Spectral Motion Alignment for Video Motion Transfer Using Diffusion Models

Geon Yeong Park*, Hyeonho Jeong*, Sang Wan Lee†, Jong Chul Ye†

Korea Advanced Institute of Science and Technology (KAIST)
{pky3436, hyeonho.jeong, sangwan, jong.ye}@kaist.ac.kr

Abstract

Diffusion models have significantly facilitated the customization of input video with target appearance while maintaining its motion patterns. To distill the motion information from video frames, existing works often estimate motion representations as frame difference or correlation in pixel-/feature-space. Despite its simplicity, these methods have unexplored limitations, including lack of understanding of global motion context, and the introduction of motion-independent spatial distortions. To address this, we present *Spectral Motion Alignment (SMA)*, a novel framework that refines and aligns motion representations in the spectral domain. Specifically, SMA learns spectral motion representations, facilitating the learning of whole-frame global motion dynamics, and effectively mitigating motion-independent artifacts. Extensive experiments demonstrate SMA's efficacy in improving motion transfer while maintaining computational efficiency and compatibility across various video customization frameworks.

1 Introduction

Given the multifaceted nature of the video, encompassing motion dynamics, appearance, etc., several studies aim to disentangle and control these signals according to user intent. Recently, diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) has played a pivotal role in video customization, owing to their superior sampling ability.

In the context of *motion* customization using diffusion models, our goal is to transfer the motion patterns from an input video to the customized output video. This necessitates the accurate estimation and extraction of motion information from the input video. While fundamental techniques such as optical flow are effective for motion estimation, integrating these into diffusion models for customization is nontrivial.

To address these challenges, recent researches suggest that motion patterns are inherently encoded in the underlying dependencies between frames or epsilon noises. For example, (Zhao et al. 2023b) have observed that videos with similar motion tend to exhibit similar connectivity between latent frames. Additionally, (Jeong, Park, and Ye 2023) utilizes residual vectors between consecutive frames as "motion vectors," in line with optical flow principles, assuming frame residuals represent motion dynamics. Specifically, they finetune pretrained VDM to align the ground-truth pixel-space residuals with their predicted denoised estimates. Thus, these works leverage the pixel-space differences between input frames as a proxy of motion reference.

While these motion representations can be obtained from off-the-shelf video diffusion models efficiently, current simple approximations have several adverse impacts. First, they may fail to capture the global context of motion. Since frame residuals may capture local motion patterns but are blind to whole-frame motion dynamics, for better motion dynamics modeling, we have to understand the whole-frame global context information during motion distillation. Furthermore, while the pixel- or feature-space residuals contain motion information, they may also contain inevitable disruptive variations that are unrelated to motion. These variations may include abrupt changes in the background, lightning, or other frame inconsistencies, leading to less reliable representation.

To address these challenges, we introduce Spectral Motion Alignment (**SMA**), a novel framework for refining and aligning motion representations in the *spectral* domain, based on intuition that the motion may be well represented by its inherent frequency components. This framework includes two primary components: First, to capture the global motion context, we propose a spectral alignment loss between predicted and ground-truth motion vectors within the wavelet domain. This facilitates the learning of multi-scale motion dynamics by leveraging rich wavelet-domain representations of video considering the global frame transitions. Moreover, to mitigate the spatial artifacts and inconsistency in motion vectors, we propose 2D FFT-based motion vector refinement that aligns the amplitude and phase spectrum of ground truth and predicted motion vectors with prioritizing low-frequency components. This is because the high-frequency components in motion representations may be associated with frame-wise motion-independent artifacts (Figure 4). In summary, we encourage accurate motion transfer via harmonized global and local levels of spectral domain alignment. Our contributions are summarized as follows:

- We introduce the Spectral Motion Alignment (SMA), a frequency-domain motion alignment framework that learns the underlying motion dynamics of input video via frequency-based regularization. SMA is orthogonal

