NewMove: Customizing text-to-video models with novel motions

Joanna Materzyńska[⋆1,2], Josef Sivic[2,3], Eli Shechtman[2], Antonio Torralba[1], Richard Zhang[2], and Bryan Russell[2]

[1] Massachusetts Institute of Technology
[2] Adobe Research
[3] CIIRC CTU [⋆⋆]

Abstract. We introduce an approach for augmenting text-to-video generation models with novel motions, extending their capabilities beyond the motions contained in the original training data. With a few video samples demonstrating specific movements as input, our method learns and generalizes the input motion patterns for diverse, text-specified scenarios. Our method finetunes an existing text-to-video model to learn a novel mapping between the depicted motion in the input examples to a new unique token. To avoid overfitting to the new custom motion, we introduce an approach for regularization over videos. Leveraging the motion priors in a pretrained model, our method can learn a generalized motion pattern, that can be invoked with novel videos featuring multiple people doing the custom motion, or using the motion in combination with other motions. To validate our method, we quantitatively evaluate the learned custom motion and perform a systematic ablation study. We show that our method significantly outperforms prior appearance-based customization approaches when extended to the motion customization task. Project webpage: *https://joaanna.github.io/customizing_motion/*.

1 Introduction

Recent advancements in text-to-video synthesis have significantly pushed the boundaries of video generation [2, 19, 28, 39, 41, 43, 55]. As synthesized videos become more realistic, there is an increasing demand to provide precise control over the output, meeting specific user needs and creative visions. Consider, for example, a creator aiming to depict characters doing the 'Carlton Dance' from "The Fresh Prince of Bel-Air" in a synthetic video with their own background and setting, as shown in Fig 1. If the pre-trained model does not know the specific dance, then capturing the subtle nuances of the dance's swinging arms and the precise timing of its poses through natural language is challenging, necessitating a method to instead learn from visual examples. However, the challenge is not to merely replicate the dance as in traditional motion transfer methods [5, 34, 35],

---

[⋆] Work done in part as an intern with Adobe Research.

[⋆⋆] Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University in Prague.
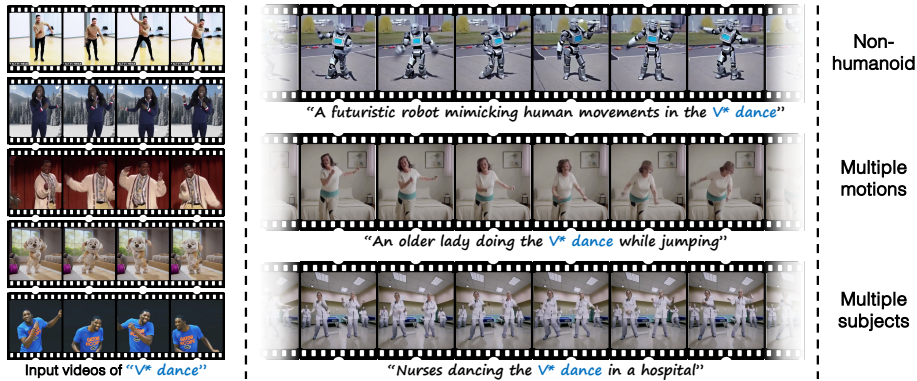
Fig. 1: (Left) Given a few examples ("Carlton dance"), our customization method learns the dynamic motion pattern common to the input examples and incorporates it into a pre-trained text-to-video diffusion model using a new motion identifier ("V* dance"). **(Right)** Our approach, *NewMove*, abstracts the motion pattern from the appearance in the input videos and enables generation of the depicted motion across a variety of novel contexts, including with a non-humanoid subject (robot, top row), multiple motions (lady, middle row), and multiple subjects (group of nurses, bottom row). **To best view the results, please view our website .**

but to teach the model about the motion from a few examples ('customization') and to apply this learned motion seamlessly across diverse scenes, without relying on an original driving video. Rather than copying a motion from a single video, we wish to capture and learn the motion's attributes while abstracting away the appearance and context of the example setting. These challenges lead us to a question: How can we leverage a pre-trained text-to-video model's priors about motion and appearance to augment them with new motions, and subsequently generate these new motions in novel settings?

Traditional methods in video generation have offered detailed control through techniques like conditional generation [10, 42, 49] and motion transfer [5, 15, 34, 36, 46], enabling the precise replication of specific motions depicted in the input video. However, these methods are limited by their reliance on the original video's scene structure and geometry, effectively locking the motion to the context of the driving video. Such an approach restricts the ability to apply the motion more broadly, as it does not truly learn and abstract the motion pattern for flexible use across varied scenes and subjects. Developing methods that can generalize motion patterns from examples for use in diverse settings presents a significant challenge, as it requires methods that go beyond simple replication to allow for the motion's adaptation to new environments. A more similar scenario to ours focuses on customization of subjects in text-to-image models [9, 11, 12, 18, 24, 33, 38, 45]. However, unlike learning a novel dynamic motion from few example videos, these methods are restricted to still subjects. In Tab. 1 we compare the design of related approaches to ours, and discuss them in more detail in Section 2.

