MotionShop: Zero-Shot Motion Transfer in Video Diffusion Models with Mixture of Score Guidance

Hidir Yesiltepe Tuna Han Salih Meral Connor Dunlop Pinar Yanardag
Virginia Tech
{hidir, tmeral, cdunlop, pinary}@vt.edu
motionshop-diffusion.github.io

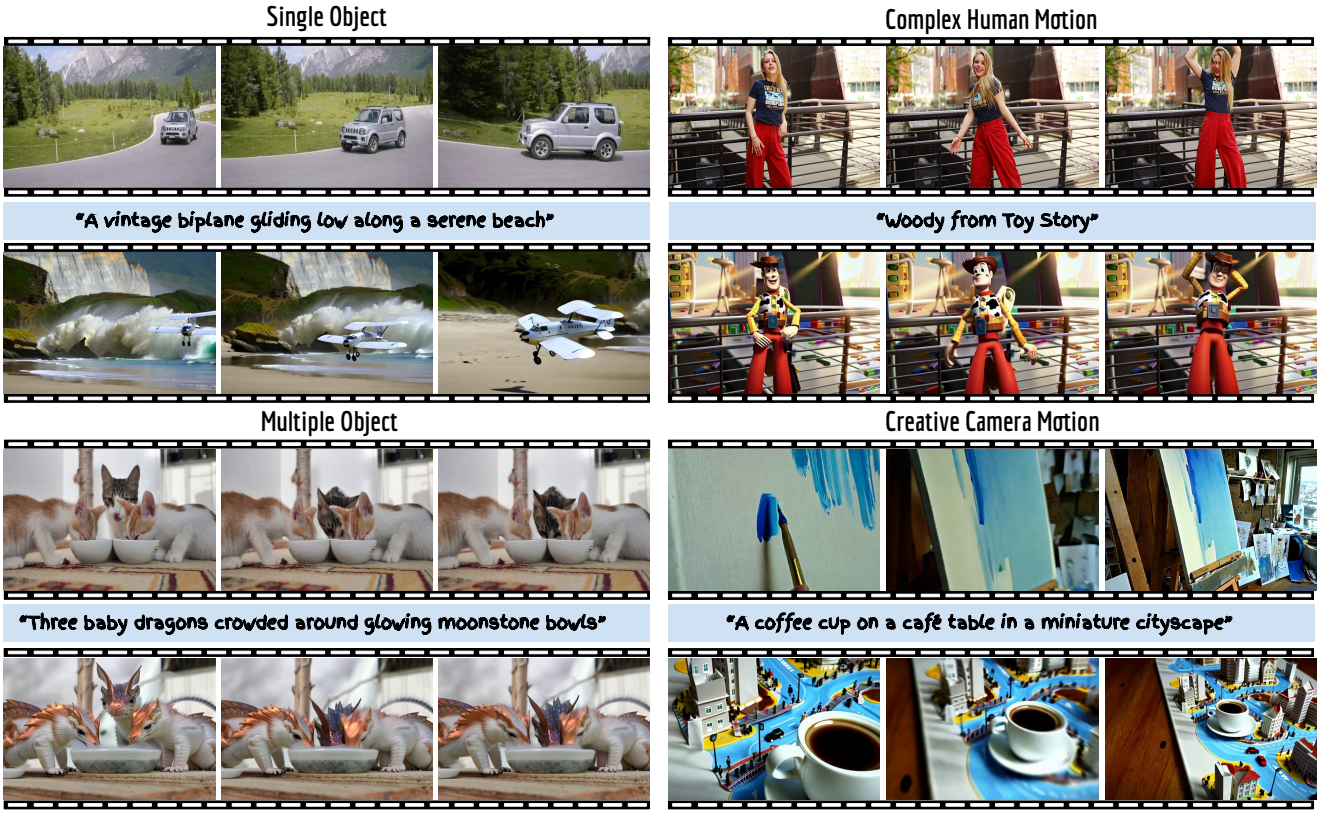

Figure 1. Mixture of Score Guidance (MSG), a novel approach for zero-shot motion transfer in diffusion models, enables high-fidelity motion synthesis across diverse scenarios. MSG successfully handles various motion patterns including complex object movements and camera trajectories. Full video results are available in the supplementary material.

Abstract

In this work, we propose the first motion transfer approach in diffusion transformer through Mixture of Score Guidance (MSG), a theoretically-grounded framework for motion transfer in diffusion models. Our key theoretical contribution lies in reformulating conditional score to decompose motion score and content score in diffusion models. By formulating motion transfer as a mixture of potential energies, MSG naturally preserves scene composition and enables creative scene transformations while maintaining the integrity of transferred motion patterns. This novel sampling operates directly on pre-trained video diffusion models without additional training or fine-tuning. Through extensive experiments, MSG demonstrates successful handling of diverse scenarios including single object, multiple objects, and cross-object motion transfer as well as complex camera motion transfer. Additionally, we introduce

MotionBench, the first motion transfer dataset consisting of 200 source videos and 1000 transferred sequences, covering single/multi-object transfers, and complex camera motions.

1. Introduction

Diffusion-based video generation models have gained substantial attention for their ability to produce high-quality, diverse video content. These models, driven by advances in text-to-video generation, open new possibilities for automated and creative video synthesis [1, 5, 8, 26, 28, 33, 33, 36]. Motion transfer in generative models [3, 10, 30, 34, 38, 40], has become a significant research area, focusing on transferring the motion from one video to another, often guided by text prompts. Consider the complex transformation depicted in Fig. 1, where a ground vehicle's trajectory is reimagined as the flight path of an aircraft. Such motion transfer involves more than merely replacing the car with a plane. For instance, translating the movement of a car into a plane gliding over a beach, as described by the text prompt (see Fig. 1) requires a significant adjustment in environmental context. This includes transforming how the car interacts with the road to how an aircraft engages with the sky. This level of control is particularly important as it enables users to create videos with motions that are challenging to describe through text prompts alone, such as complex camera motions (see Fig. 1).

Recent video generation and editing methods have focused on disentangling motion and appearance characteristics. Various approaches have emerged: MotionDirector [40] uses an appearance-debiased temporal loss with dual-path LoRA architecture, while DreamVideo [31], Customize-A-Video [22], and MotionCrafter [38] employ dedicated processing branches. VMC [10] combines fine-tuning and inversion techniques targeting temporal layers, and DMT [34] leverages space-time feature loss using DDIM inversion and UNet activations. Motion Inversion [30] uses motion embeddings trained from reference videos for temporal dynamics control. Despite these advancements, motion control in video generation remains challenging because of the complex interplay between spatial and temporal dimensions in video. Controlling motion is essential for applications in entertainment, advertising, and virtual reality, where specific and consistent movements are crucial to communicate a narrative or aesthetic vision.

However, while these methods are effective in straightforward motion transfer tasks involving single objects without significant background or object transformations, they struggle with more challenging motion transfer tasks. They often fail to adequately transform the scene, merely replacing one object with another without aligning the scenery with the changes specified in the text prompt. Other methods may dramatically alter the scene without preserving the original motion.On the other hand, video editing methods [2, 4, 11, 20] utilize structural similarities between source and target videos. However, their performance is limited when it comes to multi-object motion transfer or managing complex camera movements. Additionally, these methods struggle with significant shape transformations, such as turning a car into a flying plane (see Fig. 1). These limitations highlight the need for more advanced motion transfer techniques that can handle a wider range of transformations and motions without being constrained by the physical similarities between the source and target videos. Such capabilities would significantly enhance the flexibility and applicability of generative models in video editing and animation, opening up new possibilities for creative and practical applications.

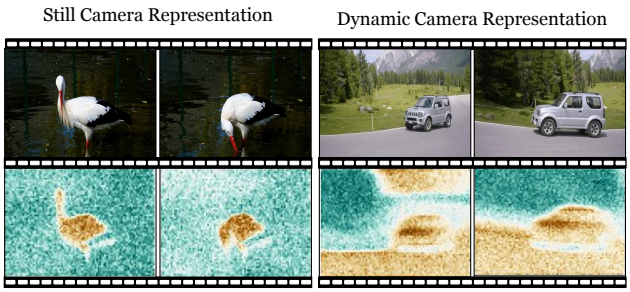

Figure 2. **Our intuition.** Visualization of motion characteristics $\mathcal{M}(z)$ extracted from early-timestep conditional scores. **(Left)** Multiple object motion representation showing the simultaneous movement of two objects. **(Right)** Combined object and camera motion representation demonstrating how our method captures both local object motion and global camera movement patterns. The visualizations are obtained from the conditional score maps $\nabla_{z_t} \log p_t(z|y)$ at early timesteps $t \ll T$.

