Video Diffusion Models Encode Motion in Early Timesteps

Vatsal Baherwani[*], **Yixuan Ren**[*], **Abhinav Shrivastava**

University of Maryland

Abstract

Text-to-video diffusion models synthesize spatial appearance and temporal motion by progressively denoising from noise, yet the distribution of these two dimensions across timesteps is not fully characterized. We present a systematic and quantitative study that measures how appearance editing and motion preservation trade off when new conditions are applied over specified ranges of timesteps. Across diverse architectures, we observe a consistent pattern in which motion is established in early steps and appearance is refined in later steps, yielding an operational boundary in timestep space that disentangles temporal and spatial factors. Building on this property, we introduce a one-shot motion customization method that restricts training and inference to early steps and achieves strong motion transfer without auxiliary debiasing modules or specialized objectives. Our spatiotemporal disentanglement property can facilitate broad applications of appearance or motion transfer and editing, and our timestep-constrained method can be easily integrated into other motion customization methods.

Introduction

Diffusion models (Ho, Jain, and Abbeel 2020) have achieved remarkable performance in image and video synthesis with high quality and extensive generalization. Large-scale pretrained foundation models have facilitated many downstream applications of controllable generation, such as editing (Zhang and Agrawala 2023; Tumanyan et al. 2023) and customization (Ruiz et al. 2023; Gal et al. 2022). In contrast to images, videos consist of additional temporal information to spatial features, and their decoupling becomes critical as different tasks demands tampering on different aspects of the media. On the other hand, the progressive process of spatial and temporal signals along the denoising process makes it challenge to decompose them.

Many effort has been made toward analyzing and extracting certain spatial attributes from image diffusion models at different timesteps (Hertz et al. 2022a; Luo et al. 2023a; Yue et al. 2024a; Qian et al. 2024; Lee et al. 2025a; Liang et al. 2025). However its temporal counterpart is less revealed. (Xiao et al. 2024; Li et al. 2024; Wu et al. 2025a) have observed that motions are usually constructed at early denoising timesteps. However, there still lack systematic and quantitative analyses. In this work, we look to decompose the appearance and motion of text-to-video diffusion models. Specifically, we explore the distribution of spatiotemporal information along the diffusion timesteps.

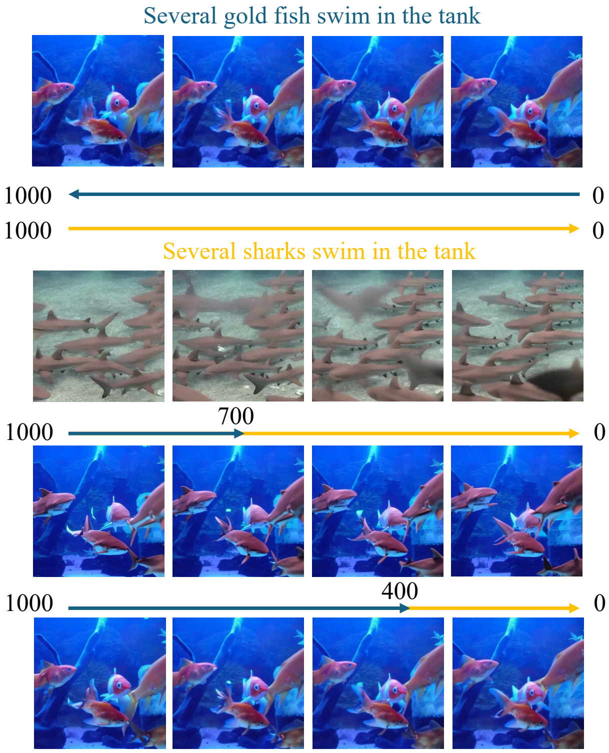

Several gold fish swim in the tank
1000
0
1000
0
Several sharks swim in the tank
1000
700
0
1000
400
0

Figure 1: Spatiotemporal disentanglement in video diffusion models. Our finding reveals that motion is primarily encoded in the early denoising timesteps. Given a reference video (top) and its ground truth caption (blue), we perform DDIM inversion and then denoise with a new prompt that modifies only the subject (yellow). The resampled videos show different subject editing and motion preservation results by applying the original or new prompts at different timesteps.

Our analysis primarily involves comparing between clean and noisy distributions across different timesteps. Given the difficulty to measure the appearance and motion similarity directly on noised samples, we design a preliminary exper-

[*] These authors contributed equally.

iment that leads to tampered RGB videos. Specifically, we invert the original video to noisy latent, and then resample it with a new text prompt whose appearance description is changed over certain denoising timesteps. While such method is not able to precisely serve as a high-quality editing tool (as demonstrated in Mokady et al.), it indicates the impact of appearance construction as distributed along the denoising process. Our findings confirm that spatial information is destroyed first when diffusing and construct last when denoising, while temporal information is on the contrary. Notably, these dimensions are decoupled with a clear threshold, enabling us to manipulate one attribute while preserving the other.

We validate the spatiotemporal disentanglement property on various denoising network architectures: ModelScope (Wang et al. 2023a), which is U-Net based with dedicated spatial and temporal attentions, Latte (Ma et al. 2024), which is transformer based with dedicated spatial and temporal attentions, and CogVideoX (Yang et al. 2024), which is transformer based with unified spatiotemporal attentions, and they display consistent behaviors. We attribute the findings to the association of most motions to non-local structure movement: as Gaussian noise first filters out high frequency and diffusion models first construct low frequency signals such as global layouts in images when denoising, motions are encoded at the early timesteps in video synthesis. Compared to prior work that explicitly control the video editing with exemplar depth, edge or optical flow (Chen et al. 2023; Yang et al. 2023), our findings enable modeling the reference motion along a continuous range of timesteps and reproducing it with temporal diversity such as movement velocity, intensity and object position, camera perspective etc.

By exploiting our spatiotemporal disentanglement property we facilitate one-shot video motion customization as a downstream application, where the reference motion is expected to be transferred to new subjects and scenarios with temporal variety. Given only a single reference video during training, this task is known to suffer from appearance overfitting when modeling the desired motion, and prior approaches usually rely on auxiliary modules or reformed losses to debias the unwanted spatial signals (Zhao et al. 2023b; Ren et al. 2024). In contrast, we propose to solely finetune the motion customization module with the vanilla diffusion loss while constraining the range of training and inference timesteps to the early denoising stage. This timestep constraint prevents appearance leakage from the reference video despite the fusion of spatial attention and full reconstruction loss and ensures the customization module captures motion information only.