Table 1: Comparison of our method across different techniques for controllable video / image generation. Here, $\mathcal{G}_I$, $\mathcal{G}_{I \to V}$ and $\mathcal{G}_V$ denote text-to-image, image-to-video and text-to-video generation models, respectively. Symbols $\mathcal{I}$ and $\mathcal{V}$ denote input images and videos, $\hat{\mathcal{I}}$ and $\hat{\mathcal{V}}$ denote output images and videos, and τ denotes a text prompt. Superscripts refer to learned or transferred concepts , either appearance† or motion$^\circ$. Our method is the only one that transforms a text-to-video motion prior ($\mathcal{G}_V$) to a model with customized motions ($\mathcal{G}_V^\circ$, $\mathcal{G}_V^{\circ\dagger}$) such that the learned concept ($\tau^\circ, \tau^{\circ\dagger}$) can be invoked at the test time.

Methods	Customization	Application
Image animation [6]	None	$\mathcal{G}_{I \to V}(\mathcal{I}) \to \hat{\mathcal{V}}$
Text-to-video generation [2, 7, 19, 28, 39, 41, 43, 55]	None	$\mathcal{G}_V(\tau) \to \hat{\mathcal{V}}$
Structure-conditioned video generation [10]	None	$\mathcal{G}_V(\tau, V) \to \hat{\mathcal{V}}$
Zero-shot text-driven motion transfer [14, 50]	None	$\mathcal{V}, \mathcal{G}_V(\tau^\circ) \to \hat{\mathcal{V}}$ or $\mathcal{V}, \mathcal{G}_I(\tau^\circ) \to \hat{\mathcal{V}}$
Image customization [11, 24, 33]	$\mathcal{G}_I \xrightarrow{\mathcal{I}} \mathcal{G}_I^\dagger$	$\mathcal{G}_I^\dagger(\tau^\dagger) \to \hat{\mathcal{I}}$
Animating text-to-image models (AnimateDiff v1) [17]	$\mathcal{G}_I^\dagger, \mathcal{G}_V \to \mathcal{G}_V^\dagger$	$\mathcal{G}_V^\dagger(\tau) \to \hat{\mathcal{V}}$
One-shot video tuning [47]	$\mathcal{G}_I \xrightarrow{\mathcal{V}, \tau} \mathcal{G}_V^\circ$	$\mathcal{G}_{\hat{V}}^\circ(\tau^\circ) \to \hat{\mathcal{V}}$
Ours (motion customization)	$\mathcal{G}_V \xrightarrow{\mathcal{V}} \mathcal{G}_V^\circ$	$\mathcal{G}_V^\circ(\tau^\circ) \to \hat{\mathcal{V}}$
Ours (motion and appearance cust.; see our website)	$\mathcal{G}_V \xrightarrow{\mathcal{V}, \mathcal{I}} \mathcal{G}_V^{\circ\dagger}$	$\mathcal{G}_V^{\circ\dagger}(\tau^{\circ\dagger}) \to \hat{\mathcal{V}}$

Incorporating new motions in text-to-video models (*motion customization*) faces unique challenges. The first issue is determining which model parameters to adjust for capturing distinct motions from few examples, balancing the need for model flexibility against the risk of overfitting due to limited data. Another challenge is maintaining the model's existing knowledge without having new customizations overshadow or alter learned concepts, such as causing a bias towards a specific dance like the Carlton. Additionally, it is critical to separate motion from appearance so that learned motions can be applied to new scenes.

In this work, we address these challenges and introduce a novel method for adapting text-to-video diffusion models to recognize and generate new motions, such as local motions (human movement) or global (camera motions), independent of the performer's appearance or scene's appearance, *i.e.*, customizing text-to-video models with novel motions. By training the model with a few example videos of a new motion, we assign this motion to a unique text token ("V*"). This assignment allows us to produce videos featuring various subjects performing the new motion, as illustrated in Fig 1. Through an ablation study, we explore customizing various model parameters, highlighting that adjustments to temporal layers and spatial cross-attention mechanisms strike an optimal balance between model adaptability and the risk of over-fitting. To preserve the model's original motion knowledge and avoid neglecting existing motion concepts, we develop a unique regularization strategy specifically for video content. To prioritize learning about motion over appearance, we introduce a new sampling strategy that focuses on samples that determine coarser visual structure.

To evaluate our method we perform two qualitative studies. The first is a user study comparing our method against two baselines across diverse scenarios, including videos with multiple people and varying subject sizes, where the cus-
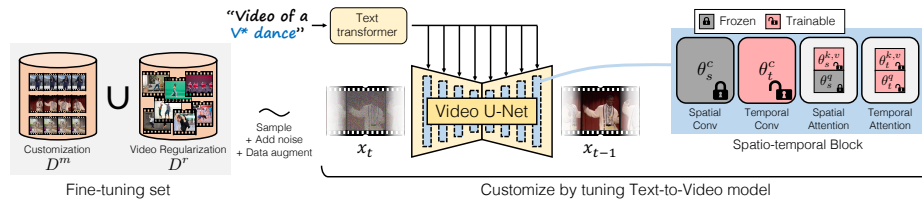
Fig. 2: Overview. Given a small set of exemplar videos, our approach fine-tunes the U-Net of a text-to-video model using a reconstruction objective. The motion is identified with a unique motion identifier and can be used at test time to synthesize novel subjects performing the motion. To represent the added motion but preserve information from the pretrained model, we tune a subset of weights – the temporal convolution and attention layers, in addition to the key & value layers in the spatial attention layer. A set of related videos is used to regularize the tuning process.

tom motions are sourced from internet videos. The second evaluation employs gestures from the Jester dataset, aiming to systematically examine our method's effectiveness in internalizing and recognizing motions via an off-the-shelf classifier. **Our contributions are**: (i) introducing an approach that leverages the rich motion and appearance prior information contained in a pretrained text-to-video generation to augment it with novel motions, (ii) demonstrating the generalization abilities of our method to diverse scenes and multiple subjects, and (iii) lastly a qualitative and quantitative evaluation of our method.

