VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models

Hyeonho Jeong[1*] Geon Yeong Park[2*] Jong Chul Ye[1,2]
[1]Kim Jaechul Graduate School of AI, [2]Bio and Brain Engineering
Korea Advanced Institute of Science and Technology (KAIST)
* Indicates co-first authors
{hyeonho.jeong, pky3436, jong.ye}@kaist.ac.kr

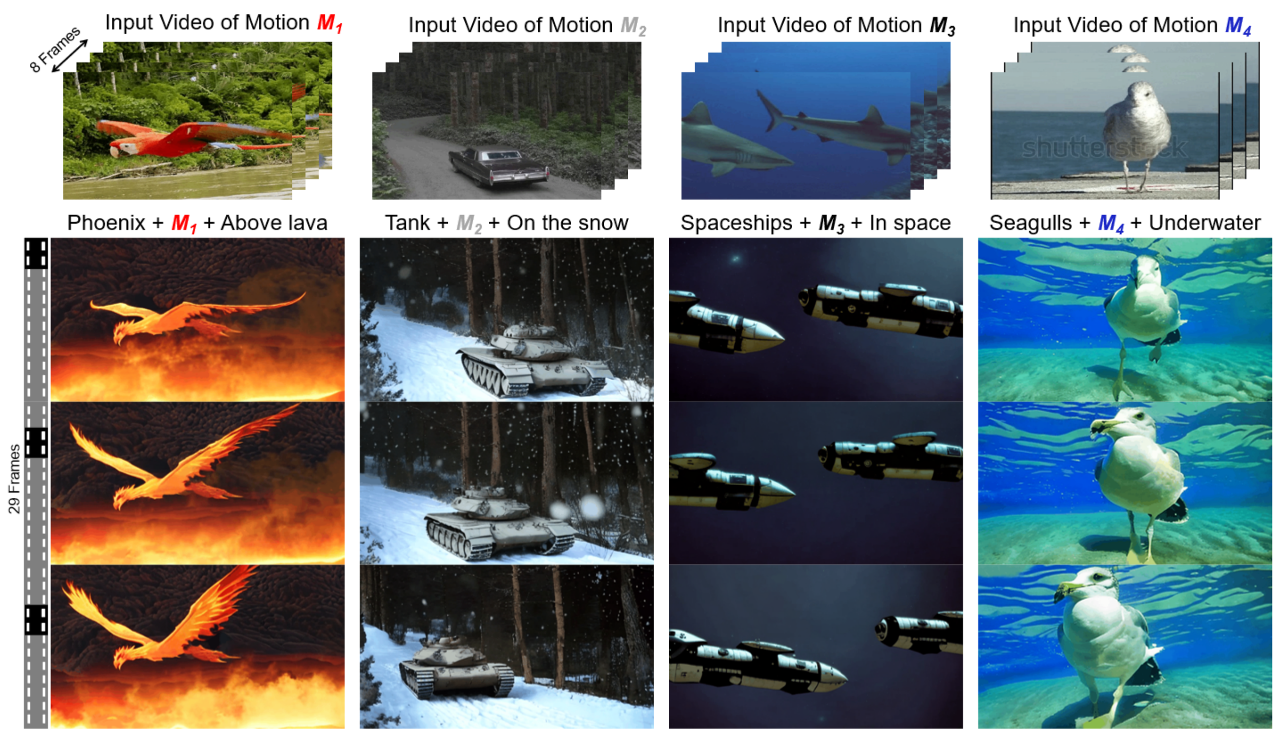

Figure 1. Using only a single video portraying any type of motion, our **V**ideo **M**otion **C**ustomization framework allows for generating a wide variety of videos characterized by the same motion but in entirely distinct contexts and better spatial/temporal resolution. 8-frame input videos are translated to 29-frame videos in different contexts while closely following the target motion. The visualized frames for the first video are at indexes 1, 9, and 17. A comprehensive view of these motions in the form of videos can be explored at our project page.

Abstract

*Text-to-video diffusion models have advanced video generation significantly. However, customizing these models to generate videos with tailored motions presents a substantial challenge. In specific, they encounter hurdles in (**a**) accurately reproducing motion from a target video, and (**b**) creating diverse visual variations. For example, straightforward extensions of static image customization methods to video often lead to intricate entanglements of appearance and motion data. To tackle this, here we present the Video Motion Customization (**VMC**) framework, a novel one-shot tuning approach crafted to adapt temporal attention layers within video diffusion models. Our approach introduces a novel motion distillation objective using residual vectors between consecutive noisy latent frames as a motion reference. The diffusion process then preserve low-frequency motion trajectories while mitigating high-*

frequency motion-unrelated noise in image space. We validate our method against state-of-the-art video generative models across diverse real-world motions and contexts. Our code and data can be found at https://video-motion-customization.github.io/.

1. Introduction

The evolution of diffusion models [12, 26, 29] has significantly advanced Text-to-Image (T2I) generation, notably when paired with extensive text-image datasets [3, 23]. While cascaded diffusion pipelines [2, 9, 13, 25, 31, 34, 36] have extended this success to Text-to-Video (T2V) generation, current models lack the ability to replicate specific motions or generate diverse variations of the same motion with distinct visual attributes and backgrounds. Addressing this, we tackle the challenge of Motion Customization [35]—adapting pre-trained Video Diffusion Models (VDM) to produce motion-specific videos in different contexts, while maintaining the same motion patterns of target subjects.