In this paper, we present Mixture of Score Guidance (MSG), a novel approach for motion transfer in diffusion-based generative video models. Our method builds upon a novel conditional score reformulation, where we formulate motion transfer as a mixture of potential energies in the score space of diffusion models. By leveraging the observation that reformulated conditional scores encode rich motion information in early diffusion timesteps, MSG successfully isolates and transfers motion patterns. We establish the mathematical connection between score mixing and Langevin dynamics, providing theoretical perspectives for stable motion transfer. Through extensive experimentation, we demonstrate that MSG enables high-fidelity motion transfer across diverse scenarios without requiring model fine-tuning or additional training data. Our work extends beyond single-motion cases to handle multiple motion sources, and complex camera motions offering a unified approach to video motion transfer. Our contributions are as

follows:

- We introduce Mixture of Score Guidance (MSG), a theoretically grounded framework for motion transfer that formulates the problem through the lens of statistical mechanics. Our method operates directly in score space without requiring additional training or fine-tuning.
- We demonstrate the relationship between conditional scores and motion information, showing that score mixing in early diffusion steps provides an effective approach to motion transfer.
- We show that MSG's theoretical foundations naturally extend to complex scenarios including multi-motion synthesis and complex camera motion transfer.
- We introduce MotionBench, a comprehensive motion transfer benchmark comprising 200 diverse source videos and 1000 transferred sequences. This benchmark spans single/multi-object transfers,and camera motion variations enabling systematic evaluation of motion transfer methods across a broad range of scenarios.

2. Related Work

2.1. Text-to-Video Generation

Transformer architectures have emerged as a powerful foundation for video generation tasks. Early research scaling transformers for T2V applications, including Sora [17], CogVideo [8], CogVideoX [33] and LATTE [16], established the viability of this approach. The introduction of Diffusion Transformers [19] further cemented transformers as core components in video diffusion models. Several works have introduced specialized conditioning inputs: ControlVideo [39] leverages depth maps, DragNUWA [35] employs motion trajectories, while VideoDirectorGPT [14] and related approaches [3, 13] utilize spatial and temporal guides. T2I-based extensions include AnimateDiff [5], ModelScope [28], and InstructVideo [37].

2.2. Video Motion Editing and Transfer

Video motion control research has developed along two primary paths: explicit control through bounding boxes and motion transfer from reference videos. Explicit control methods include AnimateAnyone [12], Boximator [29], Peekaboo [9], and Trailblazer [15].

Another significant line of work focuses on transferring motion from reference videos. MotionDirector (MD) [40] made a significant advancement with its innovative dual-path LoRA architecture, effectively separating motion and appearance characteristics through specialized components that enable precise control over temporal dynamics. DreamVideo [31] and Customize-A-Video [22] further refined this separation using distinct branches for appearance and motion learning. Motion Inversion [30] introduced a novel approach by learning motion embeddings through temporal attention layers trained directly on the original video.

Video Motion Customization (VMC) [10] introduced a novel approach combining fine-tuning with inversion through adaptive temporal layer adjustments, achieving superior motion transfer results while maintaining temporal consistency. TokenFlow [4], ReRender-A-Video [32], and RAVE [11] explored various approaches to temporal consistency. The field has further advanced with MotionInversion (MI) [30] that enable precise control over temporal dynamics while maintaining visual quality through sophisticated motion embeddings trained from a reference video.

A persistent challenge in motion transfer is the assumption of feature similarity between reference and target videos. DMT [34] addresses this limitation through a novel space-time feature loss, leveraging internal UNet activations for improved motion fidelity. This approach achieves superior results in maintaining temporal consistency while allowing for more diverse edited outputs compared to traditional feature-matching methods.

3. Background

Diffusion Process. Consider a video sequence as a high-dimensional random variable $z \in \mathcal{Z}$ following an unknown data distribution $p(z)$. The diffusion process gradually transforms this distribution to a known prior distribution through a forward process defined by the following stochastic differential equation:

$$dz = [f(z,t) - \frac{g(t)^2}{2}\nabla_z \log p_t(z)]dt + g(t)d\bar{w}_t \quad (1)$$

where the drift coefficient $f(z,t)$ is characterized by:

$$f(z,t) = -\dot{\sigma}(t)\sigma(t)\nabla_{z_t} \log p_t(z)dt \quad (2)$$

and the diffusion coefficient $g(t)$ takes the form:

$$g(t) = \sigma(t)\sqrt{2\beta(t)} \quad (3)$$

The stochastic process is driven by the standard Wiener process $d\bar{w}_t$, while $p_t(z_t)$ represents the probability distribution of the noisy samples at time t. The boundary conditions of this distribution are given by the data distribution at the initial time, $p_0(z_0) = p_{\text{data}}(z)$, and a normal distribution with specified variance at the terminal time, $p_1(z_1) = \mathcal{N}(0, \sigma^2_{\text{max}}\mathbf{I})$. The time-reversed stochastic process for variance-preserving (VP) conditional diffusion is formulated in [27] as:

$$dz = -\frac{1}{2}\beta_t z dt - \beta_t \nabla_z \log p_t(z|y)dt + \sqrt{\beta_t}\bar{w}_t \quad (4)$$

By indicating directions of increased probability, the score naturally serves as a mechanism to undo the forward

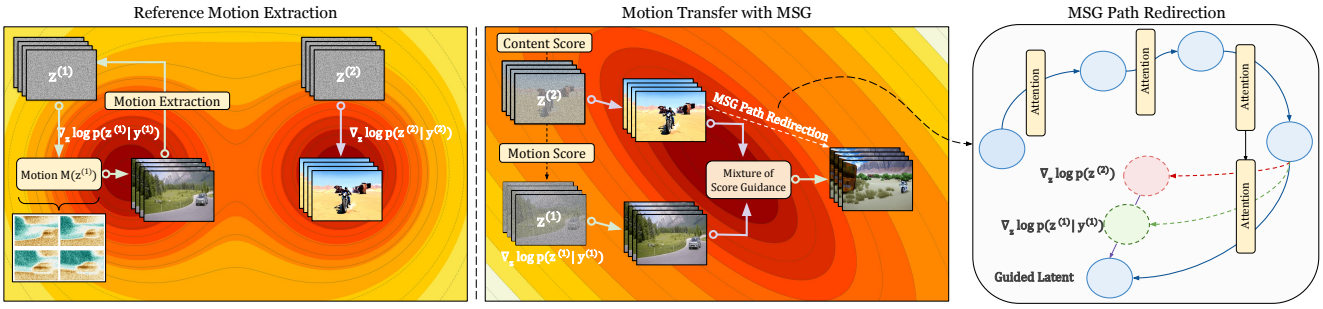

Figure 3. **Method Overview.** Framework of our Mixture of Score Guidance (MSG) for zero-shot motion transfer in diffusion models. **Left:** Reference motion extraction stage captures motion characteristics $M(z)$ from early-timestep conditional scores $\nabla_z \log p(z^{(1)}|y^{(1)})$ and $\nabla_z \log p(z^{(2)}|y^{(2)})$. **Middle:** Motion transfer combines content and motion scores through our MSG formulation $s_{\text{MSG}}(z_t, z_t^*) = \nabla_z \log p_t(z|y) + w_{\text{MSG}}(\nabla_z \log p_t(z^*|y^*) - \nabla_z \log p_t(z))$. **Right:** MSG path redirection mechanism showing attention-guided dynamics that enable stable motion transfer by exploring the correct motion manifold while preserving content through modified Langevin dynamics governed by our mixture of potential energies $U_{\text{MSG}}(z_t) = U_{\text{content}}(z_t) + w_{\text{MSG}}[U_{\text{motion}}(z_t, z_t^*) - U_{\text{prior}}(z_t)]$.

diffusion process.

