model.

3.1 Preliminary

Text-to-Video Diffusion Models. A text-to-video (T2V) diffusion model trains a 3D UNet 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 $x^{1\dots F}$ of a video, the 3D UNet θ is trained by minimizing

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

at every denoising step $t = T, ..., 0$. ϵ is Gaussian noise and ϵ_θ is the UNet prediction, τ is the text encoder with token sets v, and y is the 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 $\theta_B \in \mathbb{R}^{d\times r}, \theta_A \in \mathbb{R}^{r\times k}$ on an attention layer, whose original weight is $\theta_0 \in \mathbb{R}^{d\times k}$, $r \ll \min(d, k)$. The new forward path is

$$\theta = \theta_0 + \alpha\Delta\theta = \theta_0 + \alpha\theta_B\theta_A. \tag{2}$$

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

3.2 Customize-A-Video

We proposed two critical modules to customize the pre-trained T2V model for a single reference video. *Temporal LoRA* is introduced to learn motion from a reference video, whereas *Appearance Absorbers* are crafted to improve the separation of spatial and temporal information within the single reference video.

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 [45] 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).

Appearance Absorbers To separate spatial signals from temporal signals within a single video, we abstract the general class of Appearance Absorbers. Its objective is to absorb the spatial information, including the identity, texture, scene, etc., out of the training video, such that the reference motion can be exclusively modeled by our T-LoRA. To achieve this, we construct the absorbers leveraging a set of image customization modules, and use them in an *inverse* manner compared to their original role:

$$\text{Image Model} + \textit{Image Customization Module} \rightarrow \text{Image Custom Model};$$
$$\underbrace{\text{Video Model} - \textit{Appearance Absorber}}_{\textit{remove spatial signals from training video}} + \text{T-LoRA} \rightarrow \text{Video Custom Model}.$$

Our appearance absorbers can be built upon including but not limited to:

- *Spatial LoRA.* We apply LoRA on only the spatial attention layers in a T2V model to adopt solely the spatial information out of the 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 (S-LoRA) to distinguish from our T-LoRA for temporal modeling.

– *Textual Inversion.* We utilize textual inversion [10] as another approach to collect 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 through the 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 [39] since they require a considerable amount of training and regularization data. All types of appearance absorbers can be employed individually or jointly.

Training and Inference Pipelines Our motion customization pipeline in Fig. 2 consists of two training stages for appearance absorbers and T-LoRA respectively, and one inference stage to finally generate output videos with novel text prompts. Our configurable pipeline has dedicated stage for each module and is universal for extensive types of its components.

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 y_S cut out of the full caption so that they focus on learning the spatial information. The training images are unordered frames of the reference video. We follow their native loss as in [10, 19] to train each type of appearance absorber. Formally, for S-LoRA $\Delta\theta_S$:

$$L_{\Delta\theta_S} = \mathbb{E}_{x,\epsilon,t}[\|\epsilon - \epsilon_{\theta_0+\Delta\theta_S}(x_t^f, t, \tau_{v_0}(y_S))\|], \tag{3}$$

and for textual inversion Δv:

$$L_{\Delta v} = \mathbb{E}_{x,\epsilon,t}[\|\epsilon - \epsilon_{\theta_0}(x_t^f, t, \tau_{v_0+\Delta v}(y_S))\|]. \tag{4}$$

Second training stage. We inject above 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 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 train T-LoRA $\Delta\theta_T$ using the standard reconstruction loss as in diffusion models [38]:

$$L_{\Delta\theta_T} = \mathbb{E}_{x^{1\dots F},\epsilon,t}[\|\epsilon - (\epsilon_{\theta'+\Delta\theta_T}(x_t^{1\dots F}, t, \tau_{v'}(y)))\|] \tag{5}$$

where $\theta' = \theta_0 + \Delta\theta_S$ and $v' = v_0 + \Delta v$ if respective AAs are employed.

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 reference motion with new appearances and scenes, the customized model generates novel videos animated by the desired motion following the standard denoising process. As a result of the customized 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 hire the ModelScope T2V model [45] as the pre-trained base model. All videos are pre-processed and generated for 2 seconds, 8 FPS and 256×256 resolution. Training hyperparameters, model size statistics and time consumption analyses are detailed in the supplementary material.

Datasets. We select videos from mixed sources, including LOVEU-TGVE-2023 [51], WebVid-10M [2] and DAVIS [36] datasets to evaluate our method. [51] provides ground truth captions and target editing prompts, while we create those for videos from other sources. We also apply our method on in-the-wild videos and demonstrate its generalization ability.

