Time-to-Move: Training-Free Motion Controlled Video Generation via Dual-Clock Denoising

Anonymous CVPR submission
Paper ID 0000

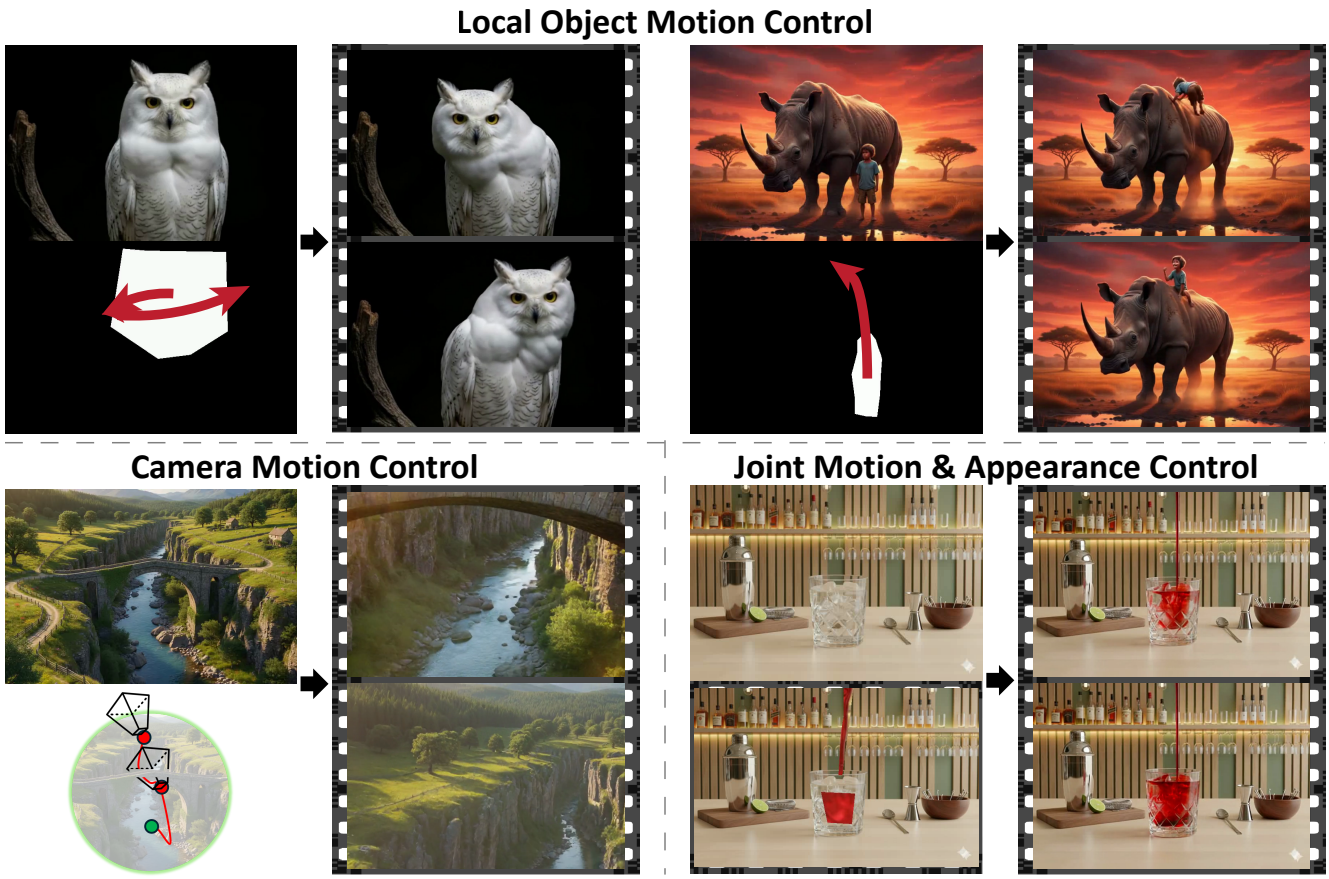

Figure 1. Qualitative results of Time-to-Move on various tasks.

