ϕ-NOISE: TRAINING-FREE TEMPORAL VIDEO CONDITIONING VIA PHASE-BASED NOISE MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent video diffusion models generate videos by progressively transforming Gaussian noise into realistic samples conditioned on text or visual inputs. However, existing conditioning methods often require additional training and computational overhead. Motivated by recent findings on the importance of frequency components in generative models, we propose a simple, training-free approach for motion-conditioned video generation by injecting low-frequency phase information from a reference video directly into the diffusion noise latents. Our method transfers motion cues without modifying the model architecture or inference pipeline. Using several applications, we demonstrate effective control over both appearance and dynamics in generated videos, while achieving competitive or superior results compared to more complex conditioning approaches.

1 INTRODUCTION

Latent diffusion models have become the dominant paradigm for visual content generation, achieving remarkable success in image synthesis Rombach et al. (2022); Podell et al. (2023); Labs (2024); Wu et al. (2025); Esser et al. (2024); Pernias et al. (2023). More recently, diffusion-based video generation models extended these capabilities to temporally coherent multi-frame synthesis HaCohen et al. (2024); Wan et al. (2025); Yang et al. (2025); Bar-Tal et al. (2024); Ho et al. (2022a); Blattmann et al. (2023a). These models progressively transform white Gaussian noise in latent space into structured visual outputs conditioned on textual prompts.

While text prompts provide high-level semantic guidance, recent works explored additional conditioning mechanisms for more precise control. In the image domain, methods for structure and style conditioning enable control beyond natural language descriptions Zhang et al. (2023); Ye et al. (2023); Mou et al. (2024); Hertz et al. (2024); Frenkel et al. (2024); Cohen et al. (2025). Extending such control to video generation is substantially more challenging, as videos require modeling both spatial appearance and temporal dynamics, including motion and camera behavior.

To address this, prior works introduced methods for conditioning either spatial Chen et al. (2023); Hu & Xu (2023); Hu (2024) or temporal Geng et al. (2025); Wang et al. (2024); Pondaven et al. (2025) aspects of generated videos. However, many approaches rely on specialized architectures, additional training, or expensive inference-time operations, increasing both complexity and runtime.

In this work, we introduce ϕ-Noise, a training-free video conditioning framework that manipulates the input noise prior to diffusion using phase information (ϕ) extracted from a reference video, without introducing significant runtime or memory overhead. While diffusion noise is typically treated as a purely stochastic source, we show that its low-frequency components strongly influence global spatial and temporal structure. Motivated by this observation, we selectively modify the low-frequency phase of the input noise to inject structural and temporal biases into the generation process without changing the model architecture or inference pipeline.

Despite its simplicity, our method enables effective motion and spatial conditioning without additional training or expensive signal analysis. We demonstrate applications including motion conditioning, spatial conditioning, and cut-and-drag generation, achieving competitive or superior results compared to recent approaches. We further show that manipulating different frequency bands provides controllable variations in the conditioning behavior.

Our contributions are summarized as follows:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

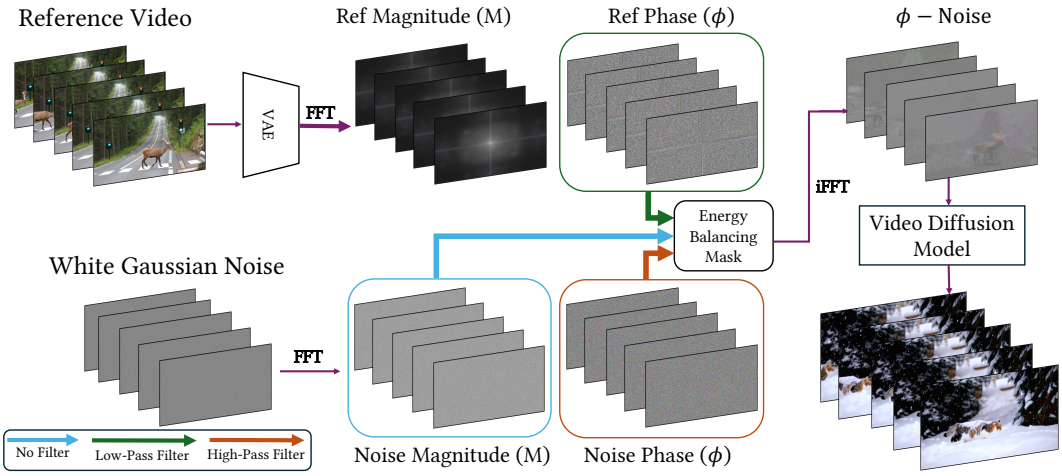

Figure 1: **Method Overview.** We calculate the frequency decomposition of both noise and signal using the Discrete-Fourier Transform to phase and magnitude. Then, we replace the low-frequencies phase information of the noise with those of the conditional input, and normalize total energy of the reconstructed noise. The output noise is then used as input to the generation model. Note that we show the frames of the original video for visualization but in practice work in latent space.

- We analyze diffusion video generation from a frequency-domain perspective, studying how temporal frequency manipulation shapes generated motion and affects latent energy evolution throughout the diffusion process.
- We propose ϕ-Noise, a training-free and low-overhead framework for spatial and temporal video conditioning through simple noise-phase manipulation.
- We demonstrate that our approach generalizes across tasks and architectures, and can also be applied to conditional image generation models.

Code and implementation details will become available on our project page.

2 Related Work

2.1 Diffusion-based Video Generation

Following the success of latent image diffusion models (Rombach et al., 2022; Podell et al., 2023; Pernias et al., 2023; Labs, 2024; Wu et al., 2025; Esser et al., 2024), video generation has naturally emerged as an extension to the temporal domain, aiming to synthesize videos by progressively denoising spatio-temporal white Gaussian noise. Early works explored unconditional or weakly conditioned video diffusion models Ho et al. (2022b); He et al. (2023), demonstrating promising visual quality but limited controllability.

Subsequent approaches leveraged pretrained text-to-image models to enable controllable video synthesis. These include methods that inject temporal information into independently generated frames Khachatryan et al. (2023), introduce trained modules that enforce temporal consistency Blattmann et al. (2023b); Singer et al. (2023); Ho et al. (2022a), or combine image-based backbones with video-specific architectures Wu et al. (2023a) to benefit from strong spatial priors learned on large-scale image datasets, enabling high-quality generation with limited video data.

More recent works further improve fidelity, scalability, and motion coherence Bar-Tal et al. (2024); Guo et al. (2024); Yang et al. (2025); Chen et al. (2024); OpenAI (2024), while still relying primarily on text conditioning or hybrid image-video training. However, text alone remains insufficient for specifying fine-grained spatial structure and complex temporal dynamics such as object motion and camera trajectories.

2.2 Visual Conditioning in Video Generation

To address the limitations of text conditioning, visual conditioning has emerged as a more expressive mechanism for controlling generated videos. Existing approaches can be broadly divided into

structural and *temporal* conditioning, which are typically addressed separately and often require task-specific designs.

Structural conditioning. Building on image-based conditioning methods Zhang et al. (2023); Ye et al. (2023); Mou et al. (2024), prior works incorporate visual signals such as edges, depth, or pose into video generation. These signals are typically injected via auxiliary networks trained to guide the generation process Chen et al. (2023); Hu (2024), or applied frame-wise using pretrained modules such as ControlNet Zhang et al. (2023) Hu & Xu (2023). Other approaches adapt pretrained image diffusion models to video generation Wu et al. (2023a); Khachatryan et al. (2023), reusing image conditioning mechanisms.

More recent methods directly condition on images or videos Geyer et al. (2023); Wan et al. (2025); HaCohen et al. (2024); Blattmann et al. (2023a), using them as structural references. Additional works explore interactive or spatial control signals, such as drag-based editing or region-level guidance Shi et al. (2023); Mou et al. (2023); Avrahami et al. (2024), extending structural conditioning to more flexible user control.

Temporal conditioning. Controlling motion dynamics remains more challenging and is often addressed separately from structure. Early methods rely on explicit motion representations such as optical-flow or camera trajectories Geng et al. (2025); Koroglu et al. (2025); Yang et al. (2024), which require additional estimation or user specification. More recent approaches instead learn implicit motion representations through latent alignment Xiao et al. (2024), inversion Wang et al. (2024); Ling et al. (2024), or motion descriptor optimization Yatim et al. (2024). Other works incorporate camera control or trajectory-guided generation He et al. (2024); Deng et al. (2023); Gokmen et al. (2025), and optimization-based temporal control Schiber et al. (2026), further highlighting the diversity of task-specific solutions.

Despite their effectiveness, most existing approaches focus on either structural or temporal control and rely on additional components such as auxiliary networks, inversion procedures, attention manipulation, or specialized representations. This often increases computational overhead and limits generality across tasks and architectures.

2.3 Latent Noise Manipulation

Diffusion models generate visual outputs by progressively transforming white Gaussian noise into structured samples. Although the input noise is typically treated as unstructured randomness, several works have explored manipulating it to guide the generation process.

FreeInit Wu et al. (2024) injects low-frequency components from inverted videos into the noise to improve temporal coherence, while FreqPrior Yuan et al. (2025) decomposes noise into frequency bands to enhance detail generation. Other approaches employ noise warping Deng et al. (2025); Chang et al. (2024); Burgert et al. (2025) for temporal consistency and motion control, and Time-to-Move Singer et al. (2025) utilizes SDEdit Meng et al. (2022) for localized motion editing. In the image domain, noise manipulation has also been used for image editing Ren et al. (2025) and conditioning on hand-drawn color maps Cohen et al. (2026). Additional works further challenge the assumption of white Gaussian noise by exploring alternative noise distributions for generative models Voleti et al. (2022); Rissanen et al. (2023); Huang et al. (2024). NeuralRemaster Zeng et al. (2026) incorporates spatial structure through phase information during training, primarily to enhance the fidelity of the input video or modify its visual style. However, its training-based formulation constrains the generated output to remain closely tied to the input, limiting its ability to perform flexible content editing or to condition motion independently of spatial structure and colors.

In contrast, our approach is entirely training-free and operates solely by manipulating the input noise prior to the generation process. This design enables a broader range of applications while disentangling structural and temporal conditioning, thereby allowing them to be controlled independently.

Our perspective. Building on these observations, we directly manipulate the frequency decomposition of the input noise by injecting low-frequency components from conditioning signals prior to diffusion. Since low frequencies capture coarse spatial structure and dominant temporal dynamics, this enables control over both appearance and motion during generation.

Unlike prior approaches, our method provides a unified conditioning framework operating over the spatial and temporal dimensions of the noise. It is entirely training-free, introduces minimal computational overhead, and remains agnostic to the underlying model architecture.

