EchoMotion: Unified Human Video and Motion Generation via Dual-Modality Diffusion Transformer

Yuxiao Yang[1,2], Hualian Sheng[2], Sijia Cai[2,*], Jing Lin[3], Jiahao Wang[4], Bing Deng[2], Junzhe Lu[1], Haoqian Wang[1,†], Jieping Ye[2,†]

[1]Tsinghua University [2]Alibaba Group [3]Nanyang Technological University [4]Xi'an Jiaotong University

*Project Lead †Corresponding Author

Overview

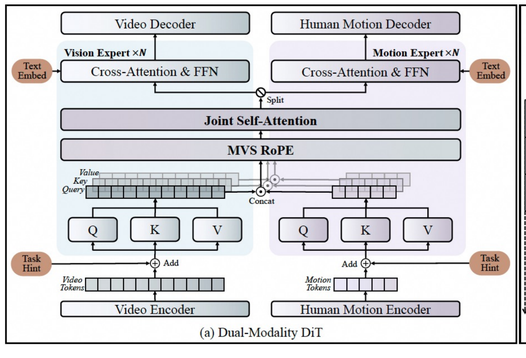
Why do state-of-the-art video generation models often fail on complex human actions?

While recent models excel at generating visually stunning scenes, they frequently struggle to synthesize plausible and coherent human movements. The reason lies in their training objective: by **focusing solely on pixel-level fidelity**, these models learn to mimic appearances but fail to grasp the underlying kinematic principles of human articulation. This leads to artifacts like floating feet and unnatural limb movements.

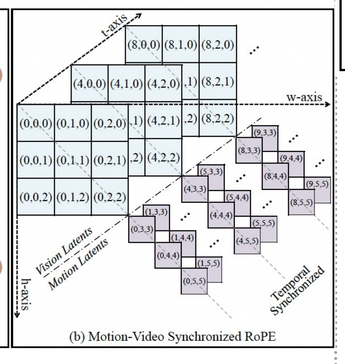
To overcome this, we introduce **EchoMotion**, a new framework that fundamentally changes the learning paradigm. Instead of treating video as just a sequence of pixels, we propose to **jointly model appearance and the explicit human motion that drives it**. Our core idea is that by providing the model with a clear understanding of kinematic laws, we can significantly improve the coherence and realism of generated human-centric videos. EchoMotion is designed to learn the joint distribution of what we see (appearance) and how it moves (motion).

We also propose a **Motion-Video Two-Stage Training Strategy** and **In-Context Classifier-Free Guidance**. These strategy enables the model to perform both the joint generation of complex human action videos and their corresponding motion sequences, as well as versatile cross-modal conditional generation tasks.

Method



(a) Dual-Modality DiT



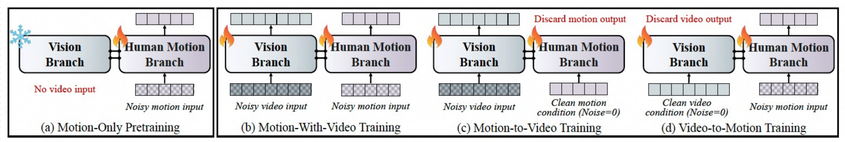
(b) Motion-Video Synchronized RoPE

Dual-Stream Architecture with Motion Latents

We first encode raw SMPL motion data into a compact latent representation. These motion tokens are then **concatenated with visual tokens from the video frames at sequence-level**. Our dual-stream DiT processes this combined sequence, enabling deep fusion of appearance and kinematic information within its attention layers.

MVS-RoPE (Motion-Video Synchronized Positional Encoding)

We introduce MVS-RoPE, a unified 3D positional encoding scheme. It provides a **shared coordinate system for both video and motion tokens**, creating a powerful inductive bias that encourages temporal alignment between the two modalities. This ensures that the generated motion perfectly syncs with the visual output.



Motion-Video Two-Stage Training Strategy

Phase 1: Motion-only Pretraining. The motion branch is trained independently using motion-only datasets, while the video branch is frozen and deactivated (inputs omitted). This stage focuses on generating motion sequences.

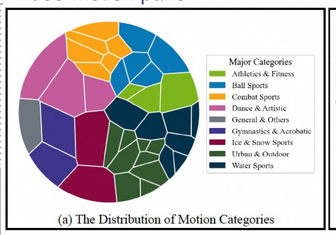
Phase 2: Motion-video Multi-task Training. Subsequently, the model is trained on motion-video paired datasets with both branches unfrozen and active, enabling the generation of both visual and motion sequences.