Abstract

Diffusion-based video generation can create realistic videos, yet existing image- and text-based conditioning fails to offer precise motion control. Prior methods for motion-conditioned synthesis typically require model-specific fine-tuning, which is computationally expensive and restrictive. We introduce Time-to-Move (TTM), a training-free, plug-and-play framework for motion- and appearance-controlled video generation with image-to-video (I2V) diffusion models. Our key insight is to use crude reference animations obtained through user-friendly manipulations such as cut-and-drag or depth-based reprojection. Motivated by SDEdit*'s use of coarse layout cues for image editing, we treat the crude animations as coarse motion cues and adapt the mechanism to the video domain. We preserve appearance with image conditioning and introduce* dual-clock denoising*, a region-dependent strategy that enforces strong alignment in motion-specified regions while allowing flexibility elsewhere, balancing fidelity to user intent with natural dynamics. This lightweight modification of the sampling process incurs no additional training or runtime cost and is compatible with any backbone. Extensive experiments on object and camera motion benchmarks show that TTM matches or exceeds existing training-based baselines in realism and motion control. Beyond this, TTM introduces a unique capability: precise appearance control through pixel-level conditioning, exceeding the limits of text-only prompting.*

1. Introduction

Recent image-to-video (I2V) diffusion models produce impressive visual quality, but prompt-only control remains too coarse for production-facing tasks such as shot design, object blocking, and appearance-directed storyboarding. In cinematic workflows, users need to specify *what* moves, *where* it moves, and *what stays consistent* as motion unfolds. Existing controls often fail to deliver this precision and frequently introduce drift, static backgrounds, or identity loss.

Most strong motion-control approaches inject trajectories, flow, or tracks through dedicated modules and substantial fine-tuning [1, 2, 18]. While effective, these methods are expensive to train, tightly coupled to a specific backbone, and difficult to transfer across rapidly evolving I2V models. For workshop settings centered on movie-grade prototyping, this limits practical use by creators who need fast adaptation rather than model retraining.

A practical assistant for AI filmmaking should support local edits (moving an actor or prop), global edits (camera trajectory changes), and style-aware evolution (controlled appearance changes) under one unified inference process. It should also preserve temporal plausibility in unconstrained regions so that secondary motion remains coherent with the primary edit. This requirement is particularly important in previsualization and iterative directing, where artists repeatedly adjust motion intent while expecting stable visual identity.

We introduce Time-to-Move (TTM), a training-free, architecture-agnostic inference method for controllable I2V generation. Our pipeline starts from a crude user-defined reference animation (e.g., cut-and-drag or depth reprojection), then uses an SDEdit-inspired initialization to inject motion while first-frame image conditioning preserves identity and scene appearance. The key novelty is *region-dependent dual-clock denoising*: masked regions receive stronger motion enforcement, while unmasked regions denoise with weaker constraints to maintain natural scene dynamics. This design directly addresses a core failure mode of single-clock sampling, which cannot simultaneously preserve strict local control and global realism.

Related methods span trajectory-conditioned training-based controllers [5, 14–17, 19], geometry/noise-based camera control [1, 11], and training-free attention manipulations in T2V [4, 7, 10]. Our method differs by enforcing region-specific control directly at inference in pretrained I2V models, combining motion fidelity with appearance consistency and supporting edits relevant to cinematic previsualization (Fig. 3).

In summary, our contributions are:

- **Training-free controllable motion from crude references.** We convert simple user guidance into realistic motion without architecture-specific retraining.
- **Dual-clock denoising for regional control.** We introduce a region-dependent schedule that balances strict local adherence with globally coherent dynamics.
- **Unified motion and appearance editing.** By conditioning on full reference frames, TTM supports joint motion and visual-attribute control beyond trajectory-only interfaces.

