MOTIVE: Motion Attribution for Video Generation








Xindi Wu

Despoina Paschalidou

Jun Gao

Antonio Torralba

Laura Leal-Taixé

Olga Russakovsky

Sanja Fidler

Jonathan Lorraine

ICML 2026 *(oral)*

https://research.nvidia.com/labs/sil/projects/MOTIVE/

Video diffusion models can now generate highly realistic videos

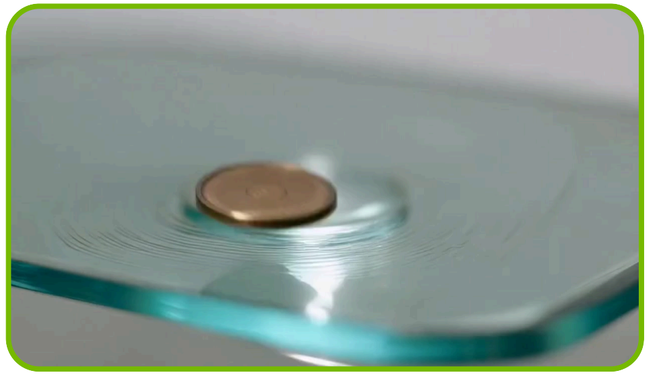
Cosmos

Veo 3

Sora 2

But, struggle with motion

Ex., unnatural physics, and temporal inconsistencies



?

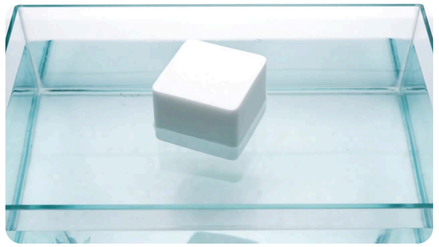
Which training clips influence the motion in generated videos?

Massive training data

Attributing performance to data?

Which training clips influence the motion in generated videos?

Query Motion



Float

Top Positive Influential Samples

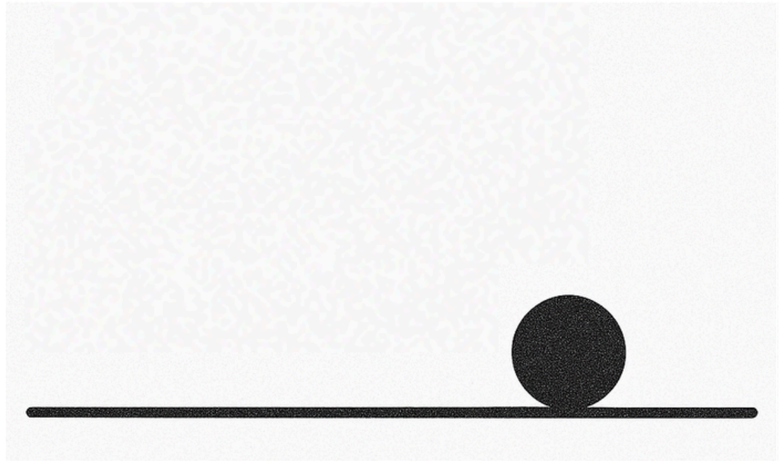
High Motion Similarity

Negative Influential Samples

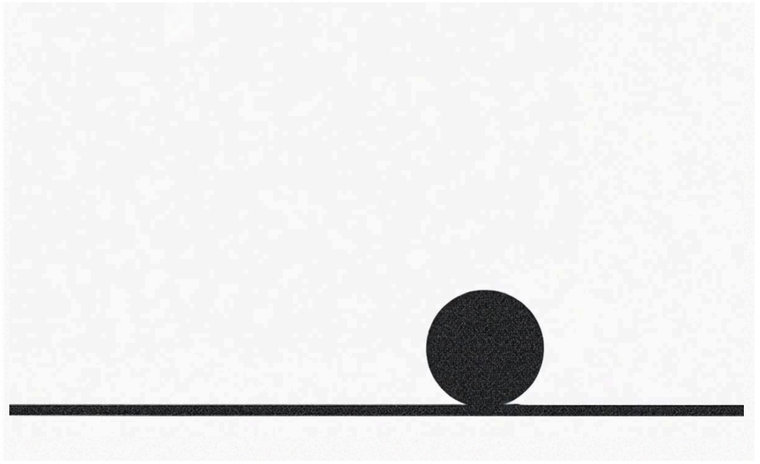
Conflicting Dynamics

Which training clips influence the motion in generated videos?