Given a few subject images for reference, appearance customization [8, 17, 21, 22, 24, 32] in generative models aims to fine-tune models to generate subject images in diverse contexts. However, these approaches, despite varying optimization objectives, commonly strive for *faithful* image (frame) reconstruction by minimizing the ℓ_2-distance between predicted and ground-truth noise. This may lead to the *entangled* learning of appearance and motion.

To tackle this, we present **VMC**, a new framework aimed at adapting pre-trained VDM's temporal attention layers via our proposed *Motion Distillation* objective. This approach utilizes residual vectors between consecutive (latent) frames to obtain the motion vectors that trace motion trajectories in the target video. Consequently, we fine-tune VDM's temporal attention layers to align the ground-truth image-space residuals with their denoised estimates, which equivalently aligns predicted and ground-truth source noise differences and motion vectors within VDM. This enables lightweight and fast one-shot training. To further facilitate the appearance-invariant motion distillation, we transform faithful text prompts into appearance-invariant prompts, e.g. `"A bird is flying above a lake in the forest"` → `"A bird is flying"` in Fig. 1. This encourages the modules to focus on the motion information and ignore others, such as appearance, structure, background, etc. During inference, our procedure initiates by sampling key-frames using the adapted key-frame generation U-Net, followed by temporal interpolation and spatial super-resolution. To summarize, VMC makes the following key contributions:

- We introduce a novel fine-tuning strategy which focuses solely on temporal attention layers in the key-frame generation module. This enables lightweight training (15GB vRAM) and fast training (< 5 minutes).
- To our knowledge, we mark a pioneering case of fine-tuning only the temporal attention layers in video diffusion models, without optimizing spatial self or cross-attention layers, while achieving successful motion customization.
- We introduce a novel motion distillation objective that leverages the residual vectors between consecutive (latent) frames as motion vectors.
- We present the concept of appearance-invariant prompts, which further facilitates the process of motion learning when combined with our motion distillation loss.

2. Preliminaries

Diffusion Models. Diffusion models aim to generate samples from the Gaussian noise through iterative denoising processes. Given a clean sample $\boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x})$, the forward process is defined as a Markov chain with forward conditional densities

$$\begin{aligned} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) &= \mathcal{N}(\boldsymbol{x}_t \mid \beta_t \boldsymbol{x}_{t-1}, (1-\beta_t)I) \\ p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) &= \mathcal{N}(\boldsymbol{x}_t \mid \sqrt{\bar{\alpha}}\boldsymbol{x}_0, (1-\bar{\alpha})I), \end{aligned} \tag{1}$$

where $\boldsymbol{x}_t \in \mathbb{R}^d$ is a noisy latent variable at a timestep t that has the same dimension as $\boldsymbol{x}_0$, and β_t denotes an increasing sequence of noise schedule where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \Pi_{i=1}^{t} \alpha_i$. Then, the goal of diffusion model training is to obtain a residual denoiser $\boldsymbol{\epsilon}_\theta$:

$$\min_\theta \mathbb{E}_{\boldsymbol{x}_t \sim p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0), \boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(0, I)} \big[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}\| \big]. \tag{2}$$

It can be shown that this epsilon matching in (2) is equivalent to the Denoising Score Matching (DSM [14, 28]) with different parameterization:

$$\min_\theta \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{\epsilon}} \Big[\big\| \boldsymbol{s}_\theta^t(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) \big\| \Big], \tag{3}$$

where $\boldsymbol{s}_{\theta^*}(\boldsymbol{x}_t, t) \simeq -\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{1-\bar{\alpha}} = -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_{\theta^*}(\boldsymbol{x}_t, t)$. The reverse sampling from $q(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t, \boldsymbol{\epsilon}_{\theta^*}(\boldsymbol{x}_t, t))$ is then achieved by

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \Big(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_{\theta^*}(\boldsymbol{x}_t, t) \Big) + \tilde{\beta}_t \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$ and $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$. To accelerate sampling, DDIM [27] further proposes another sampling method as follows:

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\boldsymbol{x}}_0(t) + \sqrt{1 - \bar{\alpha}_{t-1} - \eta^2 \tilde{\beta}_t^2} \boldsymbol{\epsilon}_{\theta^*}(\boldsymbol{x}_t, t) + \eta \tilde{\beta}_t \boldsymbol{\epsilon}, \tag{5}$$

where $\eta \in [0, 1]$ is a stochasticity parameter, and $\hat{\boldsymbol{x}}_0(t)$ is the denoised estimate which can be equivalently derived using Tweedie's formula [6]:

$$\hat{\boldsymbol{x}}_0(t) := \frac{1}{\sqrt{\bar{\alpha}_t}} (\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_{\theta^*}(\boldsymbol{x}_t, t)). \tag{6}$$

