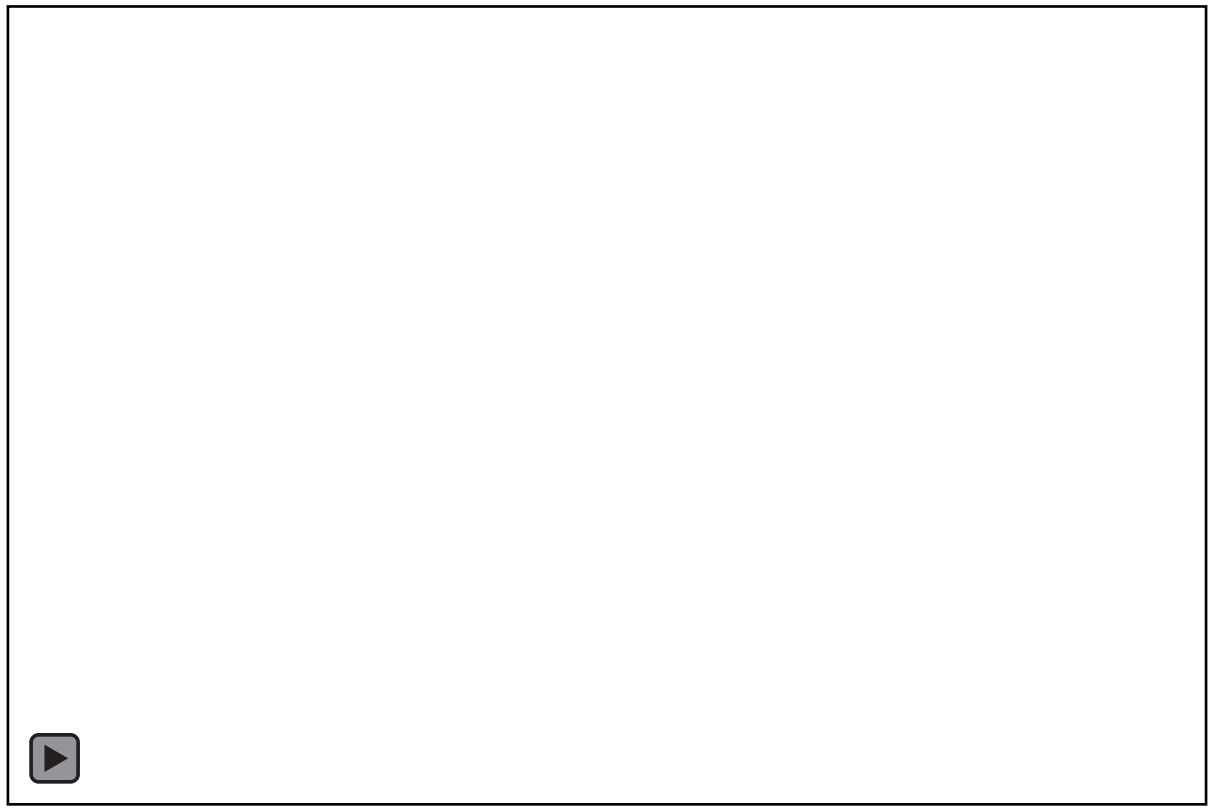
FlowMo: Variance-Based Flow Guidance for Coherent Motion in Video Generation