Image



Video



[Problem]

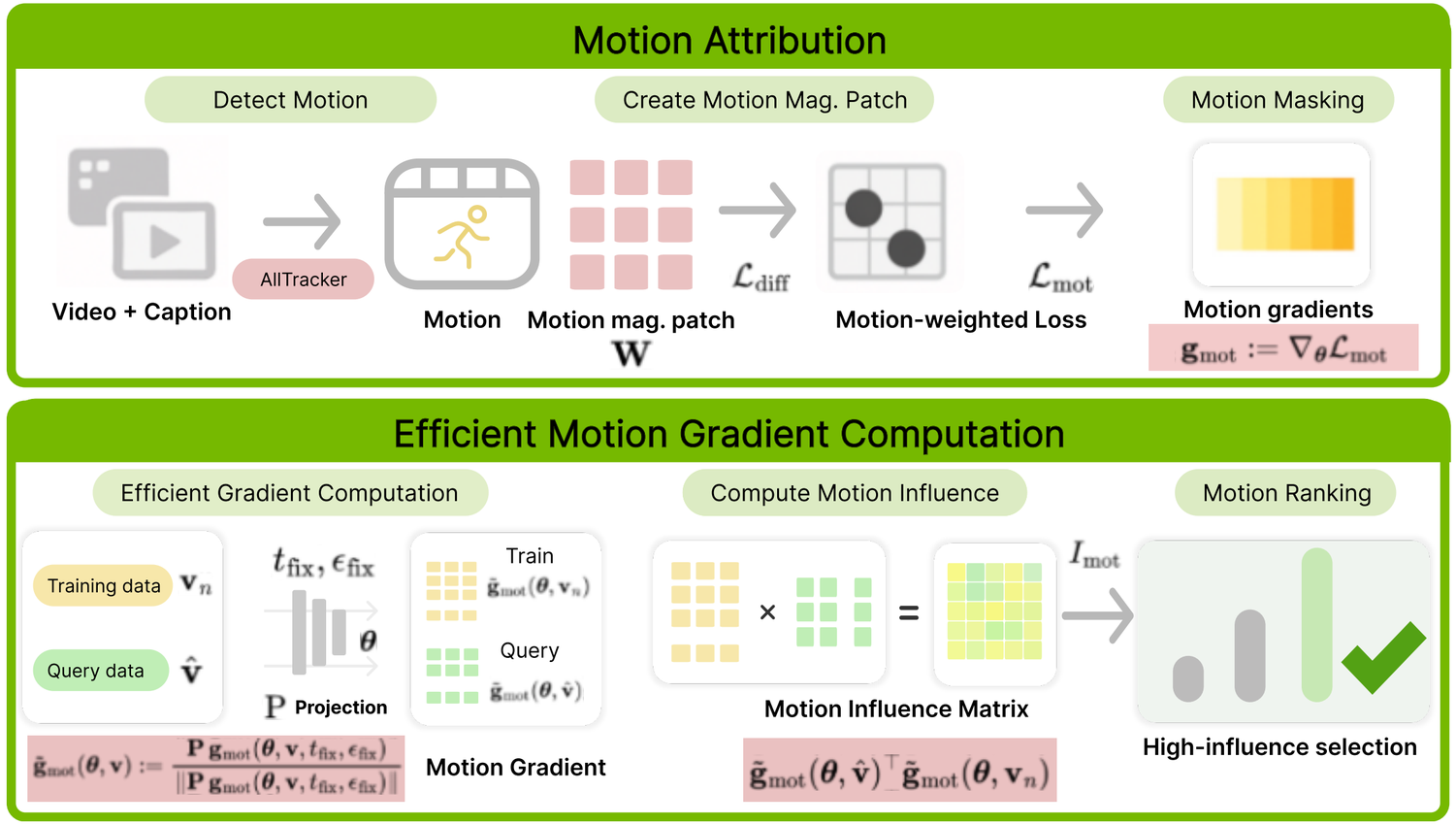
Attributing motion in video generative models faces two core challenges:

- **Motion localization:** Localize & attribute to motion
- **Scalability**: Scale to large-scale video generation models & datasets

[Introducing MOTIVE]