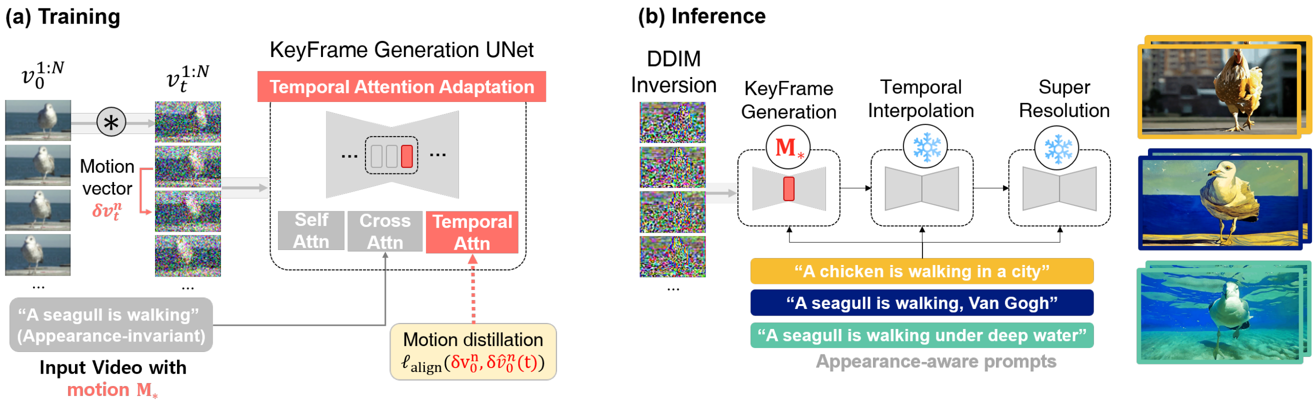

Figure 2. **Overview**. The proposed Video Motion Customization (**VMC**) framework distills the motion trajectories from the residual between noisy latent frames, namely motion vector $\delta \boldsymbol{v}_t^n$. Specifically, we fine-tune only the temporal attention layers of the key-frame generation model by aligning the ground-truth and predicted motion vectors. After training, the customized key-frame generator is leveraged for target motion-driven video generation with new appearances context, e.g. `"A chicken is walking in a city"`.

For a text-guided Diffusion Model, the training objective is often given by:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{\epsilon}, c} \Big[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, c) - \boldsymbol{\epsilon}\| \Big], \tag{7}$$

where c represents the textual embedding. Throughout this paper, we will often omit c from $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, c)$ if it does not lead to notational ambiguity.

Video Diffusion Models. Video diffusion models [11, 13, 34] further attempt to model the video data distribution. Specifically, Let $(\boldsymbol{v}^n)_{n \in \{1,\ldots,N\}}$ represents the N-frame input video sequence. Then, for a given n-th frame $\boldsymbol{v}^n \in \mathbb{R}^d$, let $\boldsymbol{v}^{1:N} \in \mathbb{R}^{N \times d}$ represents a whole video vector. Let $\boldsymbol{v}_t^n = \sqrt{\bar{\alpha}_t}\boldsymbol{v}^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t^n$ represents the n-th noisy frame latent sampled from $p_t(\boldsymbol{v}_t^n|\boldsymbol{v}^n)$, where $\boldsymbol{\epsilon}_t^n \sim \mathcal{N}(0, I)$. We similarly define $(\boldsymbol{v}_t^n)_{n \in 1,\ldots,N}$, $\boldsymbol{v}_t^{1:N}$, and $\boldsymbol{\epsilon}_t^{1:N}$. The goal of video diffusion model training is then to obtain a residual denoiser $\boldsymbol{\epsilon}_\theta$ with textual condition c and video input that satisfies:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{v}_t^{1:N}, \boldsymbol{v}^{1:N}, \boldsymbol{\epsilon}_t^{1:N}, c} \Big[\big\|\boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t^{1:N}, t, c) - \boldsymbol{\epsilon}_t^{1:N}\big\| \Big], \tag{8}$$

where $\boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t^{1:N}, t, c), \boldsymbol{\epsilon}_t^{1:N} \in \mathbb{R}^{N \times d}$. In this work, we denote the predicted noise of n-th frame as $\boldsymbol{\epsilon}_\theta^n(\boldsymbol{v}_t^{1:N}, t, c) \in \mathbb{R}^d$.

In practice, contemporary video diffusion models often employ cascaded inference pipelines for high-resolution outputs. For instance, [34] initially generates a low-resolution video with strong text-video correlation, further enhancing its resolution via temporal interpolation and spatial super-resolution modules.

In exploring video generative tasks through diffusion models, two primary approaches have emerged: leveraging foundational Video Diffusion Models (VDMs) [7, 16, 30, 35] or pre-trained Text-to-Image (T2I) models [4, 10, 15, 32]. To extend image diffusion models to videos, several architectural modifications are made. Typically, U-Net generative modules integrate temporal attention blocks after spatial attentions [11]. Moreover, 2D convolution layers are inflated to 3D convolution layers by altering kernels [11].

3. Video Motion Customization

Given an input video, our main goal is to (**a**) distill the motion patterns M_* of target subjects, and (**b**) customize the input video in different contexts while maintaining the same motion patterns M_*, e.g. `Sharks w/ motion` M_* → `Airplanes w/ motion` M_*, with minimal computational costs.

To this end, we propose a novel video motion customization framework, namely **VMC**, which leverages cascaded video diffusion models with robust temporal priors. One notable aspect of the proposed framework is that we perform fine-tuning *only* on the key-frame generation module, also referred to as the T2V base model, within the cascaded VDMs, which guarantees computational and memory efficiency. Specifically, within the key-frame generation model, our fine-tuning process *only* targets the temporal attention layers. This facilitates adaptation while preserving the model's inherent capacity for generic synthesis. Notably, we *freeze* the subsequent frame interpolation and spatial super-resolution modules as-is (Fig. 2).

