FlowMo: Variance-Based Flow Guidance for Coherent Motion in Video Generation

Ariel Shaulov*, Itay Hazan*, Lior Wolf, Hila Chefer

The Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University

NEURAL INFORMATION PROCESSING SYSTEMS

The Challenge: Coherent Motion in Video Generation

Text-to-video diffusion models are notoriously limited in their ability to model temporal aspects such as motion, physics, and dynamic interactions. Existing approaches address this limitation by retraining the model or introducing external conditioning signals to enforce temporal consistency.

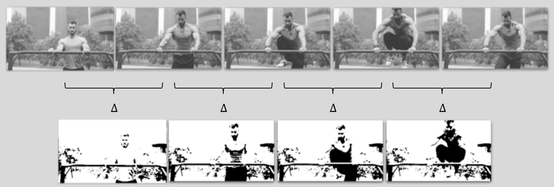
Motivation

We define a measure called **Patch-Wise Variance** and observe that videos with coherent motion have lower patch-wise variance than incoherent ones.

Our New Measure: Patch-Wise Variance

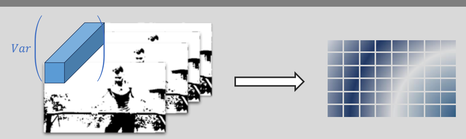
Step 1: Appearance Debiasing on Each Latent Channel

$$\forall f \forall w \forall h \forall c$$
$$\left(\Delta u_{\theta,t}\right)_{f,w,h,c} =$$
$$\left\| \left(\Delta u_{\theta,t}\right)_{f+1,w,h,c} - \left(\Delta u_{\theta,t}\right)_{f,w,h,c} \right\|_1$$



Step 2: Patch-Wise Variance on Each Latent Channel

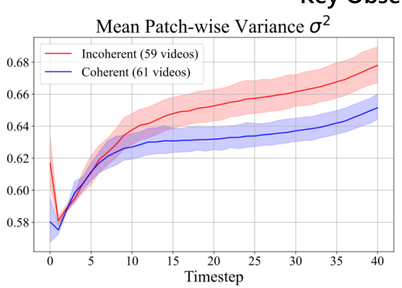
$$\forall w \forall h \forall c$$
$$\sigma^2_{w,h,c} = \mathbb{V}\mathrm{ar}_f\left[\left(\Delta u_{\theta,t}\right)_{f,w,h,c}\right]$$



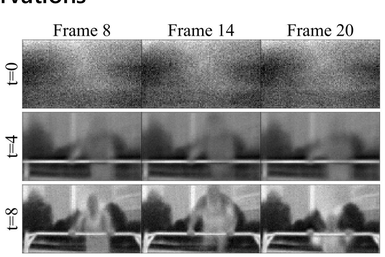
Step 3: Mean Across Latent Channels

$$\forall w \forall h$$
$$\mathrm{PWV}_{w,h} = \mathbb{E}_c\left[\sigma^2_{w,h,c}\right]$$

Key Observations



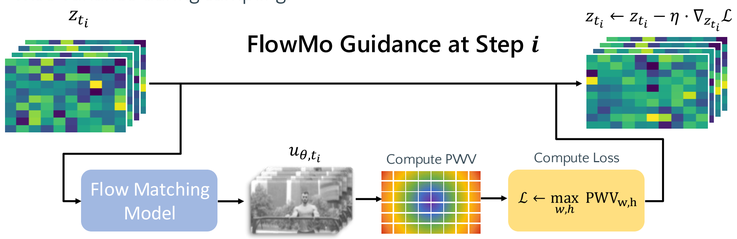
Coherent motion corresponds to lower Patch-Wise Variance



Coarse structure forms at steps 0–4 and motion emerges around steps 5–8

Our Method: FlowMo

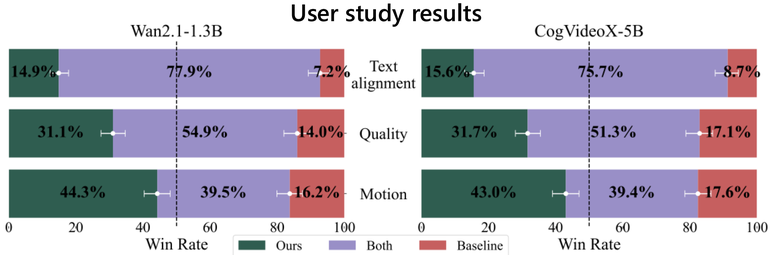
FlowMo is a training-free guidance method that enhances motion coherence using only the model's own predictions in each diffusion step. We use our newly-defined variance measure to steer the model toward coherent motion by reducing patch-wise variance during sampling.



We apply FlowMo guidance in the first 12 timesteps since these are responsible for coarse structure and motion

Quantitative Results

User study results



VBench metrics

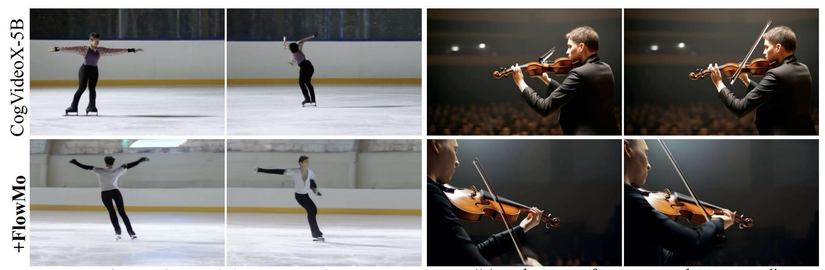
Models	Motion Metrics		Aggregated Scores		
	Motion Smoothness	Dynamic Degree	Semantic Score	Quality Score	**Final Score**
Wan2.1-1.3B	96.43%	**83.21%**	84.70%	65.58%	75.14%
+ FlowMo	**98.56%**	81.96%	**89.11%**	**73.58%**	**81.34%** (+6.20%)
CogVideoX-5B	95.01%	**65.29%**	**70.03%**	60.83%	65.43%
+ FlowMo	**97.29%**	63.92%	69.26%	**72.11%**	**70.69%** (+5.26%)

Qualitative Results

"A woman performing a challenging exercise" *"A dolphin jumping out of ocean waves"*

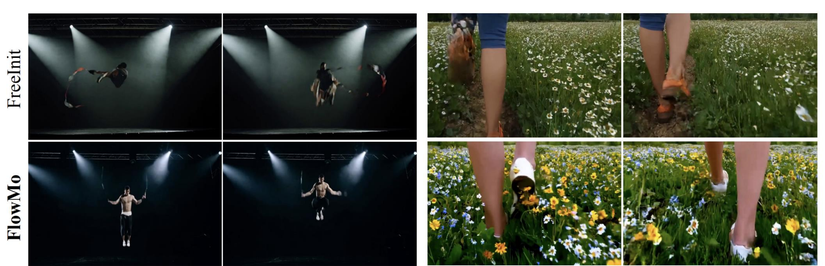
Qualitative Results

"A roulette wheel in a dimly lit room or casino floor" *"A boy with glasses flying a kite in a grassy field"*



"A figure skater gliding across the ice" *"A violinist performing a solo on stage"*

"A pair of flamingos wading through shallow water" *"A ballerina leaping through the air"*



"A man jumping rope on a dark stage" *"A close-up of a person's feet as they walk"*

References

[1] Huang et al. (2024). "**VBench: Comprehensive benchmark suite for video generative models.**". In: CVPR 2024
[2] Chefer et al. (2025). **"Videojam: Joint appearance-motion representations for enhanced motion generation in video models."**. In: ICML 2025