*: Co-first authors
†: Co-corresponding authors

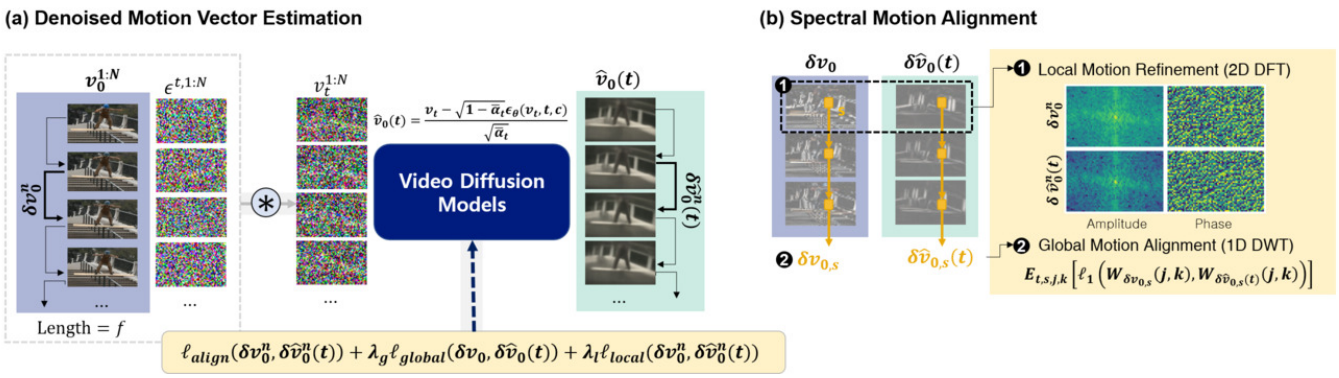

Figure 1: **Overview.** The proposed Spectral Motion Alignment (**SMA**) framework distills the motion information in frequency-domain. Considering the (latent) frame residuals as motion vectors, we first derive the denoised motion vector estimates. Then, the motion vector $\delta \boldsymbol{v}_0^n$ and its estimate $\delta \hat{\boldsymbol{v}}_0^n$ are aligned in both pixel-domain and frequency-domain. Our regularization includes (1) global motion alignment based on 1D wavelet-transform, and (2) local motion refinement based on 2D Fourier transform.

and compatible to most of existing motion customization models as they often only rely on either pixel or feature space representations.

- SMA imposes negligible memory and computational burdens, as most off-the-shelf VDMs can readily compute motion vectors estimates. For instance, VMC (Jeong, Park, and Ye 2023) with SMA demonstrates lightweight (15GB vRAM) and rapid (< 5 min) training.
- We validate the efficacy of SMA across diverse motion patterns, subjects, and various video motion transfer frameworks including Video Diffusion-based (Zhao et al. 2023b), Cascaded Video Diffusion-based (Jeong, Park, and Ye 2023), T2I Diffusion-based (Wu et al. 2023), and ControlNet-based models (Chen et al. 2023).

2 Preliminaries

2.1 Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) generate samples from the Gaussian noise through reverse denoising processes. We denote a clean sample $\boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x})$, a noisy latent $\boldsymbol{x}_t \in \mathbb{R}^d$ at time t, β_t as an increasing sequence of noise schedule, $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \Pi_{i=1}^t \alpha_i$. Then the goal of diffusion model training is to optimize a denoiser $\boldsymbol{\epsilon}_{\theta *}$:

$$\theta^* := \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{\epsilon}} \big[\| \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon} \| \big]. \tag{1}$$

The reverse sampling from $q(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t, \boldsymbol{\epsilon}_{\theta *}(\boldsymbol{x}_t, t))$ is then achieved by

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \Big(\boldsymbol{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_{\theta *}(\boldsymbol{x}_t, t) \Big) + \tilde{\beta}_t \boldsymbol{\epsilon}, \tag{2}$$

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 (Song, Meng, and Ermon 2020) further proposes another sampling method as follows:

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

where $\eta \in [0, 1]$ controls stochasticity, and $\hat{\boldsymbol{x}}_0(t)$ is the denoised estimate which can be equivalently derived using Tweedie's formula (Efron 2011):

$$\hat{\boldsymbol{x}}_0(t) := \frac{1}{\sqrt{\bar{\alpha}_t}} (\boldsymbol{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_{\theta *}(\boldsymbol{x}_t, t)). \tag{3}$$

For a text-guided generation, diffusion models are often trained with the textual embedding c. 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 (Ho et al. 2022b,a; Zhang et al. 2023) further attempt to model the video data distribution. Specifically, Let $(\boldsymbol{v}^n)_{n \in \{1, \dots, N\}}$ 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 $(\boldsymbol{v}_t^n)_{n \in 1, \dots, N}$, $\boldsymbol{v}_t^{1:N}$, and $\boldsymbol{\epsilon}^{1:N}$. 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:

$$\min_\theta \mathbb{E}_{\boldsymbol{v}_t^{1:N}, \boldsymbol{v}^{1:N}, \boldsymbol{\epsilon}^{1:N}, c} \big[\big\| \boldsymbol{\epsilon}_\theta(\boldsymbol{v}_t^{1:N}, t, c) - \boldsymbol{\epsilon}^{1:N} \big\| \big], \tag{4}$$

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

2.2 Fourier and Wavelet Analysis

Spectral analysis techniques transform time-domain or pixel-domain signals (such as video frames) into the frequency domain, revealing the frequency components and their intensities.

Fourier Transform. Let $\boldsymbol{v}^n \in \mathbb{R}^{H \times W}$ represents the n-th 2D video frame. Then, its frequency spectrum at coordinate (a, b) is given as follows:

$$\mathcal{F}_{\boldsymbol{v}^n}(a, b) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \boldsymbol{v}^n(x, y) e^{-i 2\pi \left(\frac{ax}{H} + \frac{by}{W} \right)}, \tag{5}$$

