Video Diffusion Models Encode Motion in Early Timesteps

Vatsal Baherwani
University of Maryland
vatsalb@umd.edu

Yixuan Ren
University of Maryland

Abhinav Shrivastava
University of Maryland

Abstract

Unlike text-to-image diffusion models, text-to-video diffusion models simultaneously process both spatial (2D) information and temporal (cross-frame) information. This presents a challenge for capturing motion dynamics without entangling spatial features as well. We demonstrate that temporal motion information is predominantly and independently learned in the early timesteps of the diffusion process, before significant spatial features emerge. Using this insight, we present a novel adaptation of existing LoRA fine-tuning techniques that selectively targets these early timesteps for precise motion learning. Our method successfully learns motion from single video examples without requiring any explicit spatial debiasing and is compatible with both LoRA and full-rank fine-tuning. We verify the isolation of motion learning in early timesteps across various diffusion model architectures, demonstrating the broad applicability of our approach for understanding video motion representations.

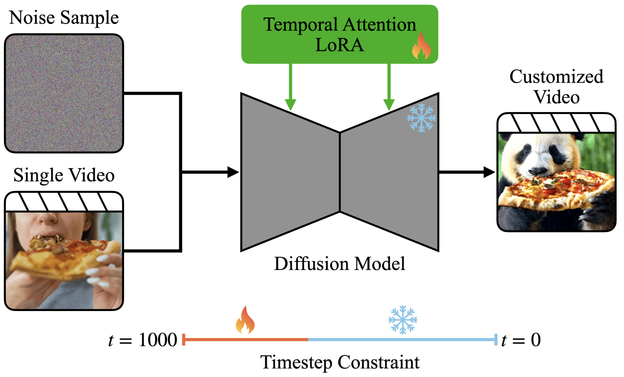

Figure 1. **Overview of our Method**. Given a pre-trained text-to-video diffusion model, we freeze the parameters and apply LoRA adapters to temporal attention layers. Unlike previous methods for decoupling spatial information, we take the novel approach of limiting the timesteps over which the diffusion model is trained. By training only on early timesteps (specifically $t \in [600, 1000]$), we can transfer the motion from a one-shot video example onto customized videos.

1. Introduction

Modern text-to-video diffusion models can generate impressive photorealistic videos given a text description, but developing precise prompts to describe a unique motion is very difficult. A more practical approach is to transfer the motion from an existing video into a synthetic generation via model fine-tuning, ideally using a single (one-shot) example. However, this is a particularly challenging task for video diffusion models as they must capture both spatial information in their individual frames and temporal information across frames. During fine-tuning, transferring motion information often leads to unintended leakage of spatial attributes (e.g., background details from the reference video), which is especially problematic when using just a single video example.

Existing approaches attempt to mitigate spatial leakage through techniques like two-stage training [24], specialized modules for capturing spatial information [46], or incorporating image conditions [35]. While these approaches are effective in decoupling motion from spatial information, they introduce trade-offs that compromise training stability, require more computational resources, or degrade motion quality. To the best of our knowledge, no prior work has adequately addressed the root issue: the entanglement of spatial and temporal information in video diffusion models.

In this paper, we provide a novel perspective by demonstrating that spatial and temporal information are independently encoded in disjoint timestep ranges of the diffusion process. Specifically, motion information is primarily learned in the early timesteps, before substantial spatial features are formed. Using this observation, we propose a method to disentangle motion from spatial information during fine-tuning, entirely preventing spatial leakage. We verify this property by applying DDIM inversion to observe the latent information encoded at different timesteps. Our experiments on multiple existing architectures consistently show that motion information is predominantly encoded in the initial stages of the diffusion process.

Building on this insight, we introduce a novel fine-tuning strategy for motion customization by applying LoRA [11] adapters to temporal attention layers. We selectively fine-tune the model only on early timesteps, and this effectively avoids the need for additional modifications to decouple the spatial information.

Given the challenges of reliably transferring motion from one-shot examples, previous works have largely relied on LoRA-based fine-tuning and avoided direct tuning due to the risk of overfitting. While direct tuning methods like DreamBooth [26] are effective for text-to-image customization, they typically require numerous examples and regularization techniques. In this paper, we address this challenge by extending our timestep restriction approach to direct tuning for text-to-video motion customization. This results in a stable method that captures the motion from a single reference video without any spatial overfitting.

By deliberately focusing training on timesteps where motion is encoded, our method not only improves the quality of motion transfer but also significantly reduces computational overhead. We further support our approach with an ablation study, minimizing the number of necessary trainable parameters to $< 1/1000$ of the base model parameters without sacrificing performance. Overall, our method offers a simpler, more efficient training process which yields better motion quality.

Our main contributions are as follows:

- We establish the disentanglement of spatial and temporal information between timestep intervals in the diffusion process, showing that temporal (motion) information is primarily encoded in the early timesteps.
- We propose a refined LoRA fine-tuning method for one-shot motion customization by selectively training only on early timesteps, avoiding spatial information leakage.
- We introduce a novel direct tuning approach that leverages timestep restriction to prevent overfitting, enabling effective motion transfer from one-shot examples.

2. Background and Related Works

2.1. Diffusion Models for Video Generation

Diffusion models [8] generate samples following a data distribution by starting with random noise and iteratively applying a stepwise denoising process. During training, the *forward* diffusion process consists of taking an existing data point x_0, and then repeatedly adding random noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ multiplied by some scaling factor. Each iteration transforms x_t to x_{t+1}, and this is repeated T times where T is referred to as the number of timesteps. Note that the forward process moves the sample x in the direction towards noise. For the purposes of our experiments, we set $T = 1000$ to match previous work. The scaling factor applied to the noise is set so that after T iterations, x_T contains almost no signal and resembles pure noise.

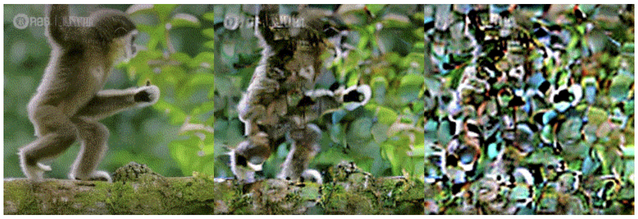

Figure 2. **Decoded Latents at Intermediate Timesteps:** $t = 0, 400, 600$**.** We apply DDIM inversion to a video and present the results of the same frame at multiple timesteps. At $t = 0$, no noise has yet been added and the latent contains all of the appearance and motion information. At $t = 400$, the appearance is mostly unchanged but it is noisy. The motion is completely intact, as the pose is the same at the selected frame. At $t = 600$ the appearance information is mostly lost to noise, but the pose at this frame still vaguely resembles the original motion. A more rigorous analysis of these claims is presented in Fig. 3

.

