MOTION ATTRIBUTION FOR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the rapid progress of video generation models, the role of data in influencing motion is poorly understood. We present MOTIVE (**MOTIon attribution for Video gEneration**), a motion-centric, gradient-based data attribution framework that scales to modern, large, high-quality video datasets and models. We use this to study which fine-tuning clips improve or degrade temporal dynamics. MOTIVE isolates temporal dynamics from static appearance via motion-weighted loss masks, yielding efficient and scalable motion-specific influence computation. On text-to-video models, MOTIVE identifies clips that strongly affect motion and guides data curation that improves temporal consistency and physical plausibility. With MOTIVE-selected high-influence data, we improve both motion smoothness and dynamic degree on VBench, achieving a 74.1% human preference win rate compared with the pretrained base model. To our knowledge, this is the first framework to attribute motion rather than visual appearance in video generative models and to use it to curate fine-tuning data.

1 INTRODUCTION

Motion is the defining element of videos. Unlike image generation, which produces a single frame, video generative models must capture how objects move, interact, and obey physical constraints (Wiedemer et al., 2025; Kang et al., 2024). Yet a fundamental question remains: *which training data influence the motion in generated videos?*

Why it matters. Diffusion models are data-driven, and their progress tracks the scaling of data and compute (Saharia et al., 2022; Ho et al., 2022; Peebles & Xie, 2023). Prior work shows that training data shapes key generative properties, including visual quality (Rombach et al., 2022) and compositionality (Favero et al., 2025). Motion is no exception, if generated motion reflects the data distribution that shaped the model, then attributing motion to influential training clips provides a direct lens on why a model moves the way it does and enables targeted data selection for desired dynamics.

Why existing methods are limited for motion. Prior diffusion attribution focuses on images and static content. Extending these methods to videos naïvely collapses motion into appearance, missing temporal structure. Three challenges drive this gap: (i) localizing motion to focus attribution on dynamic regions, (ii) scaling to sequences where gradients integrate across time, and (iii) capturing temporal relations that single-frame attribution cannot measure.

Our method. We introduce MOTIVE, a motion attribution framework that isolates motion-specific influence via motion-aware gradient masking. Our key contributions are:

- A scalable gradient-based attribution for video generation that is computationally efficient at modern scales (§2).
- Addressing video-specific bias by correcting frame-length effects in gradient magnitudes.
- Motion-weighted attribution that emphasizes temporal dynamics over static appearance.
- Showing motion improvements on VBench and human evaluation, matching full-dataset fine-tuning with only 10% of data (Tab. 1).

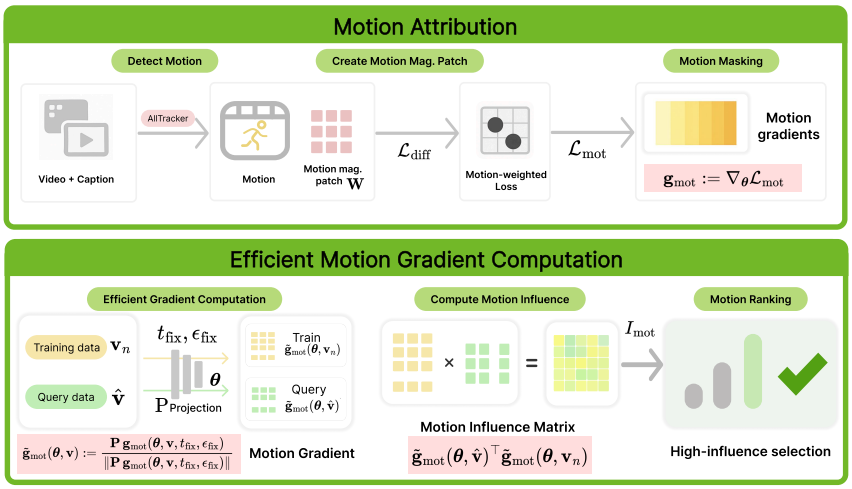


Figure 1: **MOTIVE.** *Top.* Motion-gradient computation (§2.3) has three steps: (1) detect motion with AllTracker; (2) compute motion-magnitude patches; (3) apply loss-space motion masks to focus gradients on dynamic regions. *Bottom.* Our method (§2.2) is made scalable via a single-sample variant with common randomness and a projection, computed for each pair of training and query data, aggregated (§D.2) for a final ranking, and eventually used to select fine-tuning subsets.

2 METHOD

We formalize the problem in §2.1 and develop a practical framework for motion attribution in video diffusion models with four components: scalable gradient computation (§2.2), frame-length bias fix (§D.1), motion-aware weighting (§2.3), and target data selection (§D.2). We provide efficiency analysis (§D.3) demonstrating scalability to billion-parameter models and large-scale video datasets.

2.1 PROBLEM FORMULATION

We study data attribution for motion in the fine-tuning setting. Let $\mathcal{D}_{\text{ft}} = \{(\mathbf{v}_n, \mathbf{c}_n)\}_{n=1}^{N}$ be the fine-tuning corpus. Given a query video $(\hat{\mathbf{v}}, \hat{\mathbf{c}})$, we assign to each training clip $(\mathbf{v}_n, \mathbf{c}_n)$ a motion-aware influence score $I(\mathbf{v}_n, \hat{\mathbf{v}}; \boldsymbol{\theta})$ that explains how it contributes to the dynamics observed in $\hat{\mathbf{v}}$. The score should satisfy: (i) *predictivity*, rankings correlate with observed changes from fine-tuning on the most influential subsets; (ii) *efficiency*, scales to modern video generators, such as forgoing explicit Hessian inversion, expensive per-data integration, or prohibitive storage. To do this, we augment the influence target defined in Eq. 17 to be (a) lower variance for stable rankings with feasible levels of compute, (b) more scalable to store, and (c) motion-centric.

Fine-tuning Subset Selection. For a budget $K \ll N$, we get a motion-influential subset by ranking scores and taking the top-K examples. For multiple query motions, we combine selections as described in §D.2. The resulting subsets serve as candidates for motion-centric fine-tuning.

2.2 SCALABLE GRADIENT-BASED ATTRIBUTION

We make attribution scalable for modern, large video datasets and models via inverse-Hessian approximations, lower-variance gradient-similarity estimators, low-cost single-sample estimators, and a Fastfood projection for tractable storage.

Approximating the inverse-Hessian. Computing exact inverse-Hessian-vector products is infeasible for modern neural networks. We estimate influence via gradient similarity, using an identity preconditioner for the inverse Hessian (Koh & Liang, 2017; Pruthi et al., 2020; Park et al., 2023).

Common randomness for stable rankings. To reduce variance without changing the target, we evaluate train and test gradients under the same $(t, \boldsymbol{\epsilon})$ pairs and average over a small set $\mathcal{T}$ (Xie et al., 2024; Lin et al., 2024). This paired averaging stabilizes rankings compared to independent draws:

$$I^1_{\text{diff}}(\mathbf{x}_n, \mathbf{x}_{\text{test}}) = \frac{1}{|\mathcal{T}|} \sum_{t,\boldsymbol{\epsilon}\in\mathcal{T}} \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}, t, \boldsymbol{\epsilon})}{\left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}, t, \boldsymbol{\epsilon})\right\|}^{\top} \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_n, t, \boldsymbol{\epsilon})}{\left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{x}_n, t, \boldsymbol{\epsilon})\right\|}. \tag{1}$$

Single-sample variant for reduced compute. We fix a single t_{fix} and a shared draw $\boldsymbol{\epsilon}_{\text{fix}} \sim \mathcal{N}(0, \mathbf{I})$ for all train–test pairs at the final checkpoint. Sharing $(t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})$ allows low enough variance for the low-cost single-sample estimator to maintain relative ordering (Xie et al., 2024; Lin et al., 2024). The estimator collapses to:

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

Structured projection for reduced storage. To operate at model scale, we apply a Johnson–Lindenstrauss projection via Fastfood (Le et al., 2014) and then normalize. Let

$$\mathbf{P} \in \mathbb{R}^{D' \times D} \;\text{ implemented as }\; \mathbf{P} := \tfrac{1}{\xi\sqrt{D'}}\, \mathbf{SQG\Pi QB}, \tag{3}$$

where $\mathbf{Q}$ is the Walsh–Hadamard matrix, $\mathbf{B}$ is a diagonal Rademacher matrix, $\mathbf{\Pi}$ is a random permutation, $\mathbf{G}$ is a diagonal Gaussian scaling, and $\mathbf{S}$ is a diagonal rescaling, and ξ normalizes the variance. The projected, normalized gradient is:

$$\tilde{\mathbf{g}}\big(\boldsymbol{\theta}, \mathbf{x}\big) \;:=\; \frac{\mathbf{P}\, \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta}, \mathbf{x}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})}{\|\mathbf{P}\, \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta}, \mathbf{x}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})\|}. \tag{4}$$

Then the influence score is the dot product of normalized projected gradients (i.e., cosine similarity) in $\mathbb{R}^{D'}$:

$$I^3_{\text{diff}}(\mathbf{x}_n, \mathbf{x}_{\text{test}}) \;=\; \tilde{\mathbf{g}}\big(\boldsymbol{\theta};\, \mathbf{x}_{\text{test}}\big)^{\top} \tilde{\mathbf{g}}\big(\boldsymbol{\theta};\, \mathbf{x}_n\big). \tag{5}$$

This keeps compute $\mathcal{O}(D' \log D')$ for projection and $\mathcal{O}(D')$ per dot product, with storage $\mathcal{O}(|\mathcal{D}|\, D')$, while staying close to the ranking behavior of full-gradient cosine similarity (Park et al., 2023).

2.3 Motion Attribution

To move beyond whole-video influence, we introduce motion attribution, which isolates the contribution of training data to temporal dynamics. Unlike video-level attribution, which treats each clip as a single unit and conflates appearance with motion, motion attribution reweights per-location gradients using motion masks, assigning influence via dynamic behavior rather than static content.