Classifier Free Guidance. Classifier-Free Guidance (CFG) [6] enhances generation quality by interpolating between conditional and unconditional score predictions, effectively balancing fidelity and diversity in the output. CFG introduces a guided score $\nabla_z \log p_{t,\lambda}(z|y)$ that replaces the conditional score $\nabla_z \log p_t(z|y)$ in (4), defined at each timestep as:

$$\nabla_z \log p_{t,\lambda}(z|y) = (1-\lambda)\nabla_z \log p_t(z) + \lambda \nabla_z \log p_t(z|y) \tag{5}$$

where $\lambda = 1$ reduces to the standard conditional generation, while $\lambda > 1$ amplifies the influence of the conditioning signal, typically leading to higher-quality but potentially less diverse outputs.

Langevin Dynamics. As a fundamental stochastic process in statistical physics, Langevin dynamics (LD) [18, 25] enables sampling from complex probability distributions through continuous-time evolution. The dynamics follow a stochastic differential equation that combines deterministic drift with random fluctuations [23]:

$$dz = \frac{\epsilon}{2}\nabla \log p(z)dt + \sqrt{\epsilon}d\bar{w}_t \tag{6}$$

When implemented with appropriate step sizes, this process naturally evolves toward its equilibrium state $p(z)$ [24], making it particularly valuable for sampling tasks. The method's practical implementation hinges on the availability of the score function $\nabla \log p(z)$, which, similar to diffusion models, can be estimated through neural networks.

4. Methodology

This section presents the theoretical foundations and formulations of Mixture of Score Guidance (MSG), a novel approach for motion transfer in diffusion models in terms of statistical mechanics and stochastic processes.

4.1. Score-Based Motion Transfer

4.1.1 Score Function Decomposition

Let $\mathcal{M} : \mathbf{Z} \to \mathbf{M}$ be a mapping from the latent space to motion characteristics. The score function $\nabla_z \log p_t(z|y)$ can be separated into motion and content components through our conditional reformulation $\nabla_z \log p_t(z, \mathcal{M}(z^*)|y)$:

$$\begin{aligned}\nabla_z \log p_t(z, \mathcal{M}(z^*)|y) = \nabla_z \log p_t(\mathcal{M}(z^*)|y) \\ + \nabla_z \log p_t(z|\mathcal{M}(z^*), y)\end{aligned}$$

where $\mathcal{M}(z^*)$ is a reference motion representation and it is a function of the reference video latent z^*. This decomposition separates the score function into two meaningful components:
(1) Motion Score: $\nabla_z \log p_t(\mathcal{M}(z^*)|y)$ which is responsible for capturing how the latent affects motion characteristics and representing the gradient of log-likelihood concerning motion. The term dominates in early timesteps due to motion's hierarchical nature.
(2) Content Score: $\nabla_z \log p_t(z|\mathcal{M}(z^*), y)$ which captures content-specific information conditioned on motion and represents the residual gradient after accounting for motion. As a result it is more prominent in later timesteps.

4.1.2 Mixture of Score Guidance

Given a reference video with desired motion characteristics characterized by the reference condition y^* with the latents z^*, we formulate MSG as:

$$\begin{aligned}s_{\text{MSG}}(z_t, z_t^*) = \nabla_z \log p_t(z|y) \\ + w_{\text{MSG}}(\nabla_z \log p_t(z^*|y^*) - \nabla_z \log p_t(z))\end{aligned}$$

This formulation can be interpreted as a statistical mixture model in score space, where each component contributes to different aspects of the generation process. The theoretical significance of MSG can be understood through its relationship with Langevin dynamics. Consider the standard Langevin equation:

$$dz_t = \nabla_z U(z_t)dt + \sqrt{2\beta^{-1}}dW_t \tag{7}$$

where $U(z_t)$ is the potential energy function and β is the inverse temperature. Our MSG formulation extends this to a mixture of potential energies:

$$U_{\text{MSG}}(z_t) = U_{\text{content}}(z_t) + w_{\text{MSG}}[U_{\text{motion}}(z_t, z_t^*) - U_{\text{prior}}(z_t)] \tag{8}$$

This leads to the modified Langevin dynamics:

$$dz_t = \nabla_z U_{\text{MSG}}(z_t)dt + \sqrt{2\beta^{-1}}dW_t \tag{9}$$

Since the proposed operation does not harm the original dynamics of the denoising process, the system explores the correct motion manifold while preserving content and the resulting trajectories are stable and free from spurious artifacts.

4.1.3 Motion Trajectory Representation

Let $\mathcal{V} \in \mathbb{R}^{F\times H\times W\times 3}$ denote an input video sequence. Our aim is to derive a training-free motion representation through the following formulation. The forward process transforms the initial frame latents of the reference z_0^* into noised latents z_t^* at timestep t according to:

$$z_t^* = \alpha_t z_0^* + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{10}$$

where α_t and σ_t are time-dependent coefficients controlling the noise schedule [7]. The noising step is controlled by the strength parameter. The conditional score function $\nabla_z \log p_t(z^*|y)$ is then computed via a pretrained denoising network. Through investigation of score distribution, we establish that the motion representation operator $\mathcal{M} : \mathcal{Z} \to \mathcal{Z}$ defined as $\mathcal{M}(z) = \nabla_{z_t} \log p_t(z|y)$ captures predominant motion patterns at early diffusion timesteps $t \ll T$. This finding is empirically validated through our analysis of reference video conditional noise patterns, as illustrated in Figure 2 and Figure 3.

5. Experiments

Experimental Setup. Our implementation utilizes the CogVideoX [33] model for video generation and editing. We conduct all experiments at a resolution of 720 x 480 pixels using 50 diffusion timesteps. Due to the absence of a dedicated DDIM inversion schedule in CogVideoX, we employ a stochastic inversion approach where we add controlled noise to the input video latents, regulated by a strength parameter (detailed analysis in Fig. 8). For motion transfer, our pipeline operates in two phases: first, we obtain conditional score estimates from the reference video in early timesteps ($t \ll T$), then we apply this guidance during the generation of motion-transferred videos at the same timestep range. Throughout all experiments presented in this paper, we consistently set t to 10% of the total timesteps, as this configuration provides an effective balance between motion preservation and generation quality.

6. Qualitative Experiments

Our experimental results demonstrate MotionShop's versatility across diverse motion transfer scenarios. As shown in Figures 1 and 4, our method successfully handles both single and multi-object transfers. For single-object scenarios, MotionShop effectively transforms a black swan into a horse and a man riding jet ski (Fig. 4.c), maintaining realistic movement patterns and contextual elements like water splashes. In multi-object cases, our method seamlessly converts cats into birds (Fig. 4.b) and robots (Fig. 4.f). MotionShop provides flexible background control—enabling both dramatic alterations (Fig. 4.a) and preservation according to the text prompt (Fig. 4.c). Our method also supports concurrent motion controls, demonstrated by simultaneously transforming a frisbee into a coin while converting a dog into an eagle (Fig. 4.d). Additionally, MotionShop handles complex camera movements including zoom-ins, zoom-outs (Fig. 1), and rotational movements (Fig. 7), showcasing its comprehensive motion transfer capabilities.