The denoising process i.e. the *reverse* process is traditionally learned by training a neural network parameterized by θ to predict x_{t-1} given x_t. This objective is equivalent to predicting the noise ϵ that transformed x_{t-1} to x_t, and subtracting the prediction $\epsilon_\theta(x_t, t)$ multiplied by some scalar from the given x_t. This represents moving x away from the noise distribution and towards the data distribution.

To generate a synthetic data point, we begin with a pure noise sample x_T and apply the denoising process iteratively to produce a synthetic x_0. We primarily experiment with latent diffusion models, which learn the diffusion process in a lower dimensional latent space defined by a pre-trained variational autoencoder. However, all of our analyses and methods seamlessly translate to pixel-based diffusion models, and we demonstrate this in our empirical experiments.

Most modern video diffusion models learn a denoising network based on either a UNet [25] or a Transformer [31]. UNets are primarily based on convolution layers, while Transformer blocks consist of attention layers and MLPs. Both architectures, however, contain temporal attention layers, which model the dependencies between frames. These temporal attention layers are the primary focus of our paper as we seek to learn unique motions.

Many diffusion models for video generation exist [1, 6, 9, 16, 18, 20, 22, 27, 32, 34, 40, 43], and in this paper we experiment with the ModelScope [33], Show-1 [44], and Latte [19] models. We select these models to investigate diverse diffusion architectures (pixel-based vs. latent-based, and UNet vs. Transformer).

Beyond diffusion models, high-quality video generation is also implemented with other architectures including autoregressive models [4, 10, 14, 23, 30, 39] and implicit neural representations [28, 42].

2.2. Diffusion Model Fine-tuning

Fine-tuning methods for diffusion models typically fall into two categories: low-rank and full-rank. Low-rank methods involve training adapters using LoRA [11], and recently DoRA [17] is also emerging as a more capable method for parameter-efficient fine-tuning. LoRA is used for a variety of image customization tasks [2, 5, 29].

Dreambooth [26] is an approach for personalizing text-to-image models using full-rank (direct) fine-tuning. Notably, the method requires multiple (3-5) images and employs additional loss functions to encourage diversity in generated samples. However, this direct training method has not been applied to the one-shot case in either image or video modalities.

Alternative approaches to fine-tuning involve learning unique text or image embedding representations [3, 15, 36], direct tuning of cross attention layers [13, 41], and training with an image condition [45].

2.3. Video Motion Customization

Recently, with the increasing capability of open-source video diffusion models, multiple methods have been proposed specifically for the task of fine-tuning models to customize motion. Video diffusion models broadly contain two types of layers: spatial layers and temporal layers. Spatial layers operate on individual frames of the image independently without facilitating any cross-frame interaction. Temporal layers process the entire video at once and are responsible for modeling dependencies between frames. Many existing methods (along with our own method) make use of this fact for effective motion fine-tuning.

DreamVideo [35] decouples the spatial and temporal information for video customization by applying a two stage process of subject learning and motion learning. Subject learning takes in images of the new desired subject and trains spatial layer adapters of the diffusion model to capture appearance details. Motion learning takes in videos of the desired motion and trains temporal layer adapters to capture the motion. At inference time, these two adapters are combined to transfer the motion onto the new subject.

Customize-A-Video [24] also employs a two-stage approach, but only takes in a one-shot example of the motion in order to transfer it to any new prompt. The first stage captures the appearance with LoRA on spatial layers, and the second stage captures the motion with additional LoRA adapters on temporal layers. During inference, only the temporal LoRA layers are activated to transfer the motion to any given new subject based on the generation prompt. The first stage can be substituted with any other spatial fine-tuning method; the goal is to ”absorb” the appearance so that the temporal LoRA layers only learn the motion.

MotionDirector [46] applies a similar approach of training both spatial and temporal LoRA adapters, but these are trained simultaneously in a dual-path architecture and with either single or multiple videos. One path trains spatial LoRA adapters on individual frames of the videos, and the other path shares the spatial LoRA weights while learning temporal LoRA adapters to capture the motion. An additional appearance-debiased temporal loss objective is applied to ensure the motion is learned independently of the videos' appearance.

VMC [12] is another approach for one-shot tuning which fine-tunes temporal attention layers and introduces an auxiliary motion distillation objective. The motion distillation ensures that the latents of consecutive frames align with the latents of the video containing the reference motion.

In comparison to these works, our method is focusing on the one-shot case. The fundamental difference in our method is that we do not require any additional training stages or new loss functions. We instead achieve motion decoupling exclusively through constraining the timesteps over which the diffusion model is fine-tuned. This leads to very efficient training along with improvements in overall video quality compared to previous methods.

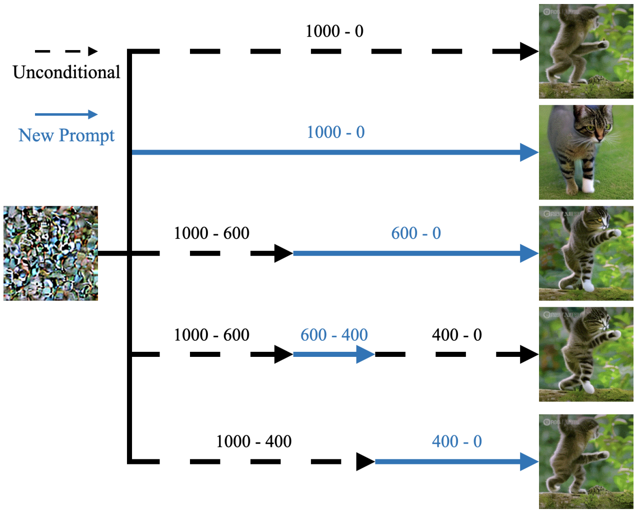

Figure 3. **Prompt Injection at Various Timestep Intervals**. These results are snapshots of the same frame across different resampled videos. Given an original video of a monkey walking, we resample its DDIM inverted latent from $t = 1000$, and selectively apply guidance with a new prompt of a cat walking. Varying the timestep interval in which guidance is applied leads to differences in whether some attributes correspond to the old video or the new prompt. The fourth example demonstrates that applying guidance over $t =\in [600, 400]$ is sufficient to edit the spatial information while preserving the original video's motion information.