2   Related Work

Motion transfer. Motion transfer is a well established task [5, 15, 34, 36, 46]. The task involves capturing motion from one video and applying it to the target subject given by an image or a video. These methods explicitly model the pose of the source motion and focus on modelling humans. More recently, the task of text-driven motion transfer has been introduced [14, 50]. Originally, Geyer *et al.* [14] leveraged a pre-trained text-to-image model to generate videos that preserve the overall motion of a source video and provide some flexibility with respect to the dynamic target subject, without tuning the original model. Yatim *et al.* [50] proposed a similar approach that instead utilizes a text-to-video model. Other approaches [3, 4, 10], ensure output videos adhere to reference videos by guiding the generation with depth information. Our approach differs from these prior works in that we learn a generalized motion pattern from a few video examples. That is, instead of directly copying the motion or the overall video motion pattern, our method allows for invoking the motion in vastly different text-defined scenarios. This setting includes creating videos where multiple individuals execute the motion simultaneously, or blending this motion with other distinct movements. **Video synthesis with image generation models.** Previous works repurposed text-to-image generation models for video generation. Wu *et al.* [47] used a pretrained image-to-video model and a source video to

stylize the video with a text condition. Similarly, several methods [8, 23, 51] can generate a video conditioned on control signals such as depth maps of edges using a pre-trained text-to-image model. Guo *et al.* [17] learn a motion prior network with text-video data that can be used together with personalized text-to-image models to generate animated images. Our work differs in that we use a video generation model directly to learn new motions and, rather than animating images, our method produces videos with complex dynamic motions.

Text-to-video models. Recently, several methods have demonstrated text-to-video generation [2, 19, 28, 39, 41, 43, 55]. An important line of research involves studying architectures and optimization for training video generation models. A high-quality model is a prerequisite for our method. In this work, we use the model from Wang *et al.* [41]. Molad *et al.* [28] show qualitative results of adding a custom subject to the base video model. They, however, do not explore customization beyond static objects [28]. The idea of multi-modal customization, however, facilitates user-controllability and creative applications, since no matter how powerful the model, there are always novel subjects and motions that cannot be all contained in the training set used to train the base video generation model.**Text-to-image and text-to-video customization.** Text-to-image customization is a related task to ours, where a text-to-image diffusion model is augmented with a custom object [9,11,12,18,24,33,38,45]. Gal *et al.* [11] showed that this mapping can be created purely by optimizing a text embedding of the textual description. Kumari *et al.* [24] proposes a method that fine-tunes selected weight matrices in the generator network. More recent approaches investigate efficient formulations for modifying an approximation of the parameters responsible for the text-image mapping in the generation network [18,38]. Other methods, alternatively to adjusting the model's parameters, introduce an image encoder that can be directly plugged into the pre-trained network [12, 45, 56]. Alternatively to customizing individual concepts, Huang *et al.* [21] propose to learn relations between two subjects from a set of images via a similar technique to Gal *et al.* [11]. Similarly, image animation techniques use a video motion prior to turn still images into videos [6,17]. The text-to-image customization is limited to static objects due to the static nature of still images. In contrast, we focus on customization with respect to a given motion depicted in a small set of videos, where the task is to learn the dynamic motion pattern. More recently, several concurrent and yet unpublished works [16, 22, 44, 48, 52, 53] highlight the growing interest in motion customization, demonstrating its importance in current research. These works explore various facets of the problem, including transferring subject motion from a video [22, 52], customizing appearance and motion from images and videos [44], animating high-resolution images [48], animating images with custom camera motion patterns [16], and separating motion from appearance [53]. While these works are exciting, our approach uniquely stands out in the context of this recent work by abstracting dynamic motion from multiple videos, enabling the creation of new, diverse videos that showcase different scenes, combine different motions, and depict various subjects.

3 Approach

3.1 Text-to-Video Diffusion Model Preliminaries

Diffusion models are probabilistic models that can approximate distributions through a gradual denoising process [20,37]. Given a Gaussian noise sample, the model learns to iteratively remove a portion of noise until a sample is within the approximated distribution by minimizing the L_2 distance of the predicted noise and the sampled noise. Latent Diffusion Models (LDMs) operate in latent rather than pixel space, encoding videos into a lower-dimensional vector via an encoder-decoder model to simplify the denoising process [32].

Specifically, a video x is represented with a latent vector, and a text condition c is encoded through a text-embedding from a pre-trained text encoder. The initial noise map ϵ is sampled from a Gaussian distribution $\epsilon \sim \mathcal{N}(0,1)$. For the diffusion timestep t sampled from a probability density $t \sim f(t)$, the noisy latent sample can be computed from the noise prediction by $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon$, where x_0 is the original video and α_t controls the amount of noise added at each diffusion timestep according to a noise scheduler [20, 37]. The model ϵ_θ with parameters θ is trained with the following weighted denoising loss,

$$\mathcal{L}_\theta(x, c) = \mathop{\mathbb{E}}_{\substack{\epsilon \sim \mathcal{N}(0,1) \\ t \sim f(t)}} [w_t \|\epsilon_\theta(x, \epsilon, c, t) - \epsilon\|_2^2], \tag{1}$$

where w_t is a user-defined variable that controls the sample quality. At inference time, a new video can be generated by sampling a Gaussian noise sample ϵ and a text prompt, denoising them with the learned model.