Motion Masking Attribution. Motion is what distinguishes video diffusion from image diffusion. Our goal is to understand how training data shapes motion in video diffusion models. Prior work has emphasized architectural or algorithmic changes for motion modeling (Peebles & Xie, 2023; Blattmann et al., 2023; Guo et al., 2023), many of the largest generative gains have instead come from scaling and curating massive video corpora, which in turn enable impressive motion synthesis results in video diffusion models (Ho et al., 2022; Wan et al., 2025; Tan et al., 2024; Yang et al., 2024). Yet we lack tools that quantify how specific training clips shape particular motion patterns. We address this by attributing motion back to data via motion-weighted gradients, which yields actionable signals for targeted data selection, artifact diagnosis, and selective fine-tuning.

Motion Detection and Latent Space Mapping. Given a video $\mathbf{v} \in \mathbb{R}^{F\times H\times W\times 3}$ with F frames of resolution $H \times W$, we first encode it into the VAE latent space as $\mathbf{h} = E(\mathbf{v}) \in \mathbb{R}^{F\times H/s\times W/s\times C}$, with downsampling factor $s = 8$ and $C = 16$ following the `Wan2.1` backbone used in our experiments.

For motion computation, we use AllTracker (Harley et al., 2025) to extract motion information in pixel space: $A = \mathcal{A}(\mathbf{v}) \in \mathbb{R}^{F\times H\times W\times 4}$, where the first two channels contain optical flow maps $A_{:,:,:,0:2}$ indicating pixel displacement between frames, and the remaining channels $A_{:,:,:,2:4}$ encode visibility and confidence scores. We extract displacement vectors at each pixel location as:

$$\mathbf{D}_f(h, w) = (A_{f,h,w,0}, A_{f,h,w,1}) = (\mathrm{d}w, \mathrm{d}h). \tag{6}$$

We then bilinearly downsample motion quantities from (H, W) to the latent grid $\left(\frac{H}{s}, \frac{W}{s}\right)$ so that our masking lives where gradients are computed.

Motion-Weighted Gradient Computation. We define the motion magnitude at each location as: $M_f(h, w) = \|\mathbf{D}_f(h, w)\|_2$. To obtain comparable motion weights across frames and pixels, we min–max normalize over all frames and pixels, ensuring values lie in $[0, 1]$:

$$\mathbf{W}(f, h, w) = \frac{M_f(h, w) - m_{\min}}{m_{\max} - m_{\min} + \zeta}, \tag{7}$$

Table 1: **VBench Evaluation.** Performance comparison on VBench (Huang et al., 2024) across different baselines (all values in %, higher is better). All selection methods use 10% of training data; our method uses majority vote aggregation (§D.2) across motion queries. MM: motion masking. Subj.: Subject Consistency, Bg.: Background Consistency, Mot.: Motion Smoothness, Dyn.: Dynamic Degree, Aesth.: Aesthetic Quality, Img.: Imaging Quality.

	Wan2.1-T2V-1.3B						Wan2.2-TI2V-5B					
Method	Subj.	Bg.	Mot.	Dyn.	Aesth.	Img.	Subj.	Bg.	Mot.	Dyn.	Aesth.	Img.
Base	95.3	96.4	96.3	39.6	45.3	**65.7**	94.9	96.4	97.5	42.0	44.4	65.5
Full FT	95.9	**96.6**	96.3	42.0	45.0	63.9	**95.3**	96.5	97.5	45.3	44.8	**66.2**
Random	95.3	96.6	96.3	41.3	45.7	65.1	94.7	96.2	97.3	41.6	44.6	65.2
Motion mag.	95.6	96.2	95.7	40.1	45.1	63.2	95.0	96.3	97.4	44.9	45.0	65.1
V-JEPA	95.7	96.0	95.6	41.6	44.9	62.7	95.2	96.4	97.3	45.6	44.9	64.8
Ours w/o MM	95.4	96.1	96.3	43.8	45.7	63.2	94.9	96.5	97.4	43.8	45.2	64.8
Ours (MOTIVE)	**96.3**	96.1	96.3	**47.6**	**46.0**	64.6	95.1	**96.6**	**97.6**	**48.3**	**45.6**	65.5

where $m_{\min} = \min_{f',h',w'} M_f(h',w')$, $m_{\max} = \max_{f',h',w'} M_f(h',w')$, and $\zeta = 10^{-6}$ ensures a positive denominator. This normalization mitigates bias from absolute motion scale, yielding weights that emphasize relative motion saliency rather than raw magnitude, following prior practice in video saliency detection (Fang et al., 2013). Let $(\tilde{h},\tilde{w})$ index the latent grid. We obtain latent-aligned weights by bilinear downsampling:

$$\tilde{\mathbf{W}}(f,\tilde{h},\tilde{w}) \;=\; \mathrm{Bilinear}\big(\mathbf{W}(\cdot,\cdot,\cdot), F, \tfrac{H}{s}, \tfrac{W}{s}\big)\,. \tag{8}$$

We compute per-location squared error at fixed $(t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})$ at each frame f and "latent pixel" $(\tilde{h},\tilde{w})$:

$$\begin{aligned}\tilde{\mathcal{L}}_{\boldsymbol{\theta},\mathbf{v},\mathbf{c}}(f,\tilde{h},\tilde{w}) \;=\; \Big(&[\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}(\mathbf{v},t_{\text{fix}},\boldsymbol{\epsilon}_{\text{fix}}),t_{\text{fix}},\mathbf{c})]_{f,\tilde{h},\tilde{w}} \\ &- [\boldsymbol{\epsilon}_{\text{target}}(t_{\text{fix}},\boldsymbol{\epsilon}_{\text{fix}})]_{f,\tilde{h},\tilde{w}} \Big)^2,\end{aligned} \tag{9}$$

and define the motion-weighted loss by averaging over frames and latent spatial locations:

$$\mathcal{L}_{\text{mot}}(\boldsymbol{\theta};\mathbf{v},\mathbf{c}) \;=\; \frac{1}{F_{\mathbf{v}}} \operatorname*{mean}_{f,\tilde{h},\tilde{w}} \Big[\tilde{\mathbf{W}}_{\mathbf{v},\mathbf{c}}(f,\tilde{h},\tilde{w}) \cdot \tilde{\mathcal{L}}_{\boldsymbol{\theta},\mathbf{v},\mathbf{c}}(f,\tilde{h},\tilde{w})\Big]. \tag{10}$$

Notably, when $\tilde{\mathbf{W}}$ is all ones, this recovers the standard objective with no motion emphasis. The $1/F_{\mathbf{v}}$ factor corrects for frame-length bias and $F_{\mathbf{v}}$ signifies how the number of frames may be video-dependent. The corresponding motion-weighted gradient for attribution is:

$$I_{\text{mot}}(\mathbf{v}_n,\hat{\mathbf{v}}) = \tilde{\mathbf{g}}_{\text{mot}}(\boldsymbol{\theta},\hat{\mathbf{v}})^{\top} \tilde{\mathbf{g}}_{\text{mot}}(\boldsymbol{\theta},\mathbf{v}_n), \tag{11}$$

$$\text{where } \mathbf{g}_{\text{mot}} := \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{mot}},\ \tilde{\mathbf{g}}_{\text{mot}}(\boldsymbol{\theta},\mathbf{v}) := \frac{\mathbf{P}\mathbf{g}_{\text{mot}}(\boldsymbol{\theta},\mathbf{v},t_{\text{fix}},\boldsymbol{\epsilon}_{\text{fix}})}{\|\mathbf{P}\mathbf{g}_{\text{mot}}(\boldsymbol{\theta},\mathbf{v},t_{\text{fix}},\boldsymbol{\epsilon}_{\text{fix}})\|}. \tag{12}$$

Loss-space masking leaves forward noising and generation unchanged and reweights only attribution, avoiding interactions between motion weighting and noise injection. In contrast, our motion-aware attribution emphasizes dynamic regions and de-emphasizes static backgrounds, so rankings identify training clips that most strongly shape the model's motion rather than appearance.

The most influenctial fine-tuning subset selection are detailed in App. §D.2 and §D.3.

3 EXPERIMENTS

More experiment setups details are in App. §E and we provide evaluation results on VBench in Tab. 1.

4 CONCLUSION

We introduced **MOTIVE**, a motion-aware data attribution framework for video diffusion. By tracing generated motion to influential training clips, our method enables targeted data curation that improves motion quality with a fraction of the data. As video models scale, such data-level understanding will be key to diagnosing failure modes and building more controllable generative systems.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger B Grosse. Training data attribution via approximate unrolling. *Advances in Neural Information Processing Systems*, 37:66647–66686, 2024.

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pp. 354–363. PMLR, 2018.

Yuxiang Bao, Di Qiu, Guoliang Kang, Baochang Zhang, Bo Jin, Kaiye Wang, and Pengfei Yan. Latentwarp: Consistent diffusion latents for zero-shot video-to-video translation. *arXiv preprint arXiv:2311.00353*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jonathan Brokman, Omer Hofman, Roman Vainshtein, Amit Giloni, Toshiya Shimizu, Inderjeet Singh, Oren Rachmil, Alon Zolfi, Asaf Shabtai, Yuki Unno, et al. Montrage: Monitoring training for attribution of generative diffusion models. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025.

RDWS Cook et al. Residuals and influence in regression. 1982.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Yuming Fang, Weisi Lin, Zhenzhong Chen, Chia-Ming Tsai, and Chia-Wen Lin. A video saliency detection model in compressed domain. *IEEE transactions on circuits and systems for video technology*, 24(1):27–38, 2013.

Alessandro Favero, Antonio Sclocchi, Francesco Cagnetta, Pascal Frossard, and Matthieu Wyart. How compositional generalization and creativity improve as diffusion models are trained. *arXiv preprint arXiv:2502.12089*, 2025.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Google DeepMind. Veo-3: Advancing controllable and physically plausible video generation. Technical report, Google DeepMind, 2025. URL `https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf`.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.