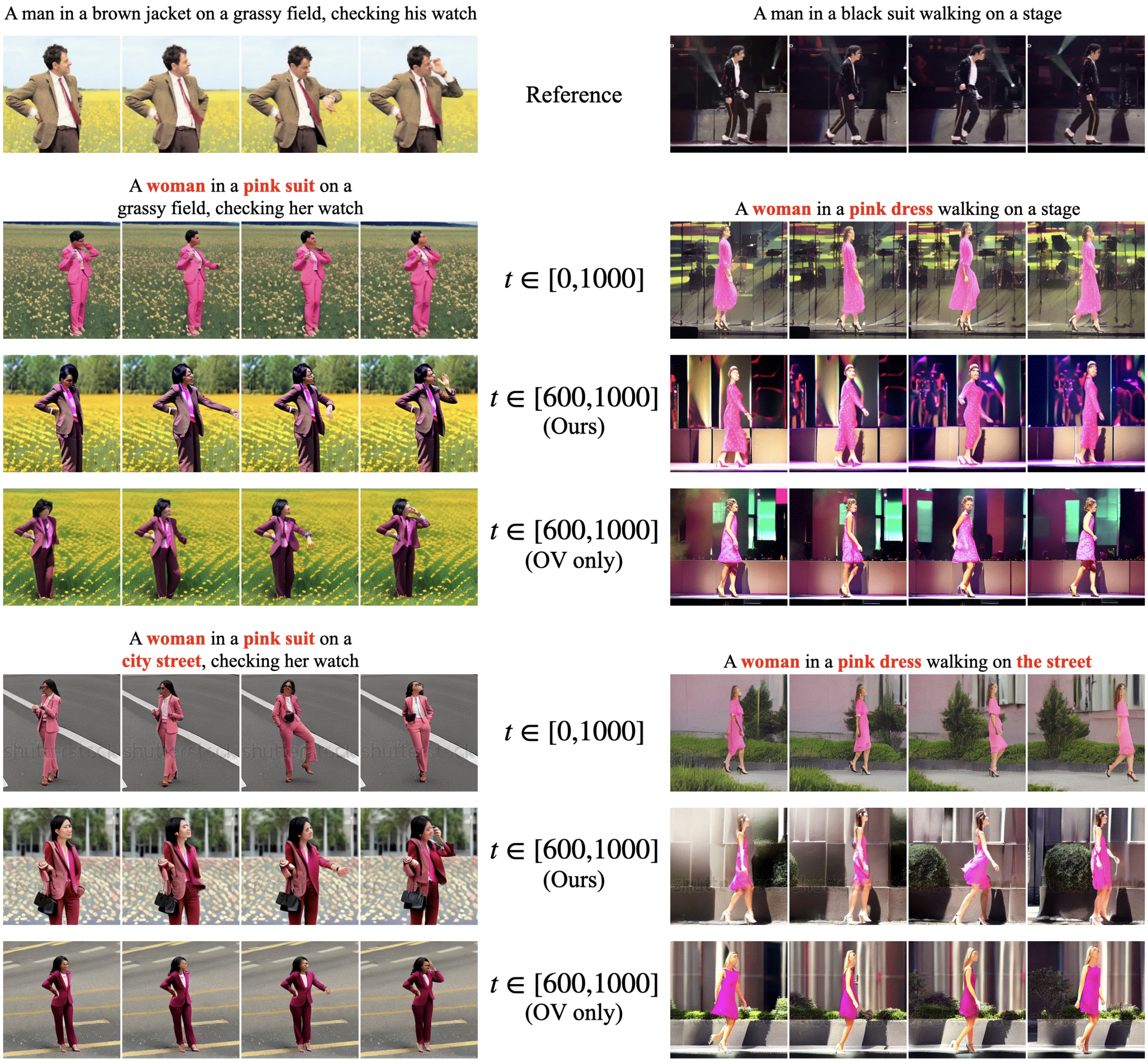
A man in a brown jacket on a grassy field, checking his watch
Reference
A man in a black suit walking on a stage
A woman in a pink suit on a grassy field, checking her watch
A woman in a pink dress walking on a stage
$t \in [0,1000]$
$t \in [600,1000]$ (Ours)
$t \in [600,1000]$ (OV only)
A woman in a pink suit on a city street, checking her watch
A woman in a pink dress walking on the street
$t \in [0,1000]$
$t \in [600,1000]$ (Ours)
$t \in [600,1000]$ (OV only)

Figure 4. **Customizing Motion by Limiting Training Timesteps**. We fine-tune the ModelScope text-to-video diffusion model by applying a LoRA adapter on all temporal attention layers. **1.** When training on $t \in [0, 1000]$ without any restriction, the model fails to completely capture the unique motion of the reference video. The partial motion that is captured does not reliably transfer onto new subjects or backgrounds. **2.** However, when we limit timesteps to $t \in [600, 1000]$, we observe the motion being learned and replicated onto a new subject. Moreover, this motion can transfer to other settings like different backgrounds. **3.** We observe that limiting the trainable parameters to just the O and V projections of attention layers is sufficient to capture the motion, and in some cases it leads to slight improvements in generalizing to new prompts. (e.g. changing the background on the left to a city street). This allows us to cut the number of required trainable parameters in half, as we freeze the Q and K projections.

3. Tracing Motion Information in Video Diffusion Models

In this section, we introduce our methodology for isolating and customizing motion information in text-to-video diffusion models. We first present an empirical analysis using DDIM inversion to demonstrate the temporal encoding of motion information at early timesteps. This is followed by a detailed discussion of our modified fine-tuning approach, which restricts training to early timesteps to avoid spatial information leakage. We conclude with a description of our direct training strategy along with a concise summary of our proposed changes to the training objective.

3.1. DDIM Inversion

DDIM inversion [21] allows for "reversing" the sampling process for a given video x_0. By iteratively moving in the direction of $\epsilon_\theta(x_t, t)$ for each timestep t, we deterministically reach a noise sample x_T. Then, sampling using this x_T will yield a video very closely resembling the original x_0. We apply DDIM inversion to latent diffusion models in our experiments, but the technique is equally applicable to diffusion models operating in pixel space.