3 ANALYSIS

In this section, we analyze the spectral behavior of white Gaussian noise in a Video Diffusion model and present experiments to build intuition for the proposed method. We define frequency extraction operators in the context of video latents, followed by an in-depth investigation into phase-based noise manipulation. We specifically examine how this manipulation affects the latent phase distribution and the critical importance of maintaining spectral energy balance to prevent generative divergence.

3.1 PRELIMINARIES

Our primary goal is motion dynamic transfer. Hence, to isolate motion dynamics as much as possible, we utilize in our analysis a neutral-background reference video featuring simple object movements (e.g., *"A ball bouncing up and down"*). We examine how temporal spectral manipulation can transfer these motion patterns to a newly generated video.

Let $V \in \mathbb{R}^{T\times W\times H\times C}$ be a sequence of T frames. A latent diffusion model encodes V into a latent tensor $\mathbf{v} \in \mathbb{R}^{t\times w\times h\times d}$. We apply the Discrete Fourier Transform (DFT) to map this latent signal to the frequency domain along either the 1D *temporal* (t) or 2D *spatial* (w, h) dimensions, denoted as $\mathcal{F}_T$ and $\mathcal{F}_S$, respectively. We first analyze temporal decomposition for motion transfer, then extend the approach to the spatial domain for structural conditioning.

Given initial white Gaussian noise $\mathbf{z}$ of the same dimensionality as $\mathbf{v}$, we define their frequency representations $\tilde{\mathbf{v}}, \tilde{\mathbf{z}}$ by applying the DFT along the temporal dimension:

$$\{\tilde{\mathbf{v}}, \tilde{\mathbf{z}}\} = \mathcal{F}_T\left(\{\mathbf{v}, \mathbf{z}\}\right). \tag{1}$$

The complex-valued representations are then decomposed into magnitudes $M^{\mathbf{v}}, M^{\mathbf{z}}$ and phases $\phi^{\mathbf{v}}, \phi^{\mathbf{z}}$. Following Parseval's Theorem Pollard (1926), the total energy of a discrete signal is preserved (up to a scaling constant) between the spatial and frequency domains. We define the energy E of the latent $\mathbf{z}$ as the sum of its squared components:

$$\mathrm{E}(\mathbf{z}) = \sum_i \mathbf{z}_i^2 \propto \mathrm{E}(\tilde{\mathbf{z}}) = \sum_i |\tilde{\mathbf{z}}_i|^2. \tag{2}$$

In this context, $\mathrm{E}(\mathbf{z})$ represents the total "strength" of the signal.

3.2 PHASE MANIPULATION OF WHITE GAUSSIAN NOISE

In signal processing, the phase spectrum typically captures structural information, whereas the magnitude spectrum captures the distribution of energy across scales Oppenheim & Lim (1981). In the temporal domain, this structure corresponds directly to motion patterns. Our hypothesis is that by substituting the low-frequency phase of $\mathbf{z}$ with the phase of a reference video $\mathbf{v}$ we can control the motion dynamics of the video generated by $\mathbf{z}$. We focus on low frequencies, as they represent the global, coarse motion trajectories most critical for temporal coherence.

Let k be the frequency cutoff point ($0 < k \leq t$). We define $\tilde{\mathbf{z}}_k$ as the latent resulting from substituting the k lowest temporal frequency phases of $\mathbf{z}$ with those of $\mathbf{v}$:

$$\tilde{\mathbf{z}}_k = M^{\tilde{\mathbf{z}}} \odot e^{i\phi^{\tilde{\mathbf{z}}_k}}, \text{ where } \phi_f^{\tilde{\mathbf{z}}_k} = \begin{cases} \phi_f^{\mathbf{v}} & \text{if } f \leq k \\ \phi_f^{\mathbf{z}} & \text{if } f > k \end{cases}, \tag{3}$$

for each frequency index f.

Next, we use $\tilde{\mathbf{z}}_k$ to generate a video conditioned on a text prompt. This operation is non-trivial, as video diffusion models are trained on i.i.d. white Gaussian noise with uniformly distributed phase, $\phi \sim \mathcal{U}(-\pi, \pi)$. In fig. 2 (left), we analyze the phase distributions by measuring the Kullback–Leibler (KL) divergence between generated outputs and reference videos. While the injected phase successfully transfers motion (orange trajectories), the resulting videos exhibit severe saturation and poor prompt alignment (fig. 2, bottom-left grid). As shown in the energy evolution plot (fig. 2-right, ▲-shaped plots), injecting low frequencies alters the latent energy throughout denoising, with larger k values causing significant energy divergence and out-of-distribution artifacts.

3.3 ENERGY EFFECT OF SPECTRAL MANIPULATION

The divergence observed in the previous section suggests that phase substitution disrupts the expected energy profile of the noise. To investigate this, we first attempted to scale the manipulated

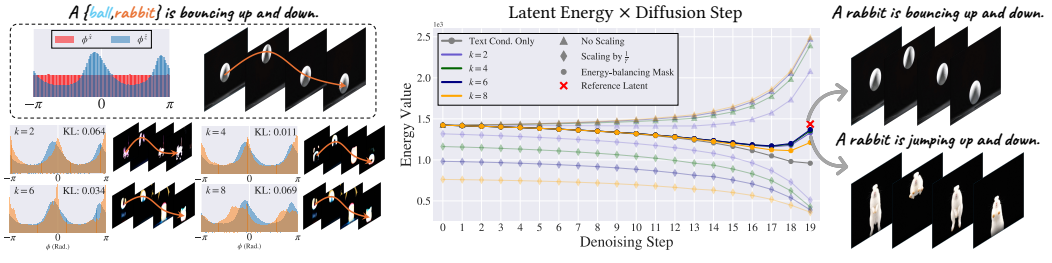

Figure 2: **Phase and Energy Analysis.** We analyze the impact of substituting k low-frequency phase components in the latent space prior to denoising. (Left) Comparison of phase distributions between the reference video (blue) and the generated outputs (orange). (Middle) Evolution of latent energy across denoising timesteps for various k values (colors) and scaling settings (markers). The red symbol ($\times$) denotes the reference energy $\mathrm{E}(\mathbf{x})$. (Right) Qualitative Comparison. Applying our energy-balancing mask Φ preserves signal energy, ensuring stable denoising and high-fidelity motion transfer that faithfully follows the reference dynamics. We recommend zooming-in for a better view.



Figure 3: **Global Structure Transfer.** In addition to motion, we propose two methods for global structure transfer: (1) Image-to-Video (I2V) Motion Transfer, by utilizing an input image to fully preserve scene characteristics, layout, and identities (left); and (2) Implicit Temporal Conditioning, where the spatial layout and dynamics are preserved from the reference video V (right).

low-frequency magnitudes by a constant factor $1/\gamma$. However, as shown in Fig. 2 (middle, ♦-shaped plots), this leads to an energy collapse–dropping to 35% of the reference level– which leads to a collapse, where no meaningful video is produced.

To stabilize the denoising process, we propose a Spectral-Temporal Energy Balancing Mask $\Phi \in \mathbb{R}^t$, which ensures that while we scale down the high-energy low frequencies (by $1/\gamma$), we compensate by scaling the remaining frequencies by a factor β to preserve the total energy $\mathrm{E}(\mathbf{z})$.

$$\Phi(\tilde{\mathbf{z}}, k, \gamma)_f = \begin{cases} 1/\gamma & \text{if } f \leq k \\ \beta & \text{if } f > k \end{cases}, \quad \text{where } \beta = \sqrt{\frac{\mathrm{E}(\tilde{\mathbf{z}}) - \frac{\mathrm{E}(\tilde{\mathbf{z}}_\mathrm{l})}{\gamma^2}}{\mathrm{E}(\tilde{\mathbf{z}}_\mathrm{h})}}, \tag{4}$$

where f is the frequency index, and $\tilde{\mathbf{z}}_\mathrm{l}, \tilde{\mathbf{z}}_\mathrm{h}$ are the k and $(t-k)$ lowest and highest frequencies of $\mathbf{z}$, respectively. For a detailed derivation of β, please refer to the Appendix C.

By construction of this mask, we ensure that energy is preserved:

$$\mathrm{E}(\tilde{\mathbf{z}}_k \odot \Phi) = \mathrm{E}(\tilde{\mathbf{z}}). \tag{5}$$

Applying Φ effectively "re-whitens" the noise energy across the spectrum. As shown in fig. 2 (middle), this stabilization keeps the denoising process in-distribution, yielding high-quality videos that follow the reference motion without the artifacts of raw phase substitution.

3.4 Structure Conditioning

In complex real-world videos, where the background contains intricate details, temporal frequencies alone may be insufficient for structure transfer and generation. The temporal domain cannot explicitly define the underlying spatial geometry. To address this, as will be shown in section 4, we must

incorporate an additional input structural condition, such as a reference image, to provide explicit spatial scene information. This serves as a "structured guideline" to anchor geometric layout and identity details, as demonstrated In fig. 3 (bottom left).

4 METHOD

Building on our findings in section 3, we formulate a general and efficient framework for manipulating specific noise latent frequencies based on a conditioning reference video. We propose two methods that leverage the frequency domain. One applies to the temporal dimension, and the other to the spatial dimensions to anchor both motion and structure, while regulating the signal's energy to ensure sampling stability and robust generation.

4.1 SPECTRAL DECOMPOSITION

Given a reference video latent $\mathbf{v}$ and a gaussian noise latent $\mathbf{z}$, we first transform both into the frequency domain using a temporal or spatial DFT– $\mathcal{F}_{D\in\{\mathrm{T,S}\}}$:

$$\tilde{\mathbf{v}} = \mathcal{F}_{\mathrm{D}}(\mathbf{v}), \quad \tilde{\mathbf{z}} = \mathcal{F}_{\mathrm{D}}(\mathbf{z}). \tag{6}$$

Next, we decompose these spectral coefficients to magnitude M and phase ϕ:

$$\tilde{\mathbf{v}} = M^{\mathbf{v}} \odot e^{i\phi^{\mathbf{v}}}, \quad \tilde{\mathbf{z}} = M^{\mathbf{z}} \odot e^{i\phi^{\mathbf{z}}}. \tag{7}$$

4.2 PHASE SUBSTITUTION AND ENERGY BALANCING

To transfer the structural motion of the reference to the noise latent, we substitute the phase of $\tilde{\mathbf{z}}$ with that of $\tilde{\mathbf{v}}$ up to a frequency cutoff k, following eq. (3), to produce $\tilde{\mathbf{z}}_k$.

