Timestep-Constrained One-Shot Video Motion Customization

Vatsal Baherwani* Yixuan Ren* Abhinav Shrivastava
University of Maryland, College Park

https://vatsal0.github.io/video-diffusion-motion/
vatsalb@umd.edu

Abstract

Video motion customization seeks to adapt a pre-trained text-to-video (T2V) model to the motion in reference videos and reproduce that motion with novel appearances. Unlike deterministic frame-wise video editing, motion customized models capture a motion concept and reinstantiate it with temporal diversity. Yet video diffusion models synthesize motion and appearance jointly through iterative denoising under a global objective, leading to entangled temporal and spatial signals. This issue is especially pronounced in the one-shot setting, where the customized model often memorizes both the reference motion and appearance, causing spatial leakage into the generated videos. In this work, we quantitatively investigate how motion and appearance are factorized across denoising timesteps through the proxy of the trade-off between appearance editing and motion preservation induced by injecting new conditions over specified timestep ranges. Across diverse architectures, we identify a consistent pattern where motion is established in early denoising steps and appearance is refined later, revealing a spatiotemporal boundary in timestep space. Motivated by this characterization, we simplify one-shot motion customization by restricting both training and inference to the motion-dominant timesteps. Our timestep-constrained recipe achieves clean motion transfer without auxiliary debiasing modules or specialized objectives, and can be readily integrated into existing motion customization frameworks regardless of model architecture.

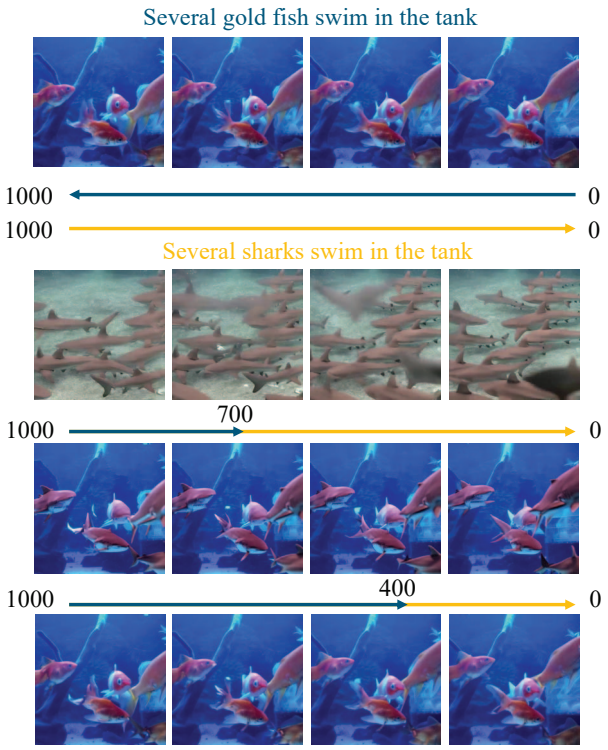


Figure 1. Spatiotemporal disentanglement in video diffusion models. Our probe reveals that motion is primarily encoded in the early denoising timesteps. Given a reference video (top) and its ground truth caption (blue), we perform DDIM inversion and then denoise with a new prompt that modifies only the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

1. Introduction

Diffusion models [13] have achieved remarkable progress in image and video synthesis, and recent foundation models support increasingly diverse forms of control [4, 7, 20, 22, 34, 42, 46, 48, 50]. Model customization [6, 30, 31, 51] further extends this paradigm by fine-tuning pre-trained models on user-provided references, so that the customized model can reproduce the target concept in novel scenarios with both fidelity and diversity. For video generation, this paradigm is particularly useful for motion control. Desired motions are often difficult to specify precisely with text alone, whereas deterministic video editing or motion transfer methods typically rely on frame-wise alignment or strong external guidance and therefore offer limited flexibil-

*Equal contribution.

ity beyond the reference layout. Video motion customization addresses this gap by adapting a pre-trained text-to-video (T2V) model to the motion in reference videos and reproducing that motion under novel appearances, subjects, and scenes.

A central challenge in video motion customization is the entanglement of motion and appearance in video diffusion models. The denoising process synthesizes temporal dynamics and spatial content jointly under a shared reconstruction objective, rather than modeling them as separable factors. This issue is especially severe in the one-shot setting, where only a single reference video is available. Instead of isolating the target motion, the customized model often memorizes the reference appearance together with it, which causes spatial leakage in generated videos and limits the freedom to synthesize new visual attributes.