In summary, our main contributions are the following:

- We demonstrated the architecture-agnostic spatiotemporal disentanglement along diffusion timesteps in pre-trained video diffusion models, that motions are synthesized in early denoising stage;
- Leveraging this property, we proposed a simple yet effective one-shot motion customization method that constrains the training and inference timestep range, obviating any extra spatial debiasing module or loss and easy to be integrated into existing pipelines;
- We further extend our method to partial attention tuning and direct tuning, enabling more efficient and flexible motion customization paradigms.

Related Works

Diffusion Attribute Disentanglement

In the image domain, a number of studies analyze what different timesteps encode along the reverse process. Luo et al. (2023b) aggregate multi-timestep and multi-scale features and show complementary geometric and semantic cues for correspondence. Stepwise spectral analyses report that low-frequency content changes dominate at earlier steps while high-frequency refinements appear later, which motivates non-uniform sampling (Lee et al. 2025b). Beyond observation, several methods make timesteps an explicit supervision axis. Yue et al. (2024b) learn timestep-aware representations grounded in how attributes vanish during the forward noising. Step-aware preference alignment allocates feedback to specific steps to better match human perception (Chen et al. 2024; Sun et al. 2024). Editing frameworks that intervene per-step further support the utility of stepwise control for separating layout from style (Hertz et al. 2022b).

In video diffusion, explicit evidence about timesteps is emerging but remains less developed than in images. Xiao et al. (2024) extract motion-aware features from pre-trained text-to-video models and operate along the denoising trajectory to guide motion without training, offering operational support that early steps are effective for shaping coarse motion and later steps can refine appearance. Personalization work emphasizes preserving native motion while injecting identity. Li et al. (2024) introduce an isolated identity adapter to maintain dynamics and semantics during customization, which is consistent with scheduling appearance-oriented updates late in sampling even though they do not quantify a step boundary. Customization with explicit early–late scheduling is exemplified by Wu et al. (2025a), who reduce subject-learning influence in the early denoising stage to preserve motion and restore it in the late stage to recover appearance details. Our study complements these efforts by providing a systematic and quantitative delineation of when temporal motion and spatial appearance dominate along the denoising timeline of pre-trained video diffusion models.

Video Motion Customization

Video motion customization is the task to learn the motion from the reference videos and adapt it to new subjects and scenarios. Prior work has been developed for deterministic video editing or motion transfer, which leverage global structure control such as edge or depth map (Chen et al. 2023; Zhang et al.; Zhao et al. 2023a), optical flow (Yang et al. 2023; Liang et al. 2024), and latent features (Geyer et al. 2023; Ling et al. 2024). In contrast, video motion customization finetunes a pre-trained text-to-video diffusion model to adopt the desired motion and transfer it with both temporal fidelity and diversity.

However, it is a common challenge that videos consist of both spatial and temporal signals and video diffusion models process them in fusion. This issue is more significant in one-shot case since multiple reference videos sharing the same motion concept provide diverse appearances that can offset each other. To capture pure motion information from the reference videos, various approaches have been proposed. Zhao et al. (2023b); Ren et al. (2024) incorporate an additional appearance debiasing module to exclude the unwanted appearance. Wu et al. (2025b) designs a temporal-only loss based on the latent frame features.

In comparison to prior work, we leverage our main finding to demonstrate a video motion customization application in one shot without auxiliary modules or special loss functions. We instead achieve motion decoupling by constraining the tuning and inference denoising timesteps to exploit our spatiotemporal disentanglement property in pre-trained video diffusion models.

Spatiotemporal Disentanglement along Denoising Timesteps

Preliminary

Diffusion Models Diffusion models (Ho, Jain, and Abbeel 2020) generate synthetic instances by sampling $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ and iteratively applying a denoising process to obtain $\mathbf{x}_0$ via

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t, c)\right) + \sigma_t \mathbf{z}, \qquad (1)$$

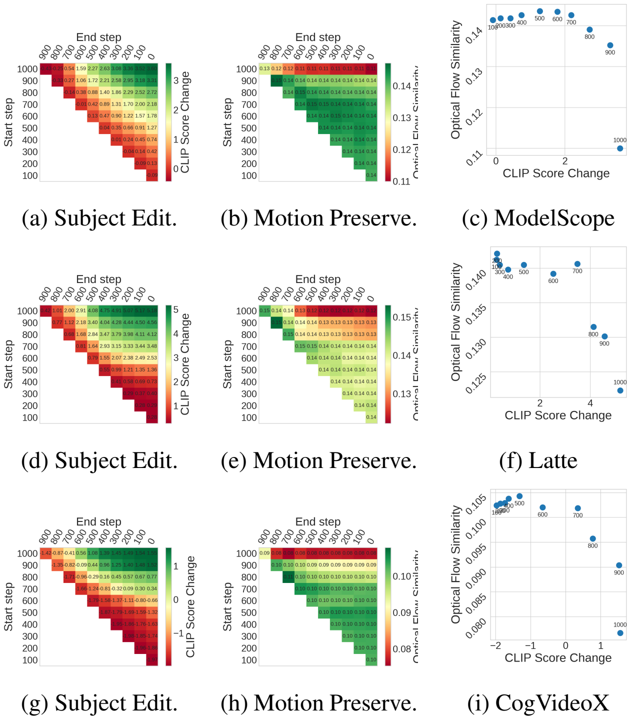
where $t = T, ..., 1$. ϵ_θ is a parameterized denoising neural network with a condition c, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ is random noise, σ_t is the variance, and $\alpha_t, \bar{\alpha}_t$ are hyperparameters defining the noise schedule.

Text-to-Video Diffusion Models In text-to-image diffusion models, c is a text prompt depicting the expected output video, and a typical ϵ_θ comprises self-attentions and cross-attentions to process the visual information with the condition incorporated. To synthesize sequential data consisting of multiple images, ϵ_θ additionally involves cross-frame attentions to regularize the temporal consistency.

DDIM Inversion In implicit diffusion models (DDIMs, Song, Meng, and Ermon), the denoising process in Eq. 1 can be made deterministic by setting $\sigma_t := 0$. Then the denoising process can be inverted by expressing x_t in terms of x_{t-1} (Mokady et al. 2022), and ultimately producing from an existing $\mathbf{x}_0$ its approximate sampling trajectory $\mathbf{x}_{\{T,...,1\}}$, which reconstructs itself following the denoising process.

Analysis Design

We aim to observe how the spatial and temporal attributes of a video are processed at various timesteps in the diffusion and denoising processes. However, this is not trivial as categorizing appearance and motion can be ambiguous in general. And understanding from noisy videos at intermediate diffusion timesteps further lifts its difficulty. Therefore, we design to leverage the inversion approach and tamper the resampling trajectory for feasible calculation and reference.