7. Qualitative Comparisons

We conducted a qualitative comparison of MotionShop against MotionInversion [30], DMT [34], VMC [10], and MotionDirector [40], as shown in Table 5. Our evaluation focused on motion transfer across single/multiple objects and complex camera movements. Our experimental results reveal key differences in background handling among the methods: MotionDirector and VMC show limitations in background preservation, introducing undesirable artifacts (Table 5). In contrast, MotionShop demonstrates two distinct advantages: it enables accurate background modification when explicitly requested in the prompt (Table 5), and maintains consistent preservation of the original scene composition while properly transforming target objects (Table 5). These results indicate MotionShop's superior ability to distinguish between intentional and unintentional scene modifications. Additionally, MotionShop excels in transferring complex camera movements, including zoom-ins, zoom-outs, and rotations and their combinations such as pan-left and zoom-out (Fig. 7 bottom right), which proved challenging for DMT and MotionDirector in creative scene

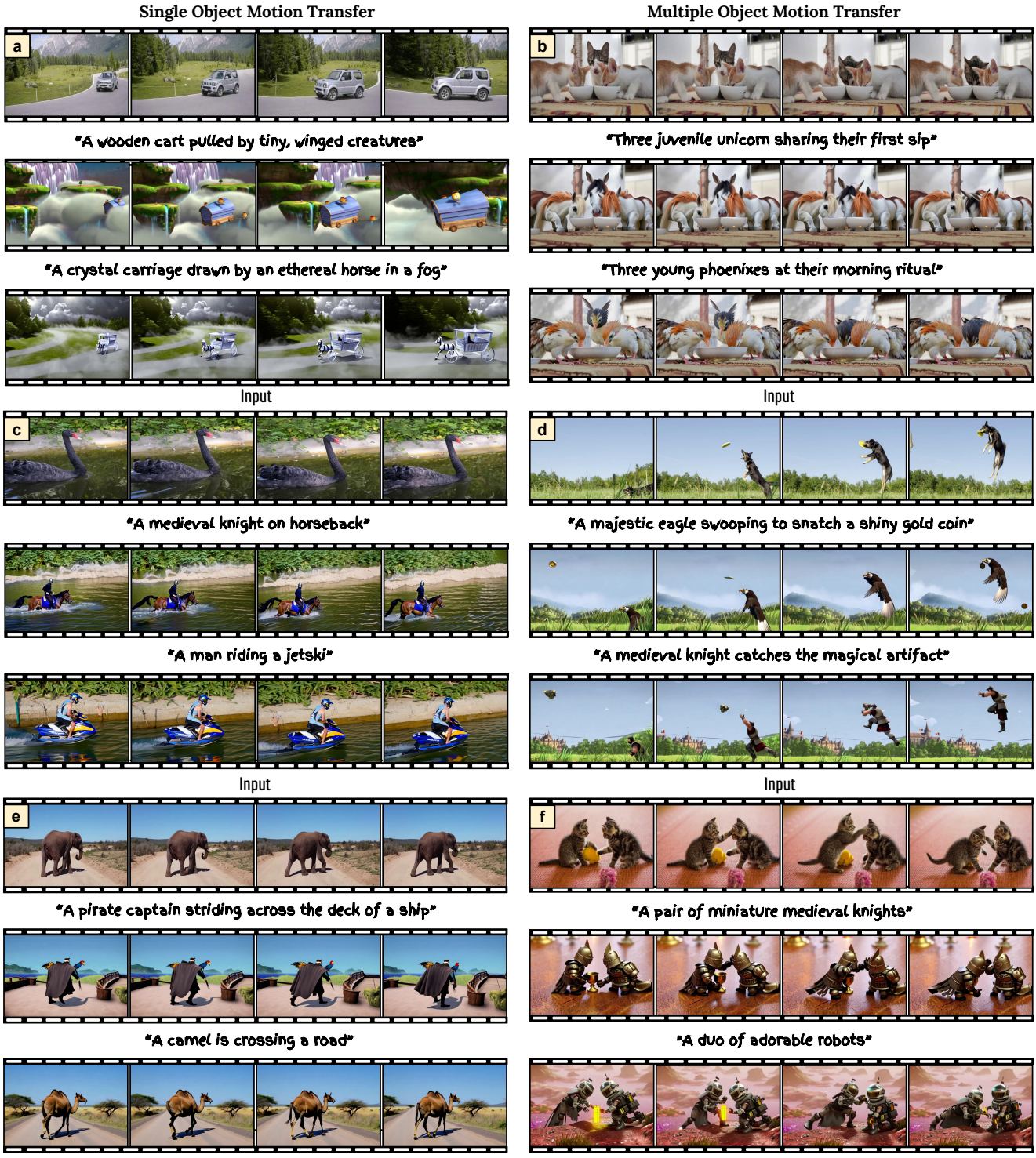

Figure 4. Qualitative results demonstrating our method's ability to preserve motion priors while generating novel content from text prompts. **(Left)** Single-object motion transfer where complex motions like mechanical movements, horseback riding sequences are accurately preserved in the generated outputs. **(Right)** Multi-object scenarios where our method successfully maintains the original motion dynamics while generating diverse subjects. Please refer to the Supplementary Material for full videos and additional examples.

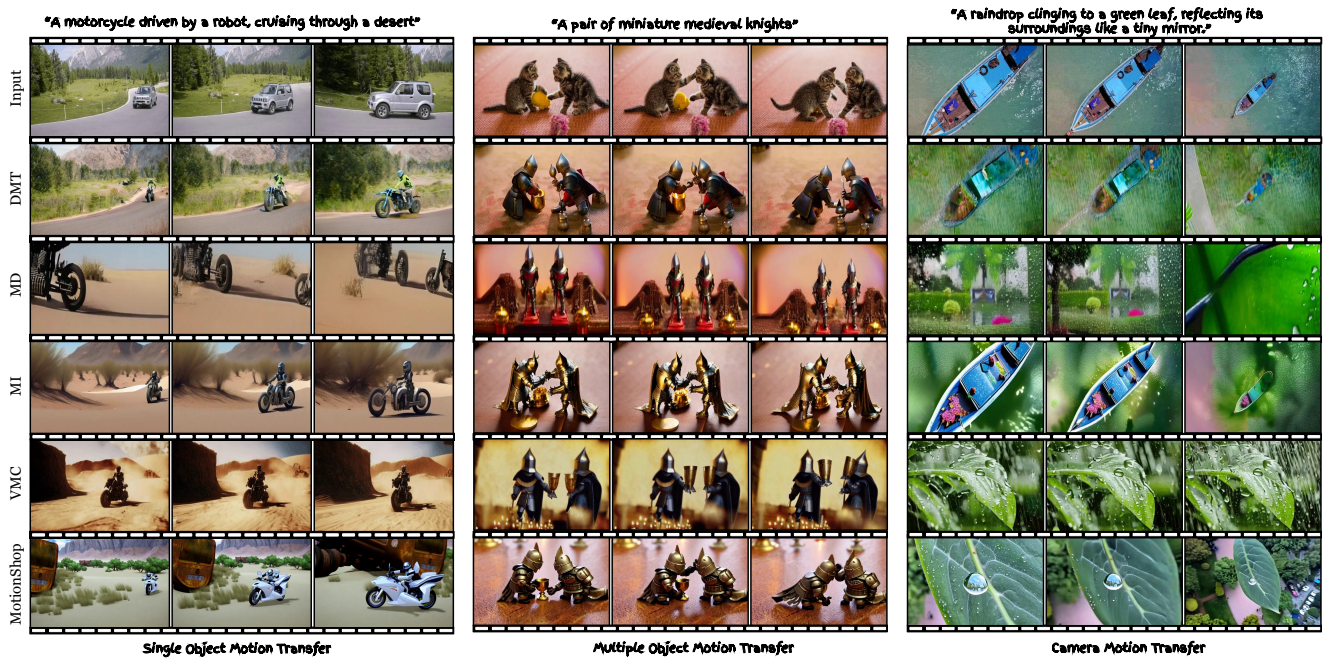

Figure 5. **Qualitative comparison of motion transfer capabilities.** We compare `MotionShop` (bottom row) with existing methods (VMC, DMT, MD, MI) on three challenging scenarios. Left: Single object motion transfer of a robot-driven motorcycle in a desert scene. Middle: Multiple object motion transfer involving miniature medieval knights, demonstrating the ability to preserve interactions between objects. Right: Camera motion transfer capturing the dynamic perspective of a raindrop on a leaf. Our method demonstrates superior motion-text alignment across all three motion transfer categories.

Method	Quantitative Metrics			User Study			
	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑	FID↓	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑
DMT [34]	0.298	0.884	0.911	**196.54**	0.21	0.19	0.20
VMC [10]	**0.328**	0.380	0.924	237.15	0.06	0.06	0.15
MD [40]	0.285	0.828	0.904	222.92	0.13	0.14	0.10
MI [30]	0.304	0.735	0.735	210.90	0.19	0.18	0.17
Ours	0.314	**0.913**	**0.928**	209.06	**0.41**	**0.43**	**0.38**

Table 1. **Comprehensive Analysis of Motion Generation Methods.** We evaluate our approach against state-of-the-art methods using both quantitative metrics (Text Similarity, Motion Fidelity, Temporal Consistency, and FID) and human evaluation. Our method achieves superior performance in most metrics, particularly showing significant improvements in user studies. Arrows (↑/↓) indicate higher/lower values are better, and best results are shown in **bold**.

camera motion transfer (Fig. 6).