As shown in section 3.3, phase manipulation can disrupt the signal's energy evolution throughout the generation process. To preserve stability, we compute the spectral energy balancing mask $\Phi(\tilde{\mathbf{z}}, \gamma, k)$ (eq. (4)) and apply it to $\tilde{\mathbf{z}}_k$ to ensure energy conservation. Finally, the latent is mapped back to the spatial domain using the Inverse DFT:

$$\mathbf{z}^{\Phi} = \mathcal{F}_{\mathrm{D}}^{-1}\left(\Phi \odot \tilde{\mathbf{z}}_k\right). \tag{8}$$

The resulting latent $\mathbf{z}^{\Phi}$ serves as the initialization for the denoising process, biasing the generation toward the motion of the reference video while remaining within the model's learned distribution. Since applying the DFT and its inverse is computationally negligible compared to even a single diffusion iteration, the proposed method introduces negligible runtime and memory overhead, as it does not intervene in the diffusion process itself.

5 APPLICATIONS

To demonstrate the capabilities of ϕ-Noise, we present three applications under a single framework: text-conditioned motion transfer, text + first-frame motion transfer, and Cut & Drag generation. Results are shown in fig. 4, Appendix D and the supplemental video. Further comparisons are provided in fig. 5. We employ WAN Wan et al. (2025) for all of the following experiments and demonstrate additional results on LTX2HaCohen et al. (2024) in Appendix D.

Text-Conditioned Motion Transfer Given a reference video and a text prompt, our goal is to generate a video that matches the prompt while preserving the input motion. As discussed in section 3, applying ϕ-Noise along the spatial dimensions yields the best performance. The results demonstrate strong alignment with both the textual content and the motion patterns. Some spatial information is also transferred, which, while not always desirable, helps maintain temporal consistency and is also observed in competing methods.

Text + First Frame Motion Transfer The goal is to align with both a text prompt and a first-frame condition while following the motion of the input video. We combine the WAN first-frame baseline with ϕ-Noise. Our method successfully transfers motion across varying subjects (e.g., replacing a human with a cat) and handles complex dynamics, such as backflips, while preserving coherence.

Cut-and-Drag Manipulations In this setting, users either cut object patches from an image or add an outside sprite on top of a given image, and animate them by dragging them rigidly across the frame. The goal is to generate a coherent video that follows the prescribed motion. We employ WAN with first-frame conditioning together with ϕ-Noise.

As shown in fig. 4, our method produces natural object motion despite the rigid inputs. In the left example, it also introduces plausible visual effects (e.g., a red line becoming fire). The right example highlights the method's flexibility: the octopus is not constrained to its rigid patch and instead moves freely while still adhering to the specified motion.

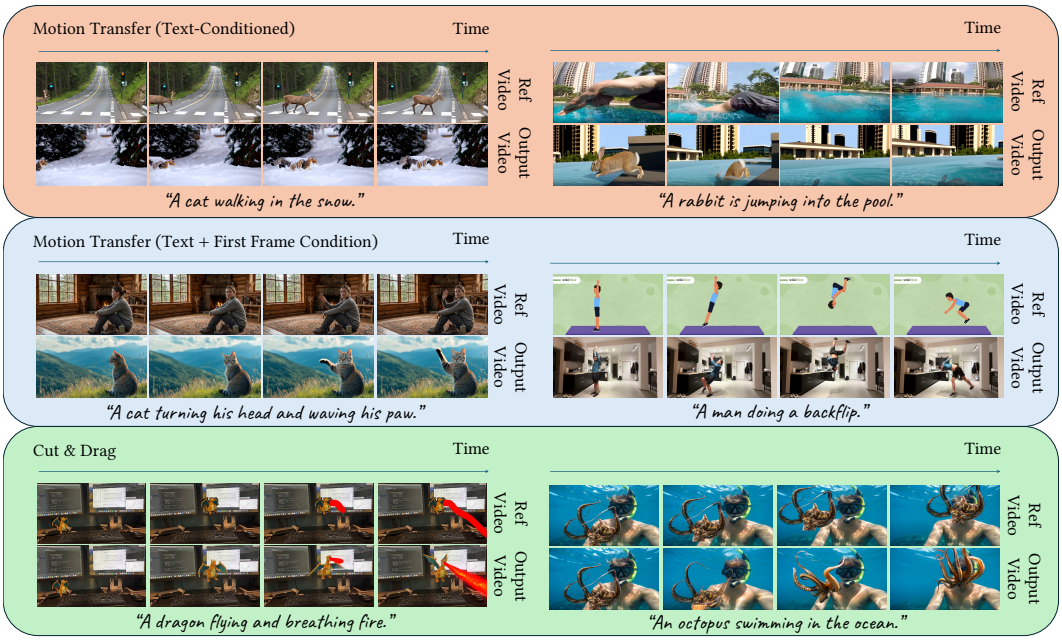

Figure 4: **Applications.** We showcase temporal conditioning under three settings: text-only conditioning (top), text combined with first-frame conditioning (middle), and Cut & Drag inputs (bottom). In the middle and bottom rows, the first-frame condition is indicated by the leftmost frame in each sequence. (We recommend zooming in for a better view).

6 Experiments

To evaluate our spectral manipulation framework, we conduct extensive experiments across our three primary applications. We benchmark our approach against state-of-the-art diffusion guidance and motion transfer techniques, providing quantitative evaluations for Text-to-Video (T2V) and Cut & Drag (CND) tasks, alongside qualitative comparisons across all settings.

6.1 Experimental Setup

Implementation Details. For all experiments, we apply our proposed ϕ-Noise manipulation directly to the initial Gaussian noise prior to the diffusion denoising process of the Wan2.2-14B model Wang et al. (2025). For Image-to-Video (I2V) and Cut & Drag tasks, we empirically select the frequency cutoff parameter $k \in [1, 5]$ and fix the scaling coefficient to $\gamma = 30$. For the implicit temporal conditioning (T2V) task, we set $\gamma = 4$ and define k as a continuous masking ratio, typically set to $\sim 5\%$.

Datasets and Evaluation Benchmark. We compile a diverse evaluation suite of 60 high-quality videos to rigorously assess motion transfer and structural preservation. This benchmark comprises 20 published examples from the Time-to-Move (TTM) dataset Singer et al. (2025), 30 videos sourced from the LOVEU-TGVE-2023 dataset Wu et al. (2023b) (utilized specifically for evaluating object replacement captions), and 10 in-the-wild videos collected to test generalization on complex real-world dynamics.

Baselines. We employ the foundational Wan2.2-14B models (Wan-T2V and Wan-I2V) as our primary unconditioned baselines to establish standard text and image capabilities. For conditional generation, we compare against recent state-of-the-art approaches. For text-based generation, we compare against DiT-Flow Pondaven et al. (2025), the T2V version of MotionClone Ling et al. (2024) DMT Yatim et al. (2024) and RoPECraft Gokmen et al. (2025). For Cut & Drag, we evaluate against Go-With-The-Flow (GWTF) Burgert et al. (2025) and Time-to-Move (TTM) Singer et al. (2025) and for TI2V Motion Transfer, we compare to IT2V MotionClone Ling et al. (2024) and I2V Wan 2.2 baseline Wan et al. (2025).

Evaluation Metrics. We evaluate the generated videos across two categories. (1) **Task-Specific Metrics:** We measure dense motion alignment using **Flow-Err** (optical flow error between the



Figure 5: **Qualitative Comparisons.** We compare ϕ-Noise with recent state-of-the-art methods for each application. In the middle and bottom rows, the first-frame condition is indicated by the leftmost frame in each sequence. (We recommend zooming in for a better view).

reference and generated video), temporal consistency via **LPIPS-Temp** Zhang et al. (2018), and semantic text alignment via **CLIP-T** Radford et al. (2021). (2) **Generative Quality:** We utilize VBench Zhang et al. (2024), a comprehensive evaluation suite for video diffusion, reporting Subject Consistency, Background Consistency, Motion Smoothness, Dynamic Degree, Aesthetic Quality, and Imaging Quality.

6.2 Qualitative Comparisons.

As shown in fig. 5, in the text-only setting (top), ϕ-Noise achieves strong motion transfer, with slight spatial leakage as the manipulation is applied along spatial dimensions. (see section 3). DiT-Flow exhibits weaker and sometimes missing motion transfer, while DMT improves alignment but introduces spatial leakage and remains inferior overall.

For text + first-frame conditioning, the Wan baseline fails to capture motion without explicit conditioning, whereas our method successfully produces complex motions aligned with the reference. MotionClone shows limited capability in this regard.

In the Cut & Drag setting, GWTF yields stiff motion and visible artifacts. Both TTM and our method perform well, but differ in behavior: constrained by its mask condition, TTM adheres closely to the input patches, while our method produces more natural and expressive motion. This is evident in the bird example, where our results include realistic wing flapping, whereas TTM remains more constrained to the patch motion.

Table 1: **Quantitative Evaluation.** We report task-specific motion and generative quality (VBench Zhang et al. (2024)) metrics across Cut & Drag and T2V Motion Transfer (T2V MT). **Bold** and underline indicate best and second-best performance among motion-based manipulation methods; Wan-T2V/I2V denote unmanipulated base models.

	Model	Image-Based Metrics			Motion-Based Metrics				
		CLIP-T ↑	Aes ↑	Img ↑	LPIPS-T ↓	Flow-E ↓	Subj-C ↑	Smooth ↑	Dyn-D ↑
Cut & Drag	*base model (Wan-I2V)*	*0.308*	*0.652*	*0.644*	*0.116*	*181.10*	*0.942*	*0.978*	*0.647*
	GWTF	**0.314**	0.620	0.637	**0.097**	152.81	0.942	**0.981**	0.647
	TTM	0.311	**0.647**	**0.653**	0.110	102.39	**0.948**	0.978	0.705
	Ours	0.313	0.637	0.627	0.171	**101.49**	0.918	0.964	**0.764**
T2V MT	*base model (Wan-T2V)*	*0.312*	*0.604*	*0.705*	*0.062*	*103.26*	*0.955*	*0.979*	*0.645*
	DiT-Flow	**0.319**	0.526	0.611	0.112	94.60	0.931	0.973	**0.935**
	DMT	0.314	0.530	0.581	0.114	67.23	0.914	0.963	0.871
	MotionClone	0.304	**0.548**	0.646	0.204	67.92	0.864	0.919	0.903
	RopeCraft	0.313	0.514	0.573	0.124	**53.03**	0.909	0.969	0.925
	Ours	0.302	0.546	**0.683**	**0.075**	61.75	**0.952**	**0.980**	0.709

6.3 Quantitative Comparisons.

Motion Transfer and Consistency. Table 1 summarizes the performance of our method against the baselines. Our approach demonstrates strong performance in dense motion alignment, achieving the lowest in Cut & Drag, second-lowest in T2V (T2V: 61.75, CND: 101.49). This supports our hypothesis from section 3 that low-frequency phase strongly influences the global motion trajectories.