To understand the encoding of information at different timesteps, we examine the decoded latents at intervals of every 100 timesteps. Initially, at $t = 0$, the latent contains a complete representation of the video. As we progress towards $t = 1000$, the latent loses its information, eventually becoming pure noise. We observe the intermediate timesteps where the latent contains a partial representation of the video. Our analysis reveals that up until $t = 400$, the latent retains most of core appearance and motion attributes from the original video, albeit with some noise. By $t = 600$, appearance details are largely lost, but the motion information remains. These observations, illustrated in Fig. 2, suggest that motion is encoded predominantly in earlier timesteps, while spatial details are encoded later.

3.1.1. Resampling with Partial Guidance

Text-to-video diffusion models typically use classifier-free guidance [7] to condition the noise prediction on a text prompt p. The guided prediction ϵ_g is given by a combination of the conditional noise prediction and the unconditional prediction:

$$\epsilon_g = \beta \epsilon_\theta(x_t, t, \tau_\phi(p)) + \epsilon_\theta(x_t, t, \mathbf{0}) \tag{1}$$

where β is the guidance scale hyperparameter and $\tau_\phi(p)$ is the text encoding of the prompt. The conditional term aligns the prediction with the text prompt, while the unconditional term reflects the general data distribution. During sampling we are simultaneously moving x in the weighted combination of both these directions.

After removing information from a video latent through DDIM inversion, we restore the information by repeating the sampling process with the intermediate latent. To verify the disentanglement of spatial and temporal information, we use "partial" guidance during the resampling process. Specifically, we perform DDIM inversion to obtain a latent at x_t and then resample using a new prompt p' from timesteps t to 0, but we only apply guidance over a subset of $[0, t]$. Applying guidance only within certain intervals allows us to narrow down which timesteps are responsible for encoding specific attributes (e.g., motion or appearance). There are two possible cases when resampling with partial guidance over a subset of $[0, t]$. The attribute could be faithfully preserved from the original video, despite applying guidance with a new prompt at this interval of timesteps. In this case, the selected timesteps were not very relevant in encoding the information related to the attribute. Conversely, if we apply guidance at some timesteps and observe changes in attributes that resemble the new p', then those timesteps are responsible for encoding the attribute.

3.1.2. Spatial and Temporal Disentanglement

We illustrate this method with a video of "a monkey walking" and a new prompt p' = "a cat walking". Note that given the subjects, the original prompt depicts a bipedal motion while the new prompt has a quadrupedal motion. In Fig. 3, the first row shows the result of unconditional resampling from $t = 1000$ to $t = 0$, which reproduces the original video with no changes. The second row shows the result of resampling with the new prompt over the entire interval, producing a completely new video of a cat walking on all four feet.

In the third row, we apply guidance only from $t = 600$ to $t = 0$. The appearance of the monkey changes to a cat, but the bipedal walking motion is preserved. This indicates that the motion information was encoded in the previous timesteps $t \in [1000, 600]$ and was not affected over later timesteps. In the fourth row, applying guidance only from $t = 600$ to $t = 400$ yields the same changes in appearance without affecting motion, suggesting that high-level spatial information is specifically restricted to this interval. Finally, stopping DDIM inversion at $t = 400$ and resampling with the new prompt from $t = 400$ to $t = 0$ preserves both motion and appearance, confirming that the later timesteps mainly influence resolution and low-level details.

Our analysis reveals that motion information is primarily encoded in $t \in [1000, 600]$, while spatial information is encoded in $t \in [600, 400]$. Note that these intervals are disjoint – this suggests that diffusion models exhibit disentanglement between spatial and temporal information across timesteps. This disentanglement allows us to exploit the early timesteps for motion learning without risking spatial leakage.

3.2. Restricting Timesteps for Motion Fine-tuning

3.2.1. Baseline Method

The baseline method in our experiments involves applying LoRA adapter modules to parameters in every temporal attention layer of the diffusion model. We then fine-tune the model with the standard DDPM noise prediction objective. This method is not sufficient to reliably capture the motion without also leaking spatial information. Spatial leakage becomes a more pressing issue with less training examples, especially in the one-shot case.

3.2.2. Our Method

Instead of introducing complex spatial debiasing techniques, we simply restrict the timesteps during fine-tuning. At each training step, we sample $t \in [600, 1000]$ instead of the full range $t \in [0, 1000]$. Our empirical analysis (see third row of Fig. 3) suggests that encoding motion information in this range effectively prevents spatial leakage, as applying a new prompt at later timesteps does not affect the motion at all. Indeed, this ends up being the case; limiting the timesteps not only allows for learning the motion but also applying it in new settings without spatial leakage. Results of our method compared to the baseline method with the ModelScope [33] model are presented in Fig. 4.

3.3. Direct Training with Restricted Timesteps

Direct tuning methods are often avoided in video customization due to the risk of overfitting. Full-rank fine-tuning significantly increases the degrees of freedom for the model to replicate the video, which leads to spatial leakage. However, we compensate for this by restricting the timesteps during training, which in turn reduces the model's degrees of freedom. This method matches the performance of LoRA-based fine-tuning for motion customization while offering increased modeling flexibility. We demonstrate an example of direct training results in Fig. 6.

3.4. Summary of Our Method

To summarize, our method focuses on restricting the timesteps during the fine-tuning of video diffusion models, leading to improved motion learning without spatial leakage. Consider the standard diffusion model training objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_t,\epsilon}[\epsilon - \epsilon_\theta(x_t, t, \tau_\phi(p))]; t \in [0, 1000] \qquad (2)$$

Our modification is simply limiting the range for sampling the timestep t:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_t,\epsilon}[\epsilon - \epsilon_\theta(x_t, t, \tau_\phi(p))]; t \in [600, 1000] \qquad (3)$$

This adjustment alone improves motion transfer and reduces redundant computation – there is no need to train on $t \in [0, 600]$ as motion is not affected in those timesteps. Our method can be applied to both direct tuning and LoRA-based approaches, allowing for a simple yet highly effective implementation.