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

Tel Aviv University, * Denotes equal contribution

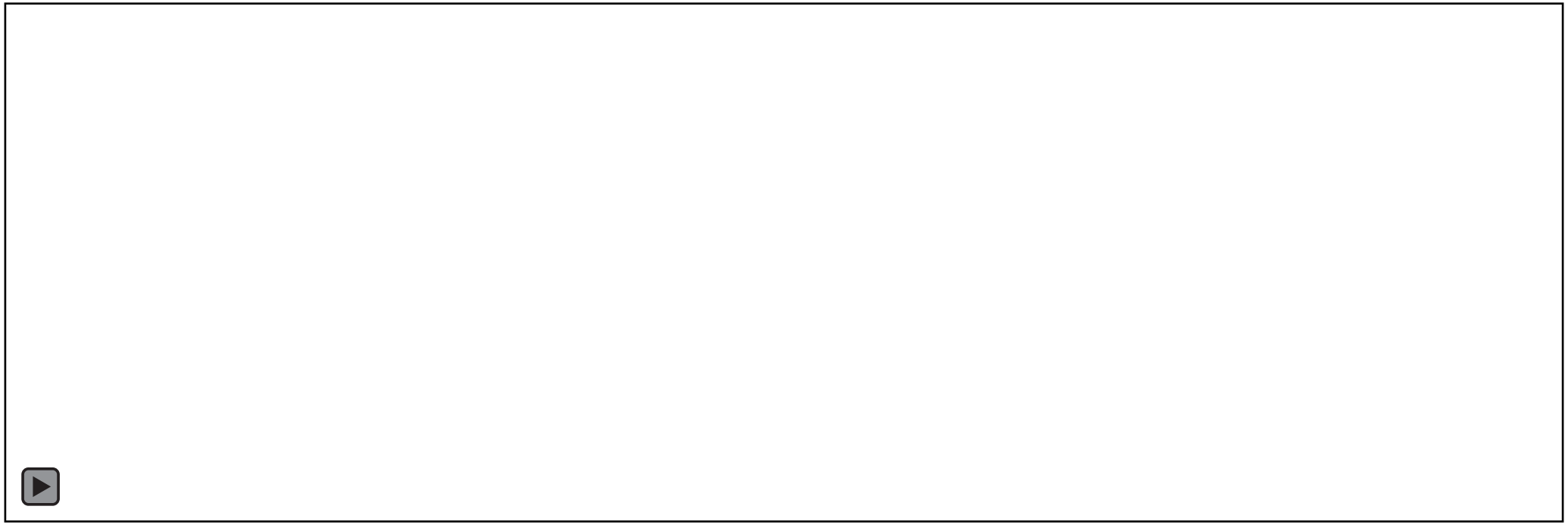
Problem Statement

Text-to-video diffusion models suffer from significant temporal artifacts that limit their ability to generate coherent motion sequences.



Motivation

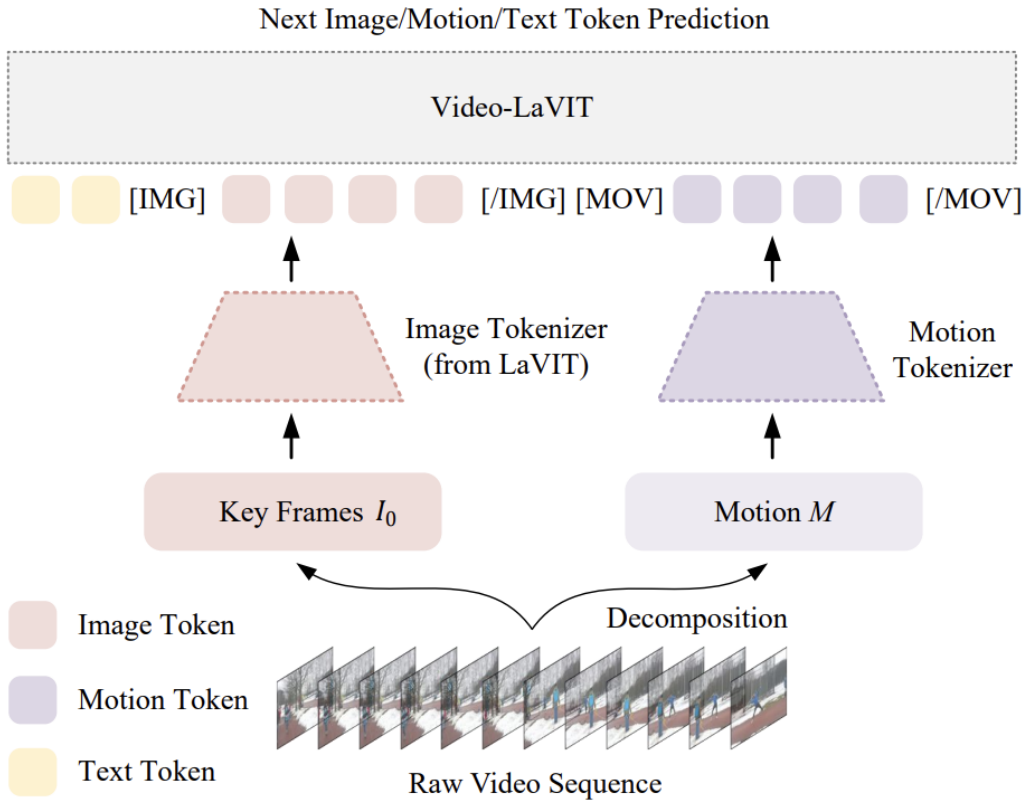
Can we extract meaningful temporal representations directly from a pre-trained model without additional training or external inputs?