Generative Quality (VBench). Since competing methods use different architectures, we compare against the unmanipulated Wan-T2V/I2V models as quality references. ϕ-Noise largely preserves the underlying generative quality while substantially improving motion alignment. In the T2V setting, it maintains comparable image quality, temporal consistency, subject consistency, and smoothness while reducing Flow-Err from 103.26 to 61.75. In the Cut & Drag setting, it achieves the highest Dynamic Degree (0.764) among motion-based methods while maintaining comparable aesthetic quality and the lowest Flow-Err (101.49).

Computational Efficiency. Unlike test-time optimization or attention-injection methods that introduce heavy per-step overhead during denoising, ϕ-Noise is highly efficient. By modifying only the initial noise $\mathbf{z}$ via a single FFT operation, it introduces near-zero latency. Consequently, our method maintains the same inference latency as the Wan2.2 baseline.

6.4 Additional Experiments

We conduct additional experiments to further evaluate the capabilities of ϕ-Noise. These include seed variation analysis, ablations over the choice of γ and k, prompt variation experiments, applying ϕ-Noise to an additional video generation model, and extending it to image generation models. We present these experiments, along with further comparisons, in Appendix D.

7 Limitations and Conclusion

In this paper, we introduced ϕ-Noise, a simple and efficient cross-task framework for motion transfer based on manipulating the low-frequency phase components of the input noise in the Fourier domain. Through extensive analysis, we showed that directly modifying the noise is non-trivial, as it disrupts the spectral balance of the latent signal. To address this issue, we proposed an energy-balancing mask that 're-whitens' the manipulated Gaussian latent prior to denoising, keeping it aligned with the expected distribution of the generative model.

Our method's primary limitation lies in its sensitivity to the parameter space, particularly the masking ratio k. While improper tuning can lead to structural artifacts or motion mistransfer, it also provides meaningful control over the degree of the transferred motion.

We evaluated ϕ-Noise across multiple motion transfer tasks, demonstrating strong performance compared to prior training- and optimization-based approaches, which are often tailored to specific settings. More broadly, our results highlight input noise as a powerful and underexplored conditioning space, suggesting that frequency-based noise manipulation can serve as a general and flexible framework for controllable video generation.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and formulate mathematical claims. We have not used generative AI tools to provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpret results, and Generate synthetic data sets, assistance with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, draft parts of the research paper, edit the research paper to improve readability, and to identify relevant literature. We have reviewed all AI-assisted work. Every text created by AI was verified for correctness and for matching the authors intentions by at least two authors of the paper. Every relevant literature suggestion was searched manually through Google search engine and citations bibliography paragraphs were sourced from project pages or arxiv pages directly and not by AI suggestions. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our method is described comprehensively in the paper, with all implementation details clearly specified. Following publication, we plan to open-source our code, including a complete implementation of the proposed method and the experimental pipeline used to obtain our results. The repository will also include the examples presented in the paper, facilitating exact reproduction of our experiments. Importantly, our approach builds on classical frequency decomposition techniques and publicly available open-source models, making it readily accessible and reproducible without requiring model retraining or prohibitively large computational resources.

REFERENCES

Omri Avrahami, Rinon Gal, Gal Chechik, Ohad Fried, Dani Lischinski, Arash Vahdat, and Weili Nie. Diffuhaul: A training-free method for object dragging in images. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687590. URL https://doi.org/10.1145/3680528.3687590.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687614. URL https://doi.org/10.1145/3680528.3687614.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023a. URL https://arxiv.org/abs/2311.15127.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 13–23, June 2025.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C. Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzElnMrgSD.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

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, 2023.

Nadav Z. Cohen, Oron Nir, and Ariel Shamir. Conditional balance: Improving multi-conditioning trade-offs in image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2641–2650, June 2025.

Nadav Z. Cohen, Ofir Abramovich, and Ariel Shamir. Colorful-Noise: Training-Free Low-Frequency Noise Manipulation for Color-Based Conditional Image Generation. *arXiv e-prints*, art. arXiv:2605.00548, May 2026.

Yitong Deng, Winnie Lin, Lingxiao Li, Dmitriy Smirnov, Ryan D Burgert, Ning Yu, Vincent Dedun, and Mohammad H. Taghavi. Infinite-resolution integral noise warping for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Y6LPWBo2HP.

Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. Dragvideo: Interactive drag-style video editing. *arXiv preprint arXiv:2312.02216*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora, 2024.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories, 2025. URL https://arxiv.org/abs/2412.02700.

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

Ahmet Berke Gokmen, Yigit Ekin, Bahri Batuhan Bilecen, and Aysegul Dundar. Ropecraft: Training-free motion transfer with trajectory-guided rope optimization on diffusion transformers, 2025. URL https://arxiv.org/abs/2505.13344.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2024.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. URL https://arxiv.org/abs/2211.13221.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4775–4785, June 2024.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022a. URL https://arxiv.org/abs/2210.02303.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022b.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8153–8163, June 2024.

Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet, 2023. URL https://arxiv.org/abs/2307.14073.

Xingchang Huang, Corentin Salaun, Cristina Vasconcelos, Christian Theobalt, Cengiz Oztireli, and Gurprit Singh. Blue noise for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657435. URL https://doi.org/10.1145/3641519.3657435.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15954–15964, October 2023.

Mathis Koroglu, Hugo Caselles-Dupré, Guillaume Jeanneret, and Matthieu Cord. Onlyflow: Optical flow based motion conditioning for video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6216–6226. IEEE, June 2025. doi: 10.1109/cvprw67362.2025.00619. URL http://dx.doi.org/10.1109/CVPRW67362.2025.00619.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28226. URL https://doi.org/10.1609/aaai.v38i5.28226.

OpenAI. Sora: Creating video from text. https://openai.com/sora, 2024. Accessed: 2026-05-02.

Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher J. Pal, and Marc Aubreville. Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

S. Pollard. On parseval's theorem. *Proceedings of the London Mathematical Society*, s2-25(1):237–246, 1926. doi: https://doi.org/10.1112/plms/s2-25.1.237. URL https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-25.1.237.

Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *CVPR*, 2025.

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*, pp. 8748–8763. PmLR, 2021.

Yufan Ren, Zicong Jiang, Tong Zhang, Søren Forchhammer, and Sabine Süsstrunk. Fds: Frequency-aware denoising score for text-guided latent diffusion image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 2651–2660, June 2025.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation, 2023. URL https://arxiv.org/abs/2206.13397.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Shira Schiber, Ofir Lindenbaum, and Idan Schwartz. Tempocontrol: Temporal attention guidance for text-to-video models, 2026. URL https://arxiv.org/abs/2510.02226.

Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023.

Assaf Singer, Noam Rotstein, Amir Mann, Ron Kimmel, and Or Litany. Time-to-move: Training-free motion controlled video generation via dual-clock denoising, 2025. URL https://arxiv.org/abs/2511.08633.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations (ICLR)*, 2023.

Vikram Voleti, Christopher Pal, and Adam Oberman. Score-based denoising diffusion with non-isotropic gaussian noise models, 2022. URL https://arxiv.org/abs/2210.12254.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningying Zhang, Pandeng Li, Ping Wu, Ruihang Chu, Rui Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wen-Chao Zhou, Wente Wang, Wen Shen, Wenyuan Yu, Xianzhong Shi, Xiaomin Huang, Xin Xu, Yan Kou,

Yan-Mei Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhigang Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *ArXiv*, abs/2503.20314, 2025. URL https://api.semanticscholar.org/CorpusID:277321639.

Luozhou Wang, Ziyang Mai, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Yingcong Chen. Motion inversion for video customization, 2024. URL https://arxiv.org/abs/2403.20193.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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 (ICCV)*, pp. 7623–7633, October 2023a.

Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, Rui He, Feng Hu, Junhua Hu, Hai Huang, Hanyu Zhu, Xu Cheng, Jie Tang, Mike Zheng Shou, Kurt Keutzer, and Forrest Iandola. Cvpr 2023 text guided video editing competition, 2023b.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models, 2024. URL https://arxiv.org/abs/2312.07537.

Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ZvQ4Bn75kN.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657481. URL https://doi.org/10.1145/3641519.3657481.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8466–8476, June 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Yunlong Yuan, Yuanfan Guo, Chunwei Wang, Wei Zhang, Hang Xu, and Li Zhang. Freqprior: Improving video diffusion models with frequency filtering gaussian noise. In *International Conference on Learning Representations (ICLR)*, 2025.

Yu Zeng, Charles Ochoa, Mingyuan Zhou, Vishal M. Patel, Vitor Guizilini, and Rowan McAllister. Neuralremaster: Phase-preserving diffusion for structure-aligned generation, 2026. URL https://arxiv.org/abs/2512.05106.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. *arXiv preprint arXiv:2412.09645*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Appendix

A HYPERPARAMETER-SENSITIVITY ANALYSIS

Table 2: Parameter sensitivity analysis for k and γ across motion similarity, motion error, text alignment, and aesthetic quality.

(a) Tracklet-Sim (↑)

$k \setminus \gamma$	1.5	2.25	3	3.75	*Avg*
3	0.76	0.77	0.78	0.70	*0.756*
6	0.68	0.73	0.69	0.59	*0.679*
9	0.58	0.58	0.59	0.60	*0.591*
12	0.55	0.50	0.58	0.58	*0.550*
Avg	*0.64*	*0.64*	*0.66*	*0.62*	–

(b) Flow-Err (↓)

$k \setminus \gamma$	1.5	2.25	3	3.75	*Avg*
3	126.66	118.74	114.66	85.38	*111.36*
6	115.23	88.93	100.97	109.48	*103.65*
9	237.35	107.94	104.26	115.36	*141.23*
12	1183.88	152.96	103.84	107.06	*386.93*
Avg	*415.78*	*117.14*	*105.93*	*104.32*	–

(c) CLIP-T (↑)

$k \setminus \gamma$	1.5	2.25	3	3.75	*Avg*
3	0.27	0.29	0.29	0.31	*0.296*
6	0.26	0.29	0.31	0.31	*0.299*
9	0.22	0.29	0.30	0.31	*0.280*
12	0.21	0.27	0.30	0.31	*0.270*
Avg	*0.24*	*0.29*	*0.30*	*0.31*	–

(d) Aes (↑)

$k \setminus \gamma$	1.5	2.25	3	3.75	*Avg*
3	0.50	0.55	0.56	0.60	*0.556*
6	0.46	0.58	0.61	0.62	*0.579*
9	0.32	0.53	0.60	0.60	*0.510*
12	0.31	0.48	0.59	0.61	*0.500*
Avg	*0.40*	*0.53*	*0.59*	*0.61*	–

To analyze the robustness of our approach, we conduct an extensive parameter sensitivity analysis for k (the fraction of low-frequency components substituted) and γ (the subdivision parameter over the lowest k frequency bins) on the T2V motion transfer (T2V-MT) task.

