ICCV HONOLULU HAWAII OCT 19-23, 2025
Workshop on Reliable and Interactable World Models

Where Is Motion From? Scalable Motion Attribution for Video Generation Models

Xindi Wu[1, 2], Despoina Paschalidou[1], Jun Gao[1], Antonio Torralba[3], Laura Leal-Taixé[1], Olga Russakovsky[2], Sanja Fidler[1], Jonathan Lorraine[1]
[1]NVIDIA, [2]Princeton University, [3]MIT CSAIL

NVIDIA

Motivation

Despite rapid progress in video generation, how data shapes motion quality remains **poorly understood**.

Key Goals

Focus on Motion
Separate motion from static appearance

Scale Efficiently
Modern, large-scale models & datasets

Guide Curation
Identify clips that improve motion quality

Our Solution: MOTIVE

***MO**tion **T**raining **I**nfluence for **V**ideo g**E**neration*

Problem Formulation

Given a query video and finetuning dataset, assign each training clip a **motion-aware influence score** to quantify its contribution to target generation.

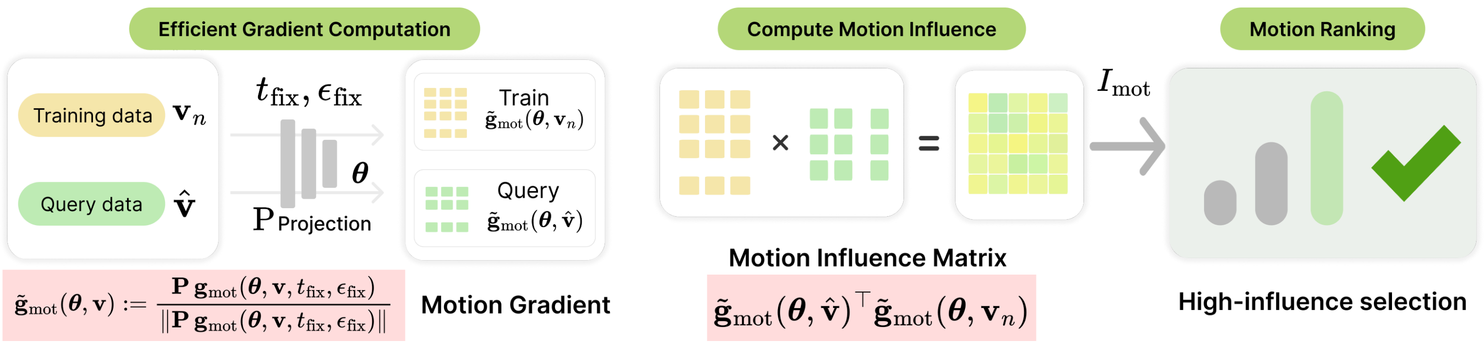
Method Components

1. Efficient Motion Gradient Computation
- Single-Sample Estimator
- Structured Projections (Fastfood)

2. Motion Attribution
- Detect motion between frames w. AllTracker
- Create motion magnitude patches highlighting dynamic areas
- Apply motion-weighted loss to focus on moving regions and compute motion-specific gradients

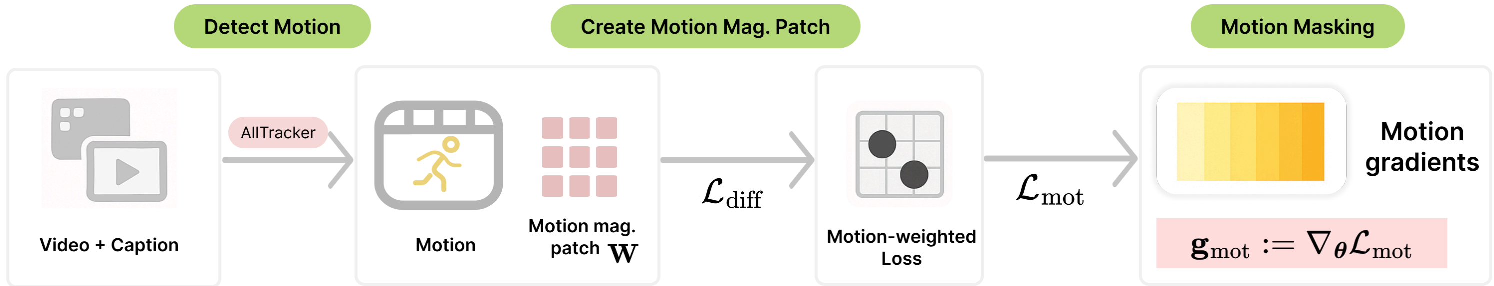
Which training clips drive the motion in a video generation sample?