(a) Subject Edit. (b) Motion Preserve. (c) ModelScope

(d) Subject Edit. (e) Motion Preserve. (f) Latte

(g) Subject Edit. (h) Motion Preserve. (i) CogVideoX

Figure 2: Subject editing and motion preservation quality of ModelScope, Latte and CogVideoX. Applying the new subject editing prompt in longer timesteps always leads to stronger new subject representation in the generated video. However, starting resampling with the new prompt at early timesteps significantly harms the motion preservation although it doesn't modify the motion description. The trade-off curves show the optimal timesteps to decompose spatial and temporal signals. This spatiotemporal property holds consistently across different model architectures.

Specifically, given a video x_0 and its ground truth caption c, we start from DDIM inversion to acquire its noise latent $\hat{x}_T$, such that the denoising network θ can faithfully recover it via the original trajectory $x_0 = \prod_{t=T}^{1} \theta(\hat{x}_t|t, c)$. Next, we tamper of c to c' by changing its subject, and perform denoising process with the edited condition $x_0' = \prod_{t=T}^{1} \theta(\hat{x}_t|t, c')$. While x_0' is ideally expected to represent the new subject with the original motion preserved as indicated by c', this process will in fact intervene the generated motion as well. For example, [show a figure or subfigure for this].

Based on this, we propose to examine how the denoising timesteps interact with the new text prompt to synthesize new appearance and original motion. To this end, we perform the resampling process with c' in a certain timestep range, and the original c is used outside. Formally, we denoise via $x_0'' = \prod_{t=T}^{1} \theta(\hat{x}_t|t, c_t'')$, where $c_t'' = c'$ when $t \in [\tau_{\text{start}}, \tau_{\text{end}}]$ and otherwise $c_t'' = c$. Then we measure the appearance editing by the CLIP score (Hessel et al. 2021) between x_0'' and c', and measure the motion preservation by the optical flow similarity between x_0'' and x_0.

In this way, we leverage the text captions as comprehensive spatiotemporal labels that are clear and easy to manipulate, and obviate direct calculations on noisy videos or compare across different noise levels via diffusion inversion and resampling in clean latent distribution. Note that although this naive resampling is not able to perfectly edit the original video reasonably and realistically, it can serve as an analytic approach to exhibit the difference in spatial and temporal impact across timesteps in our evaluation.

Experiment Setup

We consider full combination of all valid $(\tau_{start}, \tau_{\text{end}})$ pairs with an interval of 100 over the whole 1000 timesteps. A visual example of this approach is shown in 1. Here we use start timestep $\tau_{start} = 600$ and end timestep $\tau_{\text{end}} = 0$. As a result, our newly generated video preserves the information from $t \in [600, 1000]$ in the original video.

To fully reflect the editing improvement, we meaure the CLIP score change where the base score between x_0 and c' is subtracted, as x_0 already have some resemblance to c' except the tampered subjects. We use the Lucas-Kanade method for optical flow estimation, and calculate the average cosine similarity between the normalized vectors of all frames. Both metrics are higher when the new video x'_0 better represents the new subject in c' or better preserves the original motion in x_0.

We conduct this experiments on three representative text-to-video models with divergent denoising network architectures: ModelScope (Wang et al. 2023a) with U-Net and dedicated spatial and temporal attentions, Latte (Ma et al. 2024) with transformer and dedicated spatial and temporal attentions, and CogVideoX (Yang et al. 2024), with transformer and unified spatiotemporal attentions. We test on all 76 videos from the Text-Guided Video Editing (TGVE) competition dataset (Wu et al. 2023), which also provides subject editing captions.

Results

In Fig. 2 we show the trade-off between CLIP score change and optical flow similarity across all $(\tau_{start}, \tau_{\text{end}})$ options. The CLIP score change consistently improves whenever the editing interval $\tau_{start} - \tau_{\text{end}}$ is longer, as this allows for more sampling steps with the new prompt c'. Notably, for any given τ_{start}, the optimal τ_{end} is always 0. However, τ_{end} does not matter as much for motion preservation. On the contrary, the optical flow similarity increases as we delay the sampling process to start from later timesteps. In other words, sampling with the new condition c' at earlier timesteps, harms much its optical flow similarity to the original video despite c''s only modification on the subject. Based on the observed effect of the subject editing prompt of motion deviation from the original video, we claim that motion signals are dominantly encoded in early denoising timesteps in video diffusion models.

We draw the heatmaps of the appearance editing and motion preservation quality in Fig. 2. We can deduce from it the dominant ranges of motion and appearance along the denoising timesteps for each pre-trained model. τ_{end} is not significant for motion preservation while being optimal for appearance editing at 0, at which we therefore fix the end timestep.

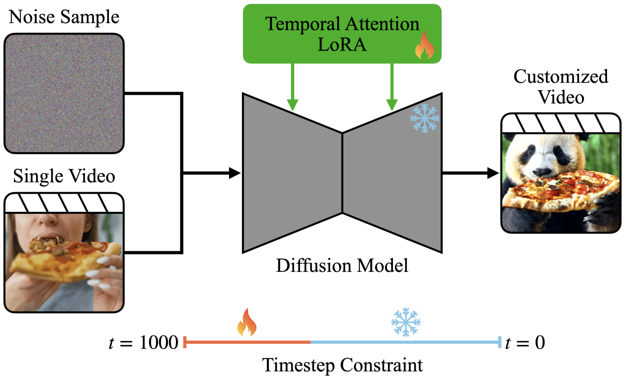


Figure 3: One-shot video motion customization via denoising timestep constraint. Leveraging our spatiotemporal disentanglement property, we train LoRAs at only early denoising timesteps to model the reference motion without appearance leakage. This single-stage fine-tuning approach achieves surpassing performance without any additional debiasing modules, stages or losses. This even works for base models with unified spatiotemporal attentions, where we add LoRA on the full spatiotemporal sequence and it is still prevented from overfitting on the reference appearance.

Given $\tau_{\text{end}} := 0$, varying the start timestep τ_{start} presents a trade-off between representing the new subject and retaining the original motion. That is, τ_{start} reflects the threshold of denoising timesteps where temporal and spatial signals are encoded. This tradeoff is also depicted in Fig. 2 for each base model. A smaller τ_{start} leads to minimal shift in optical flow similarity, while CLIP score improves significantly. A bigger τ_{start} results in drastic loss in the motion information from the original video. The threshold timestep for spatiotemporal disentanglement thus lies somewhere along the Pareto frontier. In following sections we denote $\tau = \tau_{start}$ as this threshold. While its exact value varies across specific models, it is consistently around 700 to 900.