Intuitively, increasing the frequency cutoff (by lowering k, defined as $\frac{1}{2^k}$) expands the spectral range replaced by the reference content, thereby increasing the reference guidance. Similarly, once k is fixed, γ controls the energy magnitude of these low frequencies, scaling them by a factor of $\frac{1}{\gamma}$ (see Figures 20 and 21).

In Table 2, we present the parameter grids across motion fidelity (Tracklet-Sim, Flow-Err), text alignment (CLIP-T), and visual quality (Aes). For this analysis, we used the same evaluation set employed throughout the manuscript.

In terms of *motion reliability* to the reference input (Table 2a), Tracklet-Sim reflects high sensitivity to k; smaller k values (which preserve more low-frequency energy) yield higher similarity to the reference. On the other hand, Table 2b shows that increasing γ results in extreme Flow-Err values, showcasing its contribution in stabilizing structural motion reliability.

In terms of text reliability (Table 2c), higher k values tightly correlate with improved CLIP-T scores. As expected, reducing the reference guidance allows the model to prioritize the text prompt, reflecting a natural trade-off in multi-condition methods.

Finally, for our aesthetics-related metric (Aes, Table 2d), higher γ values consistently improve Aes scores, as over-constraining the low frequencies via a small k or extreme γ can introduce visual artifacts.

We further illustrate the effect of varying the γ and k parameters in Figures 20 and 21. As observed, different parameter combinations produce varying tradeoffs between visual fidelity and motion alignment.

B Ablation Study

To evaluate the contribution of each individual component, we conduct an ablation study on the core operations of ϕ-Noise: phase substitution, γ-division, and spectral energy normalization.

As seen in Table 3, omitting each key operation of our proposed method results in significant degradation in most metrics, compared to both the baseline and ours, and surprizingly often suppresses the baseline in non-motion metrics (CLIP-T, Aes)– a common trade-off in conditional generation, where quality-related metrics are slightly degraded). While the baseline is quantitatively competitive, it relies only on the first frame and therefore utilizes no motion information from the reference. As a result, despite its high visual quality, it is not sufficient for reproducing the reference motion.

Table 3: **Ablation study.** We conduct an extensive quantitative analysis across seven different metrics on two different tasks, ablating each of the core operations of our proposed method.

	Method	CLIP-T ↑	Aes ↑	LPIPS-T ↓	Flow-E ↓	Subj-C ↑	Smooth ↑	Dyn-D ↑
Cut & Drag	Baseline (random phase)	*0.308*	*0.652*	*0.116*	*181.100*	*0.942*	*0.978*	*0.647*
	+ Phase Substitution	0.247	0.308	0.193	265.193	0.542	0.937	**0.857**
	+ γ-Division	0.243	0.345	0.194	295.456	0.627	0.951	0.800
	+ Energy-Norm (Ours)	**0.313**	**0.637**	**0.171**	**101.490**	**0.918**	**0.964**	0.764
I2V MT	Baseline (random phase)	*0.237*	*0.661*	*0.188*	*235.466*	*0.936*	*0.973*	*0.571*
	+ Phase Substitution	0.235	0.301	0.181	114.881	0.883	0.945	0.857
	+ γ-Division	**0.312**	0.308	0.143	64.291	0.542	0.937	**0.857**
	+ Energy-Norm (Ours)	0.307	**0.543**	**0.086**	**33.518**	**0.957**	**0.980**	0.703

C Derivation of the Energy-Balanced Compensation Factor

To maintain spectral consistency during phase mixing, we define a Balanced-Energy Mask. Given the latent noise $\tilde{\mathbf{z}}$ and parameters γ, k, we scale the k lowest frequency components by $1/\gamma$ while preserving the total signal energy $\mathrm{E}(\tilde{\mathbf{z}})$.

Energy Decomposition

We partition the total energy $\mathrm{E}(\tilde{\mathbf{z}})$ into low-frequency ($\mathrm{E_l}$) and high-frequency ($\mathrm{E_h}$) components based on the threshold index k:

$$\mathrm{E_l} = \sum_{i=0}^{k} |\tilde{\mathbf{z}}_i|^2, \quad \mathrm{E_h} = \sum_{i=k+1}^{t-1} |\tilde{\mathbf{z}}_i|^2 \tag{9}$$

where $\mathrm{E}(\tilde{\mathbf{z}}) = \mathrm{E_l} + \mathrm{E_h}$ by the additivity of the squared Frobenius norm.

Conservation Constraint

Let $\tilde{\mathbf{z}}_k$ denote the energy-balanced noise. We scale the low-frequency components by $\frac{1}{\gamma}$ and the high-frequency components by a compensation factor β. We require:

$$\mathrm{E}(\tilde{\mathbf{z}}) = \mathrm{E}(\tilde{\mathbf{z}}_k) \tag{10}$$

Expanding the energy of the modified signal:

$$\mathrm{E}(\tilde{\mathbf{z}}_k) = \sum_{i=0}^{k} |\frac{1}{\gamma} \cdot \tilde{\mathbf{z}}_i|^2 + \sum_{i=k+1}^{t-1} |\beta \cdot \tilde{\mathbf{z}}_i|^2 \tag{11}$$

$$= \left(\frac{1}{\gamma}\right)^2 \cdot \sum_{i=0}^{k} |\tilde{\mathbf{z}}_i|^2 + \beta^2 \cdot \sum_{i=k+1}^{t-1} |\tilde{\mathbf{z}}_i|^2 \tag{12}$$

$$= \frac{1}{\gamma^2}\mathrm{E_l} + \beta^2 \mathrm{E_h} \tag{13}$$



Figure 6: **ϕ-Noise for Image Generation.** We apply ϕ-Noise to SDXL by injecting spatial phase information into the input noise. As shown, the biased noise enables the generated images (middle row) to spatially align with the reference image (top row), whereas generation with unbiased noise (bottom row) exhibits different spatial arrangements and alignment patterns.

Closed-form Expression for β

Equating the terms to satisfy the energy conservation constraint:

$$\mathrm{E}(\tilde{\mathbf{z}}) = \frac{1}{\gamma^2}\mathrm{E_l} + \beta^2 \mathrm{E_h} \tag{14}$$

Rearranging to isolate β:

$$\beta^2 \cdot \mathrm{E_h} = \mathrm{E}(\tilde{\mathbf{z}}) - \frac{\mathrm{E_l}}{\gamma^2} \tag{15}$$

Which yields the final expression:

$$\beta = \sqrt{\frac{\mathrm{E}(\tilde{\mathbf{z}}) - \frac{\mathrm{E_l}}{\gamma^2}}{\mathrm{E_h}}}. \tag{16}$$

D Additional Experiments

D.1 ϕ-Noise for Image Generation

As shown in Section 3 of the main manuscript, ϕ-Noise can be applied in the spatial domain to preserve motion cues from an input video. In this section, we demonstrate that the same principle can also be extended to image generation models. Specifically, we employ SDXL Podell et al. (2023) and spatially bias its input noise using $\mathcal{F}_\mathrm{S}$, comparing the resulting outputs to those generated by vanilla SDXL across a variety of prompts and reference images. Results are presented in Figure 6.

As observed, the biased outputs (middle row) exhibit spatial structures similar to those of the reference images (top row), effectively enabling loose pose conditioning without requiring additional training or significant inference overhead, unlike prior approaches such as Zhang et al. (2023). Notably, our method generalizes across a wide range of object categories, including animals, humans, and even inanimate objects such as boxes, whereas many existing pose-conditioning methods are primarily limited to humans due to data and supervision constraints. In contrast, generation with unbiased noise fails to preserve the spatial alignment of the reference image, as expected.

D.2 Model Generalization

To evaluate the generality of ϕ-Noise across different architectures, we apply our method to LTX2 HaCohen et al. (2024). We present results in Figure 7 for text-conditioned motion transfer (top), text + first-frame conditioning (middle), and Cut & Drag inputs (bottom). As demonstrated, ϕ-Noise can be effectively integrated with LTX2 and produces pleasing results across all evaluated applications.

D.3 Seed Variations

Since ϕ-Noise conditions the noisy input prior to the diffusion process, the model retains the ability to modify and refine the output throughout generation. As a result, the generated samples exhibit slight variations across different random seeds, enabling exploration of diverse outputs while preserving the conditioning signal. We showcase seed variations in Figures 8 and 9.

D.4 Prompt Variations

To demonstrate our method's ability to adapt a single reference video to diverse prompts, we generate multiple videos with varying subjects and environments using the same reference. Results are shown in Figure 10.

D.5 Comparisons

We extend our comparisons presented in the main manuscript with additional samples. We present these comparisons in Figures 11 to 19.



Figure 7: **Applications with LTX-based video generation.** We demonstrate multiple applications using LTX text-to-video and image-to-video models. For each example, the first row shows the input and the second row shows the generated output conditioned on the corresponding prompt.



Figure 8: **Seed Variation.** Showcasing seed variation on a single video input for Cut & Drag generation. Each row denotes a different random seed. Prompt: *"An octopus swimming in the ocean."*

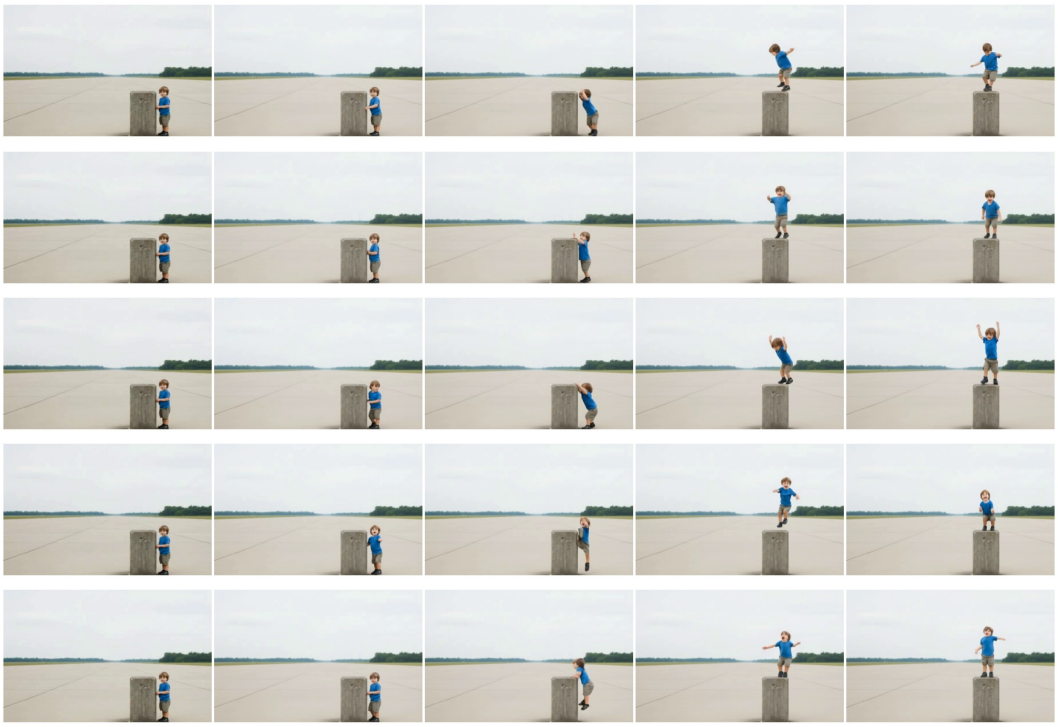

Figure 9: **Seed Variation.** Showcasing seed variation on a single video input for Cut & Drag generation. Each row denotes a different random seed. Prompt: *"A little boy jumping on a pillar"*.