8. Quantitative Experiments

In our quantitative evaluation, we compared MotionShop with MotionInversion [30], DMT [34], VMC [10], and MotionDirector [40] with 100 data-prompt pairs using four metrics: (1) *Text Similarity*, measuring frame-to-text alignment using CLIP [21], (2) *Motion Fidelity* [34], evaluating motion preservation using tracklet similarity between input and output videos, (3) *Temporal Consistency*, measuring frame-to-frame coherence via CLIP feature similarity, and (4) *FID*, assessing visual quality against DAVIS dataset. As shown in Table 1, MotionShop achieves state-of-the-art performance in both *Motion Fidelity* (0.913) and *Temporal Consistency* (0.928). While VMC shows marginally higher *Text Similarity* (0.328 vs. 0.314), our method achieves a better balance between text alignment and motion quality metrics (Table 1).

9. Discussion on Quantitative Experiments

The quantitative evaluation results presented in Table 1 demonstrate the superior performance of our approach.

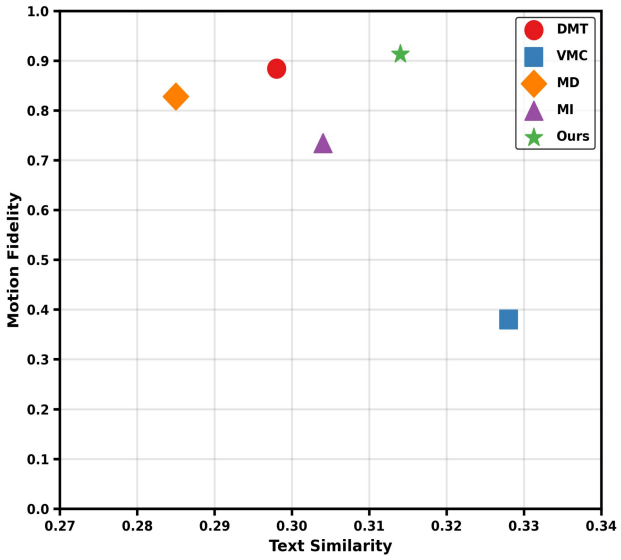

Figure 6. **Trade-off Analysis between Text Similarity and Motion Fidelity.** Comparison of our method against baselines shows superior performance in both metrics, with our approach (green star) achieving higher motion fidelity (0.913) while maintaining competitive text similarity (0.314).

Specifically, MotionShop achieves state-of-the-art performance in Motion Fidelity, surpassing the previous best method (DMT [34]) by a significant margin of 2.9%. In terms of Text Similarity metrics, our quantitative analysis reveals an interesting trade-off between text alignment and motion preservation. While VMC [10] achieves marginally higher Text Similarity scores (surpassing MotionShop by 1.4%), our experimental results indicate that this advantage comes at a significant cost to motion fidelity. In contrast, MotionShop maintains competitive text alignment capabilities (second-best among all methods) while simultaneously achieving superior motion preservation, demonstrating a more balanced approach to the inherent text-motion trade-off in video generation tasks. Analysis of the Fréchet Inception Distance (FID) reveals that our method achieves competitive performance, ranking second with a margin of *12.52* compared to DMT [34]. However, deeper examination of both Text-Similarity metrics and broader quantitative results provides critical context for these FID scores. While DMT exhibits lower FID values, this advantage appears to stem from its conservative approach to scene modification, predominantly preserving original scene layouts and compositions rather than implementing creative transformations as specified in the prompts. This characteristic leads to numerically favorable FID scores but potentially limits the method's utility for more ambitious motion transfer applications requiring significant scene modifications. Our approach, in contrast, demonstrates a more balanced capability, successfully executing substantial scene transformations while maintaining reasonable FID scores, thus offering greater practical utility for diverse motion transfer scenarios.

User Study. We conducted a user study with $N = 50$ participants on Prolific.com, evaluating 30 sets of videos. Participants assessed three metrics by selecting the top two results for each: Motion Preservation, Temporal Consistency, and Text Alignment. Results in Table 1 demonstrate MotionShop's consistent superiority across all metrics, outperforming existing approaches in motion preservation, temporal coherence, and text-guided modifications (see Appendix for more details).

9.1. Ablation Studies

We analyze three key components: motion extraction strength, guidance timestep ratio, and guidance mechanisms. Fig. 8 (left) shows the impact of motion extraction strength parameters. At 0.6, the horse's jumping motion is insufficiently transferred; at 0.8, over-stylization distorts the motion; 0.7 achieves optimal balance between motion preservation and visual quality. Fig. 8 (right) demonstrates that applying Mixture of Score guidance at different timesteps preserves generative priors, enabling diverse yet natural jumping motions.

Fig. 9 compares three guidance approaches: Classifier-Free Guidance (CFG), Unconditional Score Guidance (USG), and our Mixture of Score Guidance (MSG). CFG struggles with motion consistency, while USG better preserves motion but lacks prompt-guided precision. MSG demonstrates superior performance, evidenced by natural medieval cat motions (left) and wolf-to-pig transformations (right) while maintaining motion characteristics. This improvement stems from our novel formulation that explicitly decomposes motion and content scores, allowing for more precise control over the transfer process.

10. MotionBench Dataset

We introduce MotionBench, a comprehensive motion transfer dataset designed for systematic evaluation of motion transfer capabilities. The dataset comprises 200 carefully curated source videos and 1,000 corresponding transferred sequences, combining real-world footage from DAVIS dataset (50 videos) and high-quality synthetic videos (150 videos) generated using CogVideoX [33].

The dataset is structured around three primary motion categories, each addressing distinct challenges in motion transfer: Single Object Motion (85 videos, 42.5%), Multiple Object Motion (65 videos, 32.5%), and Camera Motion (50 videos, 25%). Single object sequences capture diverse motion patterns from rigid mechanical movements to complex articulated motions. Multiple object scenarios evaluate preservation of spatial relationships and interaction dy-

