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, 4], Jonathan Lorraine[1]
[1]NVIDIA, [2]Princeton University, [3]MIT CSAIL, [4]University of Toronto

ICML International Conference On Machine Learning

Project page

Motivation

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

Key Goals

Focus on Motion	Scale Efficiently	Guide Curation
Separate motion from static appearance	Modern, large-scale models & datasets	Identify clips that improve motion quality

Our Method

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.

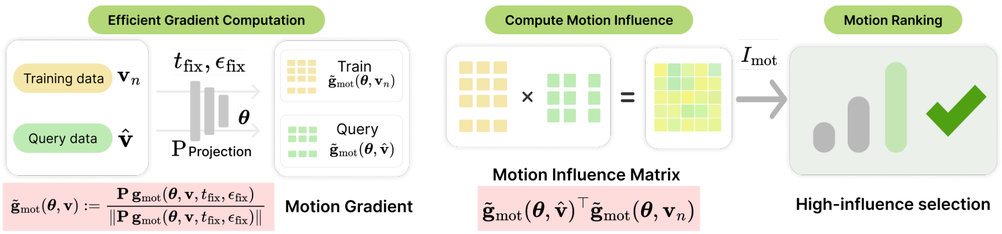
Method Components

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

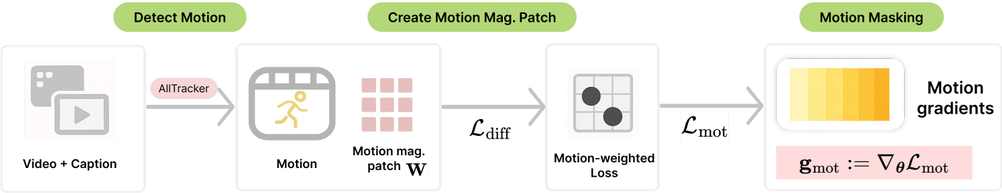
Which training clips drive the motion in a video generation sample?

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



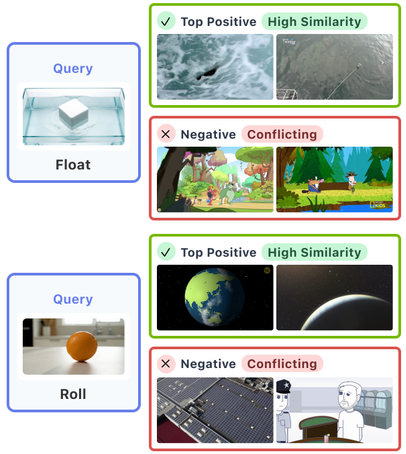
Motion-gradient computation has four steps: (1) detect motion with AllTracker, (2) compute motion magnitude patches, (3) apply motion-weighted loss, (4) compute motion-aware gradients.

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

Three Key Components

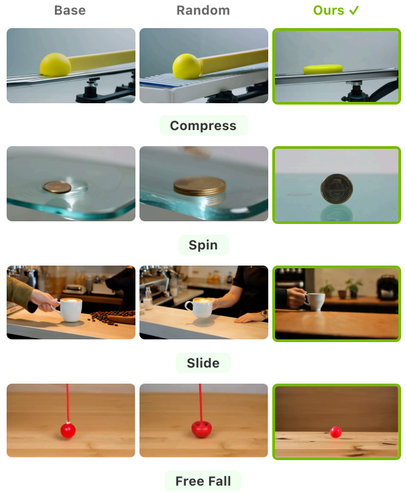
① Scalable Gradients	② Motion Attribution	③ Data Curation
Single-sample estimator + projection	Motion-weighted loss	Top-K Majority voting

Motion Attribution Samples



Qualitative Results

Generated Videos



Quantitative Results

VBench Evaluation *Wan2.1-T2V-1.3B*

Method	Subj.	Bg.	Mot.	Dyn.	Aesth.	Img.
Base	95.3	96.4	96.3	39.6	45.3	65.7
Full FT	95.9	96.6	96.3	42.0	45.0	63.9
Random selection	95.3	96.6	96.3	41.3	45.7	65.1
Motion magnitude	95.6	96.2	95.7	40.1	45.1	63.2
V-JEPA	95.7	96.0	95.6	41.6	44.9	62.7
Ours w/o MM	95.4	96.1	96.3	43.8	45.7	63.2
Ours (MOTIVE)	96.3	96.1	96.3	47.6	46.0	64.6

✓ Maintains smoothness, improves dynamics with only 10% data. All selection methods use 10% of training 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=751 achieves 66% agreement.

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

Conclusion

First motion-centric attribution framework for video generation

Scalable via **single-sample estimator & projection**

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