Comparison methods. As of a new task of one-shot video motion customization, we mainly compare to Tune-A-Video [50] and Video-P2P [28], which append raw temporal layers to pre-trained T2I models and finetune them on a single reference video. 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 fixed frame layout and view angle, so we also evaluate their variants removing this condition. We also compare our method against the pre-trained T2V model, i.e. ModelScope, to prove that our method enhances the base foundation model to produce faithful motions following the reference video that are not trivial to depict via prompt engineering. Besides, we run the concurrent work MotionDirector [60] using their released training code with the same configuration as ours, including the same base T2V model and LoRA hyperparameters for fair comparison.

Quantitative metrics. We measure the performance quantitatively over a subset containing 53 videos out of [51] of 2-3 seconds with standard original and editing captions. We consider comparisons in terms of three metrics: **text alignment** between the generated video frames and the inference prompt gauges both generated appearance and motion accuracy, in the form of CLIPScore [15] that associates text and image in a unified space; **temporal consistency** between consecutive frames of the generated video indicates the generated motion quality, in the form of LPIPS [55] that measures deep feature distance; **diversity** among multiple generated videos with the same prompt and different random

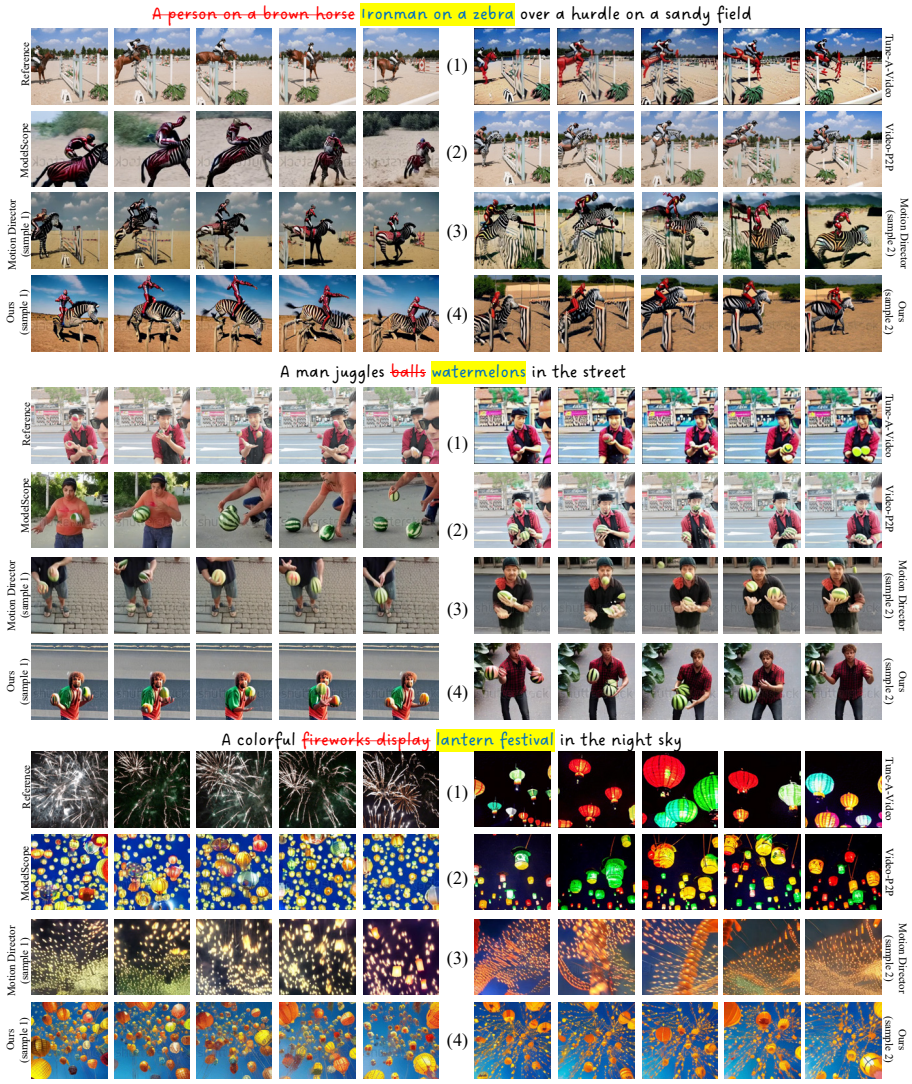

Fig. 3: Results of one-shot motion customization. **(1-left)** Reference video. **(2-left)** ModelScope [45] fails to transfer the reference motion faithfully with only text guidance. **(1-right & 2-right)** Tune-A-Video [50] and Video-P2P [28] rely on DDIM inverted latent input and duplicate the original frame structure deterministically. **(3)** Concurrent work MotionDirector [60] also generates various output following the reference motion while there exist some appearance and motion artifacts especially for hard examples with complex or intensive movements. **(4)** Our methods generate motion with both accuracy and variety in details such as view perspective and frame layout. Two variants generated with random noise are shown for MotionDirector and Ours.

Table 1: Quantitative comparisons on [51] dataset. ∼ *w/o DDIM Inversion* represents the above method without DDIM inverted latent input. Video-P2P outputs video clips of 4 FPS with 512×512 resolution. MotionDirector is a concurrent work to ours and is tested with either the same LoRA rank or comparable amount of parameters to ours.

Method	Text Alignment↑	Temporal Consistency↓	Diversity↑
ModelScope [45]	31.705	0.175	**0.636**
Tune-A-Video [50]	31.149	0.185	-
∼ w/o DDIM Inversion	30.304	0.206	0.348
Video-P2P [28]	31.001	<u>0.162</u>	-
∼ w/o DDIM Inversion	30.876	0.251	0.469
MotionDirector [60]	<u>32.500</u>	0.163	0.606
∼ w/ comparable #params	31.842	0.166	0.595
Ours No AA	31.687	0.166	0.613
Ours S-LoRA AA	31.913	0.163	0.618
Ours TextInv AA	**32.632**	**0.160**	0.621
Ours Dual AA	32.193	0.164	<u>0.631</u>