Existing Approaches and Their Limitations

Training-Based Methods:

- training with temporal objectives improving consistency but demands significant compute and access to training data

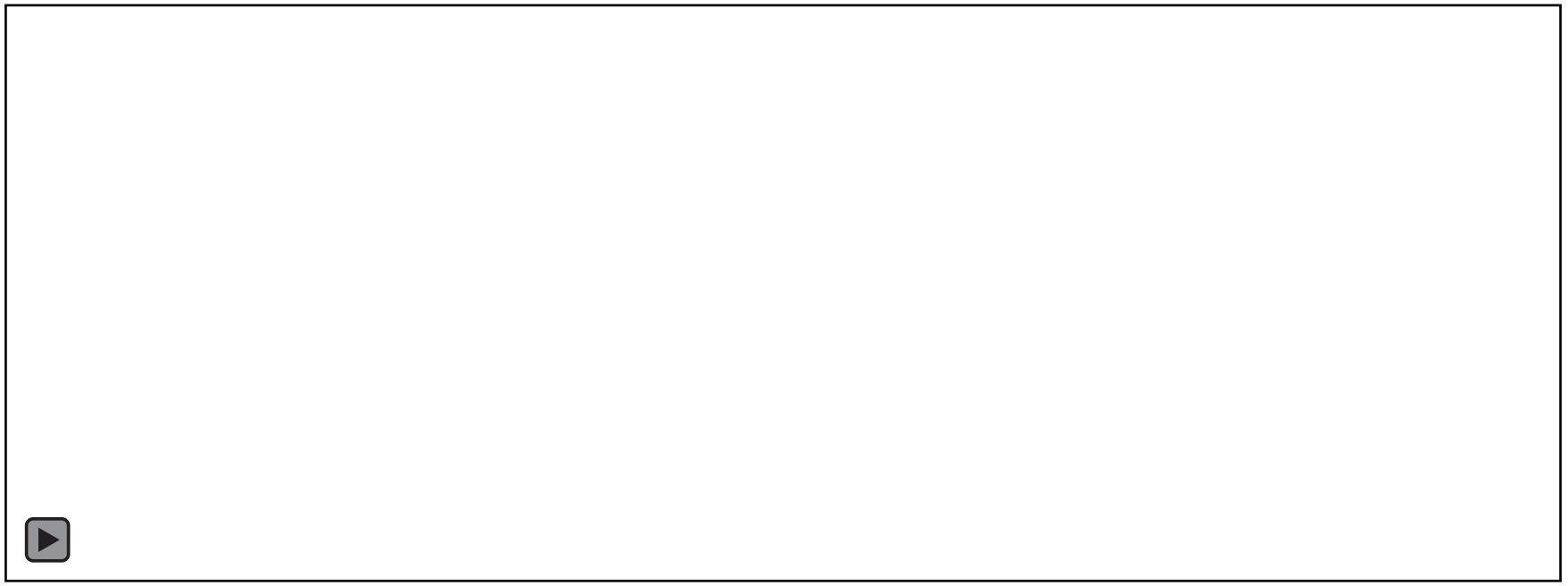


Video-LaVIT, Y. Jin et al - ICML 2024

Existing Approaches and Their Limitations

External Conditioning Methods:

- Rely on external motion signals (optical flow, trajectories)
- Require additional input preprocessing and computation
- Limited flexibility in motion generation