3.1. Temporal Attention Adaptation

In order to distill the motion M_*, we first propose a new objective function for temporal attention adaptation using residual cosine similarity. Our intuition is that residual vectors between consecutive frames may include information about the motion trajectories.

Let $(\boldsymbol{v}^n)_{n \in \{1,\ldots,N\}}$ represents the N-frame input video

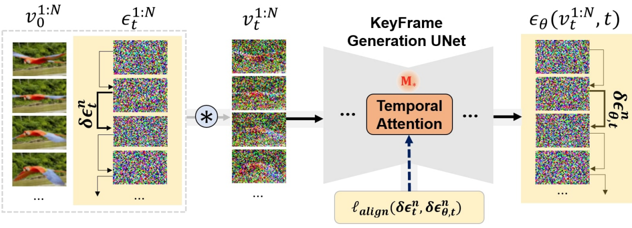

Figure 3. **Training.** The proposed framework aims to learn motion by $\delta\epsilon_t^n$-alignment using (16) or (17). Note that we only fine-tune the temporal attention layers in the key-frame generation U-Net. The blue circle represents the diffusion forward process.

sequence. As defined in Section 2, for a given noisy video latent vector $\boldsymbol{v}_t^{1:N}$ with $\boldsymbol{\epsilon}_t^{1:N}$, let $\boldsymbol{v}_t^n$ represents the n-th noisy frame latent sampled from $p_t(\boldsymbol{v}_t^n|\boldsymbol{v}^n)$ with $\boldsymbol{\epsilon}_t^n$. We will interchangeably use $\boldsymbol{v}^n$ and $\boldsymbol{v}_0^n$ for notational simplicity. Likewise, $\boldsymbol{v}_t^{n+c}$ is defined as $\boldsymbol{v}_t^n$, with $c > 0$ representing the fixed frame stride. Then, we define the frame residual vector at time $t \geqslant 0$ as

$$\delta\boldsymbol{v}_t^n := \boldsymbol{v}_t^{n+c} - \boldsymbol{v}_t^n, \tag{9}$$

where we similarly define the epsilon residual vector $\delta\boldsymbol{\epsilon}_t^n$. In the rest of the paper, we interchangeably use frame residual vector and *motion vector*.

We expect that these motion vectors may encode information about motion patterns, where such information may vary depending on the time t and its corresponding noise level. The difference vector $\delta\boldsymbol{v}_t^n$ can be delineated as:

$$\begin{aligned}\delta\boldsymbol{v}_t^n &= \sqrt{\bar{\alpha}_t}(\boldsymbol{v}_0^{n+c} - \boldsymbol{v}_0^n) + \sqrt{1-\bar{\alpha}_t}(\boldsymbol{\epsilon}_t^{n+c} - \boldsymbol{\epsilon}_t^n)\\ &= \sqrt{\bar{\alpha}_t}\delta\boldsymbol{v}_0^n + \sqrt{1-\bar{\alpha}_t}\delta\boldsymbol{\epsilon}_t^n,\end{aligned} \tag{10}$$

where $\delta\boldsymbol{\epsilon}_t^n$ is normally distributed with zero mean and $2I$ variance. In essence, $\delta\boldsymbol{v}_t^n$ can be acquired through the following diffusion kernel:

$$p(\delta\boldsymbol{v}_t^n \mid \delta\boldsymbol{v}_0^n) = \mathcal{N}(\delta\boldsymbol{v}_t^n \mid \sqrt{\bar{\alpha}_t}\delta\boldsymbol{v}_0^n, 2(1-\bar{\alpha}_t)I). \tag{11}$$

In light of this, our goal is to transfer motion information to the temporal attention layers by leveraging the motion vectors. For this, we first simulate the motion vectors using video diffusion models. Specifically, as similarly done in (6), the denoised video vector estimates $\hat{\boldsymbol{v}}_0{}^{1:N}(t)$ can be derived by applying Tweedie's formula:

$$\hat{\boldsymbol{v}}_0^{1:N}(t) := \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{v}_t^{1:N} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t^{1:N}, t)\right), \tag{12}$$

where $\hat{\boldsymbol{v}}_0^{1:N}(t)$ is an empirical Bayes optimal posterior expectation $\mathbb{E}[\boldsymbol{v}_0^{1:N} \mid \boldsymbol{v}_t^{1:N}]$. Then, the denoised motion vector estimate $\delta\hat{\boldsymbol{v}}_0^n$ can be defined in terms of $\delta\boldsymbol{v}_t^n$ and $\delta\boldsymbol{\epsilon}_\theta^n(\boldsymbol{v}_t^{1:N}, t)$ by using (12):

$$\delta\hat{\boldsymbol{v}}_0^n(t) := \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\delta\boldsymbol{v}_t^n - \sqrt{1-\bar{\alpha}_t}\delta\boldsymbol{\epsilon}_{\theta,t}^n\right), \tag{13}$$

where $\delta\boldsymbol{\epsilon}_\theta^n(\boldsymbol{v}_t^{1:N}, t) := \boldsymbol{\epsilon}_\theta^{n+c}(\boldsymbol{v}_t^{1:N}, t) - \boldsymbol{\epsilon}_\theta^n(\boldsymbol{v}_t^{1:N}, t)$ is abbreviated as $\delta\boldsymbol{\epsilon}_{\theta,t}^n$ for notational simplicity. Similarly, one can obtain ground-truth motion vector $\delta\boldsymbol{v}_0^n$ by using (10):

$$\delta\boldsymbol{v}_0^n = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\delta\boldsymbol{v}_t^n - \sqrt{1-\bar{\alpha}_t}\delta\boldsymbol{\epsilon}_t^n\right). \tag{14}$$