2. Method

Given an input image and a user motion intent, our goal is to generate a realistic video that follows the intended motion while preserving identity and scene coherence. TTM combines three components: (1) a crude reference animation as explicit motion guidance, (2) image-conditioned sampling for appearance anchoring, and (3) dual-clock denoising for region-dependent control strength. Conceptually, the method treats user edits as structural guidance rather than as final appearance targets: the reference only specifies coarse motion and placement, while the diffusion model reconstructs realistic details under image conditioning. This makes the approach robust to low-quality warps while remaining faithful to intended trajectories.

Problem Formulation.

Inputs are: an image $I \in \mathbb{R}^{3\times H\times W}$, a coarse warped reference video $V^w \in \mathbb{R}^{F\times 3\times H\times W}$, and a binary mask video $M \in \{0,1\}^{F\times H\times W}$ marking regions with explicit user guidance. The objective is a video x_0 that follows the prescribed motion in masked regions while remaining globally plausible and faithful to the source appearance.

2.1. Motion Signal

We represent user intent as a coarse animation V^w. In object-motion control, this is produced by cut-and-drag style warping of selected regions across time; in camera-motion control, it is produced via depth-based reprojection. Although crude and artifact-prone, these references encode the desired trajectory and spatial-temporal layout, making them effective control carriers during diffusion sampling. Forward warping naturally introduces tearing and disocclusions, but these artifacts are acceptable because the reference serves as a control signal rather than a photorealistic target. The key requirement is geometric alignment with user intent. In practice, this representation is flexible: users can define motion through trajectories, geometric transforms, or depth-driven camera changes without retraining the generator.

2.2. SDEdit Adaptation for Motion Injection

Following SDEdit intuition [8], we initialize denoising from a noisy version of the reference so coarse dynamics are injected early [12]. To avoid identity drift, we use an *image-conditioned* I2V model anchored to the clean first frame I,

CVPR
#0000

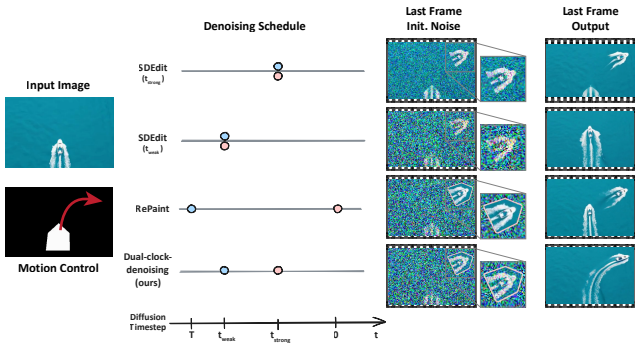

Figure 2. **Single-clock vs. dual-clock denoising.** A single timestep cannot jointly enforce strict local motion and natural global dynamics. Dual-clock denoising applies stronger guidance in masked regions and weaker guidance elsewhere, reducing drift and artifacting.

allowing motion transfer from V^w while preserving appearance. This design separates responsibilities: the noisy reference carries motion layout, and the image condition preserves content identity. Empirically, this is crucial for cinematic scenes where subject consistency and background continuity must be maintained under strong local edits.

2.3. Region-Dependent Dual-Clock Denoising

Single-clock noising creates a hard trade-off: low noise overconstrains the scene (e.g., frozen backgrounds), while high noise weakens motion adherence. We therefore apply different effective schedules across regions. Let $t_{\text{weak}} > t_{\text{strong}}$. We initialize from V^w noised to t_{weak}. For steps $t_{\text{strong}} \leq t < t_{\text{weak}}$, we override masked regions using the reference at the matching noise level, while unmasked regions follow standard denoising. After $t = t_{\text{strong}}$, we stop overriding and denoise jointly to the final sample. Intuitively, masked regions need strong guidance because they encode explicit user intent, whereas unmasked regions should retain flexibility to accommodate physically plausible secondary motion. A single global timestep cannot satisfy both goals simultaneously. Dual-clock denoising enforces this asymmetry directly at inference time, avoiding model changes or additional training.