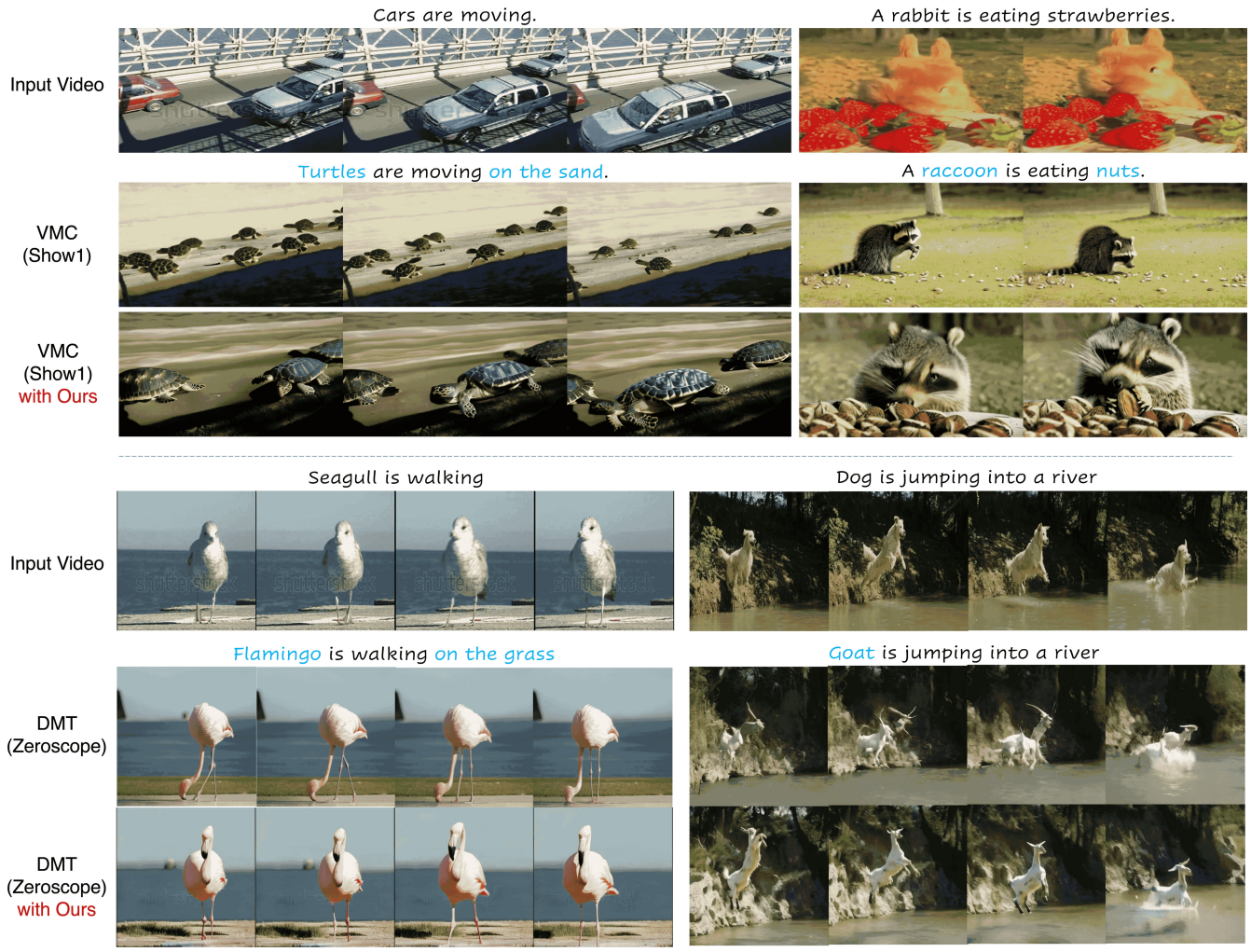

Figure 2: Comparison within VMC framework using Show-1 video model (*top*) and DMT framework using Zeroscope video model (*bottom*). Each demonstrate the compatibility of SMA in pixel-space and feature-space, respectively.

where $\boldsymbol{v}^n(x,y)$ means the pixel value at coordinate (x,y). The output frequency spectrum is represented as $\mathcal{F}_{\boldsymbol{v}^n}(a,b) = R(a,b) + I(a,b)i$, where $R(a,b), I(a,b) \in \mathbb{R}$ represents real and imaginary part, respectively. Then, the amplitude and phase is derived as follows:

$$|\mathcal{F}_{\boldsymbol{v}^n}(a,b)| = \sqrt{R(a,b)^2 + I(a,b)^2}, \\ \angle\mathcal{F}_{\boldsymbol{v}^n}(a,b) = \arctan\left(\frac{I(a,b)}{R(a,b)}\right). \tag{6}$$

Wavelet Transform. Wavelet frames, renowned for capturing multi-resolution scale features, are among the most prevalently utilized frame representations in signal processing. Let $\psi(t)$ represent a mother wavelet that can be shifted and scaled. For a function $\boldsymbol{v}(t) \in L^2(\mathbb{R})$, the wavelet transform can be expressed as:

$$\mathcal{CW}_{\boldsymbol{v}}(a,b) = \frac{1}{\sqrt{\alpha}} \int \boldsymbol{v}(t)\psi^*\left(\frac{t-b}{a}\right) dt = \langle \boldsymbol{v}(t), \psi_{a,b}(t) \rangle, \tag{7}$$

which serves as an expansion coefficient. In the case of discrete wavelet transform (DWT), it uses a finite set of wavelet and scaling functions derived from a chosen wavelet family. Specifically, the mother wavelet is shifted and scaled by powers of two as follows:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \psi(2^{-j}t - k). \tag{8}$$

Then, the DWT of a signal $\boldsymbol{v}[n]$ is given by:

$$\mathcal{W}_{\boldsymbol{v}}(j,k) = \langle \boldsymbol{v}(t), \psi_{j,k}(t) \rangle. \tag{9}$$

The original signal can be recovered from inverse DWT. In practice, this discrete wavelet transform can be implemented by convolution using an appropriate choice of filter bank.