Adam W Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Pavel Tokmakov, et al. Alltracker: Efficient dense point tracking at high resolution. *arXiv preprint arXiv:2506.07310*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3): 185–203, 1981.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Ruoxi Jia, Fan Wu, Xuehui Sun, Jiacen Xu, David Dao, Bhavya Kailkhura, Ce Zhang, Bo Li, and Dawn Song. Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894. PMLR, 2017.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. *arXiv preprint arXiv:2310.00902*, 2023.

Quoc Viet Le, Tamás Sarlós, and Alexander Johannes Smola. Fastfood: Approximate kernel expansions in loglinear time. *arXiv preprint arXiv:1408.3060*, 2014.

Jinxu Lin, Linwei Tao, Minjing Dong, and Chang Xu. Diffusion attribution score: Evaluating training data influence in diffusion models. *arXiv preprint arXiv:2410.18639*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jonathan Lorraine and David Duvenaud. Stochastic hyperparameter optimization through hypernetworks. *arXiv preprint arXiv:1802.09419*, 2018.

Jonathan Lorraine and Safwan Hossain. Jacnet: Learning functions with structured jacobians. *arXiv preprint arXiv:2408.13237*, 2024.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pp. 1540–1552. PMLR, 2020.

Jonathan Lorraine, Paul Vicol, Jack Parker-Holder, Tal Kachman, Luke Metz, and Jakob Foerster. Lyapunov exponents for diversity in differentiable games. *arXiv preprint arXiv:2112.14570*, 2021.

Jonathan Lorraine, Nihesh Anderson, Chansoo Lee, Quentin De Laroussilhe, and Mehadi Hassen. Task selection for automl system evaluation. *arXiv preprint arXiv:2208.12754*, 2022a.

Jonathan P Lorraine, David Acuna, Paul Vicol, and David Duvenaud. Complex momentum for optimization in games. In *International Conference on Artificial Intelligence and Statistics*, pp. 7742–7765. PMLR, 2022b.

Jonathan Peter Lorraine. *Scalable nested optimization for deep learning*. PhD thesis, University of Toronto (Canada), 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pp. 674–679, 1981.

Nikhil Mehta, Jonathan Lorraine, Steve Masson, Ramanathan Arunachalam, Zaid Pervaiz Bhat, James Lucas, and Arun George Zachariah. Improving hyperparameter optimization with checkpointed model weights. In *European Conference on Computer Vision*, pp. 75–96. Springer, 2024.

Bruno Mlodozeniec, Runa Eschenhagen, Juhan Bae, Alexander Immer, David Krueger, and Richard Turner. Influence functions for scalable data attribution in diffusion models. *arXiv preprint arXiv:2410.13850*, 2024.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Yonghyun Park, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, Naoki Murata, Wei-Hsiang Liao, Woosung Choi, Kin Wai Cheuk, Junghyun Koo, and Yuki Mitsufuji. Concept-trak: Understanding how diffusion models learn concepts through concept-level attribution. *arXiv preprint arXiv:2507.06547*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aniruddh Raghu, Jonathan Lorraine, Simon Kornblith, Matthew McDermott, and David K Duvenaud. Meta-learning to improve pre-training. *Advances in Neural Information Processing Systems*, 34: 23231–23244, 2021.

Jessie Richter-Powell, Antonio Torralba, and Jonathan Lorraine. Score distillation sampling for audio: Source separation, synthesis, and beyond. *arXiv preprint arXiv:2505.04621*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35, 2022.

Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. *arXiv preprint arXiv:2503.17350*, 2025.

Yanke Song, Jonathan Lorraine, Weili Nie, Karsten Kreis, and James Lucas. Multi-student diffusion distillation for better one-step generators. *arXiv preprint arXiv:2410.23274*, 2024.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018.

Paul Vicol, Jonathan P Lorraine, Fabian Pedregosa, David Duvenaud, and Roger B Grosse. On implicit bias in overparameterized bilevel optimization. In *International Conference on Machine Learning*, pp. 22234–22259. PMLR, 2022.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiachen T Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. *arXiv preprint arXiv:2406.11011*, 2024a.

Jiachen T Wang, Tianji Yang, James Zou, Yongchan Kwon, and Ruoxi Jia. Rethinking data shapley for data selection tasks: Misleads and merits. *arXiv preprint arXiv:2405.03875*, 2024b.

Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *Advances in Neural Information Processing Systems*, 37:96541–96565, 2024c.

Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023a.

Sheng-Yu Wang, Aaron Hertzmann, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Fast data attribution for text-to-image models. *arXiv preprint arXiv:2511.10721*, 2025.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023b.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*, 2024.

Tong Xie, Haoyu Li, Andrew Bai, and Cho-Jui Hsieh. Data attribution for diffusion models: Timestep-induced bias in influence estimation. *arXiv preprint arXiv:2401.09031*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023.

Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. Intriguing properties of data attribution on diffusion models. *arXiv preprint arXiv:2311.00500*, 2023.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2535–2545, 2024.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. *arXiv preprint arXiv:2512.08931*, 2025.

A NOTATION

Table 2: Glossary and notation.

Symbol	Description
Acronyms and Basic Notation	
VAE	Variational Autoencoder
DiT	Diffusion Transformer backbone
$\mathbf{I}$	Identity matrix
Video Generation	
$p_{\boldsymbol{\theta}}(\mathbf{v} \mid \mathbf{c})$	Conditional video generator with parameters $\boldsymbol{\theta}$
$\mathbf{v} \in \mathbb{R}^{F \times H \times W \times 3}$	Video clip with frames F, height H, width W
$\mathbf{c}$	Conditioning signal such as text or multimodal metadata
$\boldsymbol{\theta}$	Trainable model parameters
$f \in \{1, \ldots, F\}$	Frame index
$h \in \{1, \ldots, H\}, w \in \{1, \ldots, W\}$	Spatial indices for height and width respectively
$\tilde{h}, \tilde{w}$	Latent grid indices
Datasets	
$\mathcal{D} = \{(\mathbf{v}_n, \mathbf{c}_n)\}_{n=1}^{N}$	Training corpus with size N and index n
$\mathcal{D}_{\text{ft}} \subseteq \mathcal{D}$	Fine-tuning dataset
$\mathcal{S} \subseteq \mathcal{D}$	Selected influential subset
$k \in \{1, \ldots, K\}$	The selected subset size
Q	Number of query data
$q \in \{1, \ldots, Q\}$	Query index
$\mathbf{x}$	Generic input data pair
$\mathbf{x}_{\text{test}}, \mathbf{x}_n$	Test/query pair and training pair
$\hat{\mathbf{v}}, \hat{\mathbf{c}}$	Query video and its conditioning
Latent Space and Diffusion Components	
E	VAE encoder
$\mathbf{h} = E(\mathbf{v}) \in \mathbb{R}^{F \times (H/s) \times (W/s) \times C}$	Latent video with spatial factor s and channels C
$\mathbf{z}$	Noisy latent variable used in diffusion or flow matching
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$	Gaussian noise
$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c}, t)$	Predicted noise network in diffusion training
$\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c}, t)$	Time-indexed vector field in flow matching
$\dot{\mathbf{z}}$	Time derivative of the latent trajectory
α_t, σ_t	Scheduler signal and noise scales at timestep t
$\boldsymbol{\epsilon}_{\text{target}}$	Target noise or velocity used for supervision
$t \in \{1, \ldots, T\}$	Diffusion or flow-matching timestep, with total timesteps T
$t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}$	Fixed timestep and shared noise draw used for low-variance gradients

B BACKGROUND

Diffusion and flow matching in latent space. Let $p_{\boldsymbol{\theta}}(\mathbf{v} \mid \mathbf{c})$ be a conditional generator with parameters $\boldsymbol{\theta}$, where $\mathbf{v} \in \mathbb{R}^{F \times H \times W \times 3}$ is a clip of height H, width W, and F frames, and $\mathbf{c}$ denotes conditioning such as text or other multimodal metadata (e.g., fps, depth, pose). We operate in VAE latents: $\mathbf{h} = E(\mathbf{v})$ and train a denoiser or velocity field on noisy latents. A noise scheduler supplies time-dependent coefficients (α_t, σ_t) controlling signal and noise scales, and the forward noising is:

$$\mathbf{z}(t, \boldsymbol{\epsilon}) \;=\; \alpha_t \, \mathbf{h} \;+\; \sigma_t \, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad t \in \{1, \ldots, T\}. \tag{13}$$

Table 3: Glossary and notation (continued).

Symbol	Description
Attribution and Influence	
$I(\mathbf{v}_n, \hat{\mathbf{v}}; \boldsymbol{\theta})$	Influence score between a train clip and a query clip
$I_{\text{mot}}(\mathbf{v}_n, \hat{\mathbf{v}}; \boldsymbol{\theta})$	Motion-aware influence score
$\mathrm{TopK}(\cdot)$	Top-K operator for selecting highest scores
$\mathrm{MajVote}(\cdot)$	Majority-vote aggregation across queries
τ	Percentile cutoff for voting
ρ	Spearman correlation coefficient
Loss Functions	
$\mathcal{L}$	Generic loss
$\mathcal{L}_{\text{diff}}(\boldsymbol{\theta}; \mathbf{v}, \mathbf{c}), \mathcal{L}_{\text{flow}}(\boldsymbol{\theta}; \mathbf{v}, \mathbf{c})$	Diffusion and flow-matching objective
$\mathcal{L}_{\text{mot}}(\boldsymbol{\theta}; \mathbf{v}, \mathbf{c})$	Motion-weighted objective used for attribution
$\tilde{\mathcal{L}}$	Per-location squared error in latent space
$\mathbf{g}, \tilde{\mathbf{g}}$	Gradient and its projected version
$\mathbf{g}_{\text{mot}}, \tilde{\mathbf{g}}_{\text{mot}}$	Motion-weighted gradient and its projection
$\mathbf{H}_{\boldsymbol{\theta}}$	Hessian with respect to $\boldsymbol{\theta}$
Motion Representations	
$\mathcal{A}(\mathbf{v}) = A$	AllTracker motion extraction
$A \in \mathbb{R}^{F \times H \times W \times 4}$	Motion tensor containing flow, visibility, and confidence
$\mathbf{D}_f(h, w)$	Displacement vector at frame f and location (h, w)
$M_f(h, w)$	Motion magnitude at a location, computed from the displacement
$\mathbf{W}(f, h, w) \in [0, 1]$	Normalized motion weights used to mask per-location losses
ζ	Small numerical bias for stability (e.g., 10^{-6})
Projections and Computational Details	
D, D'	Full and projected gradient dimensions
$\mathbf{P} \in \mathbb{R}^{D' \times D}$	Projection matrix used for Fastfood-style JL projection
ξ	Variance normalization constant for projection
$\mathcal{T}$	Set of sampled $(t, \boldsymbol{\epsilon})$ pairs for gradient estimation
B	Unit compute cost used in complexity accounting