Next, we demonstrate our spatiotemporal disentanglement property in the downstream application of one-shot video motion customization task.

One-Shot Video Motion Customization

Task Settings

Video motion customization is the task to customize a pre-trained text-to-video diffusion model with specific motions from given reference videos. Given the ambiguity of text prompt control of temporal movements, motion customization is the optimal way to replicate the exemplar motions with new subjects and scenes. Previous methods of video editing and motion transfer aim at generating deterministic movements with precise frame-wise alignment, losing temporal diversities such as motion velocity, intensity, subject count and position, and camera perspective etc. In contrast, motion customization demands tuning-based modeling of the desired motions and leads to reproducing them with temporal varieties, and thus achieves broader generalization to

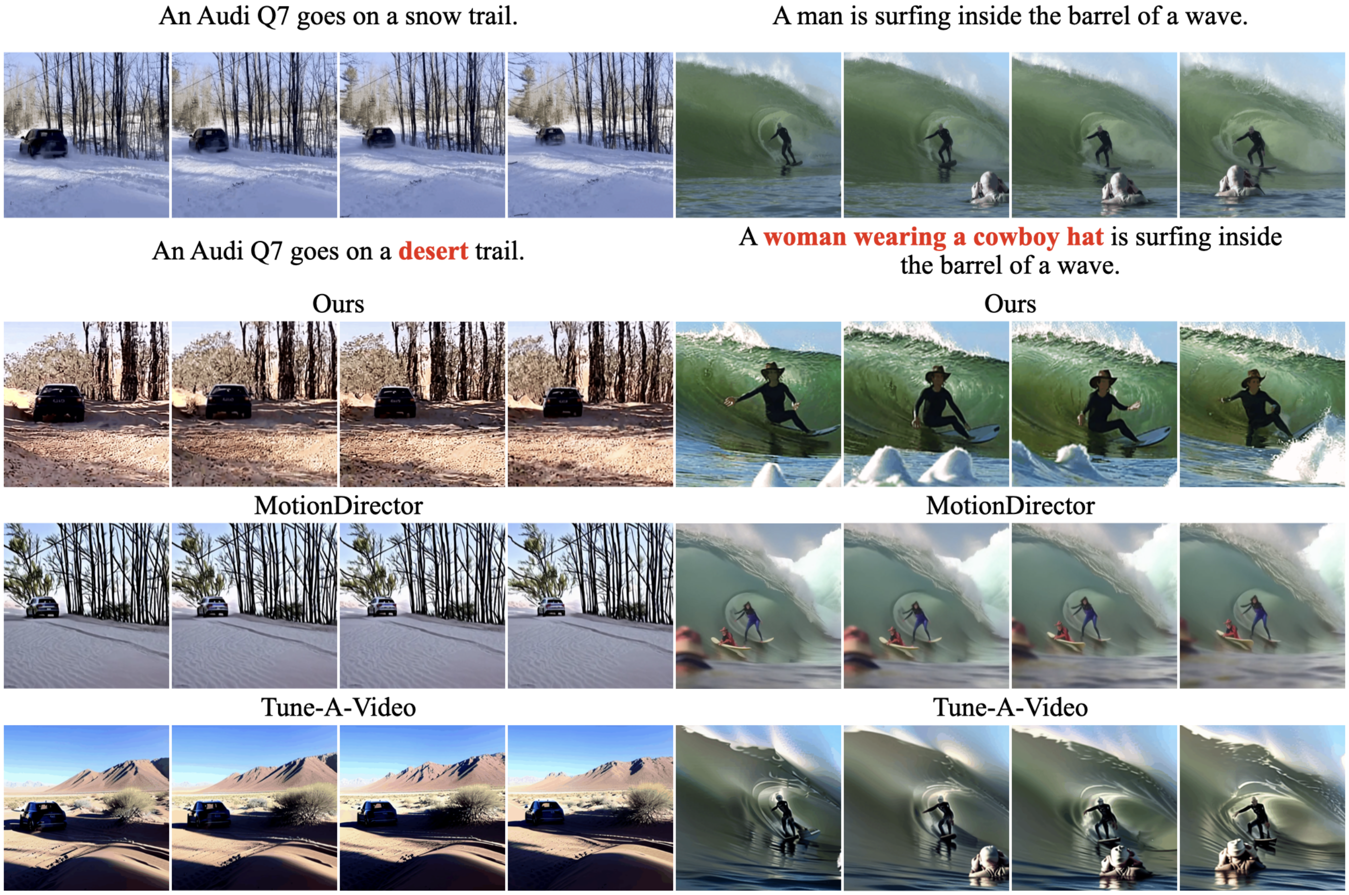

An Audi Q7 goes on a snow trail.
A man is surfing inside the barrel of a wave.
An Audi Q7 goes on a desert trail.
A woman wearing a cowboy hat is surfing inside the barrel of a wave.
Ours
Ours
MotionDirector
MotionDirector
Tune-A-Video
Tune-A-Video

Figure 4: Qualitative comparison of our motion disentanglement method to MotionDirector and Tune-A-Video. Our method faithfully replicates the motion of the reference video while also editing the subject and background with superior quality to other approaches. Without any additional spatial debiasing modules or stages, our method is stable and robust with minimal semantic discrepancy (e.g. the snow ground and hat-like reef by MotionDirector).

fit on more diverse new subjects and scenes, similar to image customization over deterministic patch stitching.

In our application, we focus on the one-shot customization case, where only one reference video is provided. The main challenge in one-shot motion customization is modeling the reference motion without overfitting on the given appearance. Tuning on multiple videos with the similar motion concept and diverse appearances, the customization module will converge fast on the common information, i.e. the motions, while it learns both spatial and temporal signals with the vanilla diffusion loss when training on a single video. Leakage of the unwanted appearances into the motion customization module will result in their deterministic reproduction in the generated videos, harming the freedom of synthesizing novel spatial attributes with new prompts. Leveraging our spatiotemporal disentanglement property of video diffusion models, we develop a targeted training method circumventing these issues to achieve high quality one-shot motion customization with largely simplified tuning modules and pipelines.