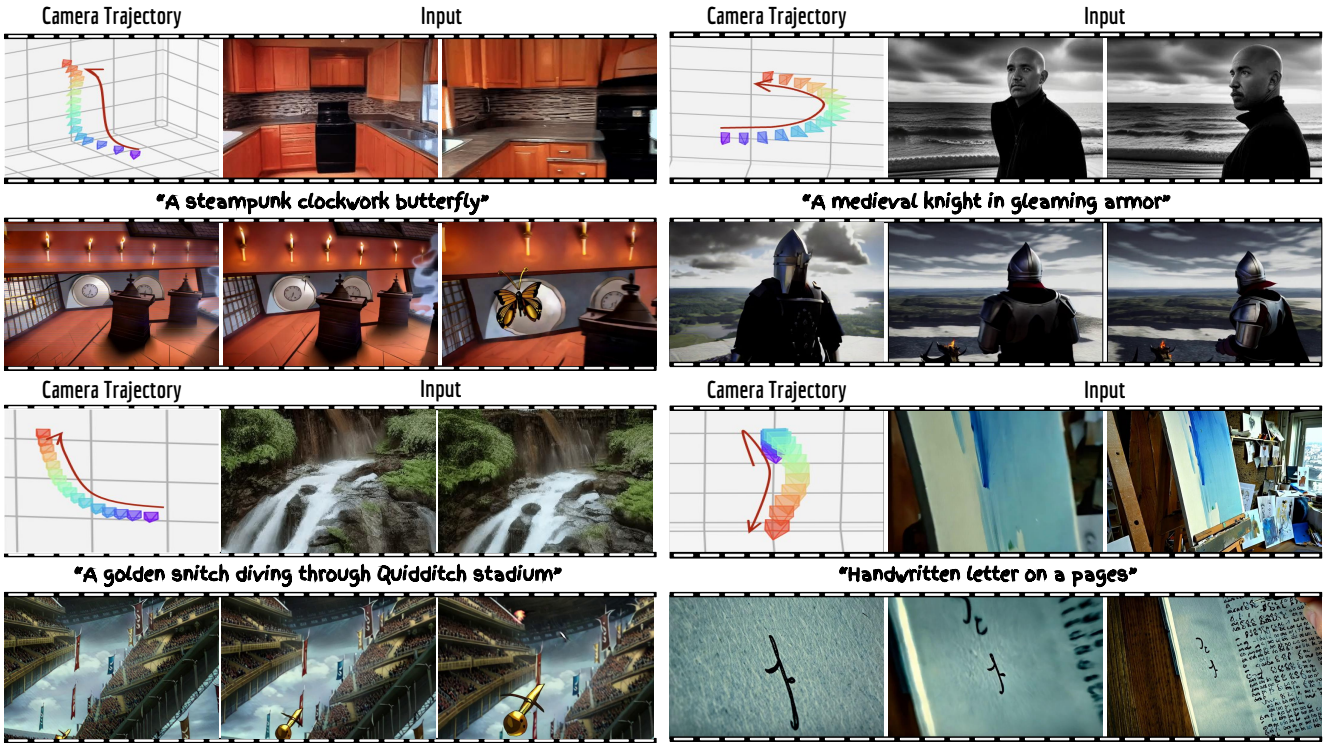

Figure 7. **Camera Motion Transfer Results Across Diverse Scenarios.** Each row shows the camera trajectory (left) and corresponding input-output image sequences. Our method can transfer camera motions while maintaining spatial consistency, as demonstrated in various cases: a steampunk clockwork butterfly animation, a raindrop on a leaf, an eagle soaring through mountain peaks, and dominos falling on a rail track. The colored trajectories represent the camera path through 3D space, with different colors indicating temporal progression.

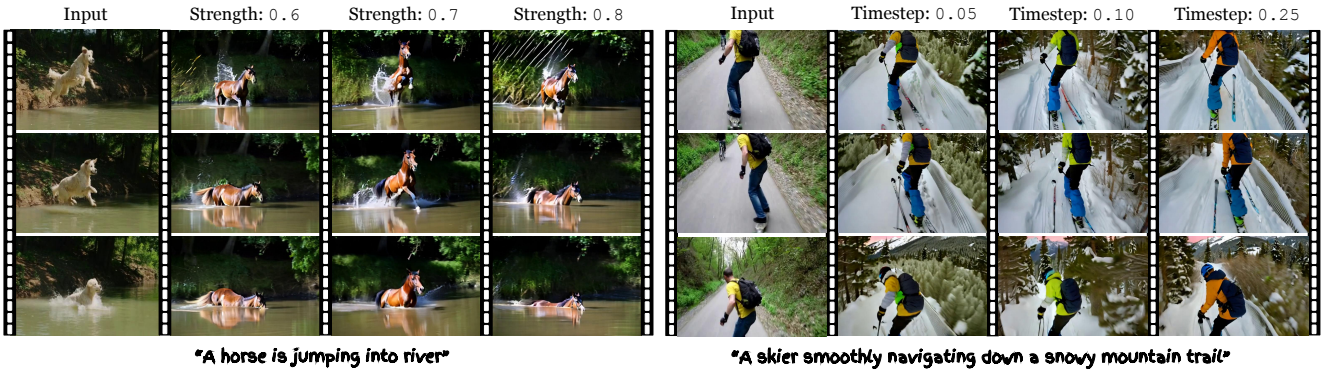

Figure 8. **Ablation study on strength and timestep parameters.** Left: We analyze the effect of noise addition in the motion extraction stage, where strength=0.7 achieves optimal motion representation - lower values (0.6) result in weak motion transfer while higher values (0.8) lead to over-stylization. Right: Impact of applying Mixture of Score guidance at different timestep ratios of total 50 timesteps on motion transfer quality.

namics between moving entities. Camera motion sequences test the handling of viewpoint changes through both simple camera operations (pan, tilt, zoom) and complex trajectories combining multiple movement types.

Each source video is paired with multiple target motion transfers, systematically exploring scenarios from straightforward object-to-object transformations to comprehensive scene-level modifications. The transfers evaluate both motion fidelity and creative adaptation capabilities, ranging from preserving precise mechanical movements to trans-

Figure 9. **Comparison of different guidance mechanisms.** Comparing our Mixture of Score Guidance (MSG) against Classifier-Free Guidance (CFG, baseline without reference) and Unconditional Score Guidance (USG, using reference video's unconditional score).

ferring organic motion patterns onto radically different targets. For example, the dataset includes challenging cases like transferring vehicle motion to flying creatures while preserving trajectory dynamics. All videos maintain consistent technical specifications (720×480 resolution) to enable standardized evaluation. We provide dataset statistics, category descriptions, and comprehensive analysis of motion transfer scenarios in the supplementary material.

11. Limitation and Societal Impact.

Our method's performance is inherently tied to the generative priors learned by the underlying T2V model. As a result, certain target concepts and motions may fall outside the model's distribution. Additionally, any biases present in the T2V model are carried over into our approach, a drawback that any zero-shot method suffers, which may influence the quality of generated outputs for specific scenarios. Since our method enables controllable video generation, there is a potential risk of it being used to create deepfake videos that spread misinformation or deceive viewers. To mitigate these risks, we emphasize the importance of ethical use of our tool.

12. Conclusion

In this paper, we presented MotionShop, the first motion transfer approach in video diffusion transformers, which reformulates conditional score to decompose motion and content scores. By treating motion transfer as a mixture of potential energies, our method enables creative scene transformations while preserving motion patterns, operating directly on pre-trained models without additional training. Extensive experiments demonstrate MSG's effectiveness across various scenarios, from single/multiple object transformations to complex camera motion transfer. Our framework provides principled guidance for balancing motion and content preservation, enabling flexible motion transfer in video generation.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[2] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 2

[3] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 2, 3

[4] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2, 3

[5] 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. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[8] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 3

[9] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 8079–8088, 2024. 3

[10] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9212–9221, 2024. 2, 3, 5, 7, 8, 1

[11] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6507–6516, 2024. 2, 3

[12] Mingxiao Li, Bo Wan, Marie-Francine Moens, and Tinne Tuytelaars. Animate your motion: Turning still images into dynamic videos. *arXiv preprint arXiv:2403.10179*, 2024. 3

[13] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 3

[14] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 3

[15] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*, 2023. 3

[16] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 3

[17] OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. [Accessed 11-11-2024]. 3

[18] Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981. 4

[19] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[20] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[22] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2402.14780*, 2024. 2, 3

[23] CP Robert. Monte carlo statistical methods, 1999. 4

[24] Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. 1996. 4

[25] Peter J Rossky, Jimmie D Doll, and Harold L Friedman. Brownian dynamics as smart monte carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978. 4

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

[27] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[28] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 3

[29] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 3

[30] Luozhou Wang, Ziyang Mai, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Yingcong Chen. Motion inversion for video customization. *arXiv preprint arXiv:2403.20193*, 2024. 2, 3, 5, 7, 1

[31] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 2, 3

[32] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 3

[33] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3, 5, 1, 4

[34] 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*, pages 8466–8476, 2024. 2, 3, 5, 7, 8, 1

[35] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[36] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. *arXiv preprint arXiv:2403.14148*, 2024. 2

[37] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: instructing video diffusion models with human feedback. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6463–6474, 2024. 3

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

[39] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[40] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2025. 2, 3, 5, 7, 1

MotionShop: Zero-Shot Motion Transfer in Video Diffusion Models with Mixture of Score Guidance

Supplementary Material

A. User Study Details

To evaluate the perceptual quality of our method, we conducted a comprehensive user study with N=50 participants recruited through Prolific.com. Following standard practices in human evaluation studies for video generation [30], we designed our study to assess three critical aspects of motion transfer quality, as illustrated in Fig. 10.