Let x_t be the current sample and $\hat{x}_{t-1}(x_t, t, I)$ the denoiser prediction. The dual-clock update is:

$$x_{t-1} \leftarrow (1 - M) \odot \hat{x}_{t-1}(x_t, t, I) + M \odot x^w_{t-1},$$

where x^w_{t-1} is the warped reference noised to step $t - 1$. This yields strong local trajectory control while preserving coherent unconstrained motion and appearance. Operationally, this update can be viewed as a controlled blending between model prediction and reference guidance. Early steps prioritize control where requested; later steps prioritize global harmonization. This staged behavior is what allows TTM to handle both local object manipulation and camera-style motion control within the same framework.

Efficiency and Applicability.

The method modifies only sampling logic, introduces no extra training, and transfers across different I2V backbones.

3. Experiments

We evaluate TTM in three settings: object motion control (Sec. 3.1), camera motion control (Sec. 3.2), and joint motion-appearance control (Sec. 3.3). Unless noted, we follow official baseline settings and each backbone's native protocol (SVD: 16-frame setup; CogVideoX: 49-frame setup), with a fixed dual-clock schedule per backbone. Extensive ablations (omitted for space) confirm that region-dependent dual-clock denoising is necessary to avoid temporal artifacts while preserving motion adherence. We emphasize evidence most relevant to production use: controllability under large edits, consistency of subject identity, and temporal plausibility in unconstrained regions.

3.1. Object Motion Control

We benchmark on MC-Bench [18] using user-provided masks and trajectories. We compare against DragAnything [15], MotionPro [18], SG-I2V [9], and GWTF [1]. Evaluation uses trajectory adherence (CTD), object-background disentanglement (BG–Obj CTD), and reference-free perceptual metrics from VBench [3].

As summarized in Tab. 1, TTM improves trajectory adherence across both SVD and CogVideoX settings relative to strong training-free alternatives, while remaining competitive with training-based systems on perceptual metrics. Qualitatively, it better preserves first-frame identity under large displacements and reduces common failures such as background co-motion and geometric distortions. These trends support the core design goal: strong local control without retraining-induced specialization. Compared to single-clock behavior, we observe more stable object placement and less temporal freezing in surrounding regions, particularly on long or curved trajectories. Relative to GWTF variants, TTM also improves robustness when trajectory-induced layout changes are large, where noise-warping approaches often drift or deform scene structure.

3.2. Camera Motion Control

Following prior work [1, 6], we evaluate controlled camera trajectories on a DL3DV subset with depth-based reprojection references. Metrics include frame fidelity (MSE, LPIPS, SSIM), flow alignment, FID, and temporal consistency.

As shown in Tab. 2, TTM yields the strongest overall trade-off between motion fidelity and visual quality, with lower pixel/flow errors and better distributional alignment than GWTF variants. In practice, dual-clock guidance better preserves the intended camera path while avoiding drift and tearing artifacts from pure warping. Quantitatively, we

Method	Training Free?	CTD$_{\downarrow}$	BG–Obj CTD$_{\uparrow}$	Dynamic Degree$_{\uparrow}$	Subject Consistency$_{\uparrow}$	Background Consistency$_{\uparrow}$	Motion Smoothness$_{\uparrow}$	Aesthetic Quality$_{\uparrow}$	Imaging Quality$_{\uparrow}$
SVD-Based Models									
DragAnything	✗	10.645	**50.885**	**0.981**	0.956	0.942	0.983	0.531	0.554
SG-I2V*	✓	**5.796**	12.042	0.803	0.976	0.953	0.991	0.553	0.621
MotionPro	✗	8.685	24.485	0.422	**0.979**	**0.975**	**0.993**	**0.559**	**0.617**
Ours	✓	7.967	35.340	0.427	**0.979**	0.967	**0.993**	0.548	**0.617**
CogVideoX-Based Models with Longer Generated Videos									
GWTF$_{\gamma=0.7}$	✗	32.548	86.614	0.736	0.963	0.965	0.989	0.517	0.539
GWTF$_{\gamma=0.5}$	✗	27.844	**87.708**	**0.764**	0.958	0.963	0.988	0.513	0.539
Ours	✓	**13.665**	70.608	0.357	**0.980**	**0.977**	**0.995**	**0.531**	**0.579**