Timestep Constrained Method

Prior diffusion-based motion customization methods typically apply LoRA on pre-trained temporal attention layers, and finetune it across all timesteps $t \in [1000, 0]$. Based on the spatiotemporal disentanglement along timesteps in video diffusion models, where the motion information is primarily processed in early denoising timesteps, we propose to train the temporal LoRA with the groud truth caption in a restricted timestep range $t \in [1000, \tau]$. τ is the aforementioned threshold between spatial and temporal signals along the denoising process. We also constrain the LoRA application during inference within the same timestep range, and at other timesteps the denoising process is proceeded with solely the base model. It worth noting that the text prompt remains the same new prompt with modified appearances and original motions throughout the inference.

The overall pipeline of our method is illustrated in Fig. 3. Compared to previous methods that have to incorporate with auxiliary modules, stages or losses to explicitly debias the appearance learning out of the temporal tuning, our method simplifies the pipeline to only one single temporal LoRA

Base Model	τ	Text Align.↑	Temp. Const.↑	Pick Score↑
ModelScope (2023a)	1000	26.05	94.88	20.13
	750	28.04	96.39	20.68
	700	**28.16**	**96.42**	20.77
	650	27.97	96.31	**20.79**
	0	27.43	96.25	20.49
Latte (2024)	1000	29.28	93.16	20.84
	750	31.85	97.12	21.65
	700	**31.96**	97.19	**21.68**
	650	31.88	**97.21**	21.66
	0	31.26	96.99	21.47
CogVideoX (2024)	1000	28.15	96.69	20.65
	950	**30.14**	**98.11**	21.09
	900	29.93	98.10	21.00
	850	29.61	97.76	20.92
	0	29.67	97.41	**21.30**

Table 1: Ablating different timestep tuning range τ for one-shot video motion customization, where the base model is tuned at $t \in [1000, \tau]$. A smaller τ corresponds to a wider range of denoising timesteps for finetuning. $\tau = 1000$ refers to the base model without tuning, and $\tau = 0$ refers to tuning the base model at all timesteps. The optimal τ for the downstream task aligns with the peak in our analysis in Fig. 2.

Method	Text Align.↑	Temp. Const.↑	Pick Score↑
Tune-A-Video (2022)	25.64	92.42	20.09
VideoComposer (2023b)	27.66	92.22	20.26
Control-A-Video (2023)	26.54	92.63	19.75
VideoCrafter (2023)	28.03	92.26	20.12
MotionDirector (2023b)	27.82	93.00	20.74
VMC† (2023)	25.53	94.58	19.92
Gen-1 (2023)	28.54	95.77	-
MotionClone† (2024)	27.23	92.88	21.07
MotionMatcher (2025b)	30.43	97.20	-
Ours-ModelScope	28.16	96.42	20.77
Ours-Latte	**31.96**	97.19	**21.68**
Ours-CogVideoX	30.14	**98.11**	21.09

Table 2: Comparison with previous SOTA motion customization methods on the TGVE benchmark. Our timestep constraining method achieves leading performance without auxiliary modules or stages, and is also compatible to be integrated with existing pipelines. † denotes the methods that were tested on other datasets and we re-evaluated on the TGVE benchmark for fair comparison.

module, one single tuning stage and the vanilla diffusion reconstruction loss. We also show that our simplified pipeline further facilitates flexible model parameter configurations with stable tuning and consistent performance with minimum appearance leakage. Furthermore, since our method only constrains the training timesteps, it is very easy to cooperate with other pipelines without any conflict of tuning models or objectives.

Experiment Setup

Base models. We implement our training method on three base T2V models: ModelScope (Wang et al. 2023a), Latte (Ma et al. 2024), and CogVideoX (Yang et al. 2024). All generate videos of 2 seconds and 16 frames, with 256×256 resolution for ModelScope, 512×512 resolution for Latte, and 480×480 for CogVideoX. They all use a DDPM noise scheduler with totally 1000 timesteps.

Datasets. To quantitatively evaluate our approach, we apply motion customization on all 76 videos in the Text-Guided Video Editing (TGVE) competition dataset (Wu et al. 2023) individually. We use the ground truth captions as the training prompts and all 4 editing captions to synthesize novel videos at inference.

Metrics. We evaluate our generated videos by the following metrics: Text alignment calculates the CLIP Score (Hessel et al. 2021) between the video frames and the new prompts to measure the fidelity of the spatial attributes following the descriptions at inference. Temporal consistency averages the pairwise CLIP embedding distances between consecutive frames. Pick Score (Kirstain et al. 2023) trained a model to emulate human preferences of prompt alignment. Every editing prompt produces 4 random samples, on which the metrics are averaged over.

Results

We experiment with choices for the temporal tuning threhshold τ in our motion customization method. We present these results in Tab. 1, using LoRA fine-tuning with a rank and alpha $r = \alpha = 4$. It displays that the optimal τ consistently align with the peak threshold of the spatiotemporal decomposition property in Fig. 2 for each base model. Meanwhile, the precise value of τ does not make a significant difference for the final motion customization performance around the optimum, demonstrating the robustness and generalization of our method for practical use.

ModelScope and Latte have separate spatial and temporal attentions in their denoising networks, while ModelScope denoises with U-Net and Latte denoises with transformer. The overall performance of Latte surpasses ModelScope due to its advanced architecture and larger model size. CogVideoX is built with unified 3D spatiotemporal attentions, which natively deepen the entanglement of appearance and motion information. Despite this, our timestep constrained method still achieves leading performance at $\tau = 950$ over all other configurations. This value is significantly larger than other base models as the core motion signals need to be decomposed with a stronger constraint.

In addition, we also list the performance of two baselines for each base model: tuning at all timesteps without a constrained range ($\tau = 0$), and the base model without any tuning ($\tau = 1000$). Their performance gaps behind our timestep constrained method indicate the effectiveness of tuning the motion customization module only at early timesteps, where

Tunable Layers	Text Alignment↑	Temporal Consistency↑	Pick Score↑
Q, K, V, O	31.69	97.19	21.68
V, O	32.64	97.16	21.62

Table 3: Ablating temporal attention layers with Latte at $\tau = 700$. By only fine-tuning value and output projections in each attention layer, we cut the number of trainable parameters in half and achieve essentially comparable results.

motion information is dominantly encoded.

Comparison to Prior SOTAs. We compare our method with various base models at their optimal τ to other one-shot motion customization approaches that have reported metrics on the TGVE dataset. Our motion customization approach yields superior quantitative results to prior SOTAs with a much simplified tuning module and pipeline. Fig. 4 exhibits a visualization of the qualitative comparison. Our method transfers the reference motion to new subjects and backgrounds with minimal semantic discrepancy compared to other approaches.