Then, our objective is to finetune θ by *aligning* the motion vector $\delta\boldsymbol{v}_0^n$ and its denoised estimate $\delta\hat{\boldsymbol{v}}_0^n(t)$:

$$\min_\theta \mathbb{E}_{t,n,\boldsymbol{\epsilon}^{t,n},\boldsymbol{\epsilon}^{t,n+c}}\left[\ell_{\text{align}}\left(\delta\boldsymbol{v}_0^n, \delta\hat{\boldsymbol{v}}_0^n(t)\right)\right], \tag{15}$$

with a loss function $\ell_{\text{align}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. By using ℓ_2-distance for ℓ_{align}, this is equivalent to matching $\delta\boldsymbol{\epsilon}_{\theta,t}^n$ and $\delta\boldsymbol{\epsilon}_t^n$:

$$\ell_{\text{align}}\left(\delta\boldsymbol{v}_0^n, \delta\hat{\boldsymbol{v}}_0^n(t)\right) = \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\left\|\delta\boldsymbol{\epsilon}_t^n - \delta\boldsymbol{\epsilon}_{\theta,t}^n\right\|^2. \tag{16}$$

Notably, aligning the ground-truth and predicted motion vectors translates into aligning epsilon residuals.

While this objective demonstrates effective empirical performance, our additional observations indicate that using $\ell_{\cos}(\delta\boldsymbol{\epsilon}_t^n, \delta\boldsymbol{\epsilon}_{\theta,t}^n)$ may further improve the distillation, where $\ell_{\cos}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ (more analysis in section 4.3). Accordingly, our optimization framework is finally defined as follows:

$$\min_\theta \mathbb{E}_{t,n,\boldsymbol{\epsilon}^{t,n},\boldsymbol{\epsilon}^{t,n+c}}[\ell_{\cos}(\delta\boldsymbol{\epsilon}_t^n, \delta\boldsymbol{\epsilon}_{\theta,t}^n)]. \tag{17}$$

In other words, the proposed optimization framework aims to *maximize* the residual cosine similarity between $\delta\boldsymbol{\epsilon}_t^n$ and $\delta\boldsymbol{\epsilon}_{\theta,t}^n$. Hence, this optimization approach can be seamlessly applied to video diffusion models trained using epsilon-matching. Practically, we exclusively fine-tune the temporal attention layers $\theta_{\text{TA}} \subset \theta$, originally designed for dynamic temporal data assimilation [32]. The frame stride remains fixed at $c = 1$ across all experiments.

3.2. Appearance-invariant Prompts

In motion distillation, it is crucial to filter out disruptive variations that are unrelated to motion. These variations may include changes in appearance and background, distortions, consecutive frame inconsistencies, etc. To achieve this, we further utilize *appearance-invariant prompts*. Diverging from traditional generative customization frameworks [21, 22, 32, 35] that rely on text prompts that "faithfully" describe the input image or video during model fine-tuning, our framework purposedly employs "unfaithful" text prompts during the training phase. Specifically, our approach involves the removal of background information. For instance, the text prompt 'a cat is roaring on the grass under the tree' is simplified to 'a cat is roaring' as presented in Fig. 4. This reduces background complexity as in Fig. 4a comapred to Fig. 4b, facilitating the application of new appearance in motion distillation.

Figure 4. **Appearance-invariant Prompt**. Comparison of input reconstruction with and without appearance-invariant prompt: (a) and (b) depict sampled low-resolution (64x40) keyframes. For (a), the training prompt used was "A cat is roaring," while for (b), the training prompt was "A cat is roaring on the grass under the tree." Our appearance-invariant prompt enables the removal of background information that can disturb motion distillation.

3.3. Inference Pipeline

Once trained, in the inference phase, our process begins by computing inverted latents from the input video through DDIM inversion. Subsequently, the inverted latents are fed into the temporally fine-tuned keyframe generation model, yielding short and low-resolution keyframes. These keyframes then undergo temporal extension using the unaltered frame interpolation model. Lastly, the interpolated frames are subjected to spatial enlargement through the spatial super-resolution model. Refer to Fig. 2 for overview.

4. Experiments

4.1. Implementation Details

In our experiments, we choose Show-1 [34] as our VDM backbone and its publicly available pre-trained weights. All experiments were conducted using a single NVIDIA RTX 6000 GPU. VMC with Show-1 demonstrates efficient resource usage, requiring only 15GB of vRAM during mixed-precision training [18], which is completed within 5 minutes. During inference, generating a single video comprising 29 frames at a resolution of 576 x 320 consumes 18GB of vRAM and takes approximately 12 minutes.