Denoising diffusion (Ho et al., 2020) trains a network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c}, t)$ to predict the injected noise:

$$\mathcal{L}_{\text{diff}}(\boldsymbol{\theta}; \mathbf{v}, \mathbf{c}) \;=\; \mathbb{E}_{t,\boldsymbol{\epsilon}}\Big[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}(t, \boldsymbol{\epsilon}), \mathbf{c}, t) - \boldsymbol{\epsilon}\|_2^2\Big]. \tag{14}$$

Flow matching (Lipman et al., 2022; Albergo et al., 2023) learns a time-dependent vector field $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}_t, \mathbf{c}, t)$ that matches the instantaneous velocity $\dot{\mathbf{z}} = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{z}$ induced by a chosen interpolant:

$$\mathcal{L}_{\text{flow}}(\boldsymbol{\theta}; \mathbf{v}, \mathbf{c}) \;=\; \mathbb{E}_{t,\boldsymbol{\epsilon}}\Big[\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}(t, \boldsymbol{\epsilon}), \mathbf{c}, t) - \dot{\mathbf{z}}(t, \boldsymbol{\epsilon})\|_2^2\Big]. \tag{15}$$

Both objectives train time-indexed predictors over the latent space by integrating over t and $\boldsymbol{\epsilon}$, thus gradient-based methods like attribution share similar challenges.

From images to video for generation. Adding a temporal axis materially changes modeling and training. Generation must capture spatial appearance and temporal dynamics such as object and camera motion, deformations, and interactions. Modern systems extend image backbones with temporal capacity, for example, 3D U-Nets or 2D U-Nets augmented with temporal attention, causal or sliding-window context, and factorized space-time blocks, often trained in a latent-video VAE that compresses frames while preserving temporal cues. Training departs from images along several axes, which we address in §2: (i) *Compute and storage.* Longer sequences multiply the cost of sampling timesteps, noise draws, and frames, motivating fixed-timestep or small-subset estimators that reduce variance without prohibitive cost (§2.2). (ii) *Variable horizon.* Clips vary in F and frame rate (§D.1). (iii) *Time-specific failure modes.* Typical artifacts include inconsistent trajectories, temporal flicker, identity drift, and physically implausible dynamics despite sharp individual frames (§2.3).

Motion representations in videos. We denote our video as $\mathbf{v} = [\mathbf{f}_f]_{f=1}^{F}$ with $\mathbf{f}_f \in \mathbb{R}^{H\times W\times 3}$ being the f-th frame. We represent motion via optical flow between consecutive frames: $\mathbf{F}_f : \{1,\dots,H\}\times\{1,\dots,W\} \to \mathbb{R}^2$, where each flow vector in $\mathbb{R}^2$ encodes the horizontal displacement $\mathrm{d}w$ and vertical displacement $\mathrm{d}h$ of a pixel. The motion magnitude is $M_f(h,w) = \|\mathbf{F}_f(h,w)\|_2$. The M_f over frames f and pixels h, w summarizes the amount and spatial layout of motion in a clip, which we use to provide masks in our motion-weighted loss in §2.

B.1 Data Attribution

Data attribution measures how individual training samples affect a model's predictions (Bae et al., 2024). A classic approach to data attribution is to use influence functions (Koh & Liang, 2017). Intuitively, the influence of a training sample measures: if we upweight this training example, how much would the model's prediction on a test datum change? Consider a loss function $\mathcal{L}(\boldsymbol{\theta};\mathbf{x})$ and a test sample $\mathbf{x}_{\text{test}}$, the influence of a training point $\mathbf{x}_n$ can be quantified as:

$$\begin{aligned} I(\mathbf{x}_n,\mathbf{x}_{\text{test}}) \;&=\; -\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta};\mathbf{x}_{\text{test}})^{\top}\,\mathbf{H}_{\boldsymbol{\theta}}^{-1}\,\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta};\mathbf{x}_n), \\ \mathbf{H}_{\boldsymbol{\theta}} &= \tfrac{1}{N}\sum\nolimits_{n=1}^{N}\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta};\mathbf{x}_n), \end{aligned} \tag{16}$$

where the inverse Hessian captures the curvature of the loss landscape, yet computing or storing it is infeasible at modern model and dataset scales. Thus, practical methods (e.g., TracIn (Pruthi et al., 2020) and TRAK (Park et al., 2023)) approximate influence via gradient inner products or gradient feature projections.

Attribution in diffusion models. Diffusion training aggregates gradients over timesteps t and noise draws $\boldsymbol{\epsilon}$, where gradient norms vary systematically with t, producing a timestep bias where examples aligned with large-norm timesteps appear spuriously influential. Diffusion-ReTrac (Xie et al., 2024) reduces this bias by normalizing gradients and sub-sampling t and $\boldsymbol{\epsilon}$ for influence. Let $\mathcal{L}_{\text{diff}}$ denote the diffusion loss, and with the sampled-timestep-and-noise set $\mathcal{T}$, we compute a cosine-style score for normalized test and train gradients:

$$\begin{aligned} I_{\text{diff}}(\mathbf{x}_n,\mathbf{x}_{\text{test}}) = &\frac{1}{|\mathcal{T}_{\text{test}}|}\sum_{t,\boldsymbol{\epsilon}\in\mathcal{T}_{\text{test}}}\frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\mathbf{x}_{\text{test}},t,\boldsymbol{\epsilon})}{\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\mathbf{x}_{\text{test}},t,\boldsymbol{\epsilon})\right\|}^{\top} \\ &\frac{1}{|\mathcal{T}_n|}\sum_{t,\boldsymbol{\epsilon}\in\mathcal{T}_n}\frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\mathbf{x}_n,t,\boldsymbol{\epsilon})}{\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\mathbf{x}_n,t,\boldsymbol{\epsilon})\right\|}. \end{aligned} \tag{17}$$

Averaging gradients over $(t,\boldsymbol{\epsilon})$ stabilizes estimates, and normalization mitigates timestep-induced scale effects. Attribution quality is also sensitive to the measurement function used to score examples, such as denoising loss versus likelihood proxies (Zheng et al., 2023).

Why vanilla attribution is insufficient for videos. Naïvely applying gradient-based attribution to video diffusion risks treating appearance and motion alike, overemphasizing appearance (objects, textures, backgrounds) while overlooking dynamics (Park et al., 2025; Tulyakov et al., 2018). Its cost grows with clip length, sampled timesteps, noise draws, and gradient dimensionality, making naïve methods impractical at modern video scales. To improve motion, we need attribution that suppresses static appearance, emphasizes motion-specific signals, and remains efficient (§2). Motion is distributed across frames and entangled with static cues, thus influence cannot be assigned frame-independently.

C Extended Related Work

C.1 Data Attribution

Understanding how individual training examples shape model behavior has been a long-standing goal. Modern data attribution methods fall into two main groups (Hammoudeh & Lowd, 2024): retraining-based methods (e.g., leave-one-out (Cook et al., 1982; Jia et al., 2021), downsampling (also known as subsampling or counterfactual influence) (Feldman & Zhang, 2020), Shapley-value (Wang et al., 2024a;b)) and gradient-based methods (influence-function family, including Influence Functions (Koh & Liang, 2017; Lorraine & Hossain, 2024), TracIn (Pruthi et al., 2020), and TRAK (Park et al., 2023)). Influence functions provide a principled framework by approximating the effect of removing a training point. TracIn (Pruthi et al., 2020) and TRAK (Park et al., 2023) make attribution feasible at scale. While effective for classification, these assume a mapping between training gradients and predictions, which becomes more complex in generative models.

Data attribution traces how individual training examples (or subsets) influence a model's predictions or behavior. Formally, it assigns an attribution score to each training sample, estimating the extent to which that sample contributes (positively or negatively) to the model's output on a given test query or behavior. Influence data attribution is an example of nested optimization (Vicol et al., 2022; Lorraine, 2024) with other examples including differentiable games (Balduzzi et al., 2018; Lorraine et al., 2021; 2022b) and hyperparameter optimization (Raghu et al., 2021; Lorraine & Duvenaud, 2018; Mehta et al., 2024; Lorraine et al., 2020). Before diffusion models, attribution methods were applied to supervised learning tasks such as classification and regression, where influence functions (Koh & Liang, 2017; Lorraine et al., 2022a) and scalable approximations such as TracIn (Pruthi et al., 2020), TRAK (Park et al., 2023), and TDA (Bae et al., 2024) quantified the impact of training examples on downstream predictions. Recent work adapted data attribution to diffusion models (Georgiev et al., 2023; Zheng et al., 2023; Wang et al., 2025; 2024a; Lin et al., 2024; Brokman et al., 2024; Kwon et al., 2023), where iterative denoising introduces timestep-dependent bias. Mlodozeniec et al. (2024) propose scalable approximations, while Xie et al. (2024) identify timestep-induced artifacts and normalization schemes. Concept-TRAK (Park et al., 2025) extends attribution to concepts by reweighting gradients with concept-specific rewards, enabling attribution to semantic factors. Wang et al. (2023a) instead design a customization-based benchmark for text-to-image models, where models are fine-tuned on exemplar images with novel tokens and attribution is evaluated by whether it can recover the responsible exemplars. However, these are limited to image diffusion, which captures static appearance but not temporal dynamics.