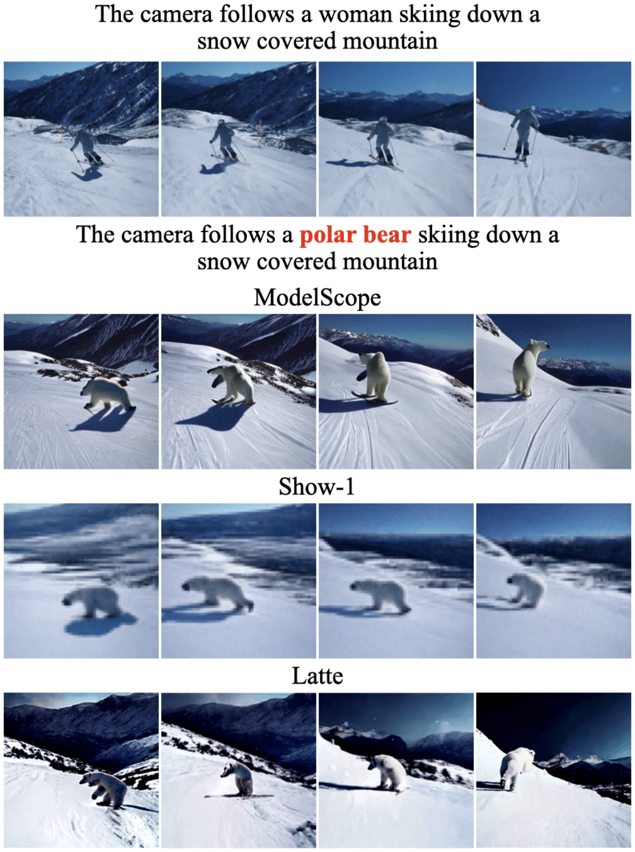
The camera follows a woman skiing down a snow covered mountain
The camera follows a polar bear skiing down a snow covered mountain
ModelScope
Show-1
Latte

Figure 5. **Motion Customization on Different Base Models**. Our motion customization method (specifically, limiting training timesteps to $t \in [600, 1000]$ and only training O and V projections of temporal attention layers) is broadly applicable to different base models. We demonstrate results on ModelScope, Show-1, and Latte. Our results suggest that this simple fine-tuning method can be plugged in to all diffusion models, regardless of their underlying architecture.

4. Evaluation

4.1. DDIM Inversion Experiments

To produce the timestep interval analysis in Fig. 3, we take a reference video with some prompt (e.g. "a monkey walking" in Fig. 3) and apply DDIM inversion at intervals of every 100 steps with the ModelScope [33] base model. At this point, we have 10 latents at $t = 100, t = 200, \ldots, t = 1000$. For one regeneration, given a latent at time t, we resample with a new prompt (e.g. "a cat walking") but only apply guidance until t'. After t' we sample unconditionally – this corresponds to setting $\beta = 0$ in Eq. (1). We regener-

	Text Alignment (↑)	Temporal Consistency (↑)	Pick Score (↑)
VideoComposer	27.66	92.22	20.26
Control-A-Video	26.54	92.63	19.75
VideoCrafter	28.03	92.26	20.12
Tune-A-Video	25.64	92.42	20.09
MotionDirector	27.82	93.00	**20.74**
Ours	**28.04**	**95.09**	**20.74**

Table 1. **TGVE Benchmark Results**. We benchmark our performance on one-shot motion customization against previous methods. We use the benchmarks of Text Alignment, Temporal Consistency, and Pick Score from the TGVE 2023 Competition. Our results are competitive with other approaches, and we especially realize significant gains in the temporal consistency of our customized videos. We achieve these results while requiring significantly less trainable parameters and training steps compared to previous approaches.

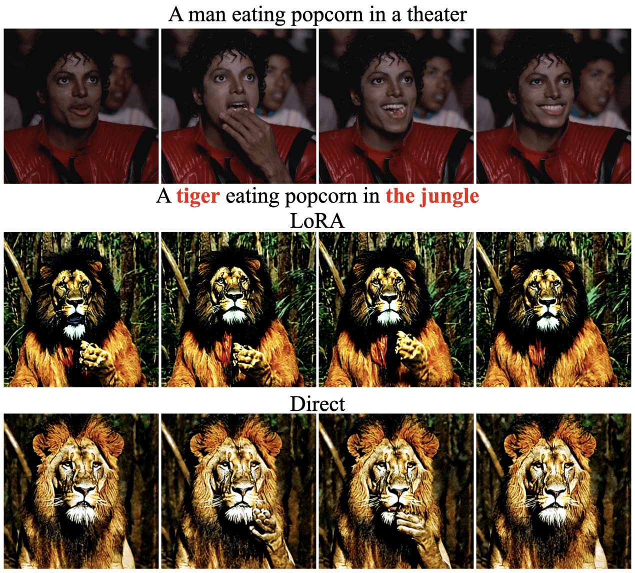
A man eating popcorn in a theater
A tiger eating popcorn in the jungle
LoRA
Direct

Figure 6. **Direct Training Without LoRA**. We present fine-tuning results on the Latte Transformer-based text-to-video model. Our method seamlessly applies to direct training and produces motion customization results without requiring LoRA modules. All the capabilities of the LoRA method are still present, such as adapting to a new subject and background without overfitting to spatial details. We believe this enables more powerful motion customization by increasing the representation capacity during fine-tuning.

ate with all possible t and t' that are intervals of 100, with $t' < t$. For example, one regeneration could use the DDIM inverted latent from $t = 500$ and apply guidance with the new prompt until $t' = 200$; there are $\binom{11}{2} = 55$ such combinations of t' and t. We evaluate each combination manually to judge which information is transferred from the new prompt into the generation.

4.2. Results on Different Models

Our results are not unique to the ModelScope text-to-video model. We apply our fine-tuning approach to the Show-1 [44] base model, along with the Latte [19] model. See Fig. 5 for motion customization results compared across each of these models. Notably, each model has a different underlying architecture: ModelScope is latent-based with a UNet, Show-1 is pixel-based with a UNet, and Latte is latent-based with a Transformer. The success of our method on all of these models suggests that this disentanglement of spatial and temporal attributes across timesteps is not just a property of a single model, but for all video diffusion models in general.

4.3. Ablating Attention Parameters

Our above methods involve fine-tuning of the cross-frame attention layers of the diffusion model. In each attention layer there are four projections: Q, K, V, O the query, key, value, and out projections respectively. We experiment with freezing all possible $2^4 - 1$ combinations of these four projections during training. We find that freezing the value or out projections independently significantly limits the fine-tuning results; freezing both the value and out projections together leads to none of the motion being learned. Moreover, freezing the query or key projections (or both) has no negative impact on the results. We hypothesize that the V and O projections are most relevant to motion customization i.e. only training V and O is sufficient for fine-tuning. This enables us to cut the number of trainable parameters in half without incurring any decrease in motion customization quality. We demonstrate the comparison between training the entire attention layer versus just the V and O projections in Fig. 4.

4.4. Quantitative Metrics

We evaluate our method on quantitive metrics using the LOVEU-TGVE-2023 [38] benchmarks. The benchmark includes 76 reference videos. Each video is labeled with an original text prompt, along with four new text prompts. The new text prompts change the style, object, background, or multiple attributes from the original prompt.