noise involves both spatial and temporal diversity by collating aligned frames at the same timestamp, in the form of LPIPS. It is calculated on 4 random samples per reference video.

User Study. We conduct a human user study among five algorithms with stochastic output: ModelScope, Tune-A-Video and Video-P2P without DDIM inverted latent, our method with both S-LoRA and textual inversion as the appearance absorbers, and MotionDirector. Every participant is presented 10 random reference videos from [51] and their output videos. Each algorithm outputs two videos, and participants are asked to assess their motion fidelity and motion diversity respectively, from 1 (worst) to 5 (best) stars. Details of the questionnaire design is provided in the supplementary materials.

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 fails to accurately replicate the specific motions as reference guided by simply the text prompt. On the other hand, Tune-A-Video [50] and Video-P2P [28] leverage DDIM inverted latents extracted from reference videos and produce 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 reference motion to new scenarios and subjects while introducing temporal variations via random noise input. 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. Concurrent MotionDirector [60] is also able to generate adapted output with variety, while for the hard examples with complex or intensive motion, such as the juggling and firework in Fig. 3, it yields less competitive visual quality than ours. Our T-LoRA is always trained with well-optimized AAs, while MotionDirector might have appearances leaked into the temporal module when their spatial path is jointly being tuned. We also notice that real-world videos usually have divergent spatial and temporal complexity, and our dedicated tuning procedures with independent steps and other schedules for each module reach their individual optimal.

Quantitative Results. The quantitative results and comparisons are listed in Tab. 1. Our methods outperform the base ModelScope [45], Tune-A-Video [50], Video-P2P [28] and concurrent MotionDirector [60] on both text alignment and temporal consistency. ModelScope [45] provides the highest comprehensive diversity as a foundation model and loses in motion fidelity with text guidance only. Our methods sacrifice it subtly to gain significant improvements in faithfully customizing the exemplar motion, as well as retaining rich varieties in motion details. Tune-A-Video [50] and Video-P2P [28] have the minimal diversity as strictly constrained by frame structures, although at the cost of which they can achieve acceptable temporal consistency. MotionDirector [60] shows comparable text alignment and temporal consistency to ours but its diversity falls behind.

User Study. We involved 20 evaluators participating our user study and collected 153 and 156 valid ratings per algorithm on each benchmark. The averaged scores are listed in Tab. 2. Our method leads on both motion fidelity and motion diversity. We asked users to rate pure motion fidelity and diversity which are hard to assess by automatic metrics in need of generic motion representations irrespective to spatial structures. These results are complementary to the text alignment and diversity measured in Tab. 1 that mix spatial and temporal signals.