The **MOTIVE** Pipeline

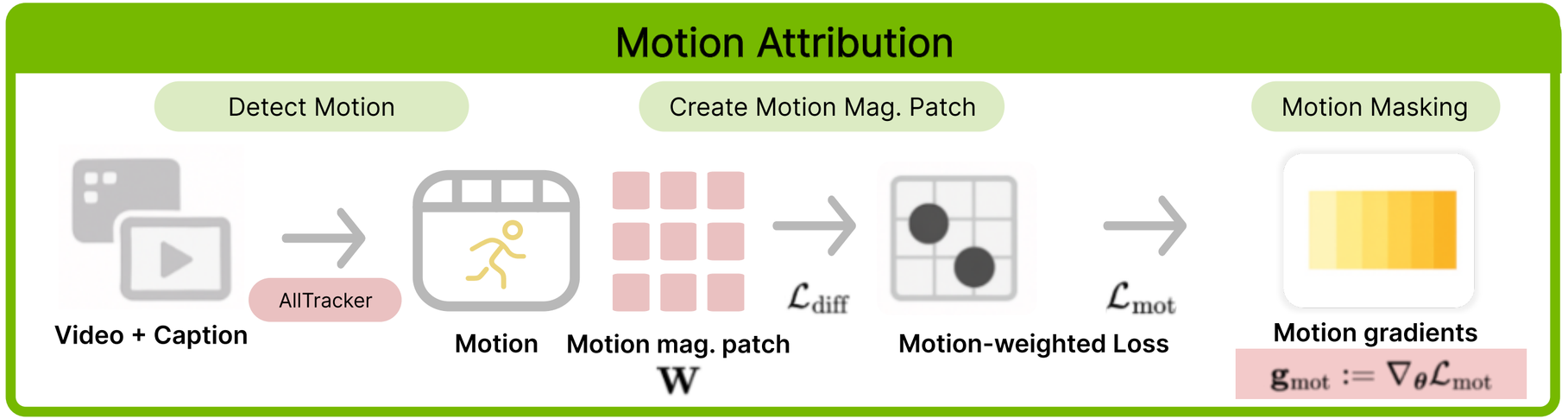
A **scalable, gradient-based, motion-centric** data attribution framework for video generation models.



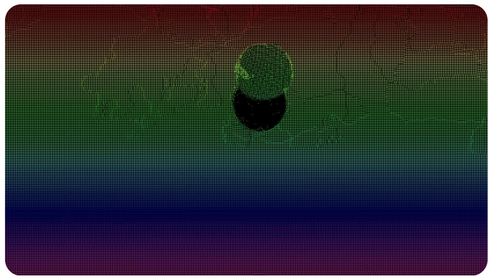
[Method]

Motion Attribution

Focusing on dynamics, not static appearance

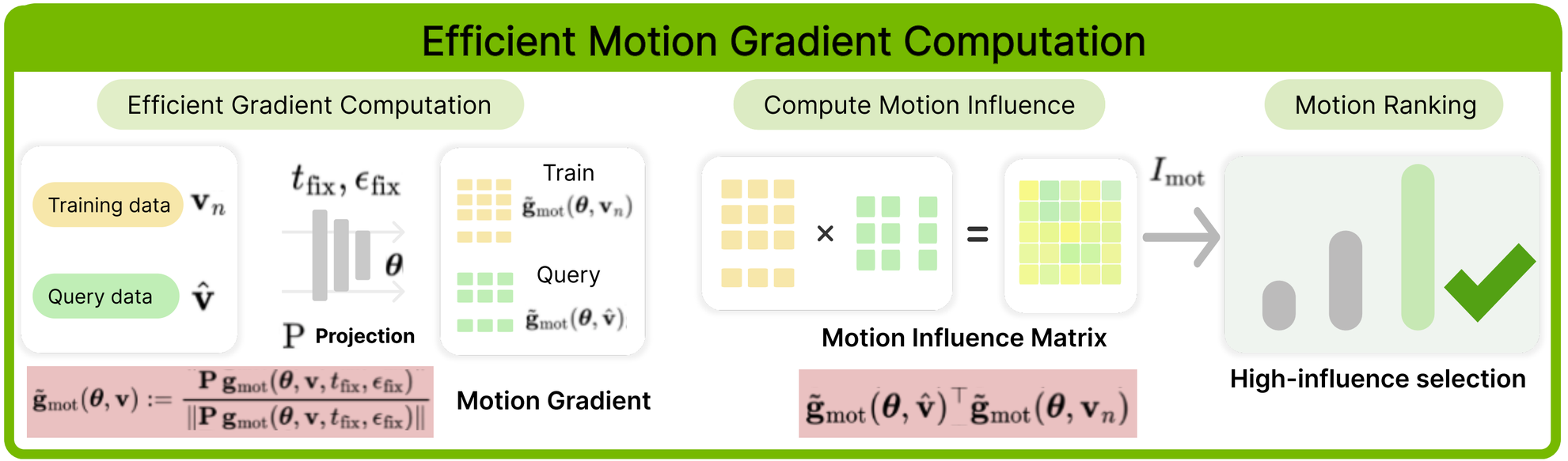


1. Detect motion with AllTracker
2. Compute motion magnitude patches
3. Apply motion-weighted loss
4. Compute motion-aware gradients



[Method]

Efficient Gradient Computation



Single-Sample Estimator

Structured Projections (Fastfood)