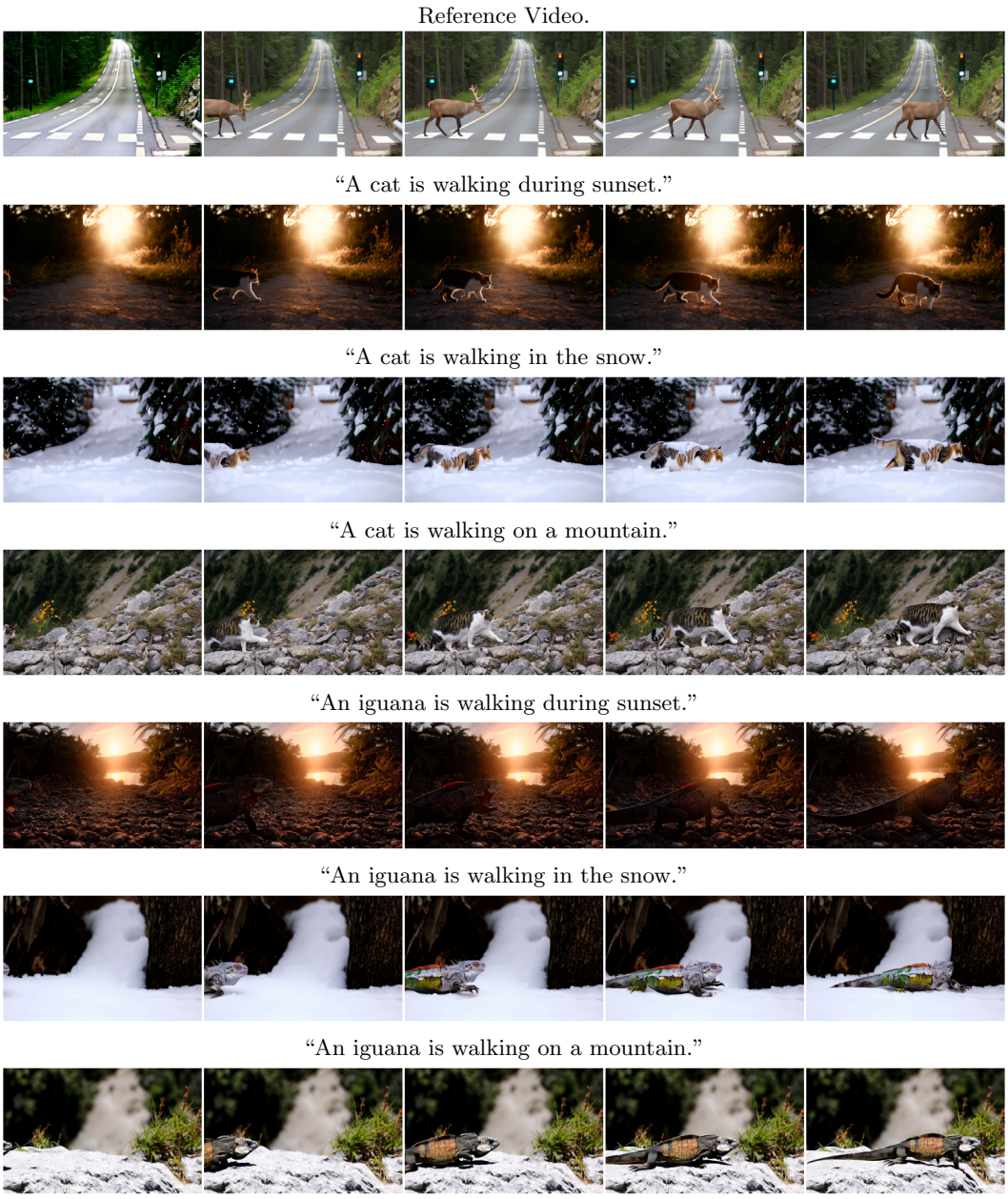

Figure 10: **Prompt Variation.** Given a reference video (top) we generate various videos depicting different animals and environments. As can be observed, all samples depict the reference video's motion.

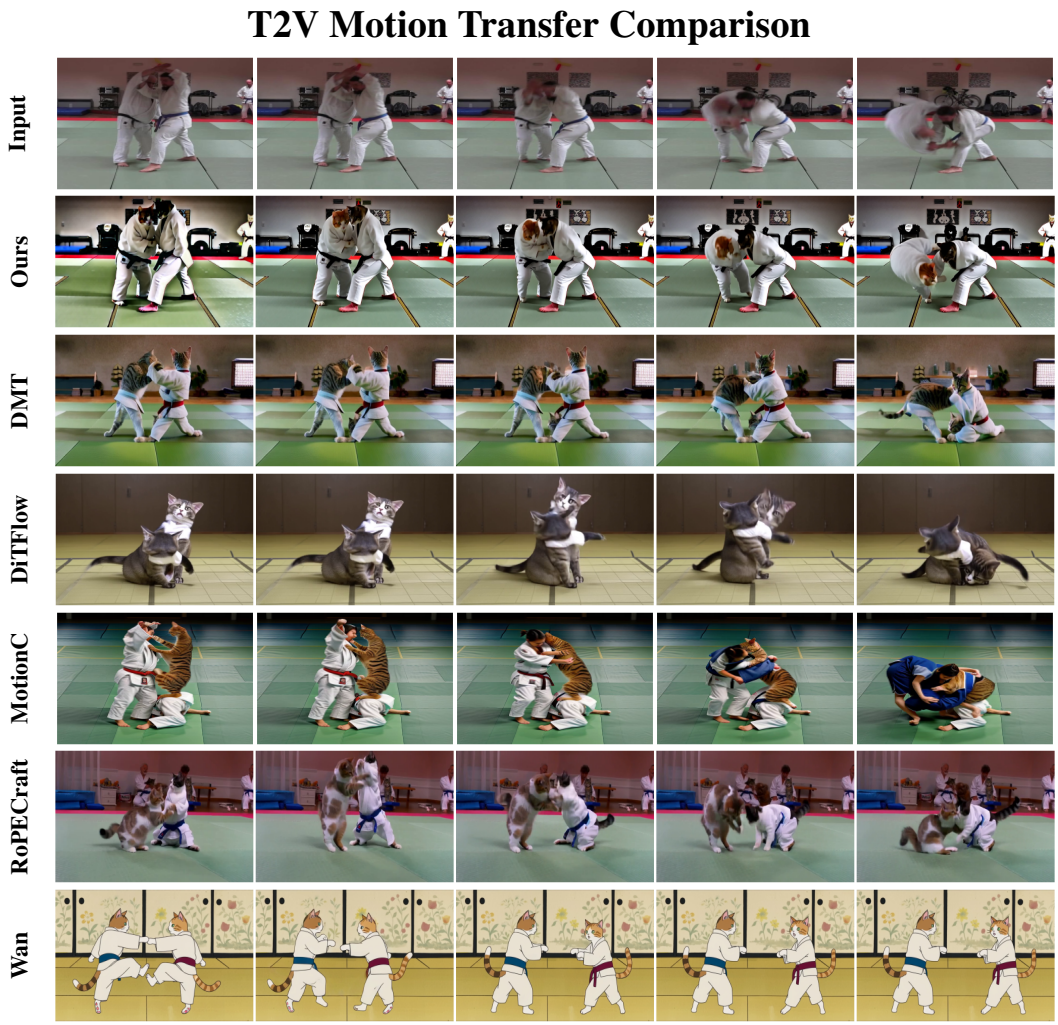

Figure 11: **T2V Motion Transfer comparison.** Qualitative comparison between different methods using the prompt "Two cats sparring in a dojo." while preserving the motion dynamics from the input video.

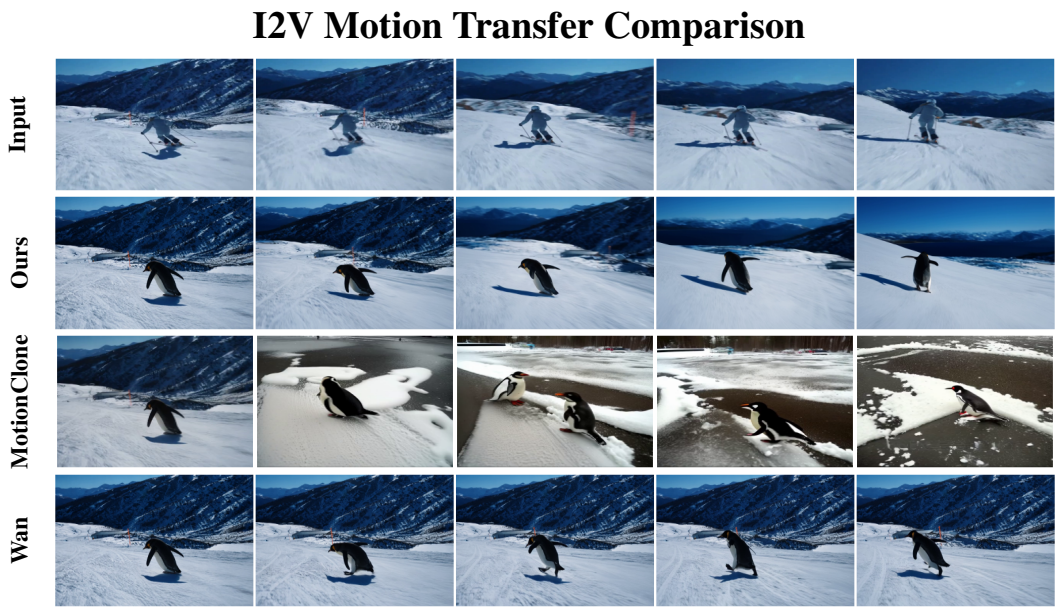

Figure 12: **I2V Motion Transfer comparison.** Results generated from an input image and the text prompt *"A penguin sliding down a snowy slope."* while preserving the transferred motion dynamics. The first frame is shown in the left column.