Table 2: Human user study results on [51] dataset. Methods are evaluated from 1 (worst) to 5 (best) stars on each benchmark.

Method	Motion Fidelity↑	Motion Diversity↑
ModelScope [45]	2.03	2.97
Tune-A-Video [50] w/o DDIM Inversion	2.29	2.23
Video-P2P [28] w/o DDIM Inversion	2.29	2.01
MotionDirector [60]	<u>3.33</u>	<u>3.50</u>
Ours (w/ dual AA)	**3.72**	**3.72**

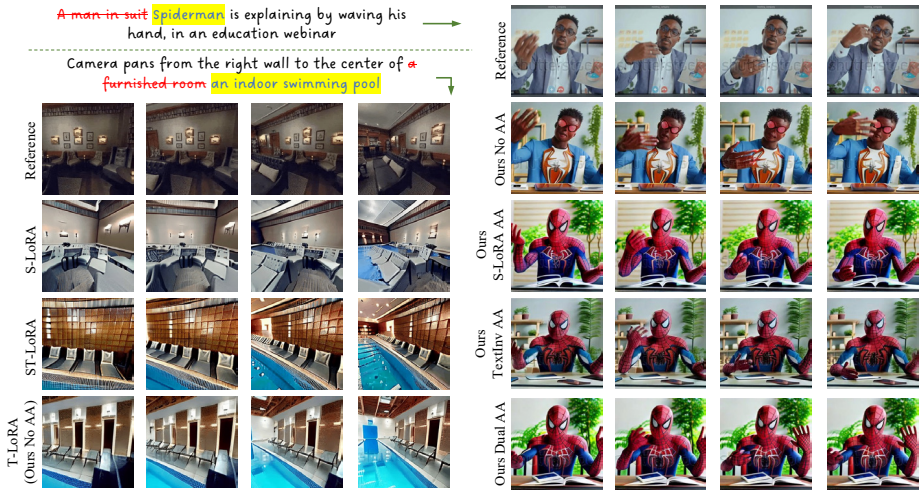

Fig. 4: Left: Ablations on applying LoRAs on different attention layers. S-LoRA memorizes the indoor furniture and wall decorations and T-LoRA converts paintings to entrances and sofas to pool benches. **Right**: Ablations on training T-LoRAs with different types of appearance absorbers. No AA adds stylish glasses and the logo but remains most the original appearance. S-LoRA AA and TextInv AA significantly boost the quality while resulting in the strips on the wall and the partially white sleeves. Dual AA reaches best spatial clearance with clear costume and background.

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 left displays the visualizations in which adding LoRA to spatial attentions significantly impairs the motion modeling. The models with spatial elements primarily memorize the video by its spatial layout, resulting in a substantial degradation of both appearance and motion adaptation.

Comparisons among Appearance Absorber Types. We explore four different configurations of Appearance Absorbers (AA). **No AA**: no appearance absorber is used, **S-LoRA AA**: a spatial LoRA based appearance absorber is used, **TextInv AA**: a textual inversion based appearance absorber is used, and **Dual AA**: two appearance absorbers of both above types are used. The comparison results are unveiled by Fig. 4. 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

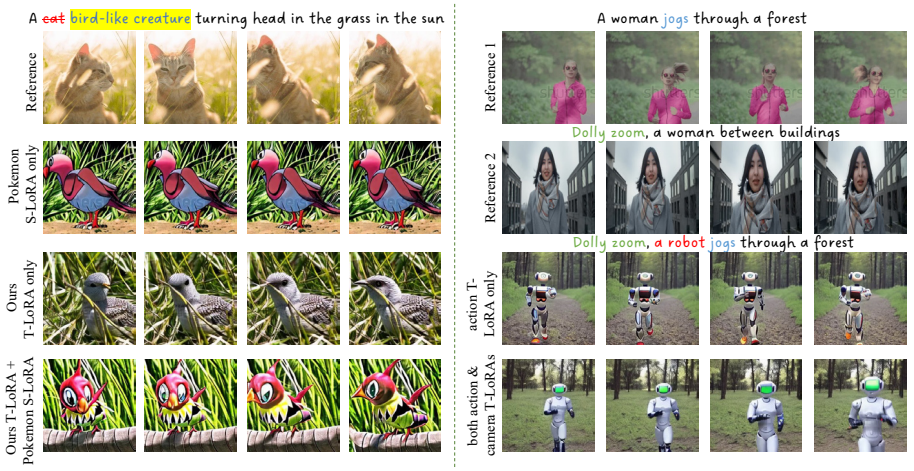

Fig. 5: Left: Video appearance customization with both T-LoRA and existing pre-trained S-LoRA (from [43]). **Right**: Multiple motion combination with two T-LoRAs loaded at the same time. When the robot is slowly jogging, it fast zooms in while the background trees rapidly zoom out (dolly zoom).