[Technical Details]

Single-Sample Estimator

Reduce variance by fixing (t, ϵ)

$$I^2_{\text{diff}}(\mathbf{x}_n, \mathbf{x}_{\text{test}}) = \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})}{\left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}) \right\|}^{\top} \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})}{\left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}) \right\|}$$

normalized test gradient

normalized training gradient

[Technical Details]

Single-Sample Estimator

Reduce variance by fixing (t, ϵ)

$$I^2_{\text{diff}}(\mathbf{x}_n, \mathbf{x}_{\text{test}}) = \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})}{\left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}) \right\|}^{\top} \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})}{\left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}) \right\|}$$

normalized test gradient

normalized training gradient

Structured Projections (Fastfood)

$$\mathbf{P} \in \mathbb{R}^{D' \times D} \quad \text{be implemented as} \quad \mathbf{P} := \frac{1}{\xi \sqrt{D'}} \mathbf{SQG\Pi QB}$$

$$\tilde{g}(\theta, \mathbf{x}) := \frac{\mathbf{P}\, \nabla_\theta \mathcal{L}_{\text{diff}}(\theta; \mathbf{x}, t_{\text{fix}}, \epsilon_{\text{fix}})}{\left\| \mathbf{P}\, \nabla_\theta \mathcal{L}_{\text{diff}}(\theta; \mathbf{x}, t_{\text{fix}}, \epsilon_{\text{fix}}) \right\|}.$$

[Technical Details]

Single-Sample Estimator

Reduce variance by fixing (t, ϵ)

$$I^2_{\text{diff}}(\mathbf{x}_n, \mathbf{x}_{\text{test}}) = \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})}{\left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}) \right\|}^{\top} \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})}{\left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}) \right\|}$$

normalized test gradient

normalized training gradient

Structured Projections (Fastfood)

$$\mathbf{P} \in \mathbb{R}^{D' \times D} \quad \text{be implemented as} \quad \mathbf{P} := \frac{1}{\xi \sqrt{D'}} \mathbf{SQG\Pi QB}$$

$$\tilde{g}(\theta, \mathbf{x}) := \frac{\mathbf{P}\, \nabla_\theta \mathcal{L}_{\text{diff}}(\theta; \mathbf{x}, t_{\text{fix}}, \epsilon_{\text{fix}})}{\left\| \mathbf{P}\, \nabla_\theta \mathcal{L}_{\text{diff}}(\theta; \mathbf{x}, t_{\text{fix}}, \epsilon_{\text{fix}}) \right\|}.$$

(Final) Influence Estimation:

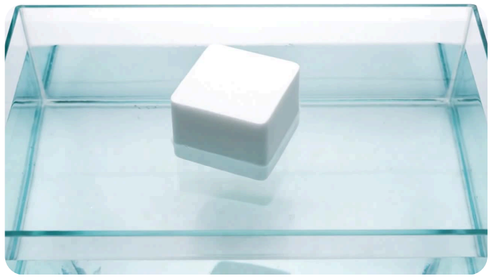
$$I^3_{\text{diff}}(\mathbf{x}_n, \mathbf{x}_{\text{test}}) = \tilde{\mathbf{g}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}})^{\top} \tilde{\mathbf{g}}(\boldsymbol{\theta};\, \mathbf{x}_n)$$

projected, normalized test gradient

projected, normalized training gradient

[Influential Examples]

Query Motion



Prompt:
"...floating..."

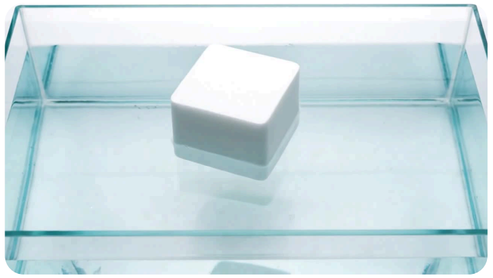
Positive Influential Samples

High Motion Similarity




[Influential Examples]

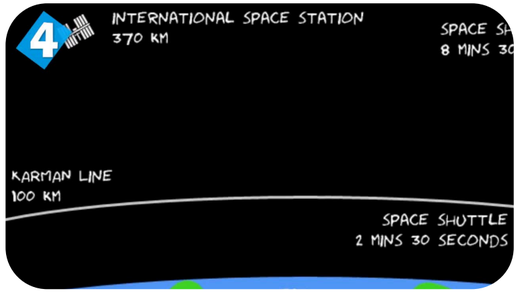
Query Motion



Prompt:
"...floating..."