3.2 Approach for Motion Customization

We illustrate our overall approach for motion customization in Figure 2. Let the motion be represented through a small exemplar set of videos and corresponding text prompts $D^m = \{(x^m, c^m)\}$. The motion can be performed by different subjects across different backgrounds, and the commonality within the videos is purely the dynamic movement. We choose a generic textual description across the videos, such as "a person doing the V*". In practice, we select rare tokens like "pll" as "V*". To customize a text-to-video model's parameters θ, we fine-tune the model's parameters by minimizing the diffusion objective $\mathcal{L}_\theta$ summed over the exemplar set D^m,

$$\min_\theta \sum_{(x,c) \sim D^m} \mathcal{L}_\theta(x, c). \tag{2}$$

At test time we can generate novel videos of the target with variations controlled by text. Customizing text-to-video generation models has, however, several open technical questions that need to be addressed: (i) Which model parameters to customize? (ii) How to prevent forgetting of already learnt concepts? (iii) How to disentangle motion and appearance? We address the above open questions with the following three technical components, as described next.

Choice of customization parameters. The quality and generalization ability of the novel motion depends on the choice of the model parameters updated during customization. A text-to-video diffusion model ϵ_θ has parameters θ that can be categorized to those operating on the temporal dimensions $\theta_t \subset \theta$ and those operating on the spatial dimensions (per-frame) $\theta_s \subset \theta$. Let $\theta_s^{k,v} \subset \theta_s$ be the keys and values parameters of the spatial cross-attention modules. The temporal layers θ_t are transformers and temporal convolutions, and are responsible for modelling temporal patterns across frames. In Section 4, we empirically show that the temporal layers alone do not effectively model a new motion pattern due to time-varying appearance changes. For example, consider an object rotating in 3D, which requires the model to generate the appearance of disoccluded surfaces. To faithfully learn a motion pattern, we also modify a subset of parameters in the spatial layers of the model. As illustrated in Figure 2, our approach fine tunes the spatial keys/values $\theta_s^{k,v}$ and temporal θ_t parameters. Note that in image customization [24, 38] and model editing [13], it has been shown that the spatial keys and values of the cross-attention modules in text-to-image models are sufficiently powerful to represent the appearance of new concepts. We show the effectiveness of tuning these parameters for motion customization.

Preventing forgetting via video regularization. Prior work has shown that directly optimizing Equation (2) leads to forgetting related concepts or the concept category [25, 26, 33]. For example, if the concept is a specific person, all people start resembling that person. To mitigate this issue, prior work has utilized a regularization set that ensures that this knowledge is preserved. For example, Ruiz *et al.* [33] proposed collecting a regularization set via generation from the original model. Kumari *et al.* [24] proposed using pairs of real images and text.

In contrast to prior work, we seek to mitigate forgetting of related motions that the model has already learned. To address this goal, we consider a video-based regularization. Let D^r be a regularization set of paired videos with text descriptions that have similar but not identical motion to the target custom motion videos D_m. For example, when learning a gesture from the Jester dataset, we chose a regularization set of real videos containing people sitting in front of a webcam. Since this type of video data might be difficult to come across, using a regularization set of real videos in-distribution to the training set is also helpful. For customization, we optimize the diffusion objective $\mathcal{L}_\theta$ over the target custom videos D^m and regularization set D^r:

$$\min_\theta \sum_{(x,c) \sim D^m \cup D^r} \mathcal{L}_\theta(x, c). \tag{3}$$

Empirically, we show in Section 4 that using real videos for the regularization set is superior to using generated videos from the model. We find that using generated videos degrades the quality of the customized model substantially.

Disentangling motion from appearance. To facilitate learning the common motion in D^m, during the training process, we aim to put emphasis on the motion pattern of the videos, rather than the appearance or subject presenting the motion. For example, in the case of the 'Carlton' dance in Figure 1 we wish to

capture the motion pattern of the dance rather than the appearance of the background or the individual performers. The denoising process in diffusion models samples a Gaussian noise and then gradually removes the noise. The initial noise as well as early denoising steps have large influence on the output overall dynamic structure of the video, whereas the later stages correspond to finer details [40]. To focus on the dynamic structure of the videos and de-emphasize the appearance of the subjects performing the motion, we define a timestep sampling strategy. In particular, we build on [21] who develop a non-uniform sampling strategy for generating still images. Here we adopt it for generating dynamic video content. In detail, instead of uniformly sampling the denoising steps in Equation (1), we define a probability distribution $f_\alpha(t) = \frac{1}{T}(1 - \alpha \cos(\frac{\pi t}{T}))$ over the timesteps that focuses more on the earlier denoising steps and hence emphasizes learning the overall dynamic motion pattern rather than the fine appearance details of the background or the individual subjects. The α parameter increases the skewness of the function. We use $\alpha = 0.5$ for all of our experiments.

4   Experiments

4.1   Qualitative Comparison and a User Study

Compared to the published prior work, as shown in Tab. 1, ours is the only method that transforms a text-to-video motion model ($\mathcal{G}_V$) to a model with customized motion ($\mathcal{G}_V^\circ$) or customized motion and appearance ($\mathcal{G}_V^{\circ\dagger}$) (see the website for our joint motion and appearance customization). We adapt prior work that learns or transfers concepts (τ°) to compare with our problem setting. In particular, we focus on "Image customization" methods that customize the appearance of the subject and reference-based ("One-shot video tuning" or "Zero-shot text-driven motion transfer") methods that transfer motion from a single input source video. We adapt these methods to closer match our problem definition as described next. We show results for customizing the base model to selected human gestures, which the base model does not know. For this purpose we use videos from the Jester Dataset [27], which contains crowd-sourced videos of diverse actors performing a gesture in front of a static camera.