C.2 Motion in Video Generation

Video diffusion extends image generation to time, requiring coherent motion across frames (Ho et al., 2022; Blattmann et al., 2023; Peebles & Xie, 2023; Wan et al., 2025; Agarwal et al., 2025). A large body of work builds temporal structure via attention layers (Wu et al., 2023), control signals (Chen et al., 2023; Zhang et al., 2023), feature correspondences (Geyer et al., 2023; Bao et al., 2023; Wang et al., 2024c), or consistency distillation (Wang et al., 2023b; Zhou et al., 2024). Recent work has highlighted the challenge of decoupling motion from appearance in video diffusion transformers, where spatial and temporal information become entangled in the model's representations (Shi et al., 2025). However, understanding which training clips influence specific motion patterns in generated videos remains an open challenge.

In parallel, motion has long been studied using optical flow and correspondence, from classical formulations (Horn & Schunck, 1981; Lucas & Kanade, 1981) to modern approaches such as RAFT (Teed & Deng, 2020), which improve accuracy and generalization. These priors are often repurposed during generation to guide dynamics, but they do not explain which training examples shaped a model's motion behavior. Our work addresses both gaps by introducing a motion-aware data attribution framework specifically designed for video diffusion. We use motion-weighted gradients that disentangle temporal dynamics from static appearance, enabling us to trace generated motion patterns back to the most influential training clips.

D Additional Details of Method

D.1 Video-specific Frame-length Bias Fix

Raw gradient magnitudes depend on the number of frames F in the video $\mathbf{v}$, thereby biasing scores toward longer videos. We correct this by normalizing for frame count before the projection–normalization step:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{v}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}) \;\leftarrow\; \frac{1}{F} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{diff}}(\boldsymbol{\theta};\, \mathbf{v}, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}). \tag{18}$$

We still apply ℓ_2 normalization in Eq. 5, further stabilizing scales across examples. Together, single-timestep, common randomness, projection, and frame-length correction form a compact, scalable estimator that we use throughout. However, naïve video-level attribution conflates appearance with motion, often ranking clips high due to shared backgrounds or objects, while offering little insight into dynamics.

D.2 Most Influential Fine-tuning Subset Selection

Goal. Given a query clip $(\hat{\mathbf{v}}, \hat{\mathbf{c}})$, we compute a motion-aware attribution value for each candidate fine-tuning sample $(\mathbf{v}_n, \mathbf{c}_n) \in \mathcal{D}_{\text{ft}}$ using $I_{\text{mot}}(\mathbf{v}_n,\, \hat{\mathbf{v}})$ (Eq. 11). We construct a fine-tuning set $\mathcal{S}$ for one or many query videos $\hat{\mathbf{v}}$.

Single-query-point fine-tuning selection. For a budget of K data points, we select the K highest-scoring examples. In practice, K is chosen as a percentile of the dataset size (e.g., top 1–10%), ensuring the subset scales consistently across datasets.

Multi-query-point fine-tuning selection: aggregating attribution scores. For Q queries, we adopt the majority voting approach from ICONS (Wu et al., 2024) and aggregate motion-aware influence scores across queries by percentile thresholding and voting. A sample receives a vote if the score is above the percentile cutoff τ for that query. The consensus score of a candidate $\mathbf{v}_n$ is the total number of queries that vote for it. We then rank all training samples by $\mathrm{MajVote}(\mathbf{v}_n)$ and select the top-K to form the fine-tuning subset. This formulation emphasizes samples that are consistently influential across multiple queries, without requiring cross-query calibration of raw scores:

$$\begin{aligned} \mathrm{MajVote}_n &= \sum\nolimits_{q=1}^{Q} \mathbb{I}\big[\, I_{\text{mot}}(\mathbf{v}_n, \hat{\mathbf{v}}_q) > \tau \,\big], \\ \mathcal{S}_{\text{vote}}(K) &= \big\{ \mathbf{v}_n \,|\, \mathbf{v}_n \text{ in top-}K \text{ by } \mathrm{MajVote} \big\}. \end{aligned} \tag{19}$$

D.3 Computational Efficiency Analysis

Gradient Compute. Naïvely averaging over timesteps and noise for every example costs $\mathcal{O}(|\mathcal{D}|\,|\mathcal{T}|\,B)$, where B is a single forward+backward cost and $|\mathcal{T}|$ is the number of sampled $t, \boldsymbol{\epsilon}$ per data. Using a single sample reduces this to $\mathcal{O}(|\mathcal{D}|\,B)$, which is key to keeping the cost reasonable for modern video datasets and models, while reusing a sample across data allows low enough variance for stable rankings. Projection adds $\mathcal{O}(D' \log D')$ per example using Fastfood (Le et al., 2014), negligible relative to a backward pass.

Gradient Storage. Storing full gradients is $\mathcal{O}(|\mathcal{D}|\,D)$. We instead store only projected vectors, $\mathcal{O}(|\mathcal{D}|\,D')$, plus the structured Fastfood state, $\mathcal{O}(D)$. Since D' is typically orders of magnitude smaller than D, this transformation makes storage tractable for billion-parameter models.

Data Ranking Compute. Influence computation in Eq. 5 is an inner product in $\mathbb{R}^{D'}$, so evaluating all train examples against a query is $\mathcal{O}(|\mathcal{D}|\,D')$, and sorting is $\mathcal{O}(|\mathcal{D}| \log |\mathcal{D}|)$.

Additional Motion-Emphasis Compute. Motion-specific overhead primarily stems from AllTracker mask extraction with complexity $\mathcal{O}(|\mathcal{D}| \cdot H \cdot W \cdot F)$ for clip length F and frame resolution $H \times W$. Masks are extracted once, cached, and negligible relative to gradient cost. We provide a detailed runtime breakdown in App. I.1.

E Additional Experiments

Setup. We evaluate on **`VIDGEN-1M`** (Tan et al., 2024) and **`4DNeX-10M`** (Chen et al., 2025) using 10k videos. Query videos span 10 motion categories (compress, bounce, roll, spin, slide, free

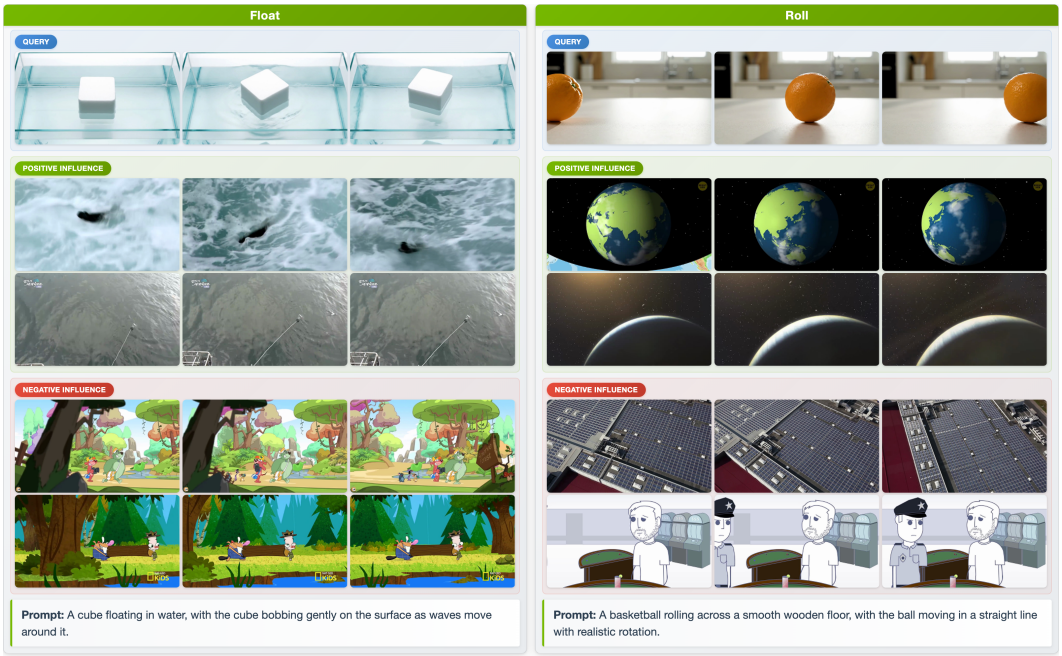


Figure 2: **Motion Attribution Samples** with Wan2.1-T2V-1.3B. *Top*: Query clips showing float (*left*) and roll (*right*) motions. *Middle*: Top-ranked positive training samples identified by MOTIVE with high influence scores. *Bottom*: Negative influence samples with minimal, camera-only motion, or cartoon-style content that conflict with target motions. Videos are included in the supplementary material.

fall, float, inflate, deflate, oscillate) with 5 videos each. We fine-tune `Wan2.1-T2V-1.3B` and `Wan2.2-TI2V-5B` models, updating only the DiT backbone with LoRA (rank 128). Baselines include random selection, motion magnitude filtering, V-JEPA embeddings (Assran et al., 2025), and whole-video attribution without motion masking. We evaluate using VBench (Huang et al., 2024) metrics: motion smoothness, dynamic degree, aesthetic quality, and imaging quality.

Quantitative Results. Tab. 1 shows MOTIVE achieves the highest dynamic degree (47.6% and 48.3%) across both models, significantly outperforming random selection (+6.3% / +6.7%) and whole-video attribution without motion masking (+3.8% / +4.5%). Notably, using only 10% of training data, we surpass full fine-tuning on dynamic degree by +5.6% / +3.0%, demonstrating superior performance of motion-specific attribution for targeted fine-tuning. Our method also maintains competitive motion smoothness (96.3% / 97.6%) and improves aesthetic quality (46.0% / 45.6%).

Human Evaluation. We conducted a human study with 17 annotators comparing 50 generated videos across methods. Results show 74.1% preference for MOTIVE over the base model and 53.1% over full fine-tuning, confirming that motion-attributed data selection produces perceptually better motion. Additional results on `LTX-2B` are in the Appendix.

Qualitative Analysis. Fig. 3 shows example attributions. Our method ranks clips with coherent physical dynamics highly, objects floating, rolling, and spinning with realistic physics, and downranks static footage, camera-only motion, and cartoon-style content with non-transferable kinematics. Positive-influence samples share motion semantics with queries (e.g., buoyant objects for "float", rotating objects for "roll"), while negative-influence samples contain minimal motion or incompatible dynamics. This pattern holds across all motion categories, aligning with the quantitative gains.