4.2. Baseline Comparisons

Dataset Selection. In our experiments, we draw upon a dataset that comprises 24 videos. These videos encompass a broad spectrum of motion types occurring in various contexts, encompassing vehicles, humans, birds, plants, diffusion processes, mammals, sea creatures, and more. This diversity provides a comprehensive range of motion scenarios for our assessment. Out of these 24 videos, 13 are sourced from the DAVIS dataset [19], 10 from the WebVid dataset [1], and 1 video is obtained from LAMP [33].

Baselines. Our method is compared against four contemporary baselines that integrate depth map signals into the diffusion denoising process to assimilate motion information. Notably, our approach operates without the necessity of depth maps during both training and inference, in contrast to these baseline methods.

Specifically, **VideoComposer** (VC) [30] is an open-source latent-based video diffusion model tailored for compositional video generation tasks. **Gen-1** [7] introduces a video diffusion architecture incorporating additional structure and content guidance for video-to-video translation. In contrast to our targeted fine-tuning of temporal attention, **Tune-A-Video** (TAV) [32] fine-tunes self, cross, and temporal attention layers within a pre-trained, but inflated T2I model on input videos. **Control-A-Video** (CAV) [5] introduces a controllable T2V diffusion model utilizing control signals and a first-frame conditioning strategy. Notably, while closely aligned with our framework, Motion Director [35] lacks available code at the time of our research.

Qualitative Results. We offer visual comparisons of our method against four baselines in Fig. 5. The compared baselines face challenges in adapting the motion of the input video to new contexts. They exhibit difficulties in applying the overall motion, neglecting the specific background indicated in the target text (e.g., "underwater" or "on the sand"). Additionally, they face difficulties in deviating from the original shape of the subject in the input video, leading to issues like a shark-shaped airplane, an owl-shaped seagull, or preservation of the shape of the ground where a seagull is taking off. In contrast, the proposed framework succeeds in motion-driven customization, even for difficult compositional customization, e.g. `Two sharks are moving.` $\rightarrow$ `Two` *`airplanes`* `are moving` *`in the sky`*.

Quantitative Results. We further quantitatively demonstrate the effectiveness of our method against the baselines through automatic metrics and user study.
Automatic Metrics. We use CLIP [20] for automatic metrics. For textual alignment, we compute the average cosine similarity between the target prompt and the generated frames. In terms of frame consistency, we obtain CLIP image features within the output video and then calculate the average cosine similarity among all pairs of video frames. For methods that generate temporally interpolated frames, we utilized the keyframe indexes to calculate the metric for a fair evaluation. To illustrate, in the case of VMC, which takes an 8-frame input and produces a 29-frame output, we considered the frames at the following indexes: *1, 5, 9, 13, 17, 21, 25, 29*. As shown in Table 1, VMC outperforms baselines in both text alignment and temporal consistency.
User Study. We conducted a survey involving a total of 27 participants to assess four key aspects: the preservation of motion between the input video and the generated output video, appearance diversity in the output video compared to the input video, the text alignment with the target prompt, and the overall consistency of the generated frames. The survey utilized a rating scale ranging from 1 to 5. For assessing motion preservation, we employed the question:

Figure 5. Qualitative comparison against state-of-the-art baselines. In contrast to other baselines, the proposed framework succeeds in motion-driven customization, even for difficult compositional customization.

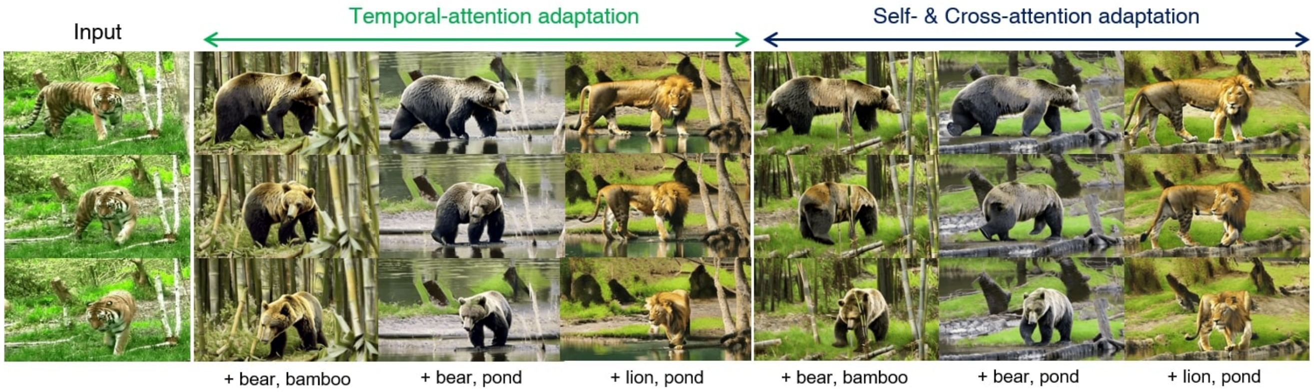

Figure 6. Comparative analysis of the proposed frameworks with fine-tuning (**a**) temporal attention and (**b**) self- and cross-attention layers.