Comparison with image customization methods. We select three well-established image customization approaches: Textual Inversion [11], Dreambooth [33] and Custom Diffusion [24], and adapt them to the motion customization task. Textual Inversion [11] optimizes a text token embedding to learn a novel subject. For Dreambooth [33], we optimize all spatial layers of the model and use a synthetic video regularization set following their prior preservation loss. Custom Diffusion trains the key and value projection matrices in the cross-attention layers of text-to-image diffusion models, optimizes a text token embedding, and uses a regularization set with real videos.

We choose at random 10 videos from the validation set from the Jester dataset [27] class "Sliding Two Fingers Up" as our training set. To learn the motion, we caption the videos with a generic text description "a person doing a V* sign".
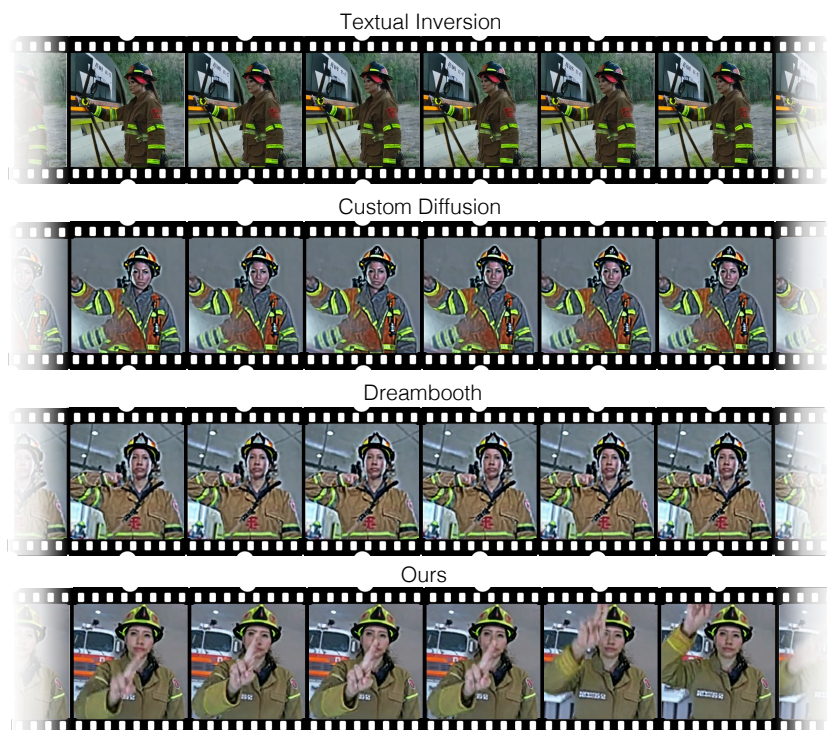
Fig. 3: Visual comparison with baseline methods. Examples of learning a customized motion *Sliding Two Fingers Up* from the Jester dataset with prompt "*A female firefighter doing the V* sign*". Baseline methods (top three rows) fail to capture the motion and produce a temporally coherent video.

For the motion learning regularization set, we choose 50 videos from each of the dataset classes {*Doing Other Things*, *No gesture*} and assign them text captions "a person doing other things" and "a person in front of a computer". We generate results with a prompt "*A female firefighter doing the V* sign*". Fig 3 shows that, in contrast to ours, the baseline methods fail to capture the motion and produce a temporally smooth video.
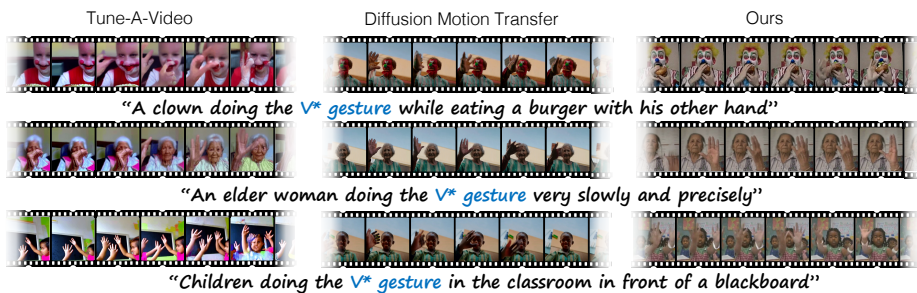
Comparison with a reference-based method. In Figure 5 we compare our approach to two reference-based methods [47, 50]. We adapt Tune-A-Video to fine-tune from multiple videos using the same training set as our setup. We test the generalization abilities of the methods, firstly combining the novel motion with a different motion: "doing the V* gesture while eating a burger with other hand". We can see that our method can generate the two motions. Next, we test the motion variation. We show that we can manipulate the execution of the motion with a natural language description; here we specify "doing the V* gesture very slowly and precisely". Lastly, we show that our method can generalize to multiple people doing the motion while preserving the temporal consistency of