In-Context Classifier-Free Guidance

During Phase 2, we apply a paradigm-specific conditional dropping strategy. (b) For joint generation, the text condition is randomly dropped. (c) For motion-to-video generation, both text and motion conditions are randomly dropped. (d) For video-to-motion generation, the text condition is always dropped, while the video condition is dropped randomly. This training strategy enables cross-modal conditional generation tasks by **replacing the noisy latents with clean conditional latents** during inference.

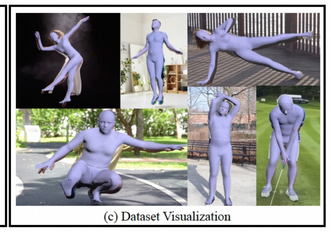
The HuMoVe Dataset

We constructed **HuMoVe**, the first dataset of its kind, containing approximately 80,000 video-motion pairs.



(a) The Distribution of Motion Categories

(b) Word Cloud Caption



(c) Dataset Visualization

Coverage: HuMoVe spans a diverse range of human activities, from daily actions and sports to complex dance performances, ensuring our model learns a robust and generalizable representation of human motion.

High-Quality Annotations: Each video is accompanied by both a detailed textual description and a precise SMPL motion sequence, extracted using state-of-the-art motion capture techniques.

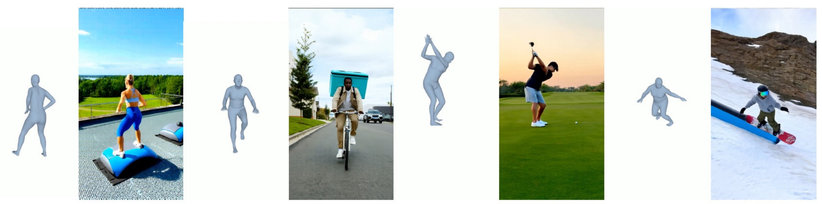
High-Fidelity Videos: We meticulously curated high-resolution, clean video clips, free from major occlusions or distracting backgrounds, providing an ideal training ground for generation models.

Result

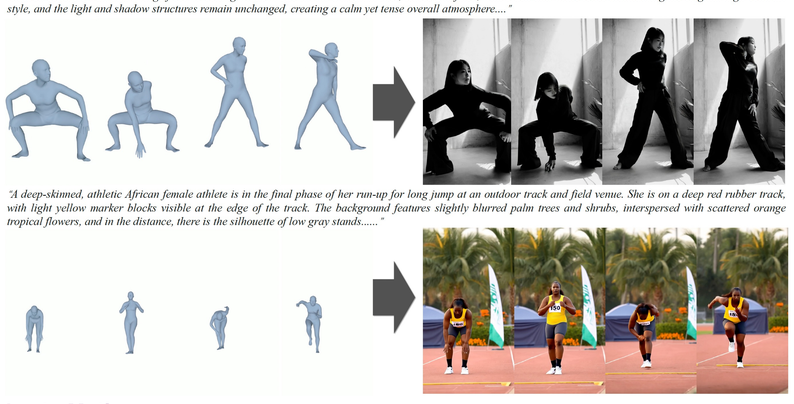
Text to Joint Video-and-Motion Generation

	Auto Metrics				*Human Eval*		
	Human Anatomy	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Video Quality	Prompt Following	Posture Plausibility
CogVideoX-2B	61.7	97.0	49.4	51.6	55.3	52.1	53.6
Wan-1.3B	78.1	98.2	60.6	**60.1**	68.2	70.3	64.0
Video Tuning(Wan-1.3B)	77.4	98.3	61.6	59.7	69.3	73.2	65.5
EchoMotion(Wan-1.3B)	79.6	98.9	61.9	60.0	71.3	73.2	66.1
CogVideoX1.5-5B	65.3	98.5	54.4	53.2	62.5	60.4	59.4
Wan-5B	83.0	98.9	62.2	58.3	72.8	78.9	68.9
Video Tuning(Wan-5B)	83.1	98.7	63.1	57.9	72.3	79.6	70.2
EchoMotion(Wan-5B)	**85.1**	**99.3**	**64.0**	58.3	**81.0**	**81.5**	**81.6**

EchoMotion jointly generates an SMPL motion sequence (left) and video (right), demonstrating a learned joint distribution:



Motion to Video Generation



Video to Motion