❌ Negative Influential Samples

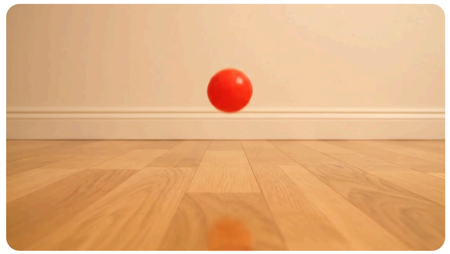
Conflicting Dynamics



Evaluation Setup

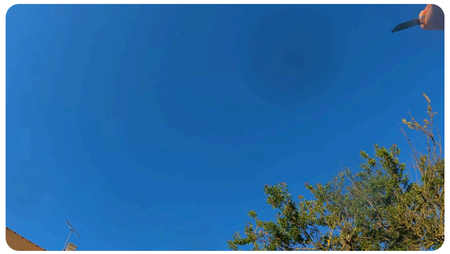
We evaluate attribution across 10 distinct motion types:

compress

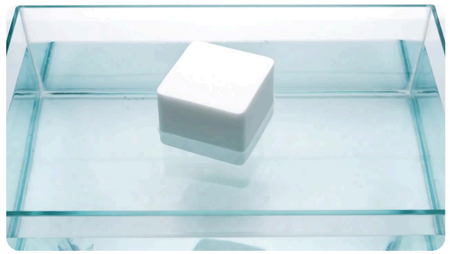


bounce

roll



explode



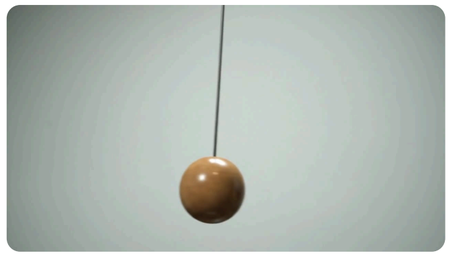
float

free fall

slide

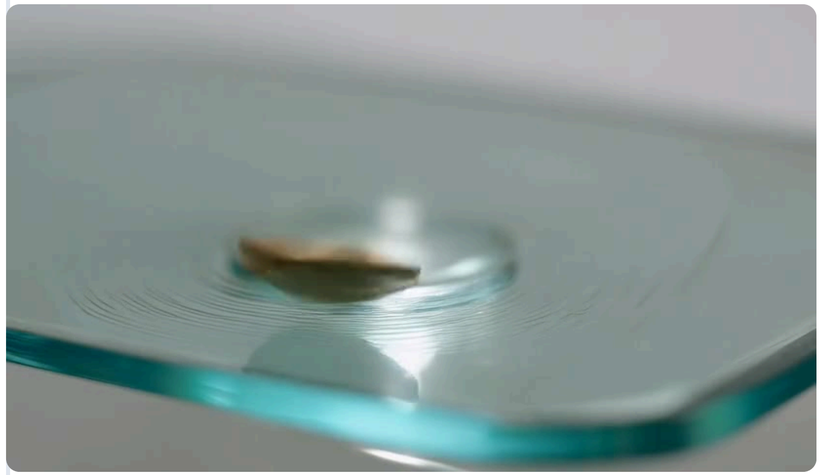
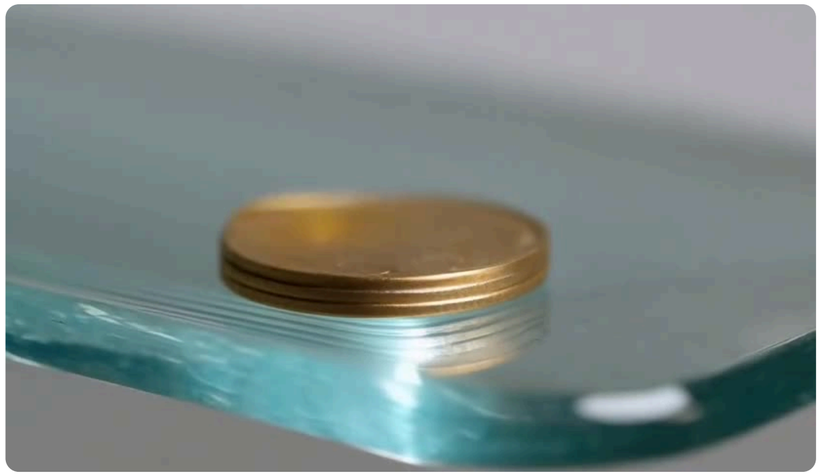
spin