Downstream Extensions

Ablating Attention Layers. Based on our findings of motion disentanglement across timesteps, we are interested in exploring whether motion control can be limited to specific model parameters as well. Given the four query, key, value, and output projections of temporal attention layers, we experiment with restricting training to all possible subsets of these parameters. From our results in Tab. 3, we see that only training the value and output projections is necessary for motion customization. In our experiments, we also observe that training only the query and key parameters yields no noticeable change in the generated videos. This suggests that the query and key parameters in temporal attention layers are not responsible for encoding motion information. This allows for cutting the number of trainable parameters in half without sacrificing generation quality.

Scaling LoRA Rank and Direct Tuning. Prior work usually suffers from increased temporal LoRA rank, as more tunable parameters will more easily overfit on unwanted appearances from the single reference video. We scale the LoRA rank up to $r = 16$. Moreover, we further extend our method to direct full-parameter fine-tuning. Previous successful approaches for direct training follow DreamBooth (Ruiz et al. 2023) and require multiple reference samples, as well as a regularization set of general data, to avoid both overfitting on the exemplar appearances or motions. We instead maintain our settings of only tuning the attention layers on a single reference video, without any additional data. The direct tuning can be viewed as a full-rank upper bound where the LoRA rank scales to the same as that in the pretrained base model.

We present the results in Tab. 4. It contradicts the trivial hypothesis that more parameters always lead to improved one-shot motion customization results. We attribute this to the limited motion information in a single video, which doesn't need many parameters to model. On the other hand, this observation also demonstrates the clear spatiotemporal disentanglement of our method, where no appearance is leaked into the tunable module even when much more than necessary parameters are being tuned with the full reconstruction denoising loss, in contrast to traditional DreamBooth pipeline.

LoRA Rank	CLIP Score↑	Temporal Consistency↑	Pick Score↑
$r = \alpha = 4$	31.69	97.19	21.68
$r = \alpha = 8$	31.61	97.17	21.63
$r = \alpha = 16$	31.34	97.12	21.57
All attentions	31.19	97.23	21.46

Table 4: Scaling up LoRA ranks and direct full-rank tuning with Latte at $\tau = 700$. While more tunable parameters contribute marginally to motion customization quality improvement due to limited temporal signals to model in a single video, our spatiotemporal disentanglement property consistently prevent additional parameters from overfitting on the appearance in the reference video.

Future Work

In this work we focus on the binary disentanglement between appearance and motion information along the timesteps. This separation facilitates downstream tasks that demands dedicated processing of either spatial or temporal signals. Given the fact that Gaussian noise gradually destroy from high to low frequency in diffusion, similar to the global and local spatial features in image models, different categories of motions may be encoded at relatively different timesteps within the overall early stage. Such finer-grained temporal processing can further enable more precise motion editing and transfer. We leave this to future exploration.

Conclusion

We investigated how temporal and spatial information are organized along the denoising trajectory of text-to-video diffusion models. By partially resampling with controlled prompt edits and evaluating appearance editing against motion preservation, we showed that motion is encoded primarily in early timesteps whereas appearance is consolidated in later timesteps. This behavior emerges consistently across different architectures, which enables a practical separation of temporal and spatial factors in timestep space. Building on this insight, we proposed a simple one-shot motion customization procedure that constrains both training and inference to early steps and attains high-quality motion transfer without auxiliary modules or tailored losses. These results demonstrate that timestep-aware scheduling is an effective and broadly applicable lever for control and adaptation in video diffusion models.

References

Chen, B.; Jiang, D.; Shi, C.; Ji, L.; Wang, Y.; Yan, S.; Wei, Z.; Lin, D.; and Zhang, H. 2024. Aligning Preference with Denoising Performance at Each Timestep. In *NeurIPS*.

Chen, W.; Wu, J.; Xie, P.; Wu, H.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. *arXiv preprint arXiv:2305.13840*.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618.

Geyer, M.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2023. VideoCrafter: A Toolkit for Text-to-Video Generation and Editing. https://github.com/AILab-CVC/VideoCrafter.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022a. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022b. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv:2208.01626*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jeong, H.; Park, G. Y.; and Ye, J. C. 2023. VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models. arXiv:2312.00845.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*.

Lee, H.; Lee, H.; Gye, S.; and Kim, J. 2025a. Beta Sampling is All You Need: Efficient Image Generation Strategy for Diffusion Models Using Stepwise Spectral Analysis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4215–4224. IEEE.

Lee, H.; Lee, H.; Gye, S.; and Kim, J. 2025b. Beta Sampling is All You Need: Efficient Image Generation Strategy for Diffusion Models using Stepwise Spectral Analysis. In *WACV*.

Li, H.; Qiu, H.; Zhang, S.; Wang, X.; Wei, Y.; Li, Z.; Zhang, Y.; Wu, B.; and Cai, D. 2024. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. *arXiv preprint arXiv:2411.17048*.

Liang, F.; Wu, B.; Wang, J.; Yu, L.; Li, K.; Zhao, Y.; Misra, I.; Huang, J.-B.; Zhang, P.; Vajda, P.; et al. 2024. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8207–8216.

Liang, Z.; Yuan, Y.; Gu, S.; Chen, B.; Hang, T.; Cheng, M.; Li, J.; and Zheng, L. 2025. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13199–13208.

Ling, P.; Bu, J.; Zhang, P.; Dong, X.; Zang, Y.; Wu, T.; Chen, H.; Wang, J.; and Jin, Y. 2024. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*.

Luo, G.; Dunlap, L.; Park, D. H.; Holynski, A.; and Darrell, T. 2023a. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36: 47500–47510.

Luo, G.; Dunlap, L.; Park, D. H.; Holynski, A.; and Darrell, T. 2023b. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. In *NeurIPS*.

Ma, X.; Wang, Y.; Jia, G.; Chen, X.; Liu, Z.; Li, Y.-F.; Chen, C.; and Qiao, Y. 2024. Latte: Latent Diffusion Transformer for Video Generation. arXiv:2401.03048.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models. arXiv:2211.09794.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6038–6047.

Qian, Y.; Cai, Q.; Pan, Y.; Li, Y.; Yao, T.; Sun, Q.; and Mei, T. 2024. Boosting diffusion models with moving average sampling in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8911–8920.

Ren, Y.; Zhou, Y.; Yang, J.; Shi, J.; Liu, D.; Liu, F.; Kwon, M.; and Shrivastava, A. 2024. Customize-A-Video: One-Shot Motion Customization of Text-to-Video Diffusion Models. arXiv:2402.14780.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Sun, H.; Feng, J.; Yue, Z.; Wang, J.; and Zhang, H. 2024. Prioritize Denoising Steps on Diffusion Model Preference Alignment via Denoised Distribution Estimation. *arXiv:2411.14871*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1921–1930.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. ModelScope Text-to-Video Technical Report. *arXiv preprint arXiv:2308.06571*.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023b. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018*.