Existing one-shot motion customization methods mainly address this issue by explicitly enforcing motion-appearance decoupling. Typical strategies include balanced reference data [26], auxiliary spatial debiasing modules [30, 49, 51], and temporal objectives that encourage motion learning without direct content copying [14, 28, 40]. Although effective, these designs increase training complexity by introducing additional modules, objectives, and hyperparameters whose balance directly affects the trade-off between motion preservation and appearance suppression. The problem becomes more delicate when motion is strongly correlated with appearance, as in actions such as walking or eating.

Recent studies on diffusion models suggest that denoising timesteps play different roles. In image generation, this behavior has been associated with coarse-to-fine synthesis and frequency-dependent refinement [11, 18, 21, 24, 29, 39, 44]. In video generation, several methods [1, 19, 38, 41] empirically exploit the observation that early denoising steps are more closely related to motion and layout, whereas later steps mainly refine appearance. For motion customization, however, this observation remains largely heuristic. It is still unclear how motion and appearance trade off across timesteps, how consistent this pattern is across architectures, and whether it can support a practical customization rule.

We address this question through a controlled probing procedure. Given a reference video and its text prompt, we resample the video while replacing appearance-related conditions only within selected timestep ranges and keeping the remaining steps unchanged. The resulting trade-off between appearance editing and motion preservation provides a quantitative proxy for how motion and appearance are distributed along the denoising trajectory. Across diverse T2V architectures, the probe reveals a consistent spatiotemporal structure: early denoising steps are predominantly motion-dominant, whereas later steps are appearance-dominant.

This characterization leads to a simple customization principle. Rather than modeling the full denoising trajectory, we restrict both training and inference to motion-dominant timesteps. Based on this principle, we develop a timestep-constrained one-shot motion customization framework that suppresses appearance leakage without auxiliary debiasing modules or specialized objectives. Despite relying only on the vanilla diffusion loss, the framework supports both efficient partial-attention tuning and direct full-rank fine-tuning across diverse text-to-video architectures.

In summary, our main contributions are as follows:

- We introduce a quantitative probe of motion-appearance factorization across denoising timesteps in text-to-video diffusion models.
- We identify a consistent timestep structure across diverse architectures, where early denoising steps are motion-dominant and later steps are appearance-dominant.
- Guided by this structure, we develop a timestep-constrained one-shot motion customization framework that achieves clean motion transfer without auxiliary debiasing modules or specialized objectives.

2. Related Works

2.1. Video Motion Customization

Video motion customization aims to adapt a pre-trained text-to-video diffusion model to the motion in reference videos and reproduce that motion under novel appearances, subjects, and scenes. It differs from deterministic video editing or motion transfer methods, which typically rely on strong external guidance, such as structural controls [4, 48, 50], optical flow [20, 42], or latent feature alignment [7, 22], to enforce frame-wise or trajectory-level correspondence. Such methods are effective when the goal is to faithfully follow a prescribed motion pattern, but they offer limited flexibility beyond the reference layout. By contrast, motion customization fine-tunes the generative model itself, so that the target motion can be re-instantiated with temporal diversity under new prompts and scenes.

For the multi-shot setting, where multiple reference videos share the same motion concept but exhibit different appearances, the appearance variation across references helps suppress instance-specific spatial signals and makes the common motion pattern easier to capture [8, 26]. The one-shot setting is substantially more challenging. Since the diffusion objective reconstructs spatial and temporal signals jointly from a single coupled example, the customized model tends to overfit not only to the target motion but also to the reference appearance, leading to appearance leakage in generated videos.

Existing one-shot methods mainly address this issue by explicitly approximating motion–appearance decoupling. Some introduce auxiliary spatial debiasing modules to steer

temporal adapters toward motion while suppressing appearance leakage [30, 49, 51]. Others design temporal objectives that distill motion without directly copying visual content [14, 28, 40]. While effective, these approaches rely on additional modules or specialized objectives, which increase training overhead and hyperparameter sensitivity. Moreover, many of them are most naturally instantiated on architectures with explicit spatial-temporal factorization. In contrast, our method does not impose motion–appearance separation through auxiliary designs. Instead, we exploit a timestep-wise spatiotemporal structure already present in the denoising process, and use it to derive a simplified one-shot motion customization framework that is architecture-agnostic and does not require auxiliary debiasing modules or specialized losses.