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. Dual 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 four downstream applications that demonstrate the collaborative potential of our proposed modules.

Video Appearance Customization Our motion customization module works on temporal layers and thus can cooperate with image customization approaches to manipulate both the temporal and spatial layers in the base T2V model at the same time. In Fig. 5 left, we inject a T-LoRA to present the reference action as well as an image spatial LoRA to reflect the comic style in one comprehensive output.

Multiple Motion Combination T-LoRA is applied to the original layers with residual connections. Therefore we can customize the base model with multiple T-LoRA modules trained on different source videos to integrate assorted motions into one outcome. Fig. 5 right demonstrates that our method merges the human action of jogging and camera movement of dolly zoom into one target scenario using two T-LoRA modules.

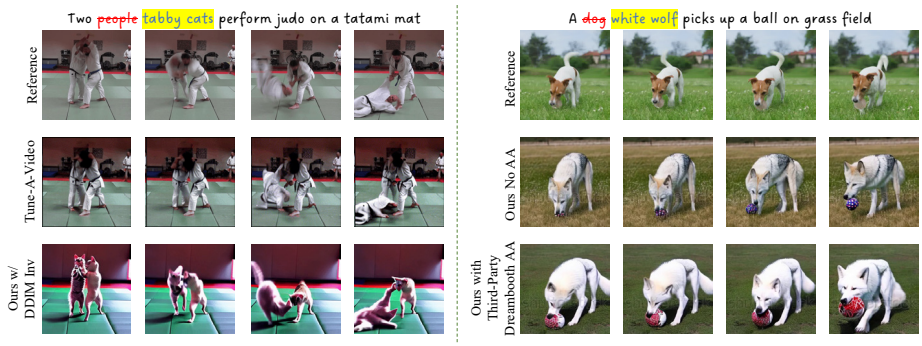

Fig. 6: Left: Precise frame-wise video editing with DDIM inverted latent as the inference input noise. **Right**: A third-party Dreambooth UNet pre-tuned on in-the-wild images of the same subject serves as the appearance absorber to train the T-LoRA.

DDIM Inverted Latent Input Our method can easily incorporate additional deterministic controls to perform precise video editing. The comparison results to Tuna-A-Video [50] are shown as Fig. 6 left. Our models prove to be also able to benefit from the DDIM inverted latent of the reference video and yield output that reproduces the exact original frame structures.

Third-Party Appearance Absorbers Our staged training pipeline enables reusing appearance absorbers across videos or loading third-party image customization modules as ready appearance absorbers when they share the similar appearance. This skips the first training stage and extends appearance absorber categories to those demand more training data. In Fig. 6 right, we finetune the spatial layers of the UNet following Dreambooth [39] and Eq. 1 on other photos of the same dog. Our T-LoRA trained with it avoids the leakage of the dog color to the white wolf, compared to that trained with the original base model.

6 Conclusion

We introduce the one-shot motion customization task that learns the motion signature from a single reference video and transfers 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's is plug-and-play nature supports various downstream tasks including precise video editing, video appearance customization, multiple motion combination as well as third-party appearance absorbers.

References

1. AILab-CVC: Ailab-cvc/videocrafter at 1f46314b6609712eea89b67f41d612557eec5b8e, https://github.com/AILab-CVC/VideoCrafter/tree/1f46314b6609712eea89b67f41d612557eec5b8e
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)
3. 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)
4. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
5. 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)
6. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
7. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
8. Chu, E., Lin, S.Y., Chen, J.C.: Video controlnet: Towards temporally consistent synthetic-to-real video translation using conditional image diffusion models. arXiv preprint arXiv:2305.19193 (2023)
9. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
10. 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)
11. Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.B., Parikh, D.: Long video generation with time-agnostic vqgan and time-sensitive transformer. In: European Conference on Computer Vision. pp. 102–118. Springer (2022)
12. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22930–22941 (2023)
13. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
14. 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
15. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
16. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)

17. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022)
18. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
19. 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)
20. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. arXiv preprint arXiv:2312.00845 (2023)
21. Kappa_Neuro: Dressed animals - sd 2.1: Stable diffusion lora, https://civitai.com/models/35733?modelVersionId=41939
22. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
23. Kim, D., Joo, D., Kim, J.: Tivgan: Text to image to video generation with step-by-step evolutionary generator. IEEE Access **8**, 153113–153122 (2020)
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
25. Li, Y., Min, M., Shen, D., Carlson, D., Carin, L.: Video generation from text. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
26. Liew, J.H., Yan, H., Zhang, J., Xu, Z., Feng, J.: Magicedit: High-fidelity and temporally coherent video editing. arXiv preprint arXiv:2308.14749 (2023)
27. Liu, B., Liu, X., Dai, A., Zeng, Z., Cui, Z., Yang, J.: Dual-stream diffusion net for text-to-video generation. arXiv preprint arXiv:2308.08316 (2023)
28. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv preprint arXiv:2303.04761 (2023)
29. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10209–10218 (2023)
30. Ma, Y., He, Y., Cun, X., Wang, X., Shan, Y., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. arXiv preprint arXiv:2304.01186 (2023)
31. Materzyńska, J., Sivic, J., Shechtman, E., Torralba, A., Zhang, R., Russell, B.: Customizing motion in text-to-video diffusion models. arXiv preprint arXiv:2312.04966 (2023)
32. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
33. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329 (2023)
34. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)

35. Pan, Y., Qiu, Z., Yao, T., Li, H., Mei, T.: To create what you tell: Generating videos from captions. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1789–1798 (2017)
36. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Computer Vision and Pattern Recognition (2016)
37. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
39. 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)
40. Shin, C., Kim, H., Lee, C.H., Lee, S.g., Yoon, S.: Edit-a-video: Single video editing with object-aware consistency. arXiv preprint arXiv:2303.07945 (2023)
41. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
43. tungdop2: tungdop2/pokemon-lora_sd2.1, https://huggingface.co/tungdop2/pokemon-lora_sd2.1
44. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
45. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
46. Wang, W., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint arXiv:2303.17599 (2023)
47. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. arXiv preprint arXiv:2306.02018 (2023)
48. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. arXiv preprint arXiv:2312.04433 (2023)
49. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: Nüwa: Visual synthesis pre-training for neural visual world creation. In: European conference on computer vision. pp. 720–736. Springer (2022)
50. 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)
51. Wu, J.Z., Li, X., Gao, D., Dong, Z., Bai, J., Singh, A., Xiang, X., Li, Y., Huang, Z., Sun, Y., et al.: Cvpr 2023 text guided video editing competition. arXiv preprint arXiv:2310.16003 (2023)
52. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)

53. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. arXiv preprint arXiv:2306.07954 (2023)
54. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
55. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
56. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
57. Zhang, Y., Tang, F., Huang, N., Huang, H., Ma, C., Dong, W., Xu, C.: Motioncrafter: One-shot motion customization of diffusion models. arXiv preprint arXiv:2312.05288 (2023)
58. Zhang, Z., Li, B., Nie, X., Han, C., Guo, T., Liu, L.: Towards consistent video editing with text-to-image diffusion models. arXiv preprint arXiv:2305.17431 (2023)
59. Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J.: Controlvideo: Adding conditional control for one shot text-to-video editing. arXiv preprint arXiv:2305.17098 (2023)
60. 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)
61. Zhao, Y., Xie, E., Hong, L., Li, Z., Lee, G.H.: Make-a-protagonist: Generic video editing with an ensemble of experts. arXiv preprint arXiv:2305.08850 (2023)
62. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)

A Implementation Details

A.1 Bypassing Temporal Layers in T2V Models

Many diffusion-based T2V models such as [3,5,14,45] 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.

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. 7. 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 [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

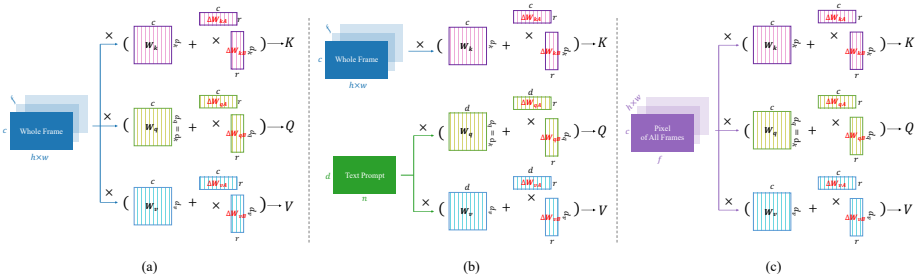

Fig. 7: (a) Spatial self-attention between each frame and itself; (b) Spatial cross-attention between each frame and the text prompt; (c) Temporal cross-frame attention among pixels of all frames in a video. Batch size is omitted for simplicity.