Ablations. We validate key design choices: (1) Single-timestep attribution ($t = 751$) achieves 66% Spearman correlation with multi-timestep ground truth while reducing compute by $10\times$. (2) Projection to 512 dimensions preserves 74.7% ranking correlation with full gradients.

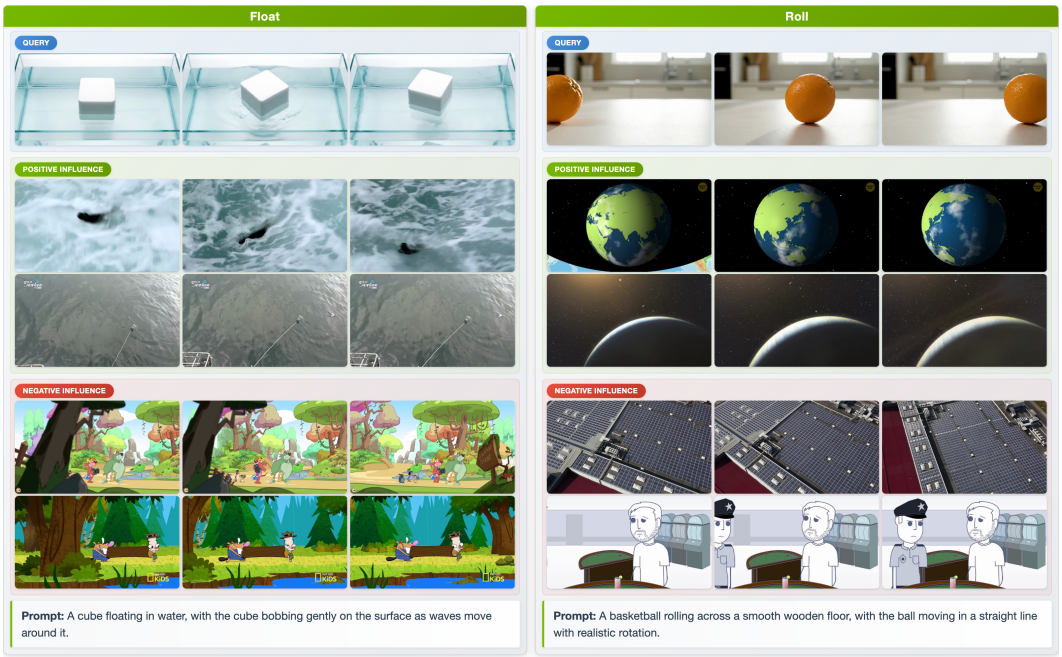


Figure 3: **Motion Attribution Examples.** *Top*: Query clips showing float (left) and roll (right) motions. *Middle*: Top-ranked positive training samples with high influence scores showing similar motion patterns. *Bottom*: Negative influence samples with static content or cartoon-style motion that conflicts with target dynamics.

Table 4: **VBench Evaluation on LTX-2B.** Performance comparison on VBench (Huang et al., 2024) across different baselines (all values in %, higher is better). All selection methods use 10% of training data; MM: motion masking.

	LTX-2B Video					
Method	Subject Consist.	Background Consist.	Motion Smooth.	Dynamic Degree	Aesthetic Qual.	Imaging Qual.
Base	93.8	95.2	94.6	36.5	32.1	48.2
Full FT	94.4	**95.7**	94.8	38.9	32.6	**49.1**
Random	93.7	95.4	94.7	37.8	**32.8**	48.4
Motion mag.	94.1	95.3	95.2	39.6	32.4	47.9
V-JEPA	94.2	95.5	95.1	40.2	31.9	47.5
Ours w/o MM	94.3	95.6	95.3	41.8	**32.8**	48.7
Ours (MOTIVE)	**94.8**	95.6	**95.5**	**45.1**	32.7	48.6

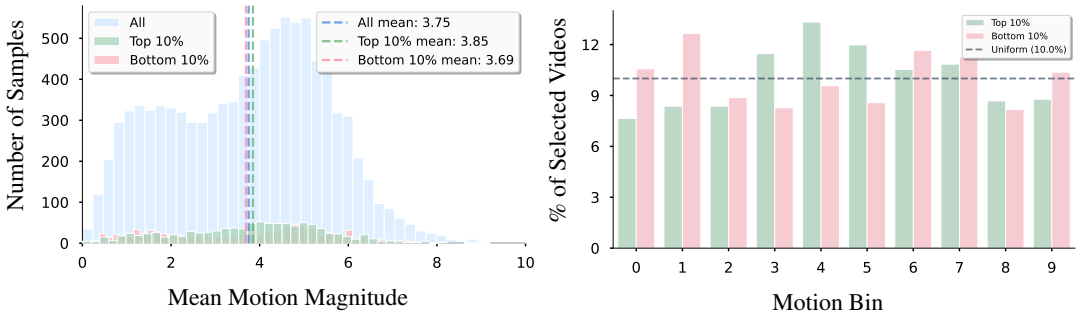
E.1 Results on Additional Video Generation Models

We further test our framework on additional video generation architectures beyond the **`Wan`** models (Tab. 1). We evaluate **MOTIVE** on **`LTX-2B`** HaCohen et al. (2024), which represents a different architectural design and training paradigm. The results in Tab. 4 demonstrate that **MOTIVE** works effectively across different model architectures and scales.

F Analysis

F.1 Motion Distribution Analysis

MOTIVE is not simply selecting "motion-rich" clips: The key distinction is that our influence scores are computed via gradients, and training videos are considered influential only when they directly allow the model to lower the loss, improving the model's ability to generate the target motion dynamics, not because they contain more motion overall.



(a) Motion Distribution (b) Distribution Across Motion Bins

Figure 4: **MOTIVE is not simply selecting "motion-rich" clips.** Our influence scores are computed via gradients, and training videos are considered influential only when they directly improve the model's ability to generate the target motion dynamics, not because they contain more motion overall.

To empirically validate this, we further analyze the distribution of motion magnitudes in our selected data. We compute the mean motion magnitude for 10k videos in the **VIDGEN** dataset and compare the distributions of the top 10% (highest influence scores) and the bottom 10% (lowest influence scores).

As shown in Fig. 4, the top 10% selected videos have a mean motion magnitude of 3.85, which is only 4.3% higher than the bottom 10% (3.69), despite representing opposite extremes of influence scores. The analysis also shows that within the moderate-motion range (bins 3, 4, and 5), the top 10% of positive-influence samples outnumber the bottom 10% of negative-influence samples. Yet, both groups also appear in low-motion bins (0-2) and high-motion bins (6-9).

This distribution pattern shows that high-influence videos selected by MOTIVE span the entire motion spectrum, not just high-motion regions. Many high-motion videos receive low influence scores, while numerous influential videos exhibit modest or even low motion magnitude. These findings show that our motion attribution approach captures training influence, focusing on motion rather than simply acting as a motion-saliency filter.

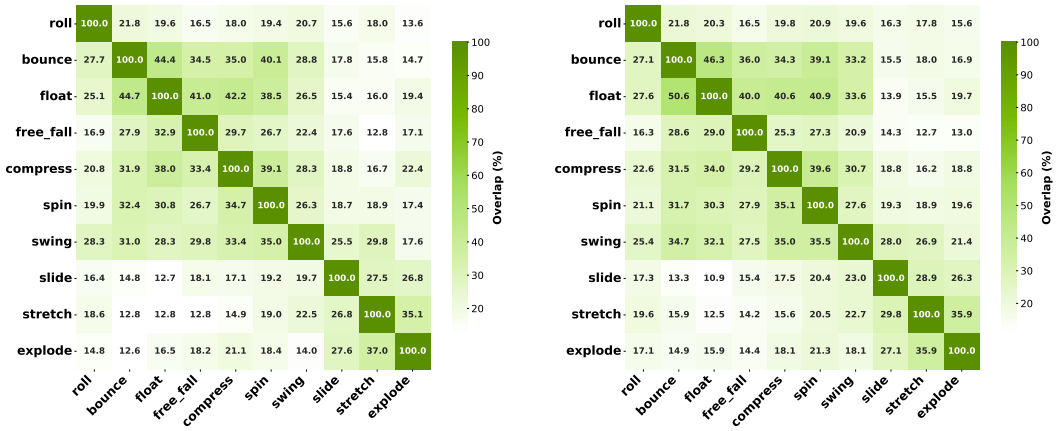
F.2 Cross-Motion Influence Patterns

To analyze cross-motion influence patterns, we examine the percentage overlap of top-100 influential training data across different motion categories in both 4DNEX and VIDGEN datasets. We use 5 query samples to identify the top-100 most influential training videos, aggregating results across queries. As shown in Fig. 5, both datasets exhibit remarkably similar patterns with mean overlaps of 24.0% and 24.3%, respectively, indicating moderate sharing of influential data across motion categories.

Both datasets consistently identify the same high-overlap pairs: bounce-float (44.4%/46.3%), compress-float (40.1%/34.0%), and compress-spin (36.9%/39.6%), suggesting these motions share fundamental characteristics that the model learns from similar training examples. Conversely, low-overlap pairs such as free fall-stretch (12.8%/12.7%) and float-slide (14.0%/10.9%) indicate more specialized influential data for mechanically dissimilar motions. The influence matrices are asymmetric because the number of unique influential samples shared across the 5 query samples differs across motion categories. The similar cross-motion influence patterns observed across both the 4DNEX and VIDGEN datasets demonstrate that these relationships are generalizable across different video datasets and reflect dynamic similarity.

G Additional Method Details

Tracker-agnostic scope. We treat the motion estimator as a pluggable source of saliency rather than a training dependency. Given displacement magnitudes, we construct latent-space weights via bilinear mapping and normalization. Our implementation supports alternative estimators (such as dense optical flow or point tracking) with identical interfaces, enabling users to swap AllTracker without modifying the attribution code.