Input videos   Our motion customizations
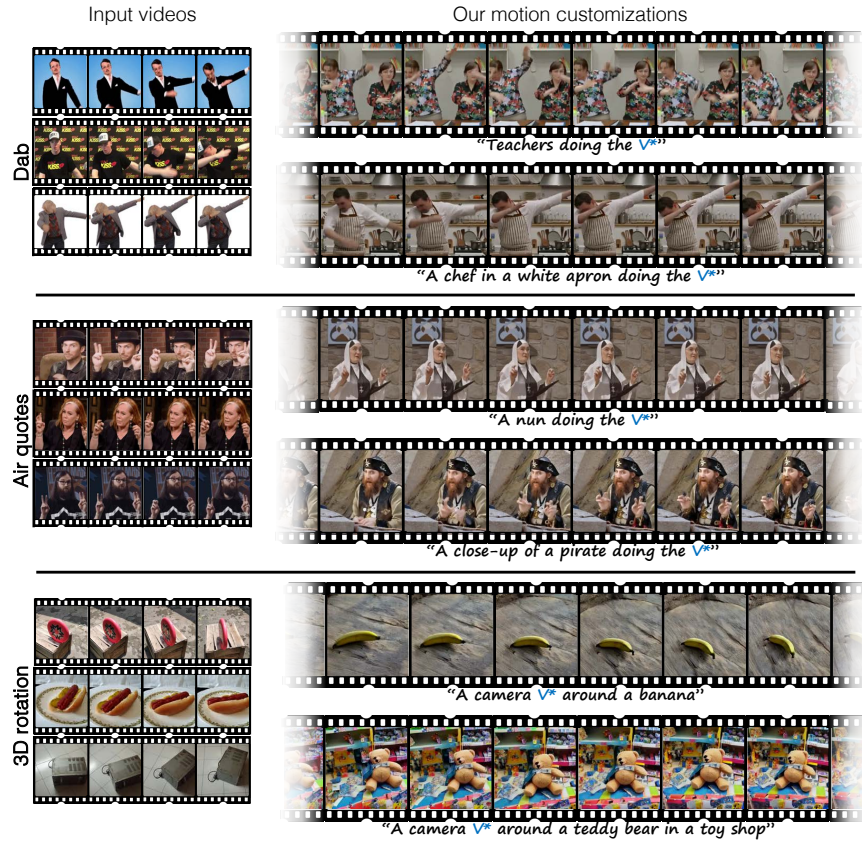
Fig. 4: Qualitative results of our method. We demonstrate two custom motions: Dab and Air quotes, trained using collected internet examples as well as a 3D camera rotation trained with examples from the CO3D dataset [31]. Our method can generalize to unseen subjects and multiple people performing the action.

the video. The Tune-A-Video and Diffusion Motion Transfer baseline produces qualitatively worse results and do not generalize to multiple people performing a motion, motion variations and combinations with other movements (Figure 5).

To further demonstrate the effectiveness of our approach on a broader class of motions, we choose the motions: *Carlton*, *Dab*, *Airquotes*, and *3D rotation*, which the original model does not know or cannot reliably perform. For the first two motions, we curate videos from the internet searching for the description of the motion. We use between 5 and 6 videos for each motion and caption the videos with a generic text description such as "A person doing the V* dance", "A person doing the V* sign" or "A person V*". To learn 3D rotation, we select 10 sequences of distinct objects from the CO3D Dataset [31] and use as captions text prompt "A camera V* around an [object]", where [object] is an object class.

Tune-A-Video   Diffusion Motion Transfer   Ours

"A clown doing the V* gesture while eating a burger with his other hand"

"An elder woman doing the V* gesture very slowly and precisely"

"Children doing the V* gesture in the classroom in front of a blackboard"

Fig. 5: Text-driven motion transfer methods versus our method trained on few examples of a custom motion "Shaking Hand". Our method seamlessly renders a custom motion in novel scenarios. Despite the training videos showing only a single actor performing one motion, our method generates the custom motion alongside another action (doing the gesture while eating a burger") and varies timing (doing the gesture slowly and precisely") or involves multiple people ("children"). In contrast, both baselines fail to generalize or produce temporally coherent videos.

Due to the lack of readily available regularization set of similar motions, we select 100 video-text pairs from the WebVid-10M Dataset [1]. The results are shown in Figure 4. For visualization, we query for objects not present in the training videos.

Fig. 6: User preference comparison.

User Study. To further demonstrate the benefits of our approach, we conducted a binary forced-choice user study on a diverse set of dynamic scenes involving complex motions. Involving 6 users, we assessed 50 prompts per motion (Carlton, Airquotes, and Dab, see Figures 1, 4), resulting in 300 votes. We presented users with videos from our method and a baseline (DreamBooth and Tune-A-Video), presented in a random order, and asked users to select the video that best matched the input motion. Users select our method more often (Figure 6), with statistical significance (p-values<0.01), further confirming the automatic quantitative evaluation performed in Section 4.2.

4.2   Automated quantitative evaluation via recognition

In this section we introduce an automated quantitative evaluation of the generated videos by employing pre-trained descriptors [29, 30], and motion-specific classifiers [54]. First, we introduce the used dataset and describe the proposed metrics. This automatic evaluation protocol then allows us to perform an exhaustive ablation study of the different components of our method. Finally, we use this automatic evaluation to compare our approach to other competing methods.
Dataset. Evaluating the motion pattern separately from the appearance is challenging. We evaluate on the Jester dataset [27], which requires gesture recogni-