Table 1. **Quantitative results on MC-Bench object motion control.**

Method	MSE $\downarrow$	FID$_{\downarrow}$	LPIPS$_{\downarrow}$	SSIM$_{\uparrow}$	CLIP Cons.$_{\uparrow}$	Optical flow$_{\downarrow}$
GWTF$_{\gamma=0.5}$	0.033	25.990	0.371	0.526	0.981	76.714
GWTF$_{\gamma=0.7}$	0.042	28.483	0.370	0.410	**0.985**	81.738
Warped	0.025	33.443	0.339	0.560	0.981	65.494
Ours	**0.022**	**21.966**	**0.332**	**0.586**	0.983	**60.558**

Table 2. **Quantitative results on DL3DV camera motion control.**

Figure 3. **Joint motion and appearance control.** TTM uses a user-provided warped reference to jointly control dynamics and per-pixel appearance in diverse cinematic editing scenarios.

observe substantial improvements over the strongest GWTF setting, including notable gains in pixel-level fidelity and FID, alongside better optical-flow consistency. These gains align with qualitative behavior: TTM follows target camera motion more faithfully while preserving scene realism over time.

3.3. Appearance Control

Beyond motion, full-frame reference conditioning enables appearance-directed control relevant to creative production tasks: controlled color evolution, object insertion with scene-consistent blending, and motion-shape coupling. These edits are difficult to realize with trajectory-only conditioning. For production-style workflows, this is important because motion and look are often co-designed. In our examples, users can manipulate both trajectory and per-pixel attributes in a single inference pass, enabling iterative art-direction loops without retraining.

3.4. Plug-and-Play Model Adaptation

TTM transfers without retraining to multiple I2V backbones, including SVD, CogVideoX, and WAN 2.2 [13]. This backbone-agnostic behavior is important for production workflows where models evolve rapidly and retraining costs are prohibitive. This portability is central for workshop scenarios where teams prototype across heterogeneous model stacks and need a consistent control interface independent of backbone internals.

4. Conclusions, Limitations, Future Directions

We presented TTM, a training-free framework for controllable I2V generation that combines crude reference animation with region-dependent dual-clock denoising. The method improves motion adherence while maintaining globally coherent dynamics and first-frame appearance, and transfers across multiple backbones without retraining. This combination directly targets practical creative workflows: users can edit local motion intent while preserving global scene plausibility, and can layer appearance changes with motion control in a unified inference procedure.

Limitations remain: the dual-clock parameters still require per-backbone tuning, identity anchoring is strongest for content visible in the first frame, and high-quality control currently assumes reasonably accurate masks. Future work includes multi-region and soft-mask scheduling, stronger long-horizon temporal coherence, and tighter integration with film-production interfaces for iterative directing and storyboarding.

References

[1] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped

noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13–23, 2025. 2, 3

[2] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1–12, 2025. 2

[3] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 3

[4] 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. 2

[5] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. Image conductor: Precision control for interactive video synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5031–5038, 2025. 2

[6] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 3

[7] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[8] 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. *arXiv preprint arXiv:2108.01073*, 2021. 2

[9] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. In *International Conference on Learning Representations*, 2025. 3

[10] Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024. 2

[11] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6121–6132, 2025. 2

[12] Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. Flowmo: Variance-based flow guidance for coherent motion in video generation. *arXiv preprint arXiv:2506.01144*, 2025. 2

[13] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4

[14] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025. 2

[15] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 3

[16] 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.

[17] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025. 2

[18] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. Motionpro: A precise motion controller for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27957–27967, 2025. 2, 3

[19] Haitao Zhou, Chuang Wang, Rui Nie, Jinlin Liu, Dongdong Yu, Qian Yu, and Changhu Wang. Trackgo: A flexible and efficient method for controllable video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10743–10751, 2025. 2