2.2. Diffusion Attribute Disentanglement

Attribute disentanglement in diffusion models has attracted increasing attention as a means to interpret internal representations and improve controllability. In image generation, several works study how information is organized across timesteps and layers. Aggregating multi-timestep and multi-scale features reveals complementary geometric and semantic cues for correspondence [23], while spectral analyses show that low-frequency content dominates early denoising steps and high-frequency details emerge later, motivating non-uniform timestep sampling and frequency-aware manipulation [17, 39]. Other methods treat timesteps as explicit supervision axes through timestep-aware representations and step-aware preference alignment [3, 33, 45], and per-step editing shows that intervening at selected timesteps can separate coarse layout from fine appearance [10]. Recent interpretability studies further suggest that semantic concepts are structured non-uniformly across layers and timesteps [15]. Taken together, these works indicate that denoising trajectories encode different attributes in a heterogeneous manner. However, they primarily focus on spatial attributes in image diffusion and do not directly characterize how motion and appearance are distributed across timesteps in video generation.

For video diffusion models, timestep-wise disentanglement has been explored much less systematically and is often used only heuristically. [19, 38] inject new appearance into reference videos while bypassing early denoising steps to reduce motion interference, implicitly assuming that motion is established earlier and appearance is refined later, but without quantifying where each factor dominates. [2] studies how camera trajectories are encoded across timesteps and separates camera motion from scene content, whereas our focus is on general object and scene motions and their coupling with appearance. [47] learns frequency-aware embeddings across all timesteps for image-to-video generation, where appearance is already largely fixed by the input image and the temporal problem is therefore different from motion customization in T2V models. [22, 41] extract motion-aware features from pre-trained T2V models and use them for motion guidance without model tuning, whereas motion customization requires adapting the model itself so that the reference motion can be reproduced under novel subjects and scenes with temporal diversity. In contrast to these directions, our goal is not disentanglement for interpretation alone or heuristic timestep selection for editing. We characterize a quantitative, architecture-agnostic motion–appearance boundary along denoising timesteps and use it as the basis for a timestep-constrained one-shot motion customization method.

3. Spatiotemporally Disentangled Diffusion

3.1. Preliminary

Diffusion Models Diffusion models [13] 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

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

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 $\alpha_t, \bar{\alpha}_t$ configure 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.

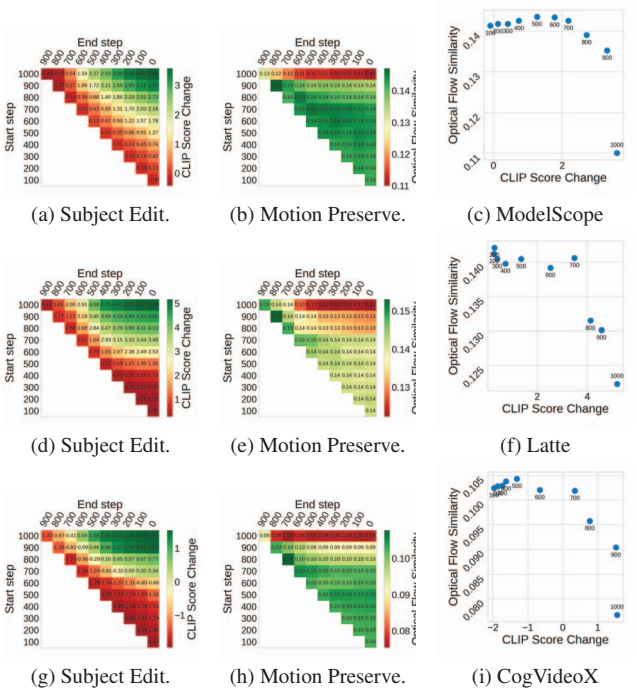
DDIM Inversion In implicit diffusion models (DDIMs, Song et al.), Eq. 1 can be made deterministic by setting $\sigma_t := 0$. The denoising process can then be inverted by deriving x_{t+1} from x_t [27] via

$$\mathbf{x}_{t+1} = \sqrt{\alpha_{t+1}}\, \mathbf{x}_t + \left(\sqrt{1-\bar{\alpha}_{t+1}} - \sqrt{\alpha_{t+1}}\sqrt{1-\bar{\alpha}_t} \right) \epsilon_\theta(\mathbf{x}_t, t, c). \quad (2)$$

Iteratively applying Eq. 2 ultimately produces the approximate sampling trajectory $\mathbf{x}_{\{T,...,1\}}$ from an existing $\mathbf{x}_0$, which reconstructs itself following the denoising process.