Wu, J. Z.; Gao, D.; Bai, J.; Shou, M.; Li, X.; Dong, Z.; Singh, A.; Keutzer, K.; and Iandola, F. 2023. The Text-Guided Video Editing Benchmark at LOVEU 2023. https://sites.google.com/view/loveucvpr23/track4.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, W.; Gu, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*.

Wu, T.; Zhang, Y.; Wang, X.; Zhou, X.; Zheng, G.; Qi, Z.; Shan, Y.; and Li, X. 2025a. Customcrafter: Customized video generation with preserving motion and concept composition abilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8469–8477.

Wu, Y.-S.; Huang, C.-P.; Yang, F.-E.; and Wang, Y.-C. F. 2025b. Motionmatcher: Motion customization of text-to-video diffusion models via motion feature matching. *arXiv preprint arXiv:2502.13234*.

Xiao, Z.; Zhou, Y.; Yang, S.; and Pan, X. 2024. Video diffusion models are training-free motion interpreter and controller. *Advances in Neural Information Processing Systems*, 37: 76115–76138.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2023. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, 1–11.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; Yin, D.; Gu, X.; Zhang, Y.; Wang, W.; Cheng, Y.; Liu, T.; Xu, B.; Dong, Y.; and Tang, J. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. arXiv:2408.06072.

Yue, Z.; Wang, J.; Sun, Q.; Ji, L.; Chang, E. I.; Zhang, H.; et al. 2024a. Exploring diffusion time-steps for unsupervised representation learning. *arXiv preprint arXiv:2401.11430*.

Yue, Z.; Wang, J.; Sun, Q.; Ji, L.; Chang, E. I.-C.; and Zhang, H. 2024b. Exploring Diffusion Time-steps for Unsupervised Representation Learning. In *ICLR*.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. ???? Controlvideo: Training-free controllable text-to-video generation. arXiv 2023. *arXiv preprint arXiv:2305.13077*.

Zhao, M.; Wang, R.; Bao, F.; Li, C.; and Zhu, J. 2023a. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2(3).

Zhao, R.; Gu, Y.; Wu, J. Z.; Zhang, D. J.; Liu, J.; Wu, W.; Keppo, J.; and Shou, M. Z. 2023b. MotionDirector: Motion Customization of Text-to-Video Diffusion Models. arXiv:2310.08465.

Experiment Details

We fine-tune the base T2V models at their original training learning rates. A higher learning rate can shorten the convergence time while taking the risk of overfitting. We sample the customized videos at the original guidance scale of each base model. Our timestep constrained method enables stable and efficient tuning with only a single stage and a single motion customization module to train. All of our experiments are conducted on a single NVIDIA A6000 GPU in less than 10 minutes.

Additional Results

We visually compare our timestep constrained method with more previous SOTA appraoches, VMC (Jeong, Park, and Ye 2023) and MotionClone (Ling et al. 2024), in Fig.

We present additional generation results of our timestep constrained method in Fig. 6.

User Study

We further conduct an user study to compare motion fidelity and motion diversity of the output videos in the motion customization task, which are ambiguous to measure with automatic metrics. We compare our method to three previous SOTA approaches under human evaluation: VMC (Jeong, Park, and Ye 2023), MotionDirector (Zhao et al. 2023b) and MotionClone (Ling et al. 2024).

In our questionnaire, we randomly select 10 reference videos and their new editing prompts, with two output videos of all 4 methods. We ask the evaluators to pick the best methods in terms of motion fidelity, which is defined as the temporal similarity between the output and reference videos, and motion diversity, which is defined as the temporal variety between the two output videos. Note that MotionClone replicate the reference motion conditioned on the latent frame features, and therefore yields deterministic results, so we only present one output video for it.

Our user study involves 30 participants, each with a random set of questions, and we collected 289 valid answers in total. The top pick rates of all methods are listed in Tab. 5. Our timestep constrained method outperforms previous SOTAs in both benchmarks.

Method	Motion Fidelity(%) ↑	Motion Diversity(%) ↑
VMC (2023)	3.8	10.7
MotionDirector (2023b)	19.4	35.6
MotionClone (2024)	31.8	0
Ours	**45.0**	**53.6**

Table 5: The top preference rates of our and previous methods in the user study. Note that MotionClone is a deterministic approach and thus results in no motion diversity.