(a) 4DNEX Influence Heatmap

(b) VIDGEN Influence Heatmap

Figure 5: **Cross-motion influence overlap across datasets.** Heatmaps showing the percentage overlap of top-100 influential training samples across motion categories for (a) 4DNEX and (b) VIDGEN datasets. Each cell (i, j) shows the percentage of motion category i's influential data (aggregated from 5 queries per category) that also appears in motion category j's top-100 influential samples. The asymmetric nature of the matrices (e.g., bounce→float $\neq$ float→bounce) arises because different motion categories have different numbers of unique influential videos, leading to directional overlap percentages. Consistent high-overlap pairs (e.g., bounce-float: $44.4\%/46.3\%$) and low-overlap pairs (e.g., free fall-stretch: $12.8\%/12.7\%$) across datasets validate that these influence patterns reflect fundamental aspects of motion representation in video generation models.

Algorithm 1 MOTIVE: Motion-Aware Data Attribution Framework

Require: fine-tuning corpus $\mathcal{D}_{\text{ft}} = \{(\mathbf{v}_n, \mathbf{c}_n)\}_{n=1}^{N}$, query video $(\hat{\mathbf{v}}, \hat{\mathbf{c}})$, fixed $(t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})$, projection matrix $\mathbf{P}$

Ensure: Motion-aware influence scores $\{I_{\text{mot}}(\mathbf{v}_n, \hat{\mathbf{v}})\}_{n=1}^{N}$

for $(\mathbf{v}_n, \mathbf{c}_n) \in \mathcal{D}_{\text{ft}}$ **do**
  $A_n \leftarrow \text{ALLTRACKER}(\mathbf{v}_n)$ {Extract per-pixel flow displacements $\mathbf{D}_f$ (Eq. 6)}
  Downsample and normalize to latent-space motion mask $\mathbf{W}_n$ (Eqs. 7–9)
  Evaluate motion-weighted loss $\mathcal{L}_{\text{mot}}(\boldsymbol{\theta}; \mathbf{v}_n, \mathbf{c}_n)$ (Eq. 10)
  Compute motion gradient $\mathbf{g}_{\text{mot}}(\boldsymbol{\theta}, \mathbf{v}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}}) = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{mot}}(\boldsymbol{\theta}; \mathbf{v}_n, \mathbf{c}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})$
  Normalize by frame length: $\mathbf{g}_{\text{mot}} \leftarrow \mathbf{g}_{\text{mot}}/F$
  Project motion gradient: $\tilde{\mathbf{g}}_{\text{mot}}(\boldsymbol{\theta}, \mathbf{v}_n) := \frac{\mathbf{P}\mathbf{g}_{\text{mot}}(\boldsymbol{\theta}, \mathbf{v}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})}{\|\mathbf{P}\mathbf{g}_{\text{mot}}(\boldsymbol{\theta}, \mathbf{v}_n, t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})\|}$ (Eq. 11)
end for
Compute query gradient: $\tilde{\mathbf{g}}_{\text{mot}}(\boldsymbol{\theta}, \hat{\mathbf{v}})$ using the same procedure for $(\hat{\mathbf{v}}, \hat{\mathbf{c}})$
for $n = 1, \ldots, N$ **do**
  $I_{\text{mot}}(\mathbf{v}_n, \hat{\mathbf{v}}) = \tilde{\mathbf{g}}_{\text{mot}}(\boldsymbol{\theta}, \hat{\mathbf{v}})^{\top} \tilde{\mathbf{g}}_{\text{mot}}(\boldsymbol{\theta}, \mathbf{v}_n)$ (Eq. 11)
end for
Rank all training clips by $I_{\text{mot}}(\mathbf{v}_n, \hat{\mathbf{v}})$ and select top-K influential samples using majority vote aggregation (Eq. 19):
$\mathcal{S} = \mathcal{S}_{\text{vote}}(K) = \{\mathbf{v}_n \mid \mathbf{v}_n \text{ in top-}K \text{ by } \text{MajVote}\}$
return $\mathcal{S}$

Model-agnostic scope. Our attribution only requires per-example gradients under matched $(t_{\text{fix}}, \boldsymbol{\epsilon}_{\text{fix}})$, and therefore applies to both diffusion and flow-matching objectives. The score reduces to a gradient inner product under a fixed preconditioner; the generator architecture affects gradient statistics but not the definition of influence. In practice, replacing the denoiser or velocity field leaves the weighting and aggregation unchanged.

Algorithm Summary. For completeness, Algorithm 1 summarizes the full MOTIVE pipeline, detailing the computation of motion-weighted gradients, projection into low-dimensional space, and the subsequent influence-based ranking and selection of training clips.

H Additional Experiment Details

H.1 Hyperparameter Settings

For reproducibility, we document the hyperparameters used throughout attribution, subset selection, and fine-tuning. Where values were not explicitly tuned, we adopted defaults from DiffSynth-Studio and the official Wan repo for Wan models, and Diffusion-Pipe for LTX models.

Attribution. Motion-aware influence estimation is computed at a single fixed timestep corresponding to the midpoint of the denoising trajectory, which strongly correlates with multi-timestep averaging. A shared Gaussian draw $\boldsymbol{\epsilon}_{\text{fix}} \sim \mathcal{N}(0, \mathbf{I})$ is used across all training–query pairs to reduce stochastic variance. Gradients are projected from dimension $D = 1418996800$ to $D' = 512$ using a Fastfood Johnson–Lindenstrauss projection $\mathbf{P}$ selected via the search to balance performance and storage. Motion weights $\mathbf{W}$ are computed from AllTracker flow magnitudes M_f, min–max normalized to $[0, 1]$ with a small bias $\zeta = 10^{-6}$. All computations use bfloat16 precision for memory efficiency.

Subset Selection & fine-tuning. For any number of query points, we select top-10% data of the datasets. We finetune the backbone while freezing both the text encoder (Raffel et al., 2020) and the VAE. The input resolution is fixed to 480×832 pixels. We use a learning rate of 1×10^{-5} and the AdamW optimizer (Loshchilov & Hutter, 2017) following the DiffSynth-Studio defaults. We train the models for 1 epoch, repeating the dataset 50 times.

Evaluation. The test set consists of the same 10 motion categories, but with different visual appearances, compared with the query set. We provide the prompt samples below.

Samples from Query Set

We illustrate representative prompts from our query set used to generate query videos with Veo-3.

[leftmargin=*, itemsep=2pt, parsep=0pt]

- **compress**: "A slice of sandwich bread flattened by a flat metal plate, steady camera, soft studio lighting, plain backdrop; emphasize air pockets collapsing."
- **bounce**: "A ping-pong ball bouncing on a white table, steady side camera, neutral light, seamless backdrop; emphasize consistent bounce height and timing."
- **roll**: "A spool of thread rolling from left to right, close-up static camera, bright studio light; highlight axle rotation and smooth travel."
- **explode**: "A single balloon bursting into fragments, captured in high-speed slow motion with a fixed camera, bright even lighting, seamless background; emphasize outward debris and air release."
- **float**: "A foam cube floating on the surface of water, static overhead camera, bright light, clean tank; emphasize buoyancy and slight rocking."

Samples from Test Set

We illustrate representative prompts from our test set that our fine-tuned models use to generate test videos.

[leftmargin=*, itemsep=2pt, parsep=0pt]

- **compress**: "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."
- **bounce**: "A basketball bouncing vertically on a wooden court plank, unmoving camera, balanced indoor lighting, plain wall background; clearly show deformation at impact."
- **roll**: "A bike tire rolling freely on a stand, static side camera, indoor neutral light; show uniform rotation without wobble."
- **explode**: "A fragile glass ornament breaking apart mid-air, fixed camera, bright controlled lighting, plain backdrop; capture shards and reflections crisply."
- **float**: "A green leaf floating gently on perfectly still water in a transparent tank, fixed top-down camera, bright even lighting; emphasize surface tension ripples."

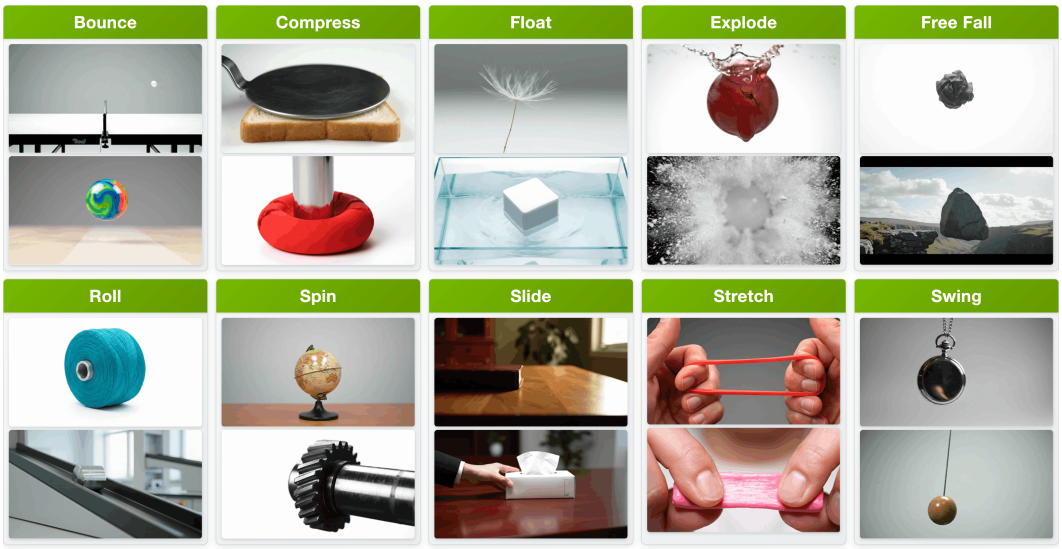


Figure 6: **Illustration of motion query set.** We generate near-realistic video queries with Veo-3 across ten motion categories. Each category contains five query videos synthesized with controlled prompts and manually screened for clarity and physical plausibility.