3 Spectral Motion Alignment

Our main goal is to develop a novel *spectral domain* motion alignment framework that capture underlying complex motion patterns across a spectrum of frequency levels that

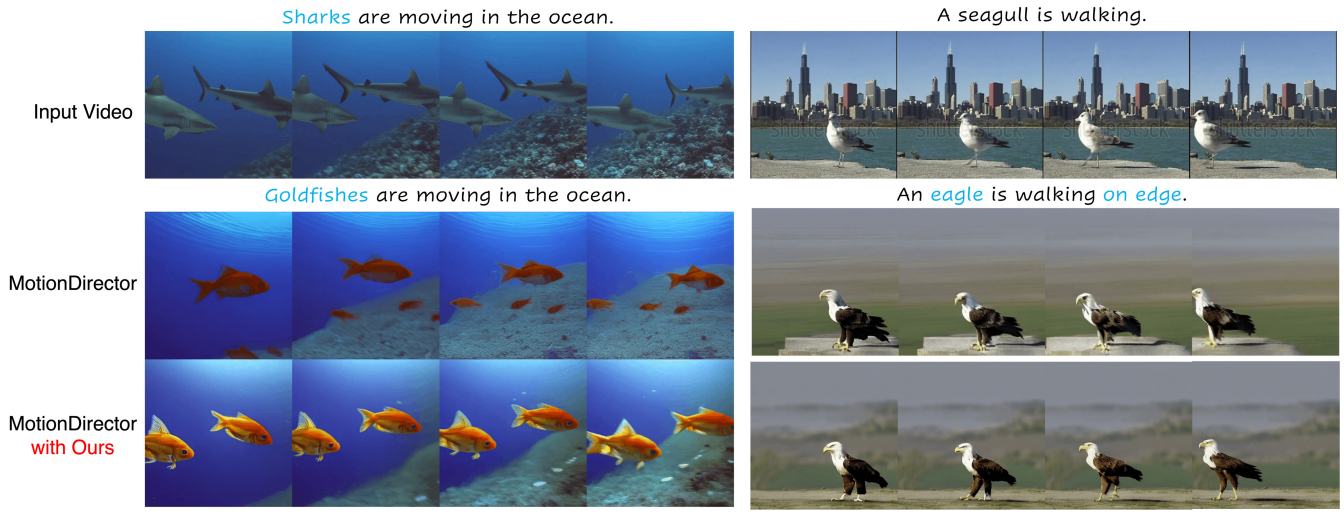

Figure 3: Comparison within MotionDirector framework.

mainly constitute motion. This is valuable in video understanding and customization as it helps in identifying repetitive motion patterns and underlying structures that may not be visible in the time or pixel domain. It is in orthogonal (and compatible) with conventional methods based on pixel- or feature-domain motion representations.

3.1 Denoised Motion Vector Estimation

To distill the motion information, we first estimate the initial motion representations (Jeong, Park, and Ye 2023; Zhao et al. 2023b) in pixel space. For this, we follow VMC (Jeong, Park, and Ye 2023) as an representative example. The intuition is that residual vectors between consecutive frames may include information about the motion trajectories. Define the n-th frame residual vector, namely motion vector at time $t \geqslant 0$ as

$$\delta \boldsymbol{v}_t^n := \boldsymbol{v}_t^{n+1} - \boldsymbol{v}_t^n, \tag{10}$$

where the epsilon residual vector $\delta \boldsymbol{\epsilon}_t^n$ is similarly defined. This $\delta \boldsymbol{v}_t^n$ can be acquired through the following diffusion kernel:

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

Given that, the ground-truth motion vector in pixel space $\delta \boldsymbol{v}_0^n$ can be derived as follows:

$$\delta \boldsymbol{v}_0^n = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\delta \boldsymbol{v}_t^n - \sqrt{1-\bar{\alpha}_t}\delta \boldsymbol{\epsilon}_t^n\right). \tag{12}$$

Similarly, one can obtain the denoised estimate version of these motion representations $\delta \hat{\boldsymbol{v}}_0^n$ by using Tweedie's formula as follows:

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

where $\hat{\boldsymbol{v}}_0^{1:N}(t)$ is an empirical Bayes optimal posterior expectation $\mathbb{E}[\boldsymbol{v}_0^{1:N} \mid \boldsymbol{v}_t^{1:N}]$.

In the context of motion transfer, we aim to align the ground-truth and estimated motion vectors by fine-tuning the pre-trained VDM (Jeong, Park, and Ye 2023; Zhao et al. 2023b):

$$\min_\theta \mathbb{E}_{t,n,\boldsymbol{\epsilon}^{t,n},\boldsymbol{\epsilon}^{t,n+1}}\left[\ell_{\text{align}}\left(\delta \boldsymbol{v}_0^n, \delta \hat{\boldsymbol{v}}_0^n(t)\right)\right]. \tag{14}$$

While these advancements in motion distillation mark significant progress, Figure 2 and 4 indicate that existing methods still has potential for further refinement.

3.2 Spectral Global Motion Alignment

One of the primary limitations in (14) is that it may not fully encapsulate the global motion dynamics. Specifically, it locally focuses on pairwise frame comparisons which may lead to overlooking the comprehensive motion dynamics of given object overall frames.