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 [10] as the appearance absorber, we set the length of new learnable tokens to 2-6 depending on the content complexity.

Table 3: 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 [60]	31.842	0.166	0.595	2	1	779.5K	274.5K
MotionDirector [60]	<u>32.500</u>	<u>0.163</u>	0.606	4	1	1559K	274.5K

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 [50] (6 minutes), Video-P2P [28] (8 minutes in fast mode; 14 minutes in full mode on A6000 for bigger VRAM) and concurrent work MotionDirector [60] (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 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. 3 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. 8. 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. 9. 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.

B.3 Training Schedule Variances

Our modules fit on each reference video individually to model its unique motion signal. Fig. 10 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. 11 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.

Fig. 8: Additional generation results of our method.

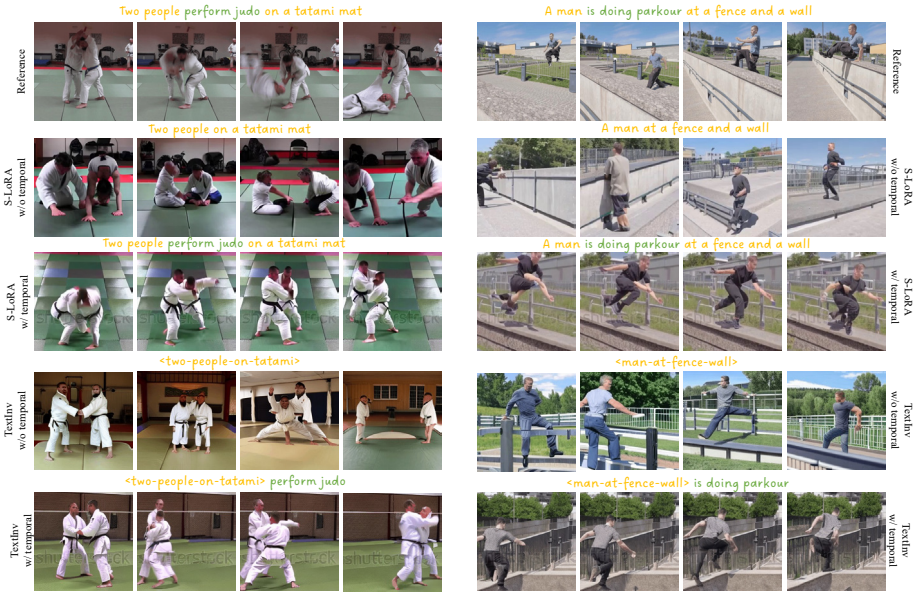

Fig. 9: Output generated with appearance absorbers loaded only and temporal layers bypassed. The 2nd and 3rd rows show S-LoRA. The 4th and 5th rows show textual inversion. The training prompts and special tokens are noted above for each sample.

C User Study Questionnaire Design

We present example questions in our user studies in Fig. 12. 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.

D Limitation Discussions

Per Instance Finetuning. Our method tunes on each reference video individually. Similar to comparing approaches [28, 50, 60], 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 [10, 39] 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.

Fig. 10: Examples where different reference videos require different tuning iterations. (1-2) Simpler motions such as camera movements usually converge faster. (3) More complex motions such as animal or human actions would demand more tuning steps.

