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

Motion Attribution for Video Generation

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

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

MOTIon attribution for VIdeo gEneration

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.

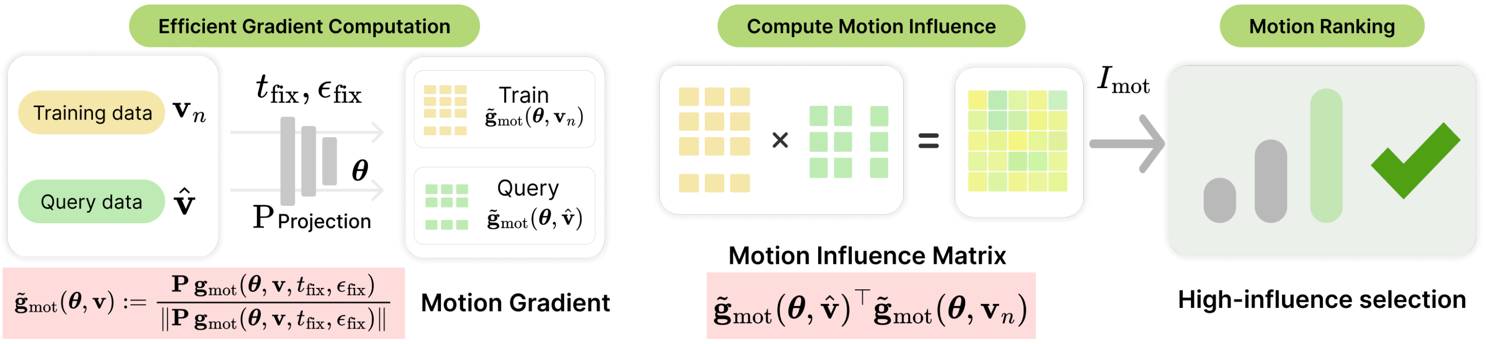
Method Components

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

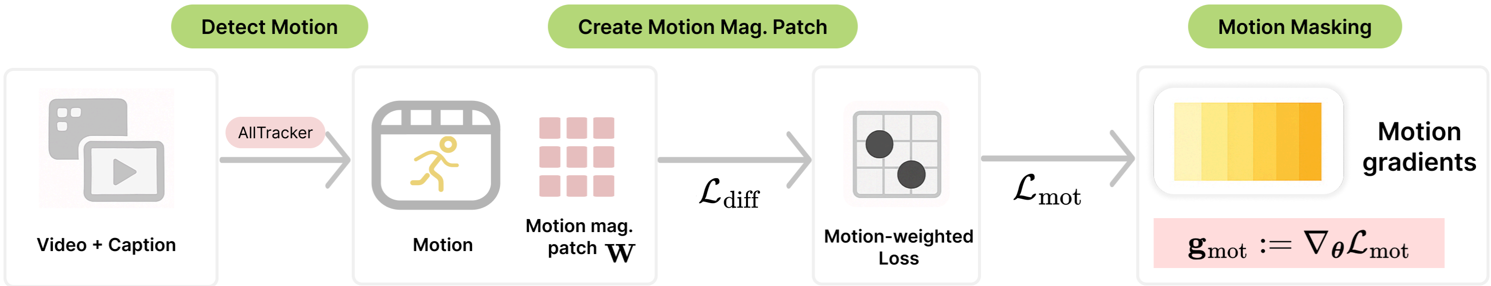
Which training clips drive the motion in a video generation sample?

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



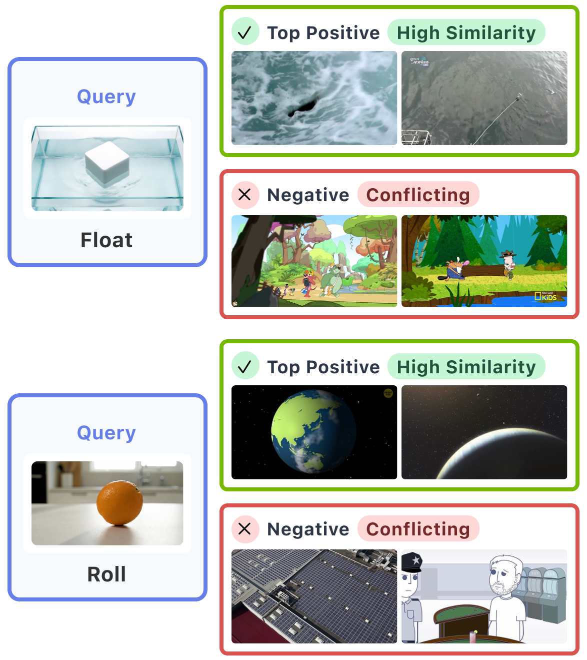
Motion-gradient computation has three steps: (1) detect motion with AllTracker; (2) compute motion-magnitude patches; (3) apply loss-space motion masks to focus gradients on dynamic regions.

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

Three Key Components

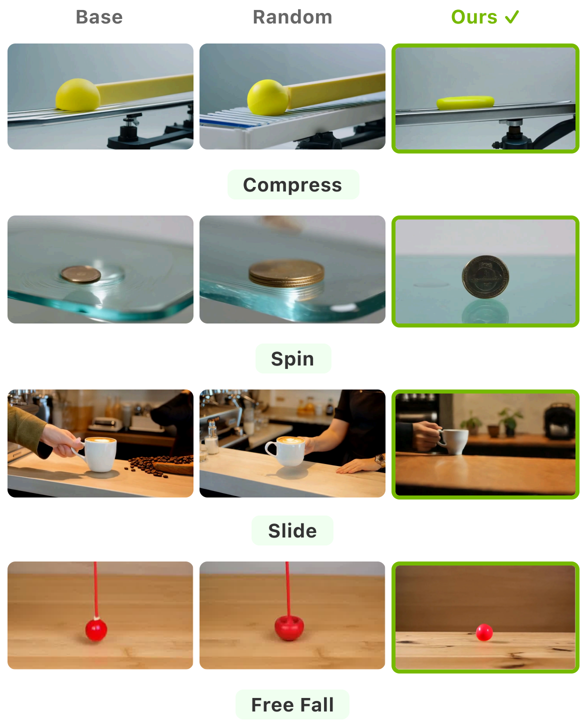
① Scalable Gradients	② Motion Attribution	③ Data Curation
Single-sample estimator + projection	Motion-weighted loss	Top-K Majority voting

Motion Attribution Samples



Qualitative Results

Generated Videos



Quantitative Results

VBench Evaluation

Method	Motion Smooth.	Dynamic Deg.
Base	96.3	39.6
Full FT	96.3	42.0
Random 10%	96.3	41.3
Ours w/o mask	96.3	43.8
MOTIVE	**96.3**	**47.6**

✓ Maintains smoothness, improves dynamics with only 10% data

Why Motion Masking? *Dynamic Degree*

Without: 43.8%

With: 47.6% (+3.8%)

Human Evaluation

vs. Base: 74.1% win
vs. Random: 58.9% win
vs. Full FT: 53.1% 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**

74.1% human preference vs. baseline with **10% data**; Motion masking: **+3.8%** Dynamic Degree