To mitigate these problems, we explore the use of wavelet transforms in motion distillation. In this paper, we use Haar wavelet, whose low and high pass filters are given as follows:

$$L[n] = \frac{1}{\sqrt{2}}[1\ 1], H[n] = \frac{1}{\sqrt{2}}[-1\ 1], \tag{15}$$

which is implemented using the multi-scale Haar filter bank. Then, given the sequence of motion vectors $\delta \boldsymbol{v}_0 = (\delta \boldsymbol{v}_0^n)_{n\in\{1,\dots,N-1\}}$ and its denoised estimates $\delta \hat{\boldsymbol{v}}_0(t) = \left(\delta \hat{\boldsymbol{v}}_0^n(t)\right)_{n\in\{1,\dots,N-1\}}$, we consider $(N-1)$-length time-dependent 1D arrays from arbitrary spatial pixel dimension $s \in \{1, \dots d\}$. The corresponding 1D array of motion vector is denoted by $\delta \boldsymbol{v}_{0,s}$ and $\delta \hat{\boldsymbol{v}}_{0,s}(t) \in \mathbb{R}^{N-1}$ (Figure 1).

Then, the frequency-matching loss between $\delta \boldsymbol{v}_0$ and $\delta \hat{\boldsymbol{v}}_0(t)$ is defined with DWT in (9) as follows:

$$\ell_{\text{global}}(\delta \boldsymbol{v}_0, \delta \hat{\boldsymbol{v}}_0(t)) = \mathbb{E}_{t,s,j,k}\left[\|\mathcal{W}_{\delta \boldsymbol{v}_{0,s}}(j,k) - \mathcal{W}_{\delta \hat{\boldsymbol{v}}_{0,s}(t)}(j,k)\|_1\right]. \tag{16}$$

Considering that the wavelet transform allows multi-resolution analysis of motion vectors, it enables us to handle

Method	Automatic Metrics		User Study		
	Text-Align	Temp-Con	Edit-Acc	Temp-Con	Motion-Acc
MotionDirector	0.7550	0.9780	3.19	3.07	2.67
MotionDirector w/ Ours	**0.8081**	**0.9784**	**4.14**	**3.89**	**3.88**
VMC (Show-1)	0.8066	0.9742	3.25	3.22	2.72
VMC (Show-1) w/ Ours	**0.8193**	**0.9776**	**3.85**	**3.84**	**3.92**
VMC (Zeroscope)	0.8223	0.9560	2.95	2.83	2.47
VMC (Zeroscope) w/ Ours	**0.8425**	**0.9578**	**4.07**	**3.84**	**4.07**

Table 1: Quantitative evaluation of SMA within text-to-video based frameworks.

motions at various scales and frequencies effectively. This could be particularly beneficial for complex scenes with varying motion speeds and types, ensuring that subtle motions are captured and transferred more accurately.

3.3 Spectral Local Motion Refinement

Another problem in (14) is that the estimated motion representations may encapsulate high-frequency local distortions, background noise, and other non-motion-related artifacts. By aligning the denoised estimates with these artifacts, the fine-tuned VDM may erroneously reproduce similar high-frequency artifacts as in Figure 4.

Accordingly, we focus on prioritizing low-to-moderate spatial frequency components particularly. Specifically, following the amplitude and phase spectrum definition in (6), we define amplitude and phase matching loss, $\ell^A_{local}(\delta \boldsymbol{v}^n_0, \delta\hat{\boldsymbol{v}}^n_0(t))$ and $\ell^P_{local}(\delta \boldsymbol{v}^n_0, \delta\hat{\boldsymbol{v}}^n_0(t))$, as follows:

$$\begin{aligned}(A)\ &\mathbb{E}_{t,n,a,b}\Big[\omega(a,b)\|\mathcal{F}_{\delta \boldsymbol{v}^n_0}(a,b)| - |\mathcal{F}_{\delta\hat{\boldsymbol{v}}^n_0(t)}(a,b)|\|_1\Big],\\ (P)\ &\mathbb{E}_{t,n,a,b}\Big[\omega(a,b)\|\angle\mathcal{F}_{\delta \boldsymbol{v}^n_0}(a,b) - \angle\mathcal{F}_{\delta\hat{\boldsymbol{v}}^n_0(t)}(a,b)\|_1\Big],\end{aligned} \tag{17}$$

where the frequency domain weighting $\omega(a,b)$ is defined as

$$\omega(a,b) = \Big[\big(\frac{H}{2}\big)^2 + \big(\frac{W}{2}\big)^2\Big]^\delta - \Big[\big(a-\frac{H}{2}\big)^2 + \big(b-\frac{W}{2}\big)^2\Big]^\delta + 1$$

for $0 < a < H, 0 < b < W$, and otherwise, set to zero. This introduces a weighting (Yang et al. 2022) that prioritizes low-frequency components for $\delta > 0$.