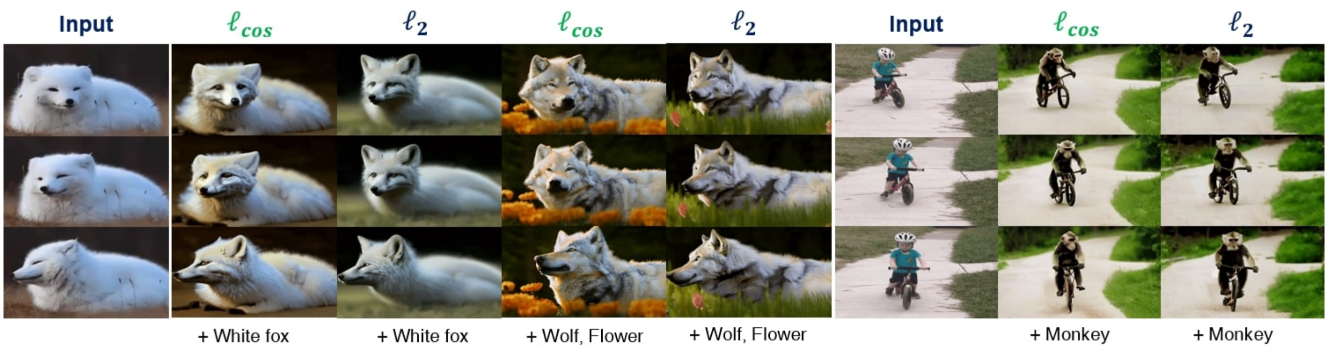

Figure 7. Comparative analysis of the proposed frameworks with (**a**) ℓ_{cos} and (**b**) ℓ_2 loss functions.

"To what extent is the motion of the input video retained in the output video?" To evaluate appearance diversity, participants were asked: "To what extent does the appearance of the output video avoid being restricted on the input video's appearance?" Tab. 1 shows that our method surpasses the baselines in all four aspects.

	Text Alignment	Temporal Consistency	Motion Preservation	Appearance Diversity	Text Alignment	Temporal Consistency
VC	0.798	0.958	3.45	3.43	2.96	3.03
Gen-1	0.780	0.957	3.46	3.17	2.87	2.73
TAV	0.758	0.947	3.50	2.88	2.67	2.80
CAV	0.764	0.952	2.75	2.45	2.07	2.00
Ours	**0.801**	**0.959**	**4.42**	**4.54**	**4.56**	**4.57**

Table 1. Quantitative evaluation using CLIP (*left*) and user study (*right*). VMC significantly outperforms the other baselines.

4.3. Ablation Studies

Comparisons on attention layers. We conducted a comparative study evaluating the performance of fine-tuning: (**a**) temporal attention layers and (**b**) self- and cross-attention layers. Illustrated in Fig. 6, both frameworks exhibit proficient motion learning capabilities. Notably, the utilization of customized temporal attention layers (**a**) yields smoother frame transitions, indicating the effectiveness of the optimization framework (17) in encouraging motion distillation, with a slight preference observed for customized temporal attention layers.

This observation stems from the premise that integrating the proposed motion distillation objective (17) may autonomously and accurately embed motion information within temporal attention layers [11, 13]. This suggests a potential application of the motion distillation objective for training large-scale video diffusion models, warranting further exploration in future research endeavors.

Choice of loss functions. In addition, we conducted a comparative analysis on distinct training loss functions in (17): the ℓ_2-distance and ℓ_{cos} as delineated in (17). As depicted in Fig. 7, the $\delta\epsilon$-matching process in (15) and (17) demonstrates compatibility with generic loss functions. While both $\ell_2(\delta\epsilon_t^n, \delta\epsilon_{\theta,t}^n)$ and $\ell_{cos}(\delta\epsilon_t^n, \delta\epsilon_{\theta,t}^n)$ are promising objectives, the marginal superiority of $\ell_{cos}(\delta\epsilon_t^n, \delta\epsilon_{\theta,t}^n)$ led to its adoption for visualizations in this study.

Importance of adaptation. To assess the importance of temporal attention adaptation, we conducted a visualization of customized generations without temporal attention adaptation, as detailed in Section 3.1. Specifically, from our original architecture in Fig. 2, we omitted attention adaptation and performed inference by maintaining the U-Net modules in a frozen state. The outcomes depicted in Fig. 9 indicate that while DDIM inversion guides the generations

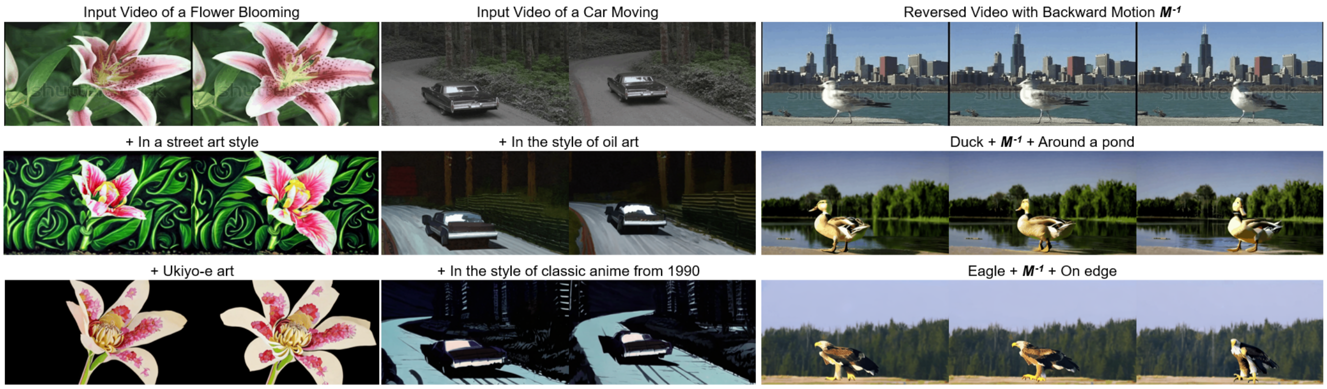

Figure 8. *Left*: Style transfer on two videos. *Right*: Motion customization results on the video of "A seagull is walking *backward*."