tion models to generalize beyond appearance and learn the specific motion pattern. The dataset contains 148,092 crowd-sourced videos of 27 human gestures by diverse actors performing a gesture in front of a static camera. We select 5 gesture classes as our target motions – {*Swiping Down*, *Drumming Finders*, *Rolling Hand Forward*, *Shaking Hand*, *Sliding Two Fingers Up*}. From each class, we choose at random 10 videos from the validation set as our training set. To learn the motion, we caption the videos with a generic text description "a person doing a V* sign". As a regularization set for the motion learning, we also choose 50 videos from each of the dataset classes {*Doing Other Things*, *No gesture*} and assign them text captions "a person doing other things" and "a person in front of a computer". We design a test set containing 100 text prompts across 3 random seeds, that detail a person's appearance, for example "A female firefighter doing the V* gesture" or "A Japanese toddler doing the V gesture". **Motion accuracy score.** The Jester dataset has been widely adopted for the task of action recognition. We leverage this for our motion accuracy score, measuring gesture recognition accuracy with a video classifier, pre-trained on the Jester training set [54].The classifier reaches 94.78% accuracy for all gesture classes, fitting our evaluation needs. Test set prompts often yield videos matching the classifier's training layout with a single person, thus being in-distribution. **Text alignment scores.** To regulate overfitting, we measure how faithfully the model adheres to the appearance described in the text prompt with an appearance score defined as the mean CLIP score between the generated frames and the part of the text prompt describing the person's appearance (*e.g.*, "A female firefighter"). We empirically find that the **appearance score** is a good proxy to determine when the model starts to overfit. For a fair comparison across all models, we fine-tune each model until this appearance score reaches a chosen value (here 0.265), determined empirically based on the visual quality of the results. To assess how well the generated video corresponds to the overall text prompt (both motion and appearance), we measure the **text alignment score** as a mean CLIP score between the video frames and a full caption written in natural language. For example, for the class *Swiping Down* and a text prompt "A female firefighter doing the V* gesture", a full caption would be "A female firefighter swiping hand down." We use this score as another criterion to compare different methods. **Copying score.** We report a copying score **("Copy")** that measures how much of the training data appearance leaks into the generated data. We use the SSCD description score [29] to detect image copying between the training videos and the generated videos. The copying score is the maximum SSCD score between generated and training frames, with the percentage of samples above a manually set threshold of 0.25, indicating significant copying of the person or background.

Ablation study. We perform an ablation study and report motion accuracy ("Accuracy") and copying score ("Copy") in Table 2.

First, we study the choice of the optimized parameters. As in the case of the image customization methods, it is clear that optimizing the spatial layers of the model is not enough to learn a novel motion pattern, as shown in Table 2b. Training the key and value projections in the cross-attention layers of the spatial

Table 2: Quantitative results of the ablation study. Each table examines the design choices of our method. We report the motion recognition accuracy ("Accuracy") obtained with a pre-trained classifier for gesture recognition. The copying score ("Copy") is the percentage of generated videos with a detection score above a set threshold.

(a) Spatial layers

Spatial layers	Accuracy ↑	Copy ↓
None	62.3	**5.9**
All	68.7	17.0
K,V (Ours)	**70.6**	8.7

(b) Temporal layers

Trans. layers	Accuracy ↑	Copy ↓
None	22.7	1.4
All (Ours)	**70.6**	8.7

(c) Text Token

Text Token	Accuracy ↑	Copy ↓
✓	**75.5**	22.7
✗ (Ours)	70.6	**8.7**

(d) Sampling Strategy

Sampling	Accuracy ↑	Copy ↓
Uniform	66.9	15.4
Coarse-noise (Ours)	**70.6**	**8.7**

(e) Regularization

Prior	Accuracy ↑	Copy ↓
None	43.9	**1.2**
Synthetic	48.3	3.7
Real (WebVid10M)	66.7	10.9
Real (Jester)	**70.6**	8.7

(f) Fine-tuning Strategy

Tuning Method	Layers	Accuracy ↑	Copy ↓
LoRA	All	17.4	**0.5**
LoRA	Ours	10.6	3.3
Full	All	68.7	17.0
Full	Ours	**70.6**	8.7

layers achieves a good balance in terms of accuracy and copying score, as shown Table 2a. We observe a two-fold reduction in the copying score when training only those parameters and an 8% improvement in motion accuracy compared to not training any spatial parameters.

Next, in Table 2c, we consider the role of optimizing the text token embedding. We notice that when jointly training the text token embedding with the model's parameters, the motion accuracy is higher than when training only the model parameters. However, it also leads to nearly three times as much memorization of the training video appearance, as evidenced by the significantly increased Copy score. We hypothesize that this indicates that the text token embedding is learning something related to the appearance of the training videos. To test our sampling strategy we compare a model trained with our motion pattern sampling strategy ("Coarse-noise") to a model that simply samples diffusion time steps uniformly ("Uniform"). As shown in Table 2d, our sampling strategy improves the motion accuracy and reduces the appearance copying. Following our intuition regarding learning motion patterns rather than fine-grained details, our model is less prone to replicating the training examples.

We study the effect of the regularization prior in Table 2e. We train four models with the same design choice as our method, yet we choose the regularization set to be either: (i) empty, (ii) a synthetic set of videos generated from the original text-to-video model, or (iii) real videos from the WebVid10m dataset (iv) containing real videos that contain similar visual scenes. We observe a significant drop in performance both without using a prior as well as using a synthetic prior. Using real data is beneficial for fine-tuning. We observe that using motions and visual scenes similar to the target motion is optimal. However, should this support set be unavailable, off-the-shelf dataset is still better than no regularization or synthetic video. Overall, our method achieves a good balance between high motion accuracy and low copying score. Lastly, we compare full parameter fine-tuning to low-rank adaptation in Table 2f. We test two variants; adding LoRA adapters to both spatial and temporal transformers (first row),

and to the same layers as our NewMove (second row). In our experiments we use rank 32 and generate videos with LoRA scale 1, we observe that full parameter fine-tuning achieves higher motion accuracy. Qualitatively, we observe that LoRA models perform relatively well on coarser gestures like *Drumming Fingers* but struggle to learn fine-grained motions like *Sliding Two Fingers Up.*

Quantitative evaluation of motion fidelity. We quantitatively evaluate our approach by computing metrics corresponding to the quality of the generated motion and the overall fidelity of the generated video with respect to the input text prompt. We compare to baseline methods that have explicitly learned the motion pattern from a few examples. We follow the evaluation protocol as defined in Section 4.2. We report the motion accuracy and text alignment metrics across different methods in Table 3 and show qualitative comparison (Figure 3).