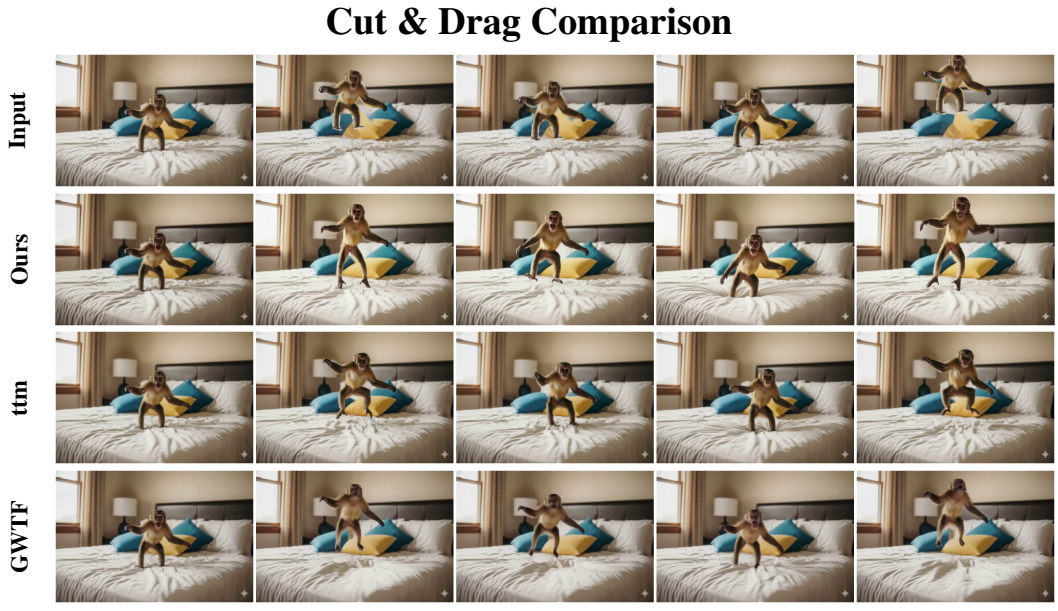

Figure 13: **Cut & Drag Comparison.** Results generated from an input image and the text prompt *"A monkey jumping on the bed."* while preserving the transferred motion dynamics.

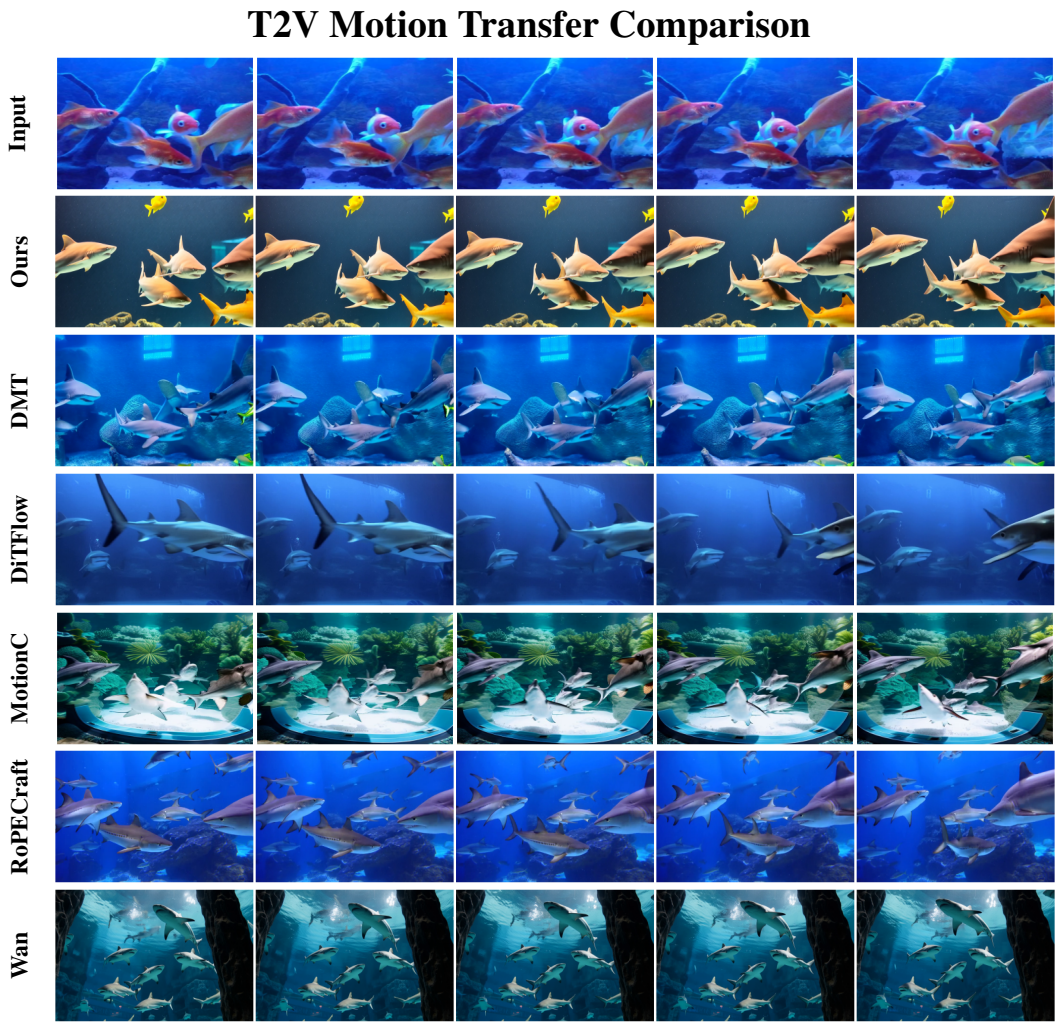

Figure 14: **T2V Motion Transfer comparison.** Qualitative comparison between different methods using the prompt *"Several sharks swim in a tank."* while preserving the motion dynamics from the input video.

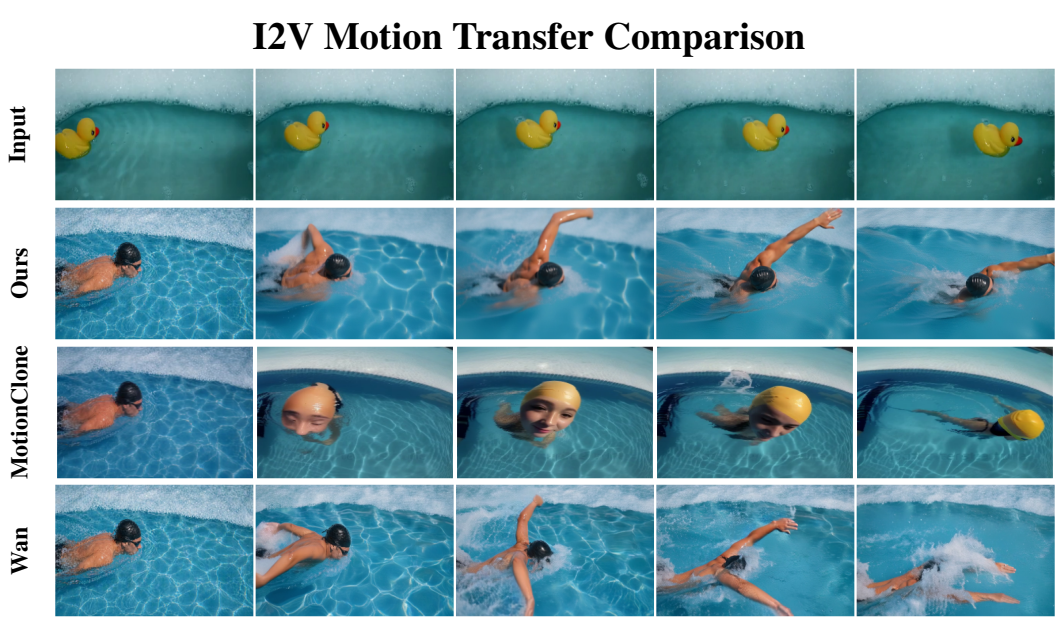

Figure 15: **I2V Motion Transfer comparison.** Results generated from an input image and the text prompt *"A swimmer is swimming in the pool."* while preserving the transferred motion dynamics. The first frame is shown in the left column.

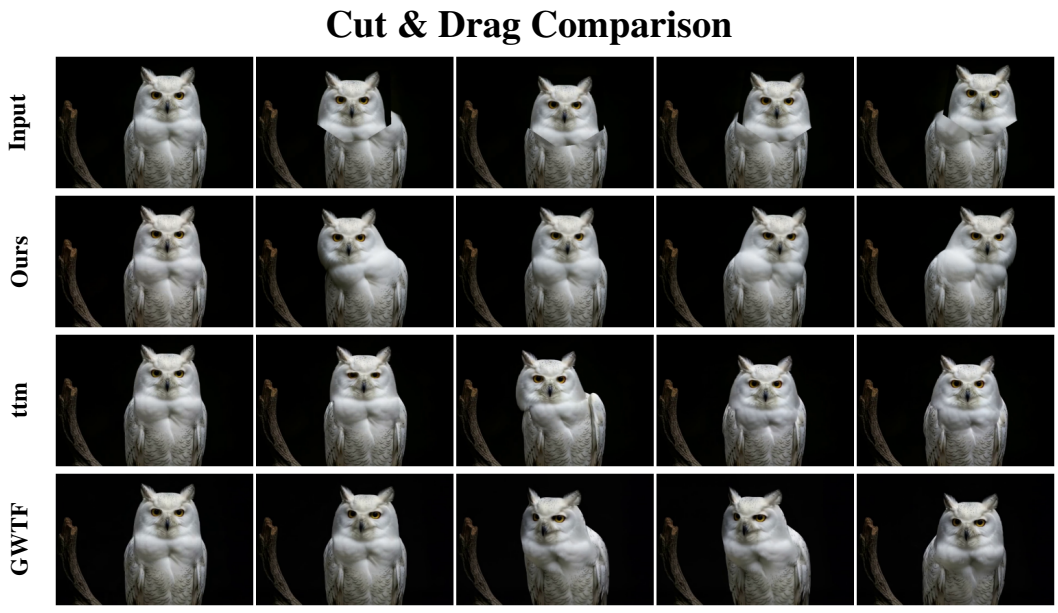

Figure 16: **Cut & Drag Comparison.** Results generated from an input image and the text prompt *"A majestic snowy owl perches gracefully on a gnarled branch, its pristine white feathers adorned with delicate black speckles. The owl's piercing yellow eyes are wide and alert, scanning the surroundings with a sense of calm authority. As a gentle breeze rustles through the leaves, the owl remains poised, its sharp talons gripping the branch securely. The dark, blurred background accentuates the owl's striking presence, creating a serene yet powerful scene in the quiet of the night."* while preserving the transferred motion dynamics.