Efficient Motion Gradient Computation



Our method is made scalable via a single-sample variant with common randomness and a projection, computed for each pair of training and query data, aggregated for a final ranking, and eventually used to select finetuning subsets.

Motion Attribution



Motion-gradient computation has three steps: (1) detect motion with AllTracker; (2) compute motion-magnitude patches; (3) apply loss-space motion masks to focus gradients on dynamic regions.

MOTIVE: A scalable, gradient-based, motion-centric data attribution framework for video generation models

Three Key Components

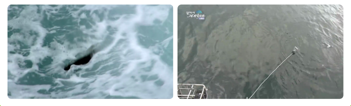
① Scalable Gradients	② Motion Attribution	③ Data Curation
Single-sample estimator + projection	Motion-weighted loss	Top-K Majority voting

Motion Attribution Samples

Query

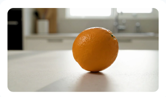
Float

✓ Top Positive High Similarity



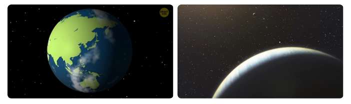
✕ Negative Conflicting

Query

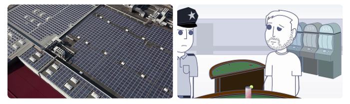


Roll

✓ Top Positive High Similarity



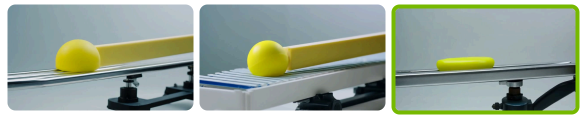
✕ Negative Conflicting



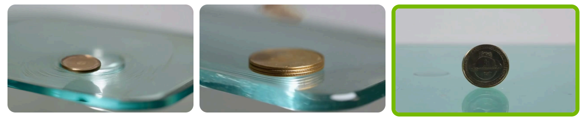
Qualitative Results

Generated Videos

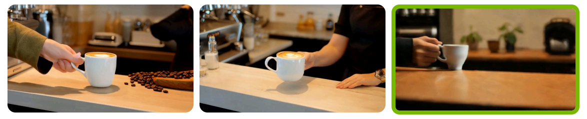
Base | Random | Ours ✓



Compress



Spin



Slide

Free Fall

Quantitative Results

VBench Evaluation

Method	Motion Smooth.	Dynamic Deg.
Base	96.3	82.3
Full FT	96.3	84.7
Random 10%	96.3	81.6
Ours w/o mask	96.3	85.3
MOTIVE	**96.3**	**89.4**

✓ Maintains smoothness, improves dynamics with only 10% data

Why Motion Masking?

Without: 85.3%

With: 89.4% (+4.1%)

Human Evaluation

vs. Base: 76.7% win
vs. Random: 66.7% win
vs. Full FT: 57.5% win

Ablation Findings

Single Timestep: t=500 achieves 68% agreement.

Projection: D'=512 reaches 74.7% Spearman ρ.

Conclusion

First motion-centric attribution framework for video generation

Scalable via **projection & majority voting**

76.7% human preference vs. baseline with **10% data**; Motion masking: **+4.1%** Dynamic Degree