We observe that Textual Inversion completely fails to learn the motion, this failure is potentially because the text encoder has been trained only on image-text pairs, and the embedding fails to generalize to the video domain. Additionally, the text embedding is injected only into spatial

Table 3: Quantitative comparison with baseline methods

	Motion accuracy ↑	Text Alignment ↑
Textual Inversion [11]	0.3	0.2733
Custom Diffusion [24]	10.5	0.2788
Dreambooth [33]	28.4	0.2796
Tune-a-Video [47]	18.9	**0.2818**
MotionDirector [53]	24.7	0.2786
Ours	**70.6**	**0.2818**

layers in the cross-attention layers, and not the temporal ones. Alternatively, because we are learning an unseen motion pattern, which is more complex than a static object, the embedding does not have enough parameters to learn it. Dreambooth and Custom Diffusion learn to adhere to the spatial structure of the training videos and some hand manipulation. However they fail to accurately reproduce the motion and produce a temporally smooth video. Our approach yields over a $2\times$ improvement in motion accuracy and more faithfully adheres to the text prompt. We also compare with the concurrent MotionDirector method [53] and find that our approach outperforms it in motion accuracy and text alignment. However, with longer training, the baseline achieves similar performance (see Appendix for details and qualitative comparison).

5   Conclusion

We present an approach for motion customization in text-to-video diffusion models. Our method can learn a new motion pattern from a set of few exemplar videos of different subjects performing the same motion. We conduct a thorough ablation study that identifies the key components of the method and evaluate them with respect to motion accuracy. We demonstrate qualitative results of our method and evaluate our method with a user study, specifically in scenarios where our method generalizes the motion to unseen actors, multiple people performing the motion, and different viewpoints.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
3. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
4. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23040–23050 (2023)
5. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5933–5942 (2019)
6. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
7. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. arXiv preprint arXiv:2401.09047 (2024)
8. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
9. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Advances in Neural Information Processing Systems **36** (2024)
10. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
12. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) (2023)
13. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
14. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
15. Gleicher, M.: Retargetting motion to new characters. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques (1998)
16. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning (v2) (2024)

17. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning (v1). arXiv preprint arXiv:2307.04725 (2023)
18. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
19. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
21. Huang, Z., Wu, T., Jiang, Y., Chan, K.C., Liu, Z.: ReVersion: Diffusion-based relation inversion from images. arXiv preprint arXiv:2303.13495 (2023)
22. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. arXiv preprint arXiv:2312.00845 (2023)
23. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the International Conference on Computer Vision (ICCV) (2023)
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
25. Lee, J., Cho, K., Kiela, D.: Countering language drift via visual grounding. arXiv preprint arXiv:1909.04499 (2019)
26. Lu, Y., Singhal, S., Strub, F., Courville, A., Pietquin, O.: Countering language drift with seeded iterated learning. In: International Conference on Machine Learning. PMLR (2020)
27. Materzynska, J., Berger, G., Bax, I., Memisevic, R.: The jester dataset: A large-scale video dataset of human gestures. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019)
28. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329 (2023)
29. Pizzi, E., Roy, S.D., Ravindra, S.N., Goyal, P., Douze, M.: A self-supervised descriptor for image copy detection. Proc. CVPR (2022)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR (2021)
31. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. 2022 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685 (2021)
33. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)

34. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
35. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems (2019)
36. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
38. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. SIGGRAPH '23 (2023)
39. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
40. Wang, B., Vastola, J.J.: Diffusion models generate images like painters: an analytical theory of outline first, details later. arXiv preprint arXiv:2303.02490 (2023)
41. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
42. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. Advances in Neural Information Processing Systems **31** (2018)
43. Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J., Liu, J.: Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. arXiv preprint arXiv:2305.10874 (2023)
44. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. arXiv preprint arXiv:2312.04433 (2023)
45. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
46. Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV) (2018)
47. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
48. Wu, R., Chen, L., Yang, T., Guo, C., Li, C., Zhang, X.: Lamp: Learn a motion pattern for few-shot-based video generation. arXiv preprint arXiv:2310.10769 (2023)
49. Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., Lin, D.: Pose guided human video generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 201–216 (2018)
50. Yatim, D., Fridman, R., Tal, O.B., Kasten, Y., Dekel, T.: Space-time diffusion features for zero-shot text-driven motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

51. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
52. Zhang, Y., Tang, F., Huang, N., Huang, H., Ma, C., Dong, W., Xu, C.: Motioncrafter: One-shot motion customization of diffusion models. arXiv preprint arXiv:2312.05288 (2023)
53. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023)
54. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. European Conference on Computer Vision (2018)
55. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)
56. Zhou, Y., Zhang, R., Sun, T., Xu, J.: Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. arXiv preprint arXiv:2305.13579 (2023)