3.2. Probe 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



(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 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.

diffusion timesteps further lifts its difficulty. Therefore, we design to leverage the inversion approach and tamper the resampling trajectory for feasible calculation and reference.

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 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 c' indicates that x_0' ideally represents the new subject with the original motion , this process will intervene the generated motion as well, as shown in Figure 1 row 2.

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, as shown in Figure 1 rows 3 and 4. 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 [12] 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.

3.3. 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 Figure 1. Here we use start timestep $\tau_{start} = 700$ and end timestep $\tau_{\text{end}} = 0$. As a result, our newly generated video preserves the information from $t \in [700, 1000]$ in the original video.

To fully reflect the editing improvement, we measure the CLIP score change where the base score between x_0 and c' is subtracted, as x_0 already has partial resemblance to c' in background. We use the Lucas-Kanade method for optical flow estimation, and calculate the average cosine similarity between the normalized vectors of all frames.

We conduct this experiments on three representative text-to-video models with divergent denoising network architectures: ModelScope [34] with U-Net and dedicated spatial and temporal attentions, Latte [25] with transformer and dedicated spatial and temporal attentions, and CogVideoX [43], with transformer and unified spatiotemporal attentions. We test on all 76 videos from the Text-Guided Video Editing (TGVE) competition dataset [37].

3.4. Results and Analyses

In Figure 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 mo-

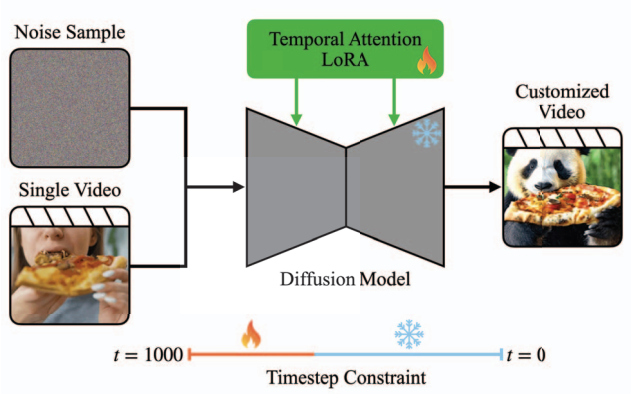


Figure 3. Single-stage one-shot video motion customization with denoising timestep constraint. Leveraging our spatiotemporal disentanglement property, we train LoRAs at only early denoising timesteps to model the reference motion without appearance leakage. It obviates any debiasing module, stage or loss, and consistently works for base models with unified spatiotemporal attentions, where we add LoRA on the full spatiotemporal sequence.

tion preservation quality in Figure 2. τ_{end} is not significant for motion preservation while being optimal for appearance editing at 0, at which we therefore fix the end timestep. 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 trade-off is also depicted in Figure 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. From now on we denote $\tau = \tau_{start}$ as this threshold. While its exact value varies across specific models, it is consistently around $[700, 900]$.

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

4. Timestep-Constrained Customization

Prior diffusion-based motion customization methods typically apply LoRAs on pre-trained temporal attention layers, and fine-tune 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. The text prompt remains the same new prompt with modified appearances and original motions throughout the inference.

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 Figure 2.

Base Model	τ	Text Align.↑	Temp. Const.↑	Pick Score↑
ModelScope [34]	1000	26.05	94.88	20.13
	750	<u>28.04</u>	<u>96.39</u>	20.68
	700	**28.16**	**96.42**	<u>20.77</u>
	650	27.97	96.31	**20.79**
	0	27.43	96.25	20.49
Latte [25]	1000	29.28	93.16	20.84
	750	31.85	97.12	21.65
	700	**31.96**	<u>97.19</u>	**21.68**
	650	<u>31.88</u>	**97.21**	<u>21.66</u>
	0	31.26	96.99	21.47
CogVideoX [43]	1000	28.15	96.69	20.65
	950	**30.14**	**98.11**	<u>21.09</u>
	900	<u>29.93</u>	<u>98.10</u>	21.00
	850	29.61	97.76	20.92
	0	29.67	97.41	**21.30**

The overall pipeline of our method is illustrated in Figure 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 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.