For each test case, participants were presented with an input video and five different edited versions, corresponding to various motion transfer methods. The evaluation criteria were as follows:

1. **Motion Fidelity:** Participants were asked to identify the two edited videos that best preserved the motion patterns from the input sequence. This assessment focused on the accuracy of transferred motion dynamics and spatial relationships. The question we asked is being *"Regarding the input video, which specific edits would you consider to be among the top two most successful regarding preserving original motion?"*
2. **Temporal Consistency:** Users selected the two results exhibiting the highest temporal coherence, evaluating frame-to-frame continuity and the absence of artifacts or jitter in the generated sequences. The question we asked is being *"Regarding the modified videos below, select the top 2 that have the smoothest motion."*
3. **Text-Motion Alignment:** Participants evaluated how well each generated video aligned with its corresponding text prompt, focusing on both semantic accuracy and motion appropriateness. The question we asked is being *"Which video best aligns with textual description (prompt) below."*

The study compared five different approaches: our proposed MotionShop method, Space-Time Features (DMT) [34], MotionDirector (MD) [40], MotionInversion (MI) [30], and Video Motion Customization (VMC) [10]. The user study interface, shown in Fig. 11, was designed to facilitate clear comparison and intuitive interaction.

B. MotionBench: A Comprehensive Motion Transfer Dataset

We introduce MotionBench, the first publicly available dataset specifically designed for evaluating motion transfer capabilities in video generation models. While existing video datasets primarily focus on general video synthesis or editing tasks, MotionBench addresses the critical gap in standardized evaluation of motion transfer capabilities. The dataset comprises 200 carefully curated source videos and 1,000 corresponding motion-transferred sequences, enabling systematic evaluation across diverse motion patterns and scene compositions.

B.1. Dataset Composition

B.1.1 Source Videos

The 200 source videos are curated from two primary sources:

- DAVIS Dataset (50 videos): Selected for their diverse real-world motions
- Synthetic Videos (150 videos): Generated using CogVideoX-5B [33] model.

The source videos are categorized into the following motion categories:

1. Single Object Motion (85 videos)

The Single Object Motion category constitutes the largest portion of our dataset (42.5% of source videos), carefully curated to capture the full spectrum of motion patterns observed in real-world scenarios. This category is subdivided into three distinct motion types:

Rigid Object Motion (35 videos): This subcategory focuses on objects that maintain their shape during motion, featuring vehicles (e.g., cars, motorcycles), toys (e.g., remote-controlled cars, mechanical toys), and mechanical objects (e.g., robotic arms, industrial machinery). These sequences are particularly valuable for evaluating a method's ability to preserve consistent object geometry while transferring motion patterns.

Non-rigid Object Motion (30 videos): This subset encompasses objects that undergo deformation during movement, primarily featuring animals (e.g., birds in flight, running quadrupeds) and deformable objects (e.g., cloth, fluid-like materials). These sequences present more complex challenges, requiring methods to handle both global motion and local deformations simultaneously. The videos capture various natural movements including galloping, flying, and elastic deformations.

Human Motion (20 videos): The human motion sequences capture a diverse range of articulated movements, including walking sequences, dance performances, and various sports activities. These videos are particularly challenging as they combine both rigid (skeletal) and non-rigid (soft tissue) motion patterns. The sequences test a method's capability to preserve complex kinematic chains and natural human dynamics while transferring motion to different target subjects or characters.

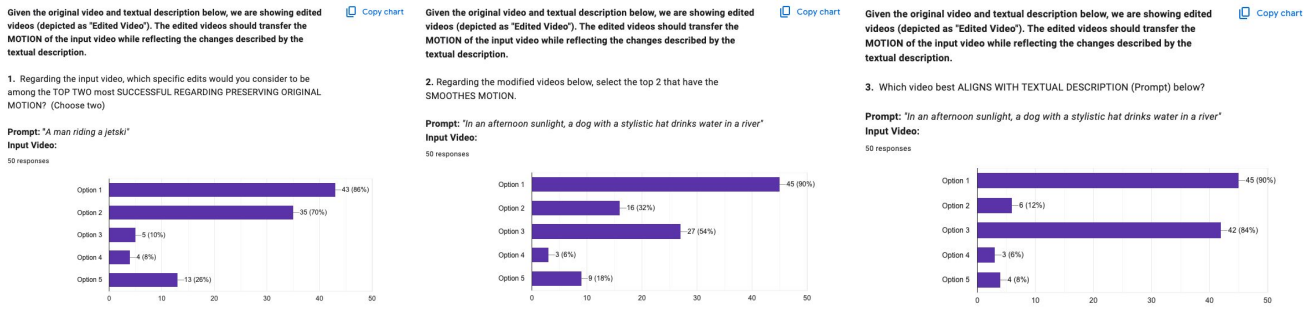

Figure 10. **Type of Questions.** We ask 3 different questions for Text Alignment, Motion Fidelity and Temporal Consistency.

Table 2. Distribution of videos across different motion categories in MotionBench. The dataset provides a balanced representation of various motion types, enabling comprehensive evaluation of motion transfer methods.

Motion Type	Source Videos	Transfer Sequences
Single Object	85 (42.5%)	400 (40%)
Multi-Object	65 (32.5%)	300 (30%)
Camera Motion	50 (25%)	300 (30%)

Each subcategory is carefully balanced to include both simple and complex motion patterns, varying speeds, and different environmental contexts. This structured approach enables systematic evaluation of motion transfer methods across a spectrum of complexity levels, from basic rigid transformations to highly articulated and deformable motion patterns.

2. Multi-Object Motion (65 videos)

The Multi-Object Motion category comprises 32.5% of our dataset, specifically designed to evaluate motion transfer capabilities in scenarios involving multiple moving entities. This category presents unique challenges in preserving spatial relationships, temporal synchronization, and complex interaction patterns. We organize these sequences into three distinct subcategories:

Interactive Motion (25 videos): These sequences capture complex interactions between multiple objects or animals, such as predator-prey chase sequences, children playing with toys, or animals engaged in social behaviors. The defining characteristic of this subset is the causal relationship between the subjects' movements, where the motion of one entity directly influences others. These videos are particularly challenging for motion transfer as they require preserving not only individual motion patterns but also the intricate timing and spatial relationships that define the interactions. Examples include dogs playing with frisbees, cats interacting with toys, and people passing objects between them.

Independent Motion (20 videos): This subcategory features scenarios where multiple objects move simultaneously but independently of each other. These sequences test a method's ability to maintain distinct motion patterns while ensuring global scene coherence. Examples include traffic scenes with multiple vehicles, scenes of birds flying in different directions, and sequences of independent mechanical systems operating simultaneously. The primary challenge lies in preserving the independence of various motion patterns while maintaining their temporal alignment and avoiding unintended interactions in the transferred results.

Group Motion (20 videos): The group motion sequences focus on coordinated movements of multiple subjects, such as synchronized dancing, flock behaviors, or team sports activities. These videos present unique challenges in maintaining both individual motion fidelity and group-level patterns. The sequences capture various forms of collective behavior, from highly structured (e.g., marching bands, synchronized swimming) to more organic patterns (e.g., school of fish, crowd movements). The key evaluation aspect is the preservation of both individual dynamics and emergent group behavior patterns during motion transfer.

Each subcategory is carefully curated to include varying levels of complexity in terms of the number of objects, spatial distribution, and temporal coordination. This structured organization enables comprehensive evaluation of how motion transfer methods handle scenarios ranging from simple multi-object scenes to complex, interdependent motion patterns, providing insights into their scalability and robustness in real-world applications.

3. Camera Motion (50 videos)

The Camera Motion category constitutes 25% of our dataset, specifically designed to evaluate motion transfer methods' capabilities in handling various camera movement patterns. This category is particularly crucial as camera motion adds an additional layer of complexity to the motion transfer task, requiring methods to maintain coherent scene composition while adapting to changing viewpoints and perspectives.

Simple Camera Movements (20 videos): This subcate-

gory encompasses fundamental camera operations that form the building blocks of cinematographic techniques. Each type presents unique challenges for motion transfer:

- **Pan (5 videos):** Horizontal camera rotations that test a method's ability to maintain consistent object appearance and motion during lateral viewpoint changes. These sequences include landscape shots, architectural surveys, and subject tracking, with varying pan speeds and ranges.
- **Tilt (5 videos):** Vertical camera rotations that challenge perspective preservation, particularly in maintaining proper scale relationships as the viewing angle changes. Examples include vertical scans of buildings, waterfalls, and ascending/descending subject movements.
- **Zoom (5 videos):** Sequences involving camera focal length changes, testing a method's capability to handle continuous scale variations while preserving motion coherence. These include both zoom-in sequences revealing fine details and zoom-out shots revealing broader context.
- **Dolly (5 videos):** Forward/backward camera translations that evaluate depth handling and parallax effects. These shots are particularly challenging as they require maintaining proper spatial relationships between foreground and background elements during motion transfer.