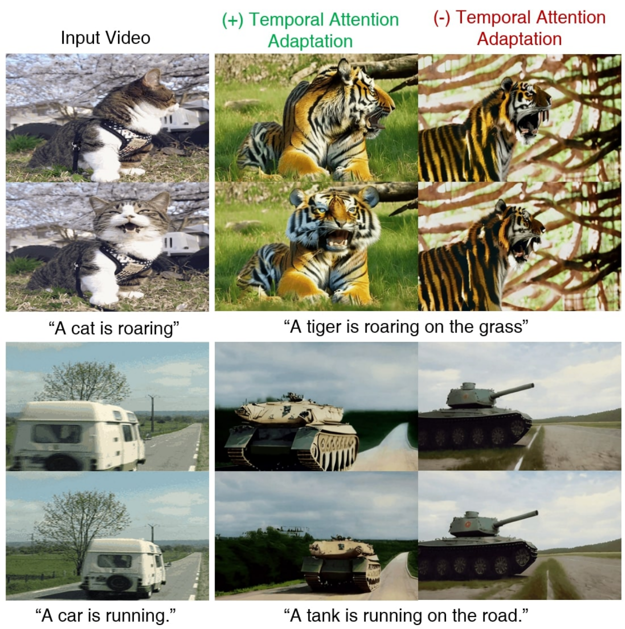

Figure 9. Ablation study on temporal attention adaptation. Without temporal attention adaptation, motion distillation fails.

to mimic the motion of the input video, it alone does not ensure successful motion distillation. The observed changes in appearance and motion exhibit an entangled relationship. Consequently, this underlines the necessity of an explicit motion distillation objective to achieve consistent motion transfer, independent of any alterations in appearance.

4.4. Additional results

Video Style Transfer. We illustrate video style transfer applications in Fig. 8-*Left*. We incorporate style prompts at the end of the text after applying appearance-invariant prompting (see Section 3.2). Target styles are fluidly injected while preserving the distilled motion of an input video.

Learning Backward Motion. To further verify our video motion customization capabilities, we present a challenging scenario: extracting backward motion from a reversed video sequence where frames are arranged in reverse order. This scenario, an exceedingly rare event in real-world videos, is highly improbable within standard training video datasets [1]. Illustrated in Fig. 8, our VMC framework showcases proficiency in learning "a bird walking backward" motion and generating diverse videos with distinct subjects and backgrounds. This capability not only enables leveraging the distilled motion but also offers prospects for further contextual editing.

5. Conclusion

This paper introduces Video Motion Customization (VMC), addressing challenges in adapting Text-to-Video (T2V) models to generate motion-driven diverse visual customizations. Existing models struggle with accurately replicating motion from a target video and creating varied visual outputs, leading to entanglements of appearance and motion data. To overcome this, our VMC framework presents a novel one-shot tuning approach, focusing on adapting temporal attention layers within video diffusion models. This framework stands out for its efficiency in time and memory, ease of implementation, and minimal hyperparameters.

Ethics Statement. Our work employs a generative model that necessitates caution due to its potential for misuse in creating deceptive content with negative societal impacts.

This research was supported by Field-oriented Technology Development Project for Customs Administration through National Research Foundation of Korea (NRF) funded by the Ministry of Science & ICT and Korea Customs Service (NRF-2021M3I1A1097938), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT, Ministry of Science and ICT) (No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation), Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711137899, KMDF_PR_20200901_0015), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 5, 8

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[3] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[4] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 3

[5] 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. 5

[6] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 2

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3, 5

[8] 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. *arXiv preprint arXiv:2208.01618*, 2022. 2

[9] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 2

[10] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3

[11] J Ho, T Salimans, A Gritsenko, W Chan, M Norouzi, and DJ Fleet. Video diffusion models. arxiv 2022. *arXiv preprint arXiv:2204.03458*. 3, 7

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

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3, 7

[14] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 2

[15] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*, 2023. 3

[16] Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similarity score distillation for zero-shot video editing. *arXiv preprint arXiv:2403.12002*, 2024. 3

[17] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 2

[18] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017. 5

[19] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[21] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 4

[22] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2, 4

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[24] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2

[25] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International confer-*

ence on machine learning, pages 2256–2265. PMLR, 2015. 2

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[28] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[29] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[30] 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. 3, 5

[31] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2

[32] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3, 4, 5

[33] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*, 2023. 5

[34] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 2, 3, 5

[35] 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. *arXiv preprint arXiv:2310.08465*, 2023. 2, 3, 4, 5

[36] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2