4.1. Experiment Setup

Base models. We implement our training method on three base T2V models: ModelScope [34], Latte [25], and CogVideoX [43]. 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.

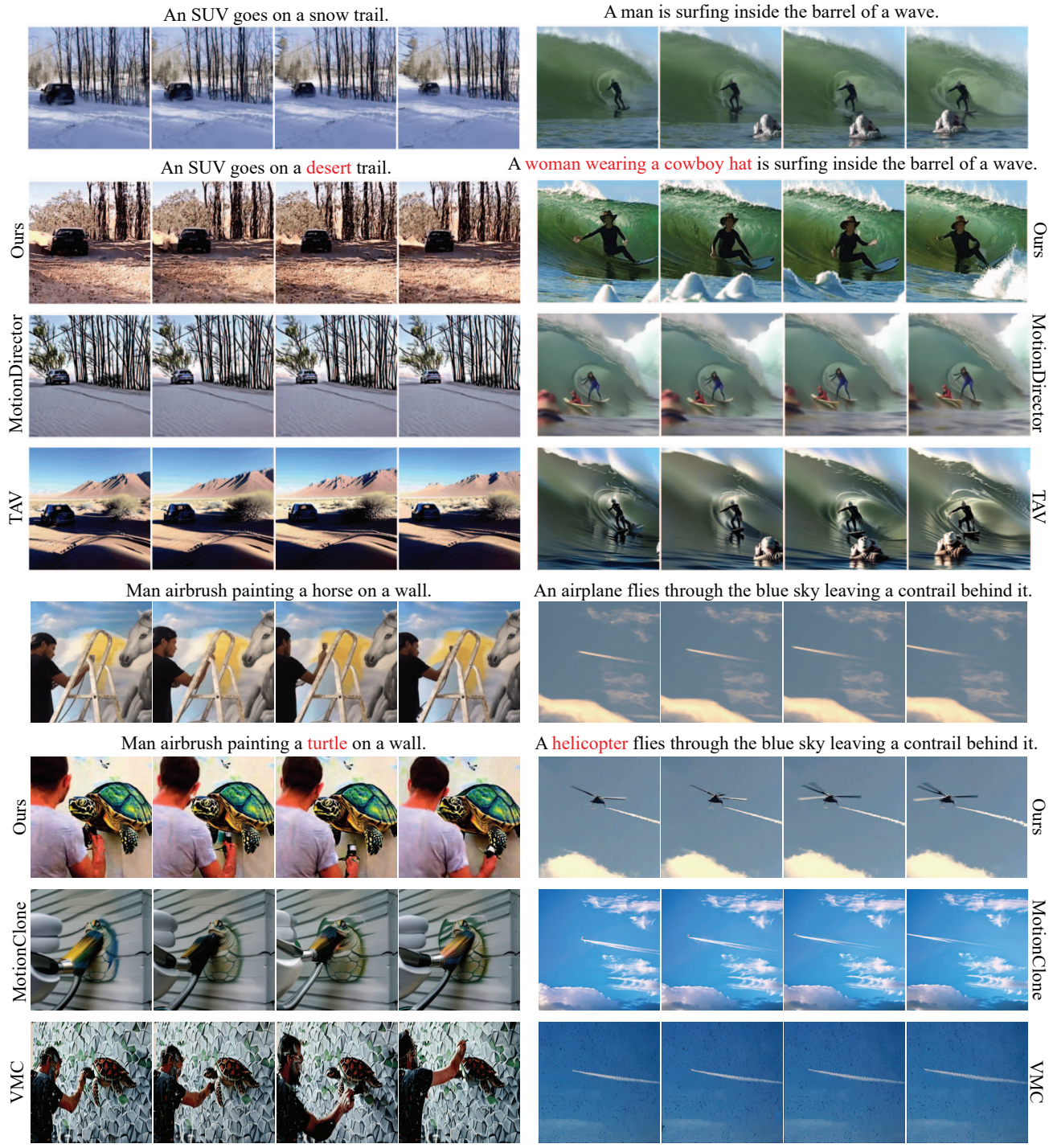


Figure 4. Qualitative comparison of our motion disentanglement method to previous SOTAs. Our method faithfully replicates the motion of the reference video while also editing the subject and background with superior quality to other approaches. Without any additional spatial debiasing modules or stages,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

Training. We fine-tune all base models at their original training learning rates. We sample new videos at their original guidance scales All of our experiments are conducted on a single NVIDIA A6000 GPU in less than 10 minutes.