Complex Camera Movements (30 videos): This subcategory features more sophisticated camera work that combines multiple basic movements, presenting higher-level challenges for motion transfer systems:

- **Combined Motion Patterns (15 videos):** These sequences feature simultaneous execution of multiple camera movements (e.g., pan-with-zoom, tilt-with-dolly). They test a method's ability to handle compound camera transformations while maintaining scene coherence and motion fidelity. Examples include aerial shots with multiple degrees of freedom, elaborate reveal sequences, and complex establishing shots.
- **Dynamic Tracking Shots (15 videos):** These sequences involve camera movements that actively follow moving subjects, requiring simultaneous handling of both camera and subject motion patterns. They present particularly challenging scenarios where the camera movement must maintain a specific spatial relationship with the tracked subject while adapting to the subject's motion. Examples include sports coverage, chase sequences, and nature documentaries.

The camera motion sequences are carefully selected to include variations in speed, acceleration, and motion smoothness. Additionally, they encompass different environmental contexts (indoor/outdoor, varying lighting conditions) and subject types, providing a comprehensive evaluation framework for testing motion transfer methods' robustness to camera movement. This category is particularly valuable for assessing a method's potential in real-world applications such as cinematography, virtual production, and automated video editing.

B.2. Motion Transfer Sequences

Our dataset includes 1,000 carefully curated motion-transferred sequences, each derived from the source videos through various transformation scenarios. These sequences are specifically designed to evaluate different aspects of motion transfer capabilities, ranging from object transformations to comprehensive scene alterations. The transfers are organized into two primary categories, each addressing distinct challenges in motion transfer tasks: **1. Cross-Category Transfers**

This category evaluates a method's capability to transfer motion patterns across different object categories while maintaining motion fidelity. The sequences are divided into three distinct transfer types:

Object-to-Object: These transfers focus on motion preservation across different object categories while handling significant shape and appearance variations. Examples include:

- Vehicle-to-creature transformations (e.g., car motion applied to a mechanical horse)
- Mechanical-to-organic conversions (e.g., robot movements mapped to flowing water)
- Rigid-to-deformable translations (e.g., toy motion adapted to cloth-like objects)

These sequences test the ability to maintain motion characteristics despite fundamental changes in object properties and physical constraints.

Human-to-Character: This subset specifically addresses the challenging task of transferring human motion to non-human characters while preserving natural movement patterns. Examples include:

- Human dance movements applied to animated characters
- Sports motions transferred to fantasy creatures
- Gesture sequences mapped to mechanical entities

These transfers test the preservation of complex articulated motion while adapting to different skeletal structures and movement constraints.

Animal-to-Object: These sequences evaluate the transfer of organic motion patterns to inorganic objects, presenting unique challenges in motion adaptation. Examples include:

- Bird flight patterns applied to flying vehicles
- Quadruped locomotion mapped to mechanical assemblies

2. Scene Transformation Transfers

This category focuses on evaluating motion preservation within dramatically altered environmental contexts, addressing two key aspects:

Environment Changes: These transfers test the ability to maintain motion fidelity while completely transforming the surrounding environment. The sequences include:

- Context shifts (e.g., street scene to underwater environment)

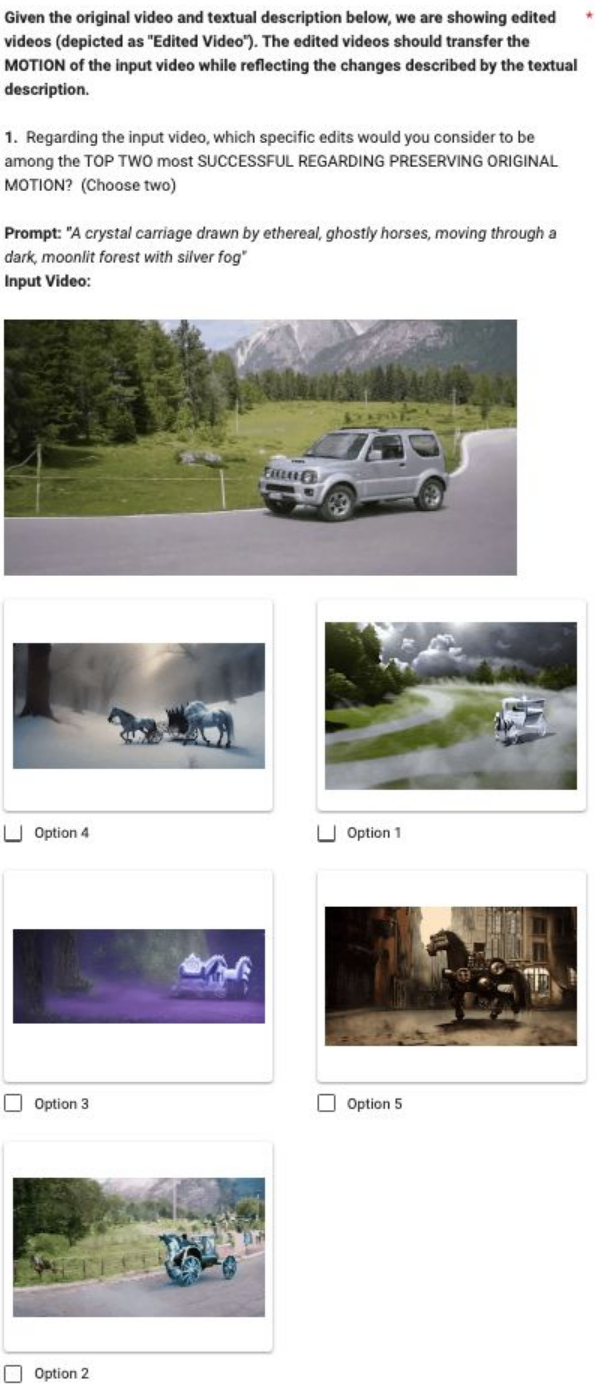

Figure 11. **User Study Interface.** Given a reference video we ask for 3 different type of questions with 5 different options including DMT, MI, MD, VMC and MotionShop results.

- Scale transformations (e.g., human-scale to miniature worlds)
- Physical domain changes (e.g., terrestrial to aerial scenarios)

These sequences evaluate how well methods handle motion transfer when environmental physics and constraints change significantly.

Style Transfers: This subset focuses on artistic and stylistic transformations while maintaining motion integrity. Examples include:

- Realistic to animated style conversions
- Contemporary to historical aesthetic adaptations
- Natural to fantastical scene transformations

Each transfer category is carefully designed to test specific aspects of motion transfer capabilities, from basic motion preservation to complex scene-level transformations. The sequences vary in complexity, duration, and transformation extent, providing a comprehensive evaluation framework for assessing motion transfer methods across different scenarios and applications. This structured approach enables systematic analysis of a method's strengths and limitations in handling various types of motion transfer challenges.

B.3. Dataset Statistics

Key characteristics of the dataset:

- Resolution: 720×480 pixels
- Frame Rate: 15 FPS
- Duration: 1-7 seconds per video (Due to the frame processing limitation of CogVideoX)
- Total Frames: ∼45,000
- Format: MP4 (H.264 codec)

Here we note that the duration limit stems from the CogVideoX [33] backbone. It can only process 49 frames at most.

B.4. Discussion

MotionBench addresses several critical requirements essential for comprehensive motion transfer evaluation. First, it achieves comprehensiveness by encompassing a diverse range of motion types and scene compositions, ensuring broad coverage of real-world scenarios. Its scalable design provides sufficient data for meaningful model training and evaluation, while maintaining standardized evaluation protocols and metrics that enable consistent comparisons across different approaches. Furthermore, the dataset's inclusion of both synthetic and real-world scenarios ensures practical applicability across various use cases. Through these carefully considered design choices, MotionBench enables researchers to systematically compare motion transfer methods, analyze motion preservation capabilities, evaluate scene composition handling, and assess temporal consistency.