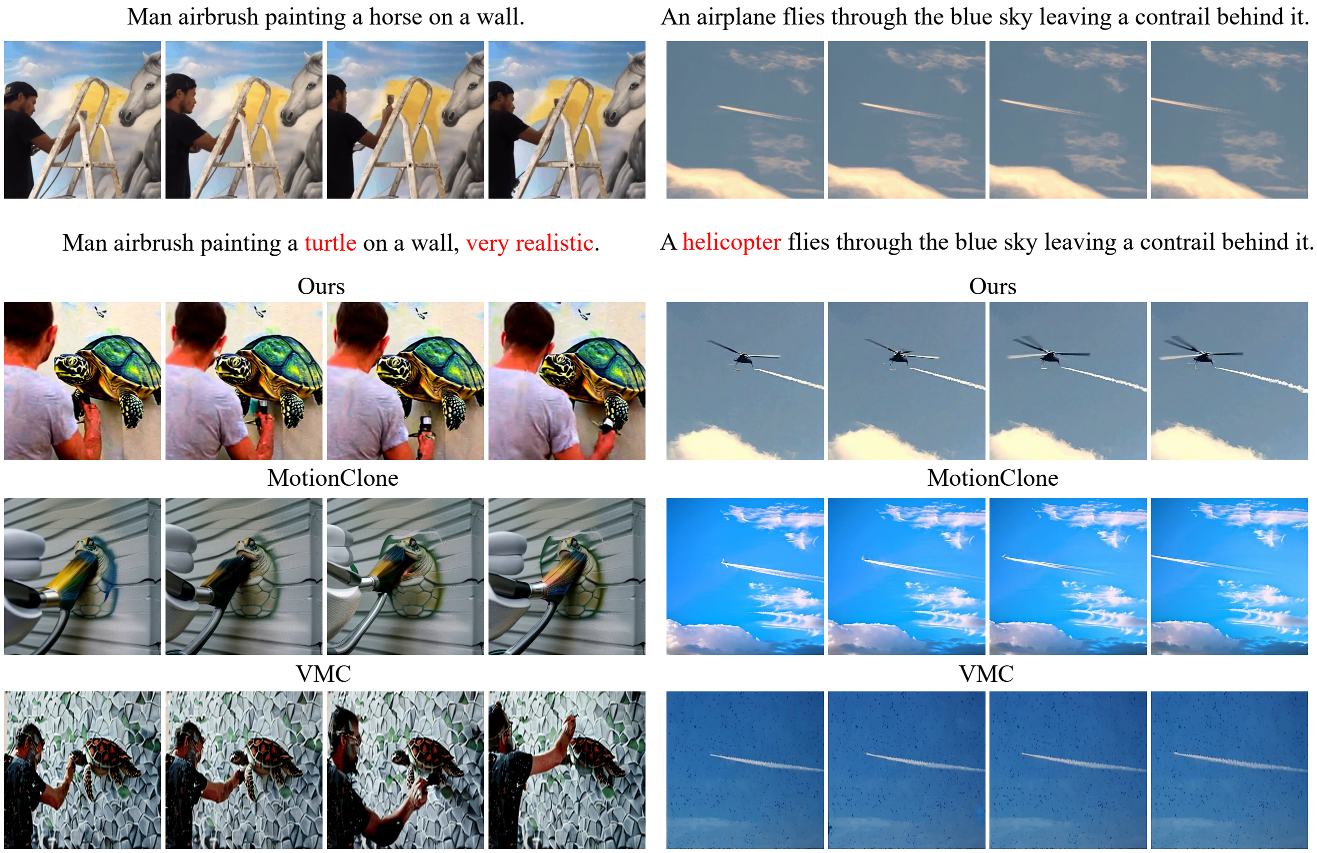

Man airbrush painting a horse on a wall.
An airplane flies through the blue sky leaving a contrail behind it.
Man airbrush painting a turtle on a wall, very realistic.
A helicopter flies through the blue sky leaving a contrail behind it.
Ours
Ours
MotionClone
MotionClone
VMC
VMC

Figure 5: Additional qualitative comparison of our motion disentanglement customization method to MotionClone and VMC.

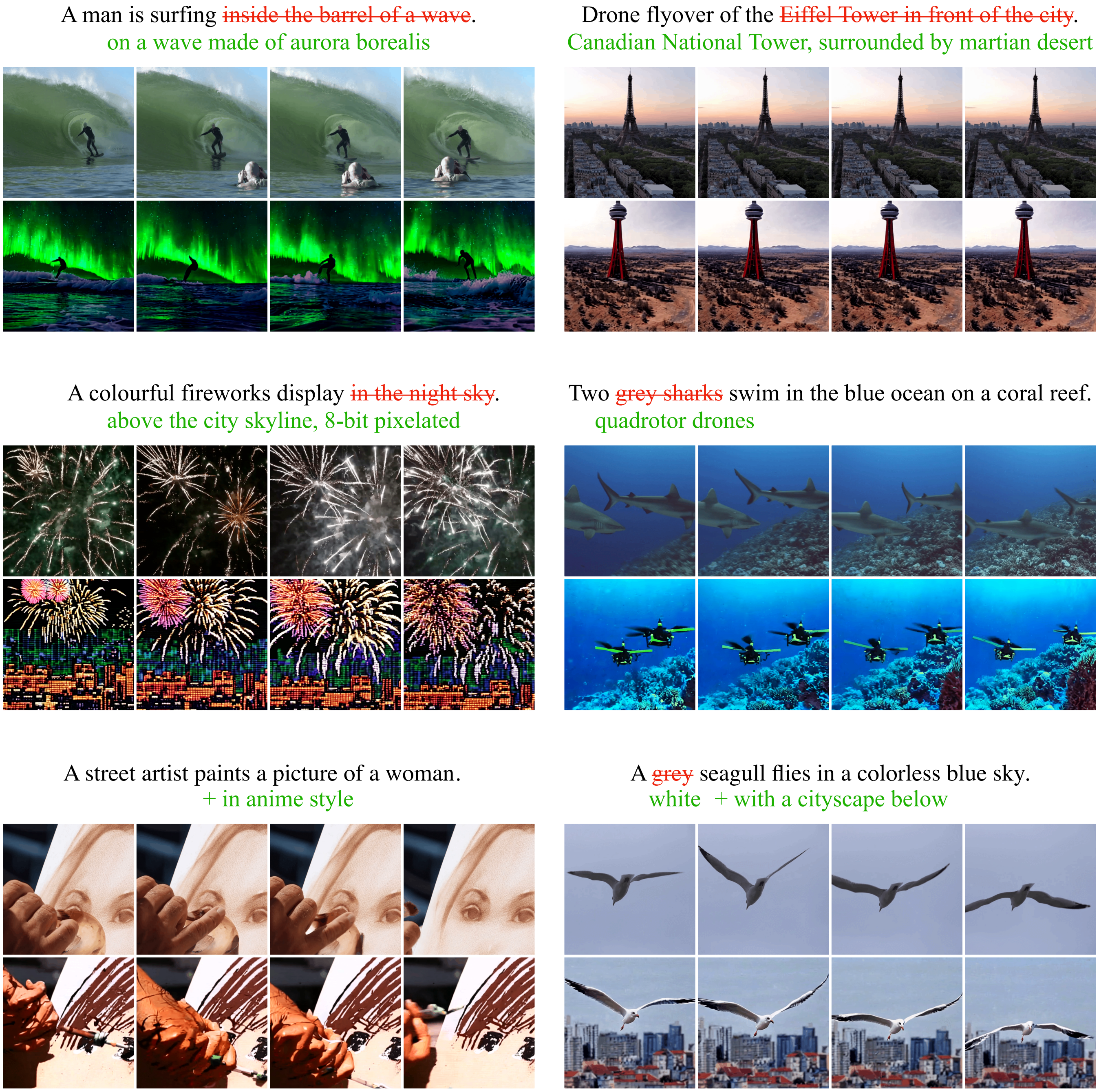

A man is surfing ~~inside the barrel of a wave~~.
on a wave made of aurora borealis
Drone flyover of the ~~Eiffel Tower in front of the city~~.
Canadian National Tower, surrounded by martian desert
A colourful fireworks display ~~in the night sky~~.
above the city skyline, 8-bit pixelated
Two ~~grey sharks~~ swim in the blue ocean on a coral reef.
quadrotor drones
A street artist paints a picture of a woman.
+ in anime style
A ~~grey~~ seagull flies in a colorless blue sky.
white + with a cityscape below

Figure 6: Additional qualitative comparison of our motion disentanglement customization method.