We use the Latte [19] text-to-video model for our evaluations, and apply LoRA on the O and V projections of all temporal attention layers with a LoRA rank of 4. This is significantly lower than other LoRA approaches like MotionDirector [46], which uses a rank of 64. We train a total of 516,000 parameters – a negligible fraction compared to Latte's 600M total parameters. We apply our fine-tuning method of restricting timesteps on the original video and train for 200 steps with a learning rate of 3×10^{-4}. On a single RTX A6000 GPU, this training only takes 7 minutes. After fine-tuning, we generate two videos for each new prompt from the TGVE dataset and average the text alignment, temporal consistency, and pick score metrics for both videos. Results for each specific prompt category are provided in Tab. 2.

We also report our average scores across all TGVE prompts and compare them to other approaches for one-shot motion customization. We either match or outperform existing methods in all quantitative video generation metrics. Our method leads to a significant increase in temporal consistency as a result of strictly limiting training to the timesteps responsible for encoding temporal information. We achieve these results with significantly less trainable parameters than the other methods, proving that our method is efficiently capturing the motion from a single video.

	Text Alignment	Temporal Consistency	Pick Score
Style	27.47	95.38	20.77
Object	27.80	94.82	20.79
Background	28.41	94.97	20.74
Multiple	28.47	95.20	20.64
Average	28.03	95.09	20.74

Table 2. **Individual Metrics on Motion Customization**. We collect the individual TGVE benchmark scores on each type of prompt, where either the style, object, background, or multiple attributes are modified. Our method's performance is consistent across all of these changes. Notably, we do not observe a dropoff in customization results from changing one attribute to multiple attributes from the original prompt. This makes our method robust in applying motion to a broad variety of new settings.

5. Discussion

5.1. Limitations

In this paper, we specifically focus on fine-tuning text-to-video diffusion models on limited timestep intervals. We specifically focus on the interval $t \in [600, 1000]$ for motion customization, and ignore the rest of the timesteps. As demonstrated in Fig. 3, there are more interesting unexplored concepts like the interval $t \in [400, 600]$ primarily encoding spatial information and $t \in [0, 400]$ having a marginal effect on the high-level contents of the video. Moreover, we do not explore the possibility of smaller intervals that may exhibit fine-grained disentanglement properties; for example, maybe timesteps in a subinterval of $[400, 600]$ encode specific properties like background or texture. We leave the investigation of these claims to future work.

Our approach is specifically limited to text-to-video pipelines based on diffusion models. The concept of timesteps is unique to the diffusion process and so our insights do not apply outside of this scope. However, there are many other works that achieve high-quality video generation through autoregressive models or implicit representations. Further work is necessary to understand the disentanglement of spatial and temporal attributes in these other models, if it exists at all.

All of our methods and results focus specifically on the one-shot setting for motion customization. This is intentional, as we look to demonstrate the efficacy of our approach with minimal examples. However, this also is intrinsically inferior to multi-shot approaches and we believe higher quality motion customization is possible with applying our method given multiple reference videos.

5.2. Future Work

The property of spatial and temporal disentanglement across timesteps can offer insight into many other areas in video generation/editing and more broadly in computer vision. The simplicity and efficiency of our method enables downstream applications including video editing, animation, and personalized content creation, where precise control over motion is necessary. A notable corollary of our findings is that in the DDIM inverted latent at $t = 600$ for an existing video, the spatial information is removed but motion information is still intact. In other words, this latent information contains a complete motion representation free from any other information. This decoupled latent information can be utilized beyond motion customization for improving motion tracking, understanding poses and gestures, predicting movement, among other tasks.

Our modified approach to fine-tuning now makes direct tuning a feasible option for one-shot motion customization. This opens the door for future work on improved methods for direct tuning for potentially capturing very complicated motions or multiple motions at a time.

We also recognize that our observations on spatial and temporal disentanglement are purely empirical. Nonetheless, this property is validated across multiple diffusion models each with different underlying architectures. We hope that future works develop a more theoretical understanding on if and why the diffusion model training process leads to this disentanglement.

5.3. Conclusion

The primary contribution of this paper is demonstrating a disentanglement of temporal and spatial attributes across timesteps for video diffusion models. We apply this insight to the task of motion customization and show that diffusion models can be fine-tuned to replicate motions without requiring any modifications to the architecture or training objective. Our experimental results verify that this method is capable of capturing a diverse range of motions and applying them to a variety of new prompts.

References

[1] 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. 2

[2] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora, 2024. 3

[3] 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, 2022. 3

[4] 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. *arXiv preprint arXiv:2204.03638*, 2022. 2

[5] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023. 3

[6] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 2

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 5

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[9] 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. 2

[10] 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. 2

[11] 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, 3

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

[13] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. 3

[14] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Ccvs: context-aware controllable video synthesis. *Advances in Neural Information Processing Systems*, 34:14042–14055, 2021. 2

[15] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023. 3

[16] Shanchuan Lin and Xiao Yang. Animatediff-lightning: Cross-model diffusion distillation, 2024. 2

[17] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation, 2024. 3

[18] 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 (CVPR)*, 2023. 2

[19] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation, 2024. 2, 7, 3

[20] Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9117–9125, 2023. 2

[21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. 5

[22] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 2

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[24] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models, 2024. 1, 3

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[26] 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

[27] 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. 2

[28] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. *arXiv preprint arXiv:2112.14683*, 2021. 2

[29] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual dif-

fusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023. 3

[30] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2

[32] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022. 2

[33] 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, 6, 3

[34] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Pe der Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Y. Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models. 2023. 2

[35] 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, 2023. 1, 3

[36] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 3

[37] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022. 2

[38] Jay Zhangjie Wu, Difei Gao, Jinbin Bai, Mike Shou, Xiuyu Li, Zhen Dong, Aishani Singh, Kurt Keutzer, and Forrest Iandola. The text-guided video editing benchmark at loveu 2023. https://sites.google.com/view/loveucvpr23/track4, 2023. 7

[39] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[40] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024. 2

[41] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 3

[42] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2021. 2

[43] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023. 2

[44] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023. 2, 7, 3

[45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

[46] 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, 2023. 1, 3, 7, 2

Video Diffusion Models Encode Motion in Early Timesteps

Supplementary Material

6. Training Details

For training LoRA with the Latte text-to-video model, we apply LoRA with rank 4 to the single temporal attention layer in each of the 28 Transformer blocks. This results in $516,096$ trainable parameters, compared to the 673 million parametes of the Latte base model. The LoRA layers are trained for 200 steps on a single video with a learning rate of 3×10^{-4}, using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay 0.01. Sampling is done with a guidance scale of 9.0 with 50 inference steps.