H.2 Details on Motion Query Data

A small, controlled set of query videos is constructed to isolate specific motion primitives while minimizing confounding factors (e.g., textured backgrounds, uncontrolled camera motion). Such clean and consistent clips are challenging to obtain from natural data sources. To address this, we synthesize the query set using Veo-3 (Google DeepMind, 2025) and apply a strict post-generation screening for physical plausibility and generation realism. We target ten motion types: *compress, bounce, roll, explode, float, free fall, slide, spin, stretch, swing*. For each category, we retain 5 query samples, yielding a total of 50 queries. This scale provides adequate coverage of the motion taxonomy used in our evaluations while maintaining tractable attribution computation. We further provide a few examples of the generation prompts and the generated video query set in Fig. 6.

Rationale for synthetic queries. The query set is not used as training data; instead, it specifies targets for attribution and for multi-query aggregation. Synthetic generation offers controllability that is difficult to achieve at scale with web videos. This design yields near-realistic yet standardized stimuli aligned with our goal of probing motion-specific influence.

I Discussion

I.1 Runtime

All training runs are conducted on 4-8 NVIDIA A100 GPUs. We provide a detailed runtime breakdown from our experiments on 10k training samples with `Wan2.1-T2V-1.3B` model to address scalability concerns. The key insight is that the dominant cost of our pipeline is computing per-training-sample gradients, which is performed once and then can be reused for all subsequent queries. Each training clip's gradient is projected into a compact 512-dimensional vector, and adding a new query requires only (i) a single backward pass to obtain its own projected 512-dimensional gradient and (ii) computing cosine similarity between the query vector and stored training vectors, which is exceptionally lightweight (on the order of seconds). Thus, the computational burden does not scale with the number of queries but only with the size of the training set. As shown in Tab. 5, the dominant cost is the one-time training data gradient computation for 10k samples ($\sim$ 150 hours on 1 A100), which is amortized across all queries. Once computed, adding a new query requires only $\sim$ 54 seconds (gradient computation) + 46ms (influence computation) = $\sim$54 seconds total. The training data gradient computation is embarrassingly parallel and can be reduced to $\sim$2.3 hours with 64 GPUs.

Table 5: **Runtime Breakdown.** Detailed computational complexity and runtime for each component of our motion attribution framework on 10k training samples with `Wan2.1-T2V-1.3B` model.

Component	Complexity	Runtime	Notes
Gradient computation	$\mathcal{O}(B)$ per sample	Query: $\sim$ 54 seconds Training: $\sim$ 150 hours	1 A100 GPU; Single forward+backward pass; training is dominant cost but amortized over all queries; embarrassingly parallel (with 64 GPUs, $\sim$ 2.3 hours)
Projection	$\mathcal{O}(\lvert\mathcal{D}\rvert \cdot D' \log D')$	$\sim$ 1.97 seconds per sample	$D' = 512$
Influence computation	$\mathcal{O}(\lvert\mathcal{D}\rvert \cdot D')$	$\sim$ 46 milliseconds per query	1 query $\times$ 10k training samples
Majority-vote aggregation	$\mathcal{O}(\lvert\mathcal{D}\rvert \cdot q)$	$\sim$ 139 milliseconds	50 queries $\times$ 10k samples

Table 6: **Runtime Comparison with Baselines.** Total computational time required for processing 10k training samples with `Wan2.1-T2V-1.3B` model on a single GPU across different data selection methods.

Method	Random	Motion Magnitude	Optical Flow	V-JEPA	Ours
Total for 10k (1 GPU)	$<$ 1 second	$\sim$ 5.5 hours	$\sim$ 5.7 hours	$\sim$ 3 hours	$\sim$ 150 hours

Runtime comparison with baselines. We compare the computational cost of our method with the baseline approaches for processing 10k training samples on a single GPU (Table 6). While our method requires more upfront computation than the baseline approaches, this cost is amortized across all queries, and the computed gradients can be reused for multiple selection queries, making it practical for large-scale data curation scenarios.

I.2 Limitations

Gradient-based attribution is computationally expensive, requiring high upfront cost for per-sample gradient computation (see §I.1), though this cost is amortized across queries. Our analysis treats each video as a whole unit, thereby avoiding collapsing motion into frame-level appearance, but it risks overlooking the fact that only certain segments may carry motion-relevant information. Highly informative intervals can be diluted when averaged with static or redundant portions of the same clip. This suggests an open direction toward finer-grained attribution at the motion segment or motion event level, which could reveal more precise insights into how different phases of a trajectory shape motion learning. Another limitation is that our motion masks may overemphasize camera-only motion; we detect this by spatial uniformity of $\mathbf{W}$ and down-weight such clips, but a full disentanglement of ego and object motion remains future work.

Additionally, our framework does not explicitly account for classifier-free guidance (CFG), which is widely used in practice to steer video generation but introduces discrepancies between training-time attribution and inference-time dynamics. As a result, our influence estimates may not fully capture how guidance alters motion behavior. In addition, while attribution-driven fine-tuning improves targeted motion quality, it may introduce trade-offs with base model capabilities. This raises the need for future work on balancing targeted motion adaptation with the preservation of broader generative capabilities.

I.3 Future Directions

Improving the attribution pipeline. Several directions can strengthen our motion attribution methodology. *Tracker-robust motion saliency*: replace or ensemble AllTracker with alternative estimators and use its confidence/visibility channels to weight masks. *Self-generated video queries*: use model-generated videos as queries to trace problematic motion patterns (e.g., unrealistic physics) back to training data, enabling iterative diagnostics and targeted motion improvement.

Scaling data curation. *Closed-loop data curation*: move from one-shot ranking to active selection: iteratively attribute, finetune, and re-attribute, or replace simple majority voting with learned query weights. *Sophisticated finetuning*: move to more sophisticated finetuning setups, such as multi-student distillation (Song et al., 2024).

Extending to new domains. *Other modalities*: extend our methodology to other modalities, including world models (Zhu et al., 2025), audio (Evans et al., 2025; Richter-Powell et al., 2025), or video+audio (Wiedemer et al., 2025).

Safety and governance. Use negative-influence filtering to suppress undesirable or unsafe dynamics, document curator choices, and audit motion behaviors exposed by our framework.

J VISUALIZATION

J.1 MOTION VISUALIZATION

To provide intuition for the behavior of our motion-weighted loss, we visualize the motion magnitude as an overlay. We compute per-pixel motion magnitude using optical flow and apply motion-based weighting that preserves the appearance of dynamic regions while attenuating static backgrounds. This motion overlay directly illustrates the spatial weighting applied by our motion loss during training.

Fig. 7 presents representative frames from our dataset, comparing original frames with their corresponding motion overlays for seven distinct video samples. These visualizations show that the motion-weighted loss preferentially emphasizes dynamic content while down-weighting static scene elements.

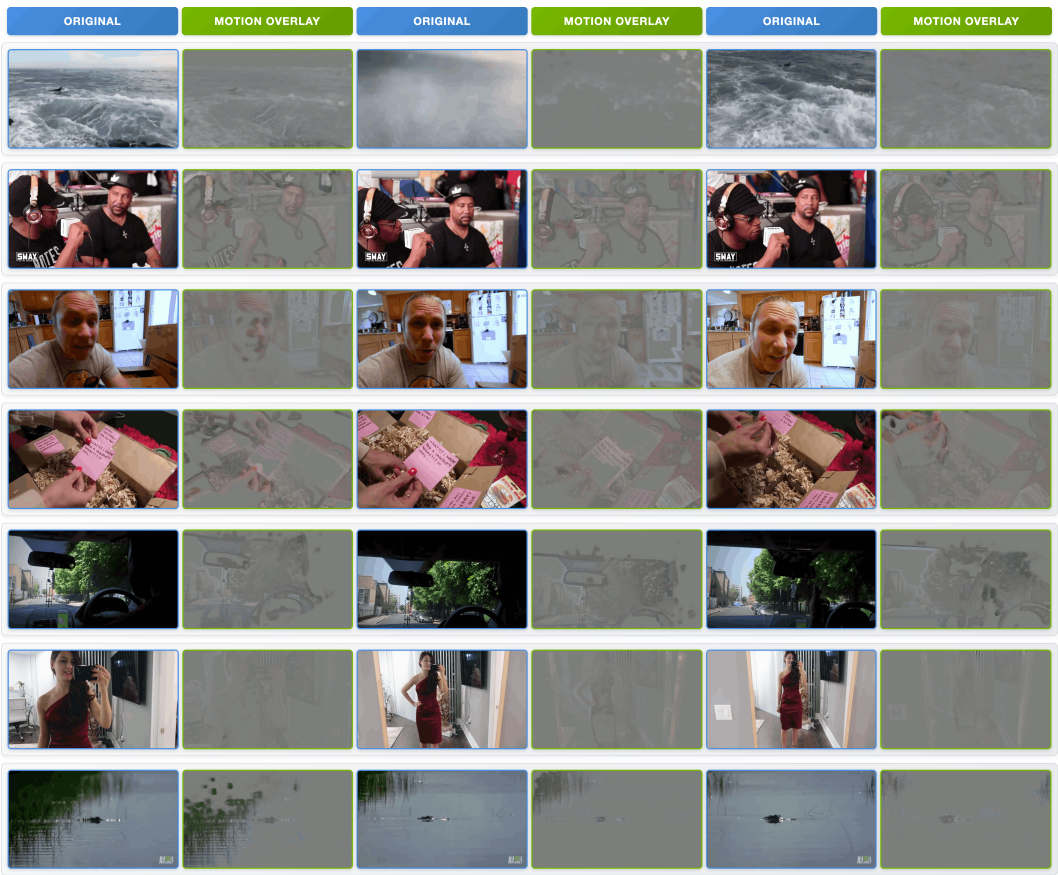


Figure 7: **Motion Overlay Visualization.** Comparison of original frames and motion overlays for seven video samples across three time points (early, middle, late). The motion overlay demonstrates the spatial weighting of our motion loss: dynamic regions remain visible, while static backgrounds are attenuated to neutral gray. *Takeaway:* This provides heuristic intuition into what information our motion attribution focuses on: the information in grayer regions, which lack motion, is down-weighted by our method.