Table 2. Comparison with previous SOTA motion customization methods on TGVE. Our timestep constraining method achieves leading performance without auxiliary modules or stages, and is also compatible to be integrated with existing pipelines. † denotes methods that were tested on other datasets and we re-evaluated on the TGVE benchmark for fair comparison. ‡ denotes methods that were tested on different datasets but not open-sourced.

Method	Text Align.↑	Temp. Const.↑	Pick Score↑
Tune-A-Video [2022]	25.64	92.42	20.09
VideoComposer [2023]	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 [2023]	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‡ [2025]	<u>30.43</u>	<u>97.20</u>	–
Ours–ModelScope	28.16	96.42	20.77
Ours–Latte	**31.96**	97.19	**21.68**
Ours–CogVideoX	30.14	**98.11**	<u>21.09</u>

Datasets. To quantitatively evaluate our approach, we apply motion customization on all 76 videos in the Text-Guided Video Editing (TGVE) competition dataset [37] individually. It is composed of videos from various sources including DAVIS, Youtube and Videovo with various editing tasks such as object, background and style editing. We use the ground truth captions as the training prompts and sample novel videos for all 4 editing captions.

Metrics. We quantitatively evaluate generated videos using the following metrics. Text alignment. We compute the frame-wise CLIP Score [12] between generated frames and the edited prompt, and average the scores over frames and samples. Temporal consistency. We measure the average pairwise CLIP embedding distance between consecutive frames, where lower values indicate better temporal smoothness. PickScore. We measure prompt alignment with the pretrained CLIP-based human preference scorer [16] finetuned on the Pick-a-Pic dataset. Frame-level scores are averaged over each video and then across samples. Each edited prompt produces 4 samples for metric averaging.

4.2. τ Ablations

We experiment with choices for the temporal tuning threhshold τ in our motion customization method. We present these results in Table 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 Figure 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 in practice.

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 module only at early timesteps, where motion information is dominantly encoded.

Table 3. 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.

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

4.3. Comparisons

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. The quantitative results are listed in Table 2. Our motion customization approach yields superior quantitative results to prior SOTAs with a much simplified tuning module and pipeline. Figure 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.

4.4. User Study

We further conduct an user study to compare motion fidelity and motion diversity of the output videos in the motion cus-

Table 4. 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.

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

tomization task, which are ambiguous to measure with automatic metrics. We compare our method to three previous SOTA approaches under human evaluation: VMC [14], MotionDirector [51] and MotionClone [22].

In each 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.

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 Table 3. Our timestep constrained method outperforms previous SOTAs on both benchmarks.

4.5. 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 Table 4, 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 Full Parameter 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 extend our method to full-parameter fine-tuning. Previous successful approaches for direct training follow DreamBooth [31] 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 additional data.

Table 5. 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.

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

We present the results in Table 5. 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 where extra balance data are necessary.

5. Conclusion

In this work, we addressed one-shot video motion customization in text-to-video diffusion models. The main difficulty of this task lies in the entanglement of motion and appearance during denoising, which causes appearance leakage when the model is tuned on a single reference video. We proposed a simple timestep-constrained customization framework that restricts both training and inference to motion-dominant timesteps, thereby enabling clean motion transfer under novel appearances. The framework is guided by a quantitative probe of appearance editing and motion preservation, which identifies a consistent motion–appearance boundary across diverse architectures. By leveraging this structure, our method avoids auxiliary debiasing modules and specialized objectives while remaining effective across different model designs. Overall, our results show that timestep-aware customization offers a practical and streamlined solution to motion adaptation in video diffusion models.