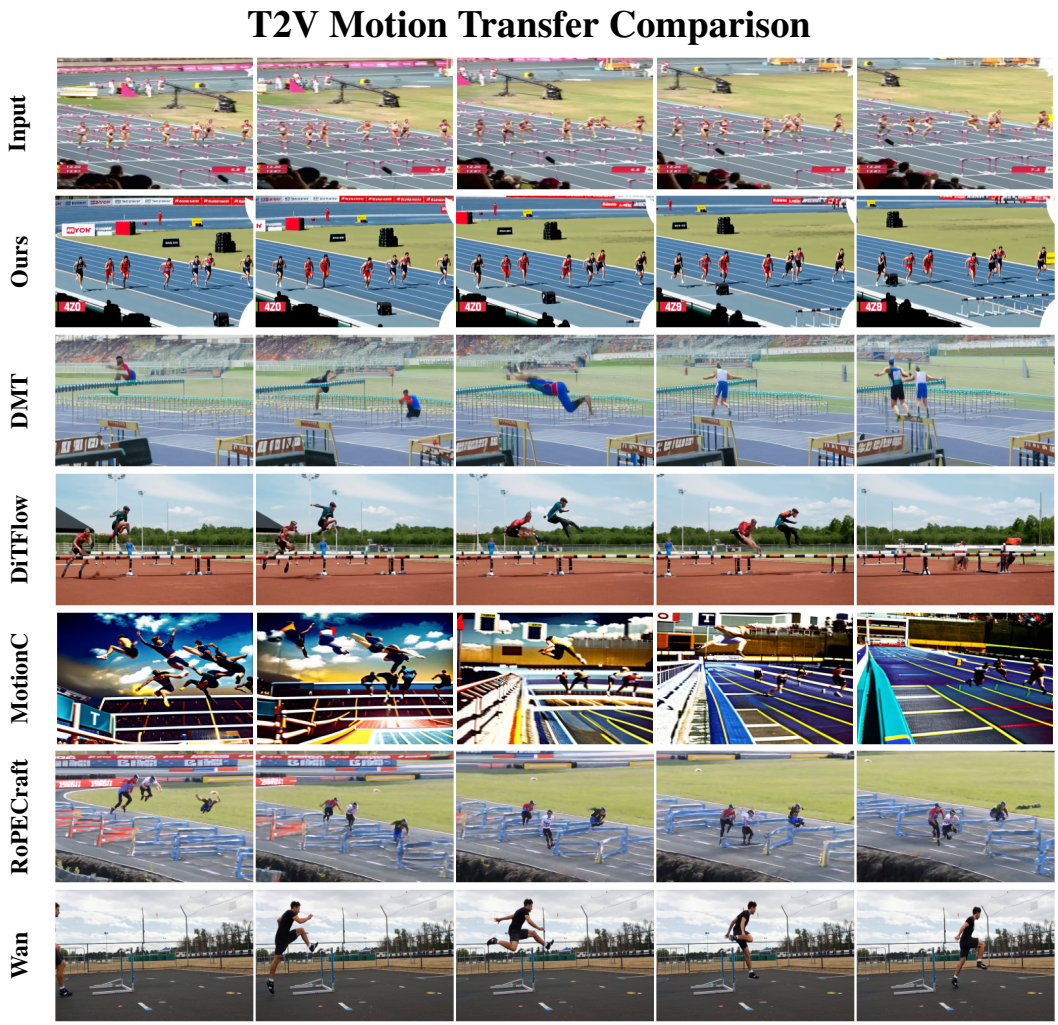

Figure 17: **T2V Motion Transfer comparison.** Qualitative comparison between different methods using the prompt *"Men jump over hurdles on a racetrack."* while preserving the motion dynamics from the input video.

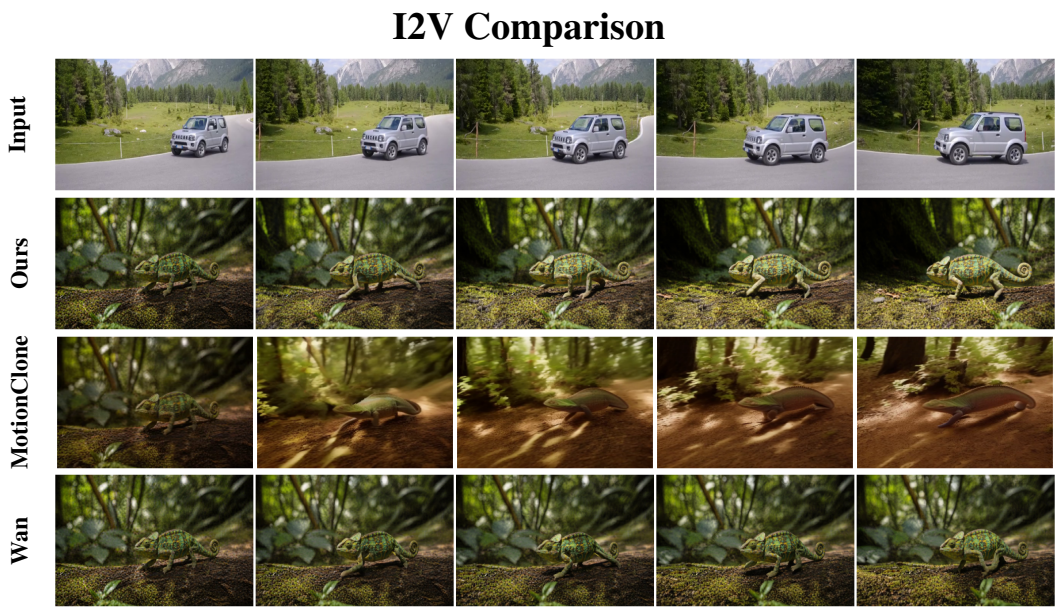

Figure 18: **I2V Motion Transfer comparison.** Results generated from an input image and the text prompt *"A chameleon is walking in the forest."* while preserving the transferred motion dynamics. The first frame is shown in the left column.

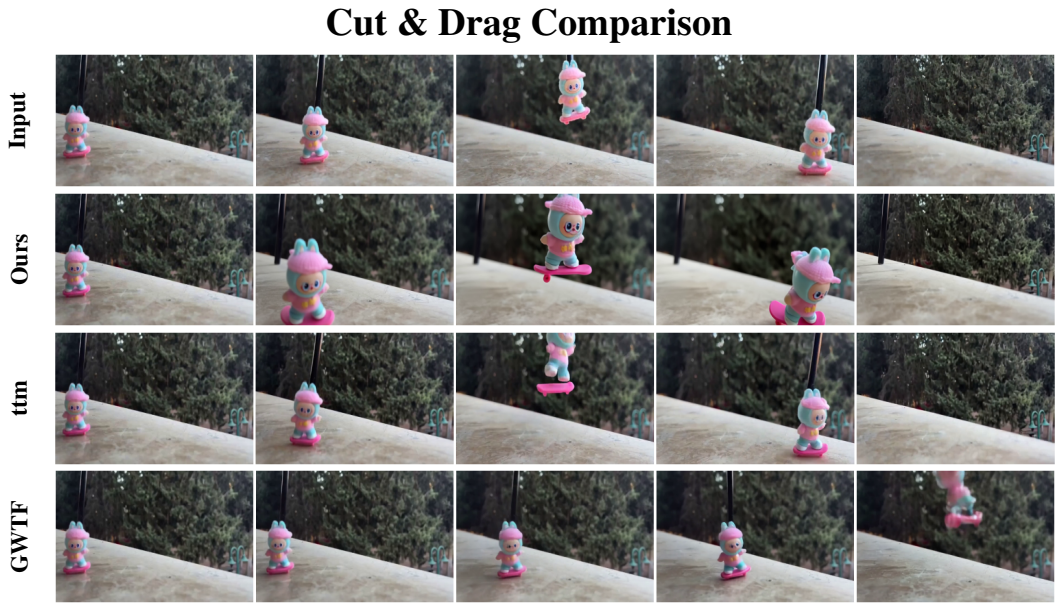

Figure 19: **Cut & Drag Comparison.** Results generated from an input image and the text prompt ‘‘The toy is riding a miniature pink skateboard along a light-colored stone ledge. Against a blurred background of green trees. Midway through the scene, the skateboard jump, before it lands back on the ledge and continues its ride out of the frame.” while preserving the transferred motion dynamics.

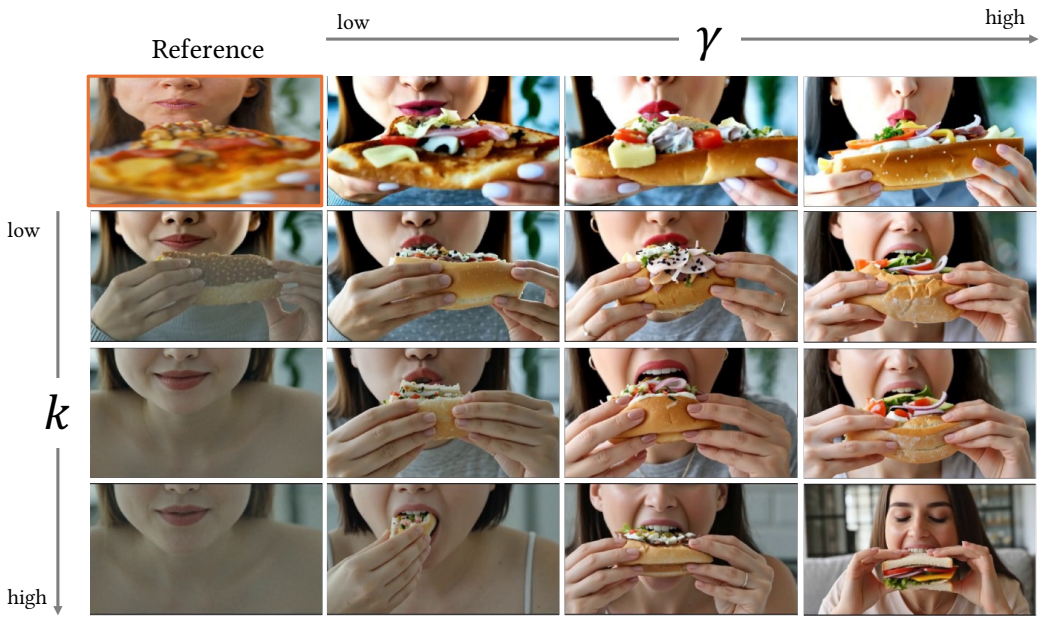

Figure 20: γ **and k Ablation.** Demonstration of the effect of different γ and k combinations. The reference image is shown in the top-left corner. Prompt: *"A woman eating a sandwich."*



Figure 21: γ **and k Ablation.** Demonstration of the effect of different γ and k combinations. The reference image is shown in the top-left corner. Prompt: *"A man riding a motorcycle."*