stretch



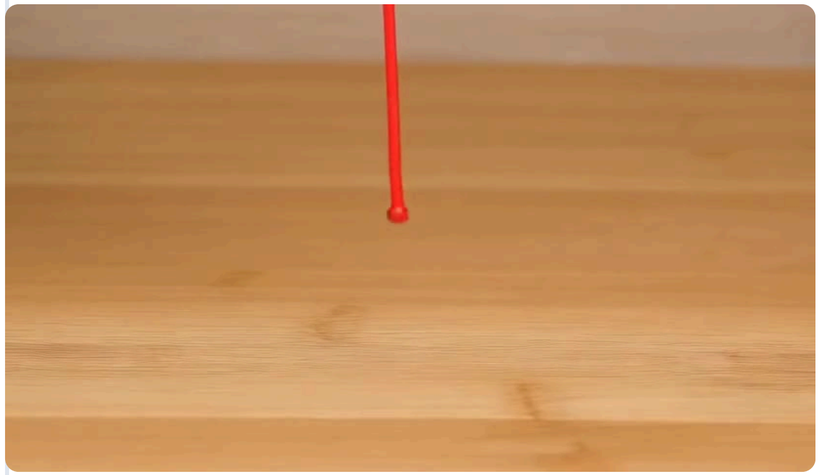
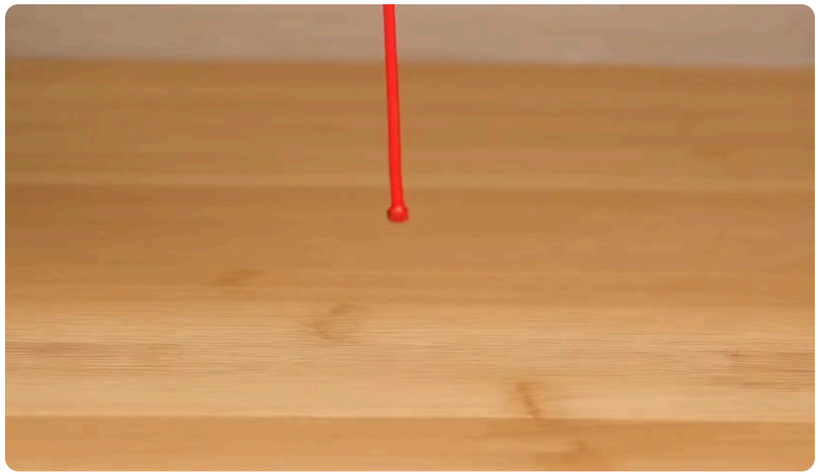
swing

Finetune with Single-Query Selected Data

Base	Random Selection	Ours
		

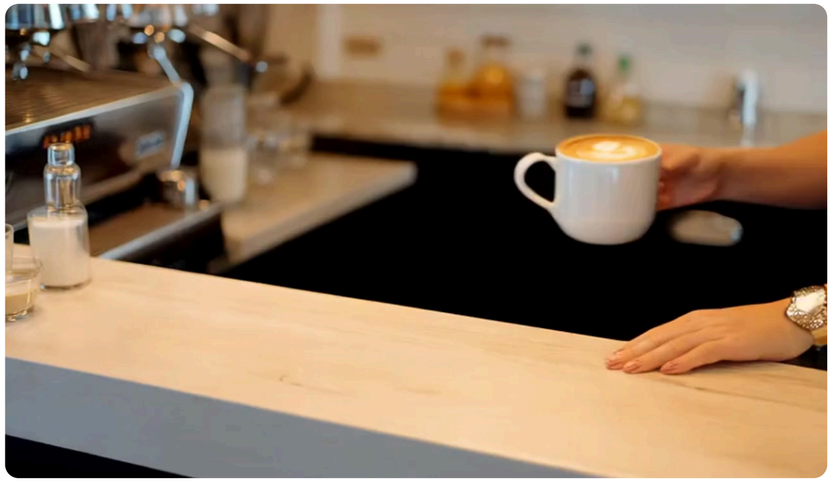
A single coin **spins** quickly on a polished glass surface, close-up fixed camera, bright even lighting, plain backdrop; capture its precession and slow wobble as it settles.