7. Figure Visualizations

Accompanied in this zip file are 6 folders corresponding to each of the figures in the main text. Each folder contains the videos that were sampled to create the respective figure.

Folder `fig1` (Fig. 1) contains two videos and a text file for the corresponding prompts.

Folder `fig2` (Fig. 2) contains latents at all decoded timesteps $t = 0$, $t = 100$, $\ldots t = 900$. Note $t = 1000$ is expected to be pure noise.

Folder `fig3` (Fig. 3) contains results from all possible intervals where guidance can be applied. Note we made an error in our original manuscript; there are actually $\binom{11}{2} = 55$ possible intervals instead of $\binom{10}{2} = 45$ because there are 11 possible interval points $t = 0, 100, \ldots, 900, 1000$. Each result is named by the start and end point of guidance; for example, the result with guidance using the new prompt from steps 600 to 400 is labeled `600_400`. While the figure shows results for ModelScope, we also provide them for the same process using Latte. While the new spatial information of a cat does not transfer as well, the main property underlying our paper of motion preservation by timestep $t = 600$ still holds.

Folder `fig4` (Fig. 4) contains the videos for both prompts tuned with all timesteps, selected timesteps, and selected timesteps only on OV layers.

Folder `fig5` (Fig. 5) contains videos for the original reference along with samples from each LoRA trained model.

Folder `fig6` (Fig. 6) contains videos for the original reference, LoRA trained sample, and direct trained sample.

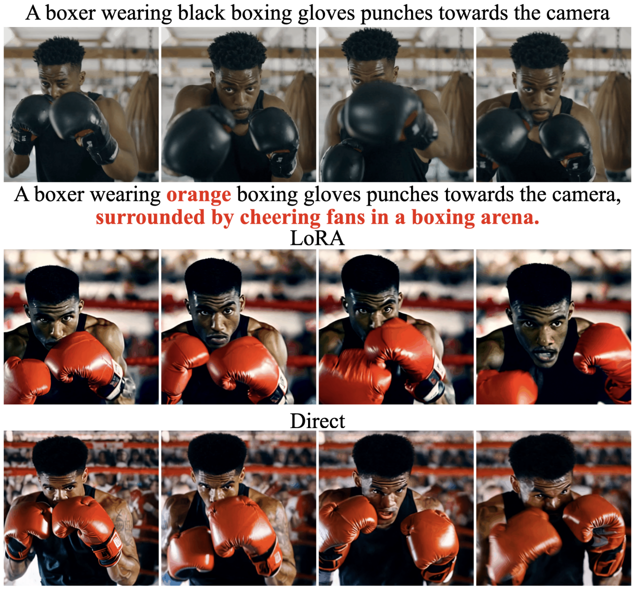
A boxer wearing black boxing gloves punches towards the camera
A boxer wearing orange boxing gloves punches towards the camera, surrounded by cheering fans in a boxing arena.
LoRA
Direct

Figure 7. Additional results of our direct tuning method in comparison to LoRA. The videos are sampled after 200 steps with LoRA and 150 steps with direct training. The direct tuning method reliably captures the motion with the same fidelity as LoRA without overfitting.

Method	Ours	MotionDirector[46]	Tune-A-Video[37]
Text Alignment(↑)			
Style	27.73	27.38	26.55
Object	26.74	28.15	27.32
Background	28.42	29.49	26.60
Multiple	28.37	27.24	28.05
Temporal Consistency(↑)			
Style	95.17	96.48	95.73
Object	95.16	96.49	94.35
Background	95.13	95.03	96.54
Multiple	95.40	95.69	95.35
Pick Score(↑)			
Style	20.05	20.19	19.71
Object	19.86	20.32	19.94
Background	20.34	20.35	19.79
Multiple	20.41	19.91	20.07

Table 3. Comparison of metrics across different prompt changes for our method, MotionDirector [46], and Tune-A-Video[37]

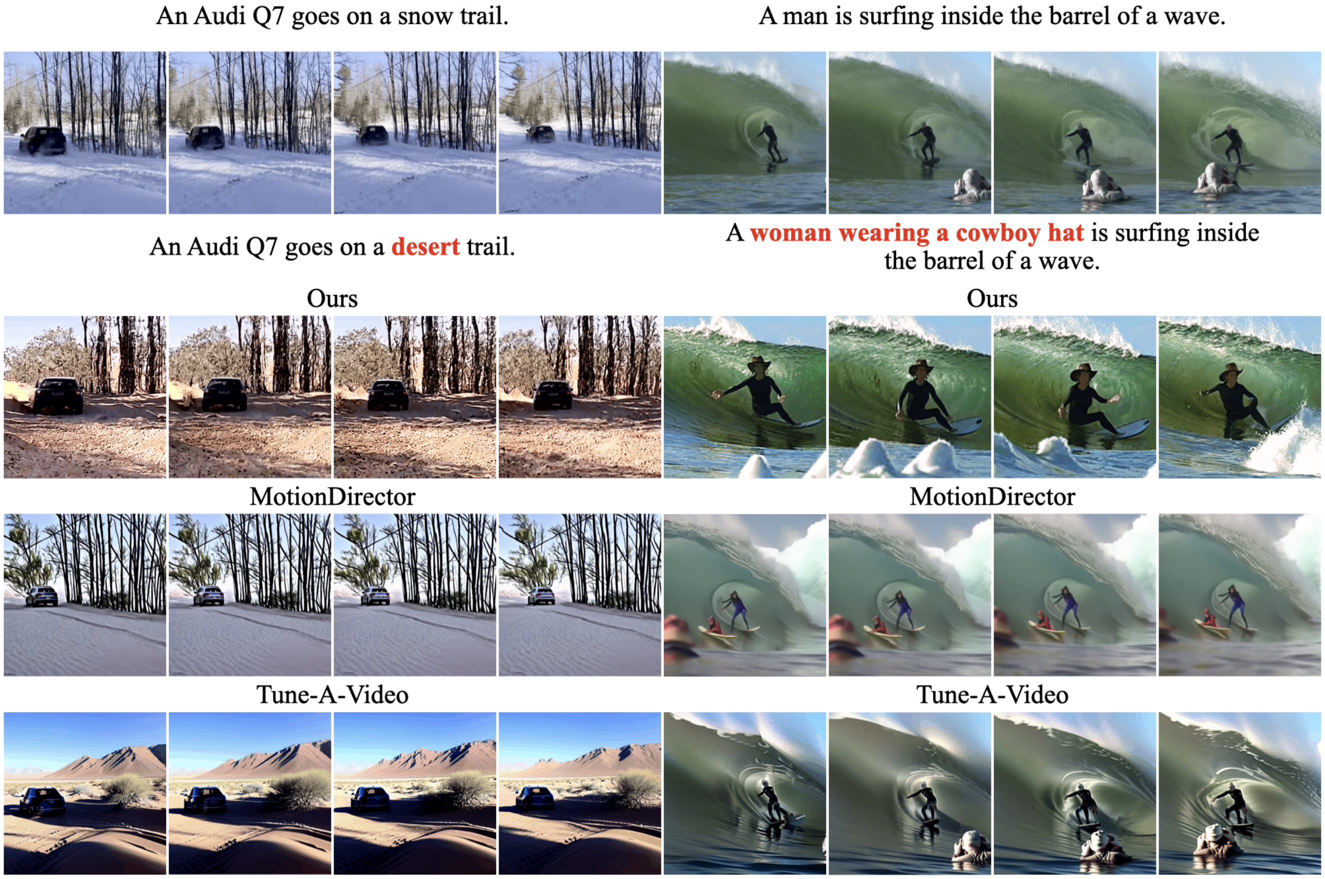
An Audi Q7 goes on a snow trail.
A man is surfing inside the barrel of a wave.
An Audi Q7 goes on a **desert** trail.
A **woman wearing a cowboy hat** is surfing inside the barrel of a wave.
Ours
Ours
MotionDirector
MotionDirector
Tune-A-Video
Tune-A-Video