References

[1] Yuval Atzmon, Rinon Gal, Yoad Tewel, Yoni Kasten, and Gal Chechik. Motion by queries: Identity-motion trade-offs in text-to-video generation. *arXiv preprint arXiv:2412.07750*, 2024. 2

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22875–22889, 2025. 3

[3] Bohan Chen, Dongjun Jiang, Chaofan Shi, Lei Ji, Yun Wang, Songyang Yan, Zhen Wei, Dahua Lin, and Hanwang Zhang. Aligning preference with denoising performance at each timestep. In *NeurIPS*, 2024. 3

[4] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 1, 2, 7

[5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 7

[6] 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. 1

[7] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1, 2

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[9] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Videocrafter: A toolkit for text-to-video generation and editing. https://github.com/AILab-CVC/VideoCrafter, 2023. 7

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv:2208.01626*, 2022. 3

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4, 7

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[14] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models, 2023. 2, 3, 7, 8

[15] Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. Also available as arXiv:2411.16725. 3

[16] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 7

[17] Haeil Lee, Hansang Lee, Seoyeon Gye, and Junmo Kim. Beta sampling is all you need: Efficient image generation strategy for diffusion models using stepwise spectral analysis. In *WACV*, 2025. 3

[18] Haeil Lee, Hansang Lee, Seoyeon Gye, and Junmo Kim. 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)*, pages 4215–4224. IEEE, 2025. 2

[19] Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. *arXiv preprint arXiv:2411.17048*, 2024. 2, 3

[20] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2024. 1, 2

[21] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13199–13208, 2025. 2

[22] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 1, 2, 3, 7, 8

[23] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*, 2023. 3

[24] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36: 47500–47510, 2023. 2

[25] 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. 4, 5

[26] Joanna Materzynska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Customizing motion in text-to-video diffusion models. *arXiv preprint arXiv:2312.04966*, 2(3):17, 2023. 2

[27] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. 3

[28] Geon Yeong Park, Hyeonho Jeong, Sang Wan Lee, and Jong Chul Ye. Spectral motion alignment for video motion transfer using diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6398–6405, 2025. 2, 3

[29] Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. Boosting diffusion models with moving average sampling in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8911–8920, 2024. 2

[30] 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, 2, 3

[31] 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. 1, 8

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 3

[33] Haoran Sun, Jiayi Feng, Zhongqi Yue, Jiankun Wang, and Hanwang Zhang. Prioritize denoising steps on diffusion model preference alignment via denoised distribution estimation. *arXiv:2411.14871*, 2024. 3

[34] 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. 1, 4, 5

[35] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 7

[36] 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. 7

[37] 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. 4, 7

[38] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8469–8477, 2025. 2, 3

[39] Wei Wu, Qingnan Fan, Shuai Qin, Hong Gu, Ruoyu Zhao, and Antoni B. Chan. Freediff: Progressive frequency truncation for image editing with diffusion models. In *European Conference on Computer Vision (ECCV)*. Springer, 2024. To appear in ECCV 2024 proceedings; also available as arXiv:2404.11895. 2, 3

[40] Yen-Siang Wu, Chi-Pin Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Motionmatcher: Motion customization of text-to-video diffusion models via motion feature matching. *arXiv preprint arXiv:2502.13234*, 2025. 2, 3, 7

[41] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *Advances in Neural Information Processing Systems*, 37:76115–76138, 2024. 2, 3

[42] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 1, 2

[43] 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. 4, 5

[44] Zhongqi Yue, Jiankun Wang, Qianru Sun, Lei Ji, Eric I Chang, Hanwang Zhang, et al. Exploring diffusion time-steps for unsupervised representation learning. *arXiv preprint arXiv:2401.11430*, 2024. 2

[45] Zhongqi Yue, Jiankun Wang, Qianru Sun, Lei Ji, Eric I.-Chao Chang, and Hanwang Zhang. Exploring diffusion time-steps for unsupervised representation learning. In *ICLR*, 2024. 3

[46] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1

[47] Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. Flexiact: Towards flexible action control in heterogeneous scenarios. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 3

[48] Y Zhang, Y Wei, D Jiang, X Zhang, W Zuo, and Q Tian. Controlvideo: Training-free controllable text-to-video generation. arxiv 2023. *arXiv preprint arXiv:2305.13077*. 1, 2

[49] Yuxin Zhang, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Motioncrafter: One-shot motion customization of diffusion models. *arXiv preprint arXiv:2312.05288*, 2023. 2, 3

[50] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2(3), 2023. 1, 2

[51] 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, 2, 3, 7, 8