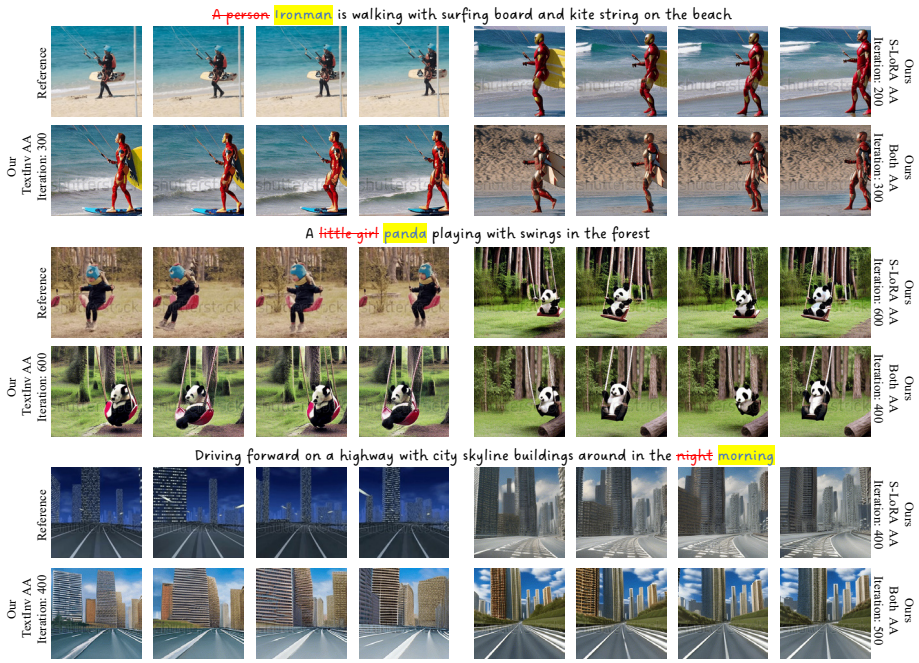

Fig. 11: Examples where different appearance absorbers exhibit different characteristics. (1) Our Both AA absorbs the original appearance more thoroughly, leading to more diverse new background generated. (2) Our Both AA may reduce the necessary convergence step compared to a single appearance absorber. (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.

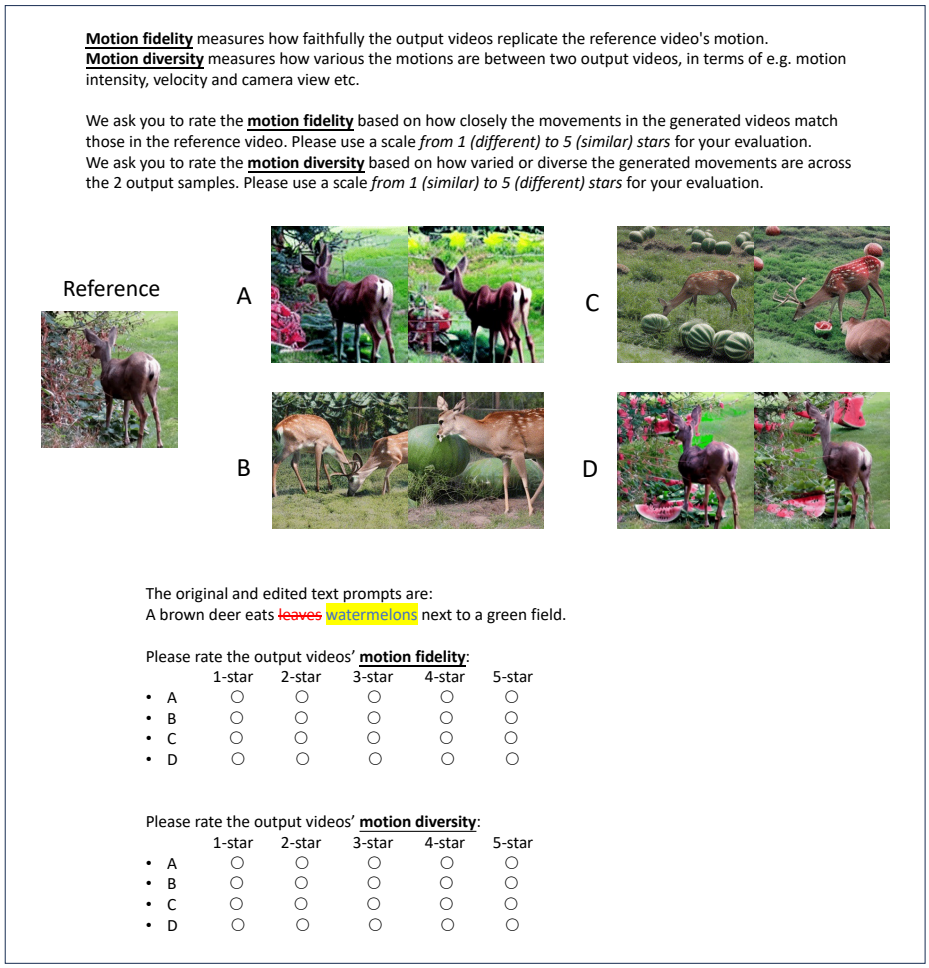

Fig. 12: An example question in the human user study. Participants are asked to rate each algorithm's output videos from 1 to 5 stars.

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.