Figure 8. Comparison of our methods to MotionDirector [46] and Tune-A-Video [37], which are concurrent one-shot motion customization methods. Our method can transfer motion to a variety of new backgrounds and subjects. Notably, in the left example our method changes the background to a desert trail without any visual artifacts. In the right example, we capture the surfing motion for a new subject without overfitting to the other person in the background.

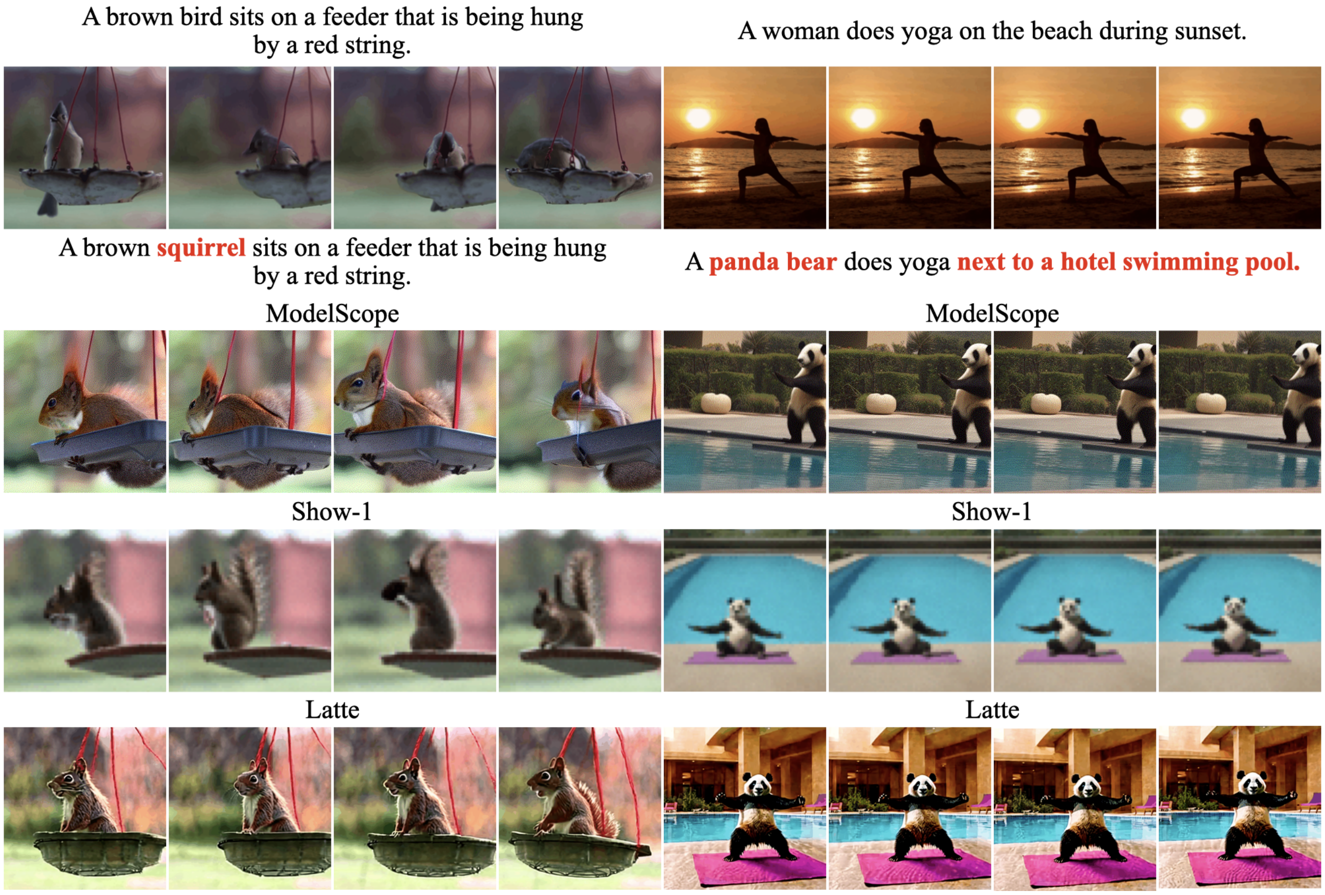
A brown bird sits on a feeder that is being hung by a red string.
A woman does yoga on the beach during sunset.
A brown **squirrel** sits on a feeder that is being hung by a red string.
A **panda bear** does yoga **next to a hotel swimming pool.**
ModelScope
ModelScope
Show-1
Show-1
Latte
Latte

Figure 9. Additional results across multiple base models. Our LoRA training method for temporal attention layers seamlessly adapts to multiple base models like ModelScope [33], Show-1 [44], and Latte [19]. In all instances, we are able to capture the motion and transfer it to new prompts.