3.4 Inference Pipeline

To sum up, the overall spectral motion alignment framework is given as follows:

$$\begin{aligned}\min_\theta \mathbb{E}_{t,n,\boldsymbol{\epsilon}^n_t,\boldsymbol{\epsilon}^{n+1}_t}\Big[&\ell_{\text{align}}\big(\delta \boldsymbol{v}^n_0, \delta\hat{\boldsymbol{v}}^n_0(t)\big) + \\ &\lambda_g \ell_{\text{global}}\big(\delta \boldsymbol{v}_0, \delta\hat{\boldsymbol{v}}_0(t)\big) + \lambda_l \ell_{local}\big(\delta \boldsymbol{v}^n_0, \delta\hat{\boldsymbol{v}}^n_0(t)\big)\Big],\end{aligned} \tag{18}$$

where $\ell_{local}(\cdot,\cdot) = \ell^A_{local}(\cdot,\cdot) + \ell^P_{local}(\cdot,\cdot)$. Upon optimization, the inference is performed using new text prompts to transform appearances, e.g. "a seagull is walking" → "a chicken is walking".

This Spectral Motion Alignment (SMA) is universally adaptable across various motion distillation frameworks. While diverse diffusion-based motion distillation frameworks adopt their pixel-domain motion learning objectives, the proposed frequency-domain alignment seamlessly integrates with these arbitrary objectives. Moreover, different motion distillation frameworks target specific parameters θ for fine-tuning, varying from temporal attention layers (Jeong, Park, and Ye 2023) to dual-path LoRAs (Zhao et al. 2023b). We empirically demonstrate the global compatibility of the proposed spectral motion alignment with diverse neural architectures and parameterizations. Pseudo-code is provided in the appendix.

3.5 Extending SMA to Feature Space

Beyond pixel-space motion representations, SMA can be further extended to semantic diffusion features (DIFT, (Tang et al. 2023)). Specifically, Diffusion-Motion-Transfer (DMT, (Yatim et al. 2023)) constructs motion vectors based on pairwise differences in space-time diffusion features, which are then utilized for latent optimization-based video motion transfer. Given input and target video latents, $\boldsymbol{v}_t$ and $\tilde{\boldsymbol{v}}_t$, the model extracts space-time features $f(\boldsymbol{v}_t)$ and $f(\tilde{\boldsymbol{v}}_t)$. Then, the feature residuals are defined as $\delta f(\boldsymbol{v}_t)^n$ and $\delta f(\tilde{\boldsymbol{v}}_t)^n$, the n-th consecutive difference between hidden feature frames. This leads to the spectral alignment objective in feature space as follows:

$$\begin{aligned}\mathbb{E}\Big[&\ell_{\text{DMT}}\big(f(\boldsymbol{v}_t), f(\tilde{\boldsymbol{v}}_t)\big) + \lambda_g \ell_{\text{global}}\big(\delta f(\boldsymbol{v}_t), \delta f(\tilde{\boldsymbol{v}}_t)\big) \\ &+ \lambda_l \ell_{local}\big(\delta f(\boldsymbol{v}_t)^n, \delta f(\tilde{\boldsymbol{v}}_t)^n\big)\Big],\end{aligned} \tag{19}$$

where ℓ_{DMT} refers to the original space-time feature loss in (Yatim et al. 2023). Note that DMT does not finetune the models, leveraging (19) for a latent optimization in sampling process. We demonstrate the effectiveness of spectral alignment in the diffusion feature space by comparing it against the original DMT framework in Fig.2-bottom.

4 Experiments using T2V Diffusion Models

4.1 Experimental Setting

To assess the capability of Spectral Motion Alignment (SMA) to capture accurate motion within contemporary diffusion-based motion learning frameworks, we curated a