Finetune with Single-Query Selected Data

Base	Random Selection	Ours
		

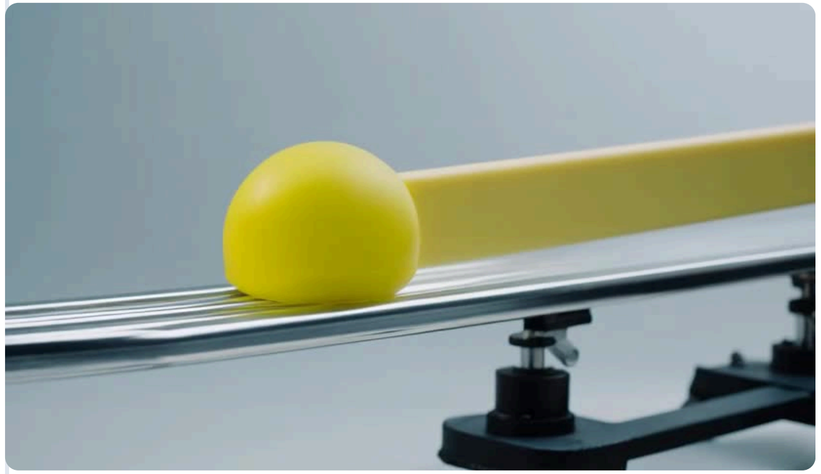
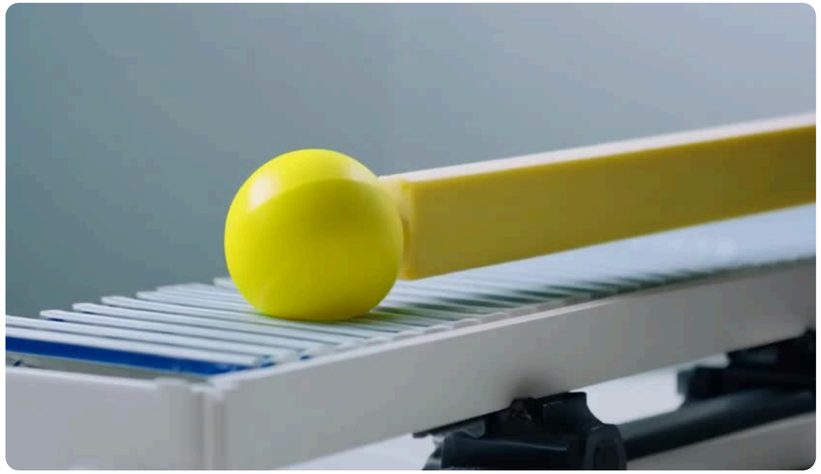
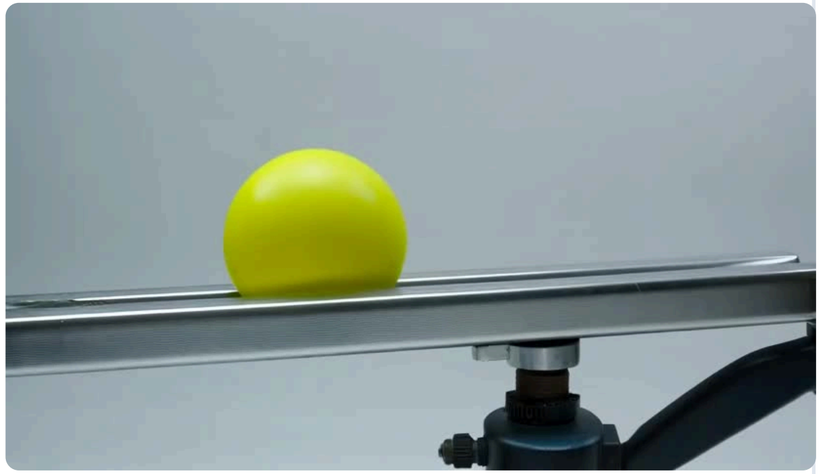
A red ball drops vertically from above and **falls** straight onto the wooden surface. The motion is quick and direct, with light motion blur showing its fall against the clean wooden background.

Finetune with Single-Query Selected Data

Base	Random Selection	Ours
		

A white mug is placed and then **slid** across a wooden kitchen counter, fixed side camera, diffuse lighting, blurred kitchen background; emphasize the sliding motion of the mug.