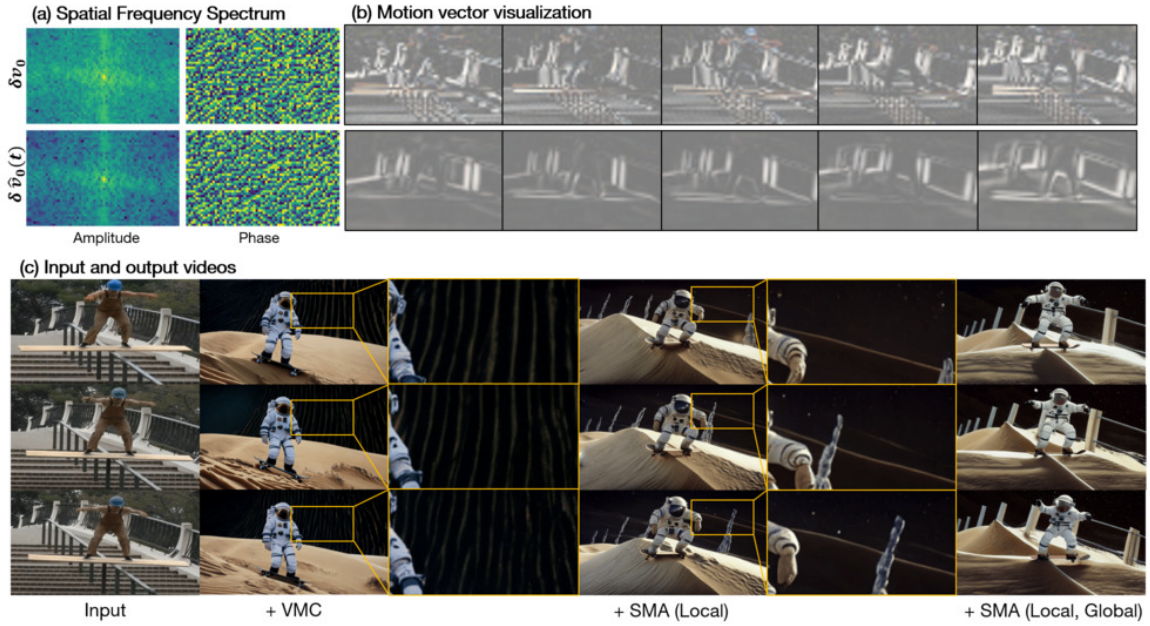

Figure 4: Visualization of (**a**) spatial frequency spectrum and (**b**) motion vectors estimated from the pre-trained Show-1 (Zhang et al. 2023) without fine-tuning. (**c**) Ablation study on spectral motion alignment based on VMC (Jeong, Park, and Ye 2023).

	Automatic Metrics		User Study		
Method	Text-Align	Temp-Con	Edit-Acc	Temp-Con	Motion-Acc
Tune-A-Video	0.8289	0.8951	2.71	2.11	2.27
Tune-A-Video w/ Ours	**0.8633**	**0.9568**	**3.41**	**2.88**	**3.25**
ControlVideo	0.8686	0.9451	2.88	2.25	2.69
ControlVideo w/ Ours	**0.8781**	**0.9590**	**3.56**	**3.16**	**3.40**

Table 2: Quantitative evaluation of SMA within text-to-image based frameworks.

dataset comprising 30 text-video pairs sourced from the publicly available DAVIS (Pont-Tuset et al. 2017) and WebVid-10M (Bain et al. 2021) collections. This dataset is deliberately designed to cover a broad spectrum of motion types and subjects, with video lengths ranging between 8 and 16 frames. For this study, we leverage two foundational text-to-video diffusion models: Zeroscope (Sterling 2023), a non-cascaded VDM, and Show-1 (Zhang et al. 2023), a cascaded VDM. More details are provided in appendix.

4.2 Baselines

MotionDirector (Zhao et al. 2023b) tailor the appearance and motion of a video by developing a unique dual-path (spatial, temporal) framework with Low-Rank Adaptation (LoRA, (Hu et al. 2021)). **VMC** (Jeong, Park, and Ye 2023) achieves state-of-the-art performance in motion customization through their novel epsilon residual matching objective, facilitating efficient motion distillation within a cascaded video diffusion. **DMT** (Yatim et al. 2023) proposes a new space-time feature loss, guiding the sampling process towards preserving the motion patterns while complying with the target object. Please see Sec 3.5 for more details.

4.3 Qualitative Comparison

Fig. 3 and 2 offer qualitative comparisons with and without SMA. The top of Figure 2 shows videos from a cascaded diffusion pipeline, while the bottom displays those from a non-cascaded model. Without SMA, videos may capture appearance to some extent but fail to replicate motion patterns accurately. In contrast, SMA significantly improves motion transfer, distinguishing dynamic from static objects. For instance, in the last example of Fig. 3, SMA produces a video where only the eagle moves from right to left, whereas without SMA, the video inaccurately depicts the ground moving alongside the eagle.

4.4 Quantitative Comparison

The results of our quantitative evaluation are presented in Table 1. To evaluate text-video alignment (Hessel et al. 2021), we measure the average cosine similarity between the target text prompt and the frames generated. Regarding frame consistency, we extract CLIP image features for each

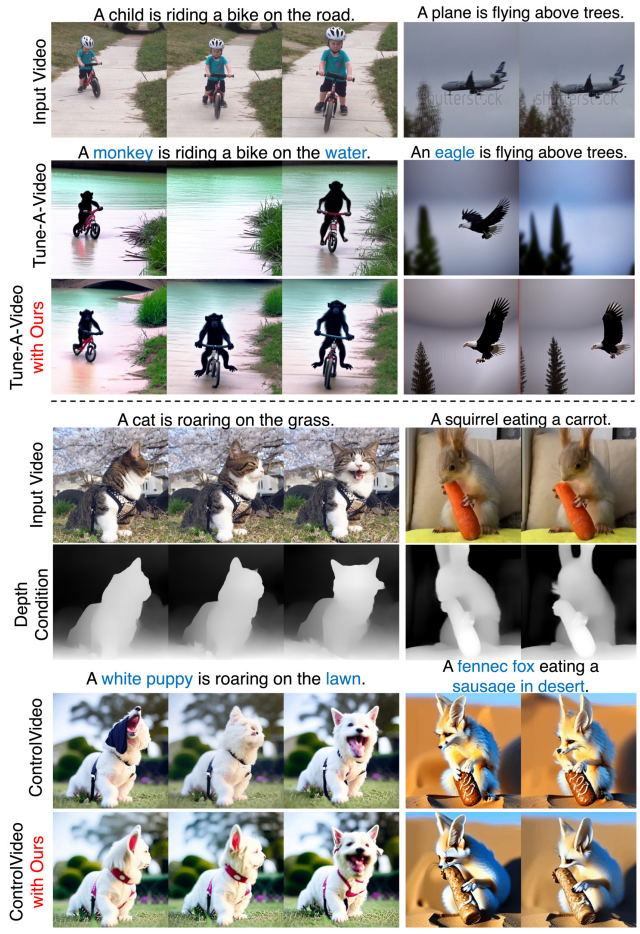

Figure 5: Comparison within Tune-A-Video (Top) and ControlVideo-Depth (Bottom) baseline.

frame in the output video and subsequently calculate the average cosine similarity among all frame pairs in the video. For human evaluation, we conduct a user study with 42 participants to assess three key aspects, guided by the following questions: (1) Editing Accuracy, (2) Temporal Consistency, (3) Motion Accuracy. Tab. 1 demonstrates that SMA enhances the performance of MotionDirector and VMC across all measured metrics.