Finetune with Single-Query Selected Data

Base	Random Selection	Ours
		

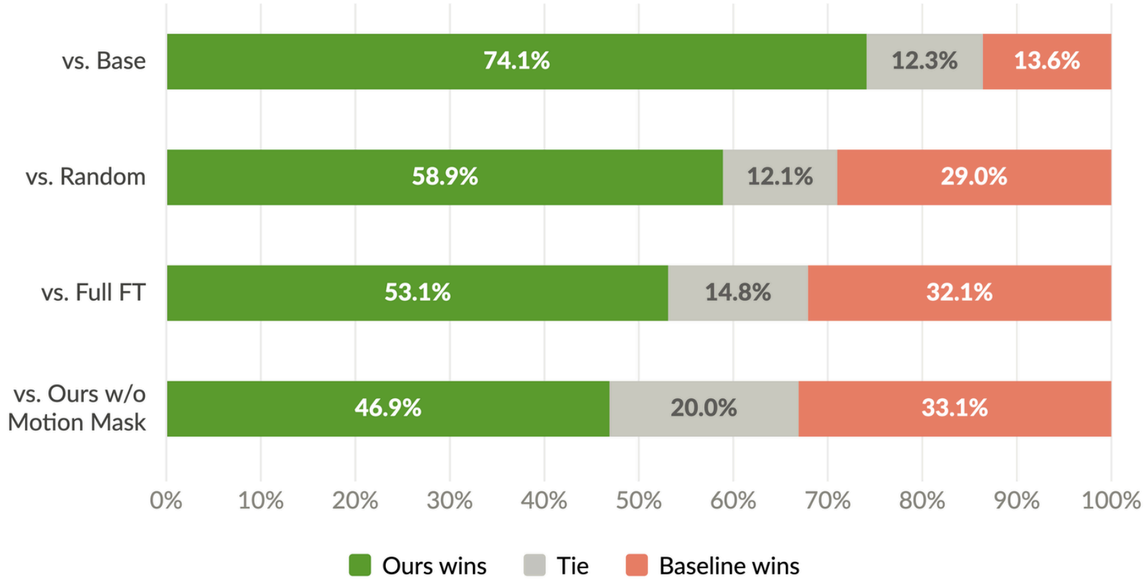
A rubber ball being **compressed** under a flat press, filmed with a stationary camera. Bright, shadow-free lighting and a clean background emphasize the deformation as it flattens.

[Experiments]

Finetune with Multi-Query Selected Data

Method	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
Base	95.3	96.4	96.3	39.6	45.3	65.7
Full fine-tuning	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 embedding	95.7	96.0	95.6	41.6	44.9	62.7
Ours w/o Motion Mask	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

Vbench Evaluation



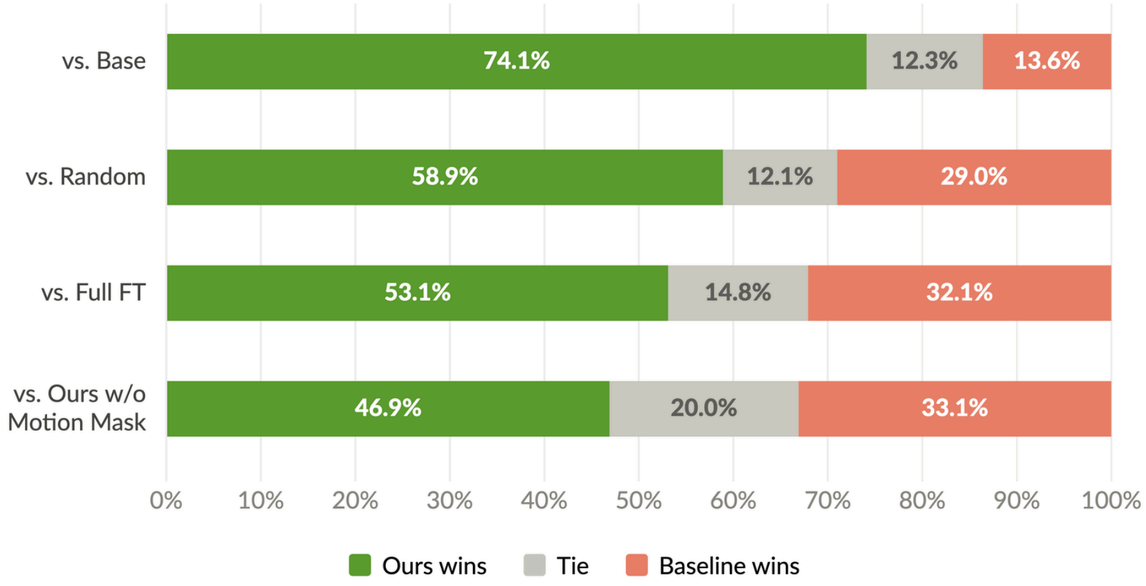
Human Evaluation

[Experiments]

Finetune with Multi-Query Selected Data

Method	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
Base	95.3	96.4	96.3	39.6	45.3	65.7
Full fine-tuning	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 embedding	95.7	96.0	95.6	41.6	44.9	62.7
Ours w/o Motion Mask	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

Vbench Evaluation



Human Evaluation

[Summary]

Arxiv Website

ICML 2026 *(Oral)*

MOTIVE: Motion Attribution for Video Generation

MOTIVE provides a scalable and effective way to attribute motion in video generation models, enabling targeted data curation that significantly improves motion quality.

Key Contributions:

- Our motion-centric attribution is scalable and effective for improving video generation models.
- Motive enables data efficiency - 10% of carefully selected data can match or exceed full dataset performance especially motion


Xindi Wu


Despoina Paschalidou


Jun Gao


Antonio Torralba


Laura Leal-Taixé


Olga Russakovsky


Sanja Fidler


Jonathan Lorraine