5 T2I Video Diffusion Models

5.1 Experimental Setting

We further evaluate the efficacy of SMA with methods based on text-to-image diffusion model. Same text-video pairs are used as in Sec. 4.1. The resolution for all produced videos is standardized to 512x512. In this experiment, Stable Diffusion v1-5 (Rombach et al. 2022) and ControlNet-Depth (Zhang, Rao, and Agrawala 2023) are utilized.

5.2 Baselines

Tune-A-Video (Wu et al. 2023) transforms a pretrained T2I diffusion model to psuedo T2V model by adding temporal attention layers and expanding spatial self-attention into spatio-temporal attention. **ControlVideo** (Zhao et al. 2023a) is another one-shot-based video editing method stems from pretrained T2I model. ControlVideo extends ControlNet (Zhang, Rao, and Agrawala 2023) from image to video to incorporate structural cues obtained from the input video.

5.3 Qualitative Comparison

Fig. 5-(*top*) demonstrates the efficacy of SMA with Tune-A-Video method, where SMA alleviates the flickering artifacts in foreground objects. Fig. 5-(*bottom*) further illustrates the improvements with the ControlVideo framework. While depth control encourages ControlVideo to maintain the structural integrity, Fig. 5 shows that it is not sufficient for motion accuracy, where SMA plays a crucial role in accurate capture of motion details.

5.4 Quantitative Comparison

Quantitative results are detailed in Tab. 2, following the metrics introduced in Sec. 4.4. Across both the Tune-A-Video and ControlVideo frameworks, the integration of SMA improves performance across all five evaluated metrics, notably achieving a substantial advantage in motion accuracy.

6 Analysis

We explore the impact of SMA by examining motion vectors ($\delta \boldsymbol{v}_0$, $\delta \hat{\boldsymbol{v}}_0(t)$) and their (amplitude, phase) spectrum in Figure 4 ($t = 700$). Figure 4b indicates that the pixel-space motion vector $\delta \boldsymbol{v}_0$ faces frame-wise distortions or inconsistencies. These are characterized as high-frequency noises in the amplitude spectrum, Fig. 4a. Thus, we prioritize low spatial frequency components which improves the overall fidelity and removes background distortions (Figure 4c).

Moreover, global motion alignment further improves motion transfer. Specifically, without considering the global motion dynamics, existing frameworks occasionally generate "reversed" motions, i.e. an astronaut skateboarding in an upward direction. In contrast, the proposed global motion alignment effectively mitigates these challenges (Table 3).

	Base	$+\mathcal{L}_{\text{local}}$	$+\mathcal{L}_{\text{global}}$	$+\mathcal{L}_{\text{local}}, \mathcal{L}_{\text{global}}$
TA	0.816	0.838	0.834	**0.843**
FC	0.950	0.961	0.959	**0.966**

Table 3: Quantitative ablation of $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$. Base refers to baseline, TA refers to text alignment and FC stands for frame consistency.

7 Conclusion

We propose Spectral Motion Alignment (SMA), a novel motion distillation framework in spectral domain. We explore the limitations of conventional motion estimation methods: (**a**) lack of global motion understanding, (**b**) vulnerability to spatial artifacts. Then, we mitigate these problems by harmonizing both local and global motion alignment and effectively distills motion patterns.

Acknowledgements

This work was supported by 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), No. RS-2023-00233251, System3 reinforcement learning with high-level brain functions), National Research foundation of Korea(NRF) (**RS-2023-00262527**, RS-2024-00336454, RS-2024-00341805).

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision*.

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

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

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

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

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

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *arXiv:2204.03458*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sterling, S. 2023. Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36: 1363–1389.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. 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*, 7623–7633.

Yang, G.; Liu, W.; Liu, X.; Gu, X.; Cao, J.; and Li, J. 2022. Delving into the frequency: Temporally consistent human motion transfer in the fourier space. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1156–1166.

Yatim, D.; Fridman, R.; Tal, O. B.; Kasten, Y.; and Dekel, T. 2023. Space-Time Diffusion Features for Zero-Shot Text-Driven Motion Transfer. *arXiv preprint arXiv:2311.17009*.

Zhang, D. J.; Wu, J. Z.; Liu, J.-W.; Zhao, R.; Ran, L.; Gu, Y.; Gao, D.; and Shou, M. Z. 2023. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhao, M.; Wang, R.; Bao, F.; Li, C.; and Zhu, J. 2023a. ControlVideo: Adding Conditional Control for One Shot Text-to-Video Editing. *arXiv preprint arXiv:2305.17098*.

Zhao, R.; Gu, Y.; Wu, J. Z.; Zhang, D. J.; Liu, J.; Wu, W.; Keppo, J.; and Shou, M. Z. 2023b. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*.