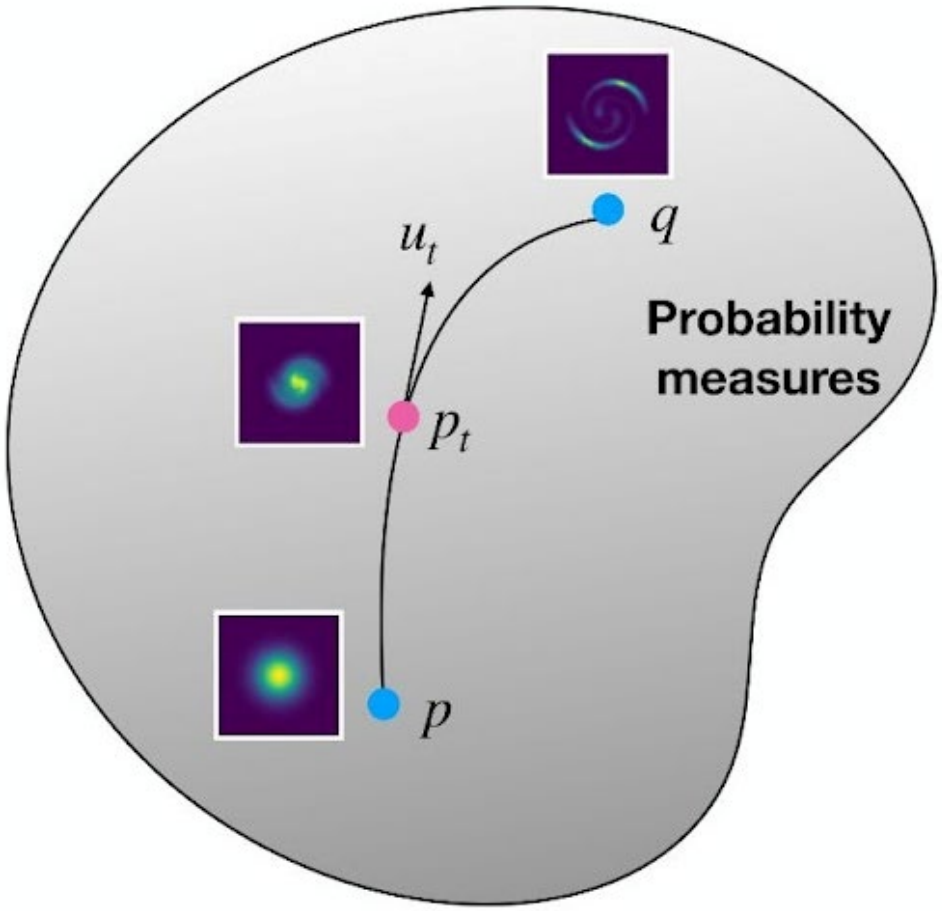


Motion prompting, D. Geng et al - CVPR 2025 (Oral)

Flow Matching

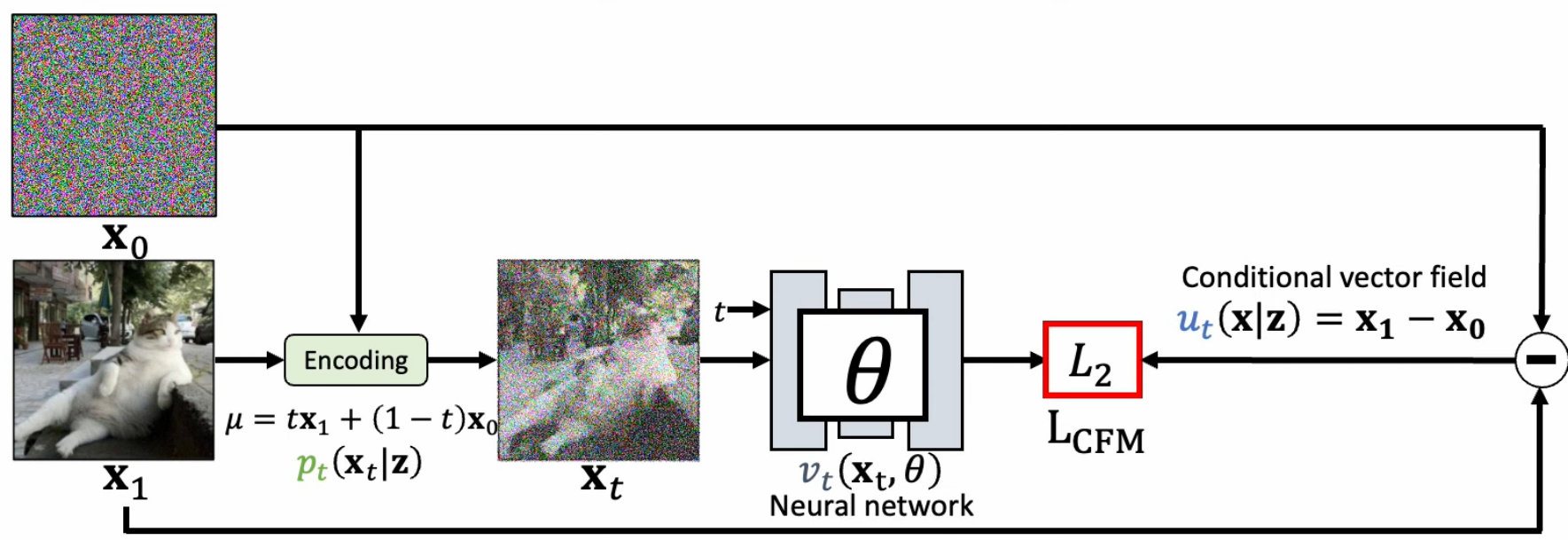
Continuity Equation:

$$\frac{dp_t}{dt} = -\text{div}(p_t v_t)$$

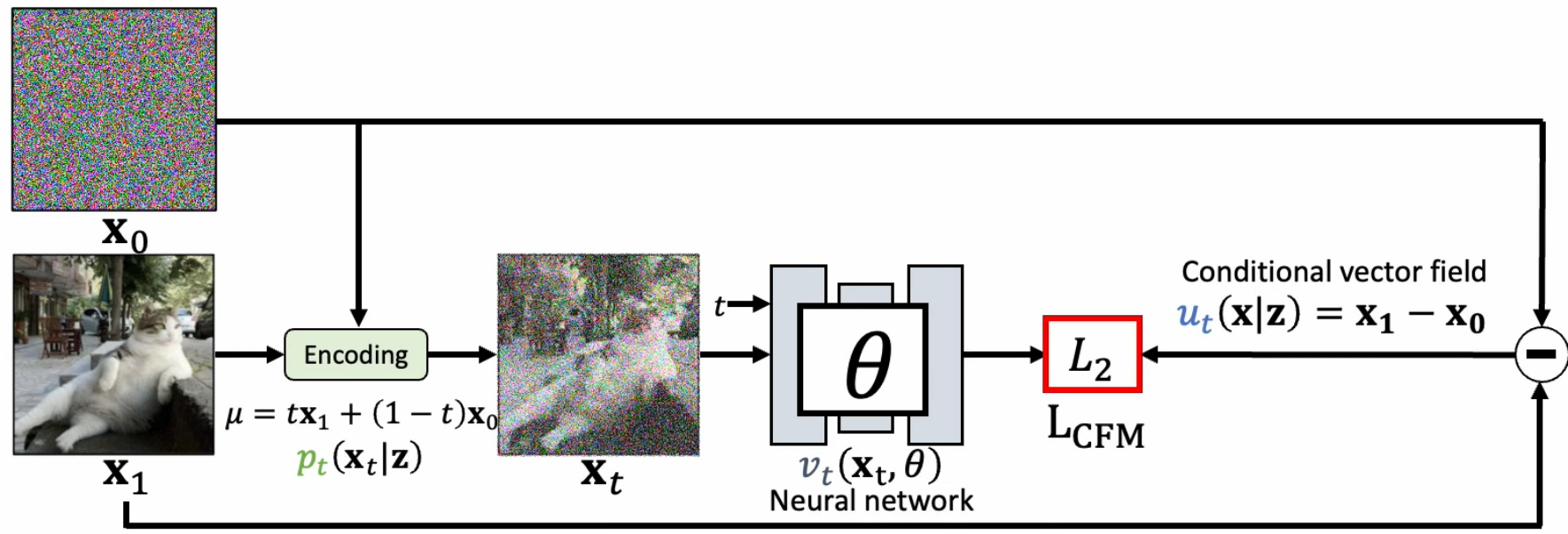


Flow Matching for Generative Modeling, Lipman. et al - 2022

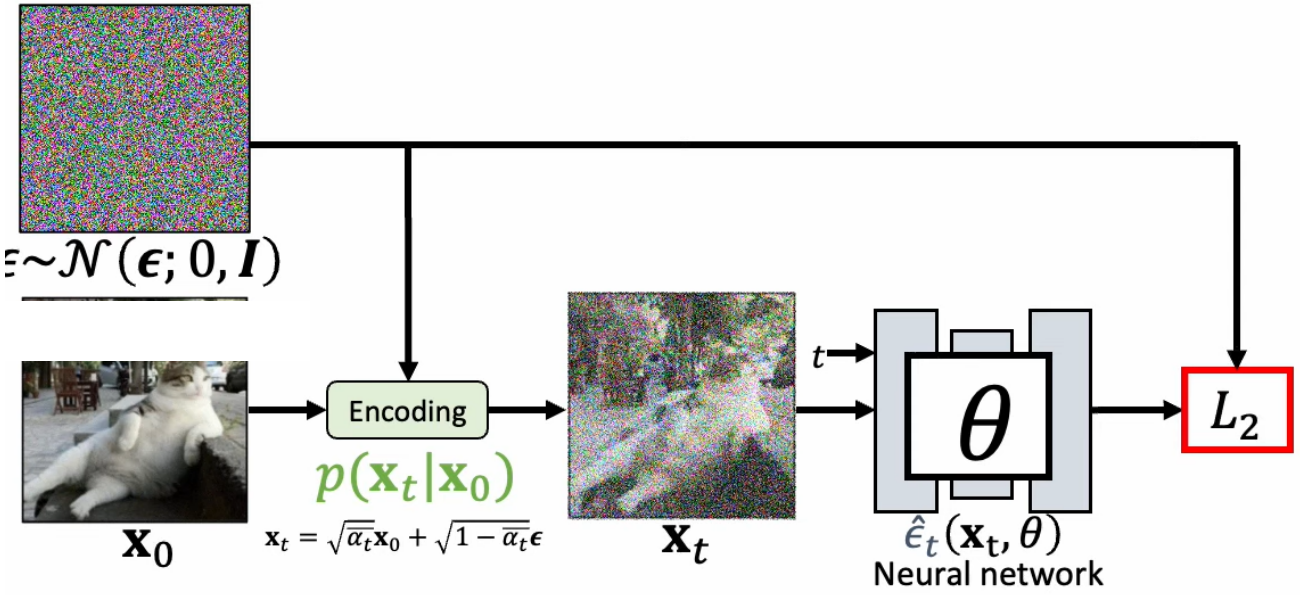
Flow Matching



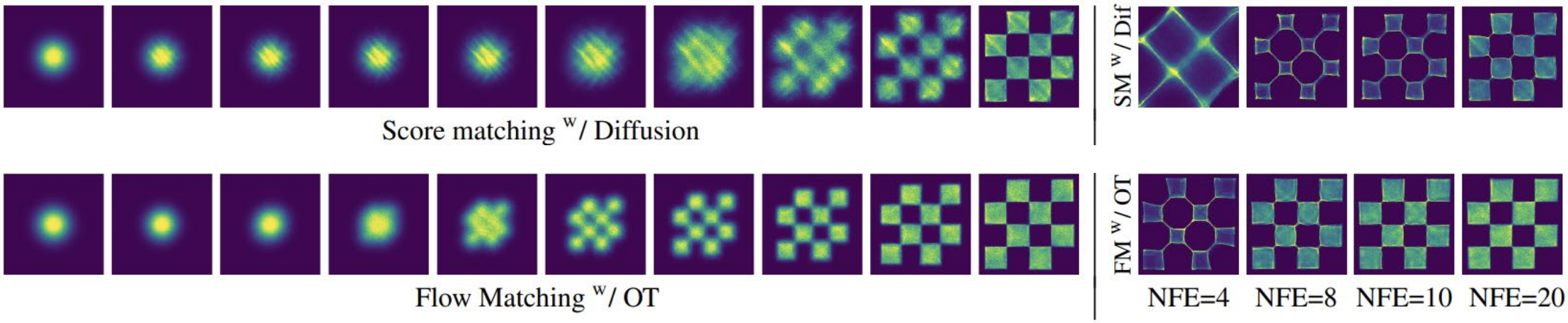
Flow Matching



Diffusion Model



Optimal Transport vs Diffusion Path



Flow Matching Training

- Sample timestep: $t \in [0, 1]$
- Latent interpolation: $z_t = (1 - t) \cdot z_1 + t \cdot z_0$
- Predict velocity: $v_t = \frac{dz_t}{dt} = z_0 - z_1$
- Optimize FM to predict v_t:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_1, t \sim \mathcal{U}(0,1), z_0 \sim \mathcal{N}(0,I)} \left[\| u_\theta(z_t, t) - (z_0 - z_1) \|^2 \right]$$

Flow Matching Sampling

- Initial noisy latent:

$$z_0 \sim \mathcal{N}(0, I)$$

- For each discrete time step t_i:

$$z_{t_{i+1}} = (1 - \sigma_{t_i}) \cdot z_{t_i} - \sigma_{t_i} \cdot u_\theta(z_{t_i}, t_i)$$

where σ_{t_i} is interpolation coefficient from scheduler.

- Denoised latent estimation:

$$\bar{z}_1 = z_t - \sigma_t \cdot u_{\theta,t}$$

Debiasing operator Δ

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

Debiasing operator Δ

$\Delta: \mathbb{R}^{F \times W \times H \times C} \rightarrow \mathbb{R}^{(F-1) \times W \times H \times C}$

$\forall f \in [F-1], \forall w \in [W], \forall h \in [H], \forall c \in [C]$

where:

- F is number of frames,
- W and H are the width and height, respectively,
- C is the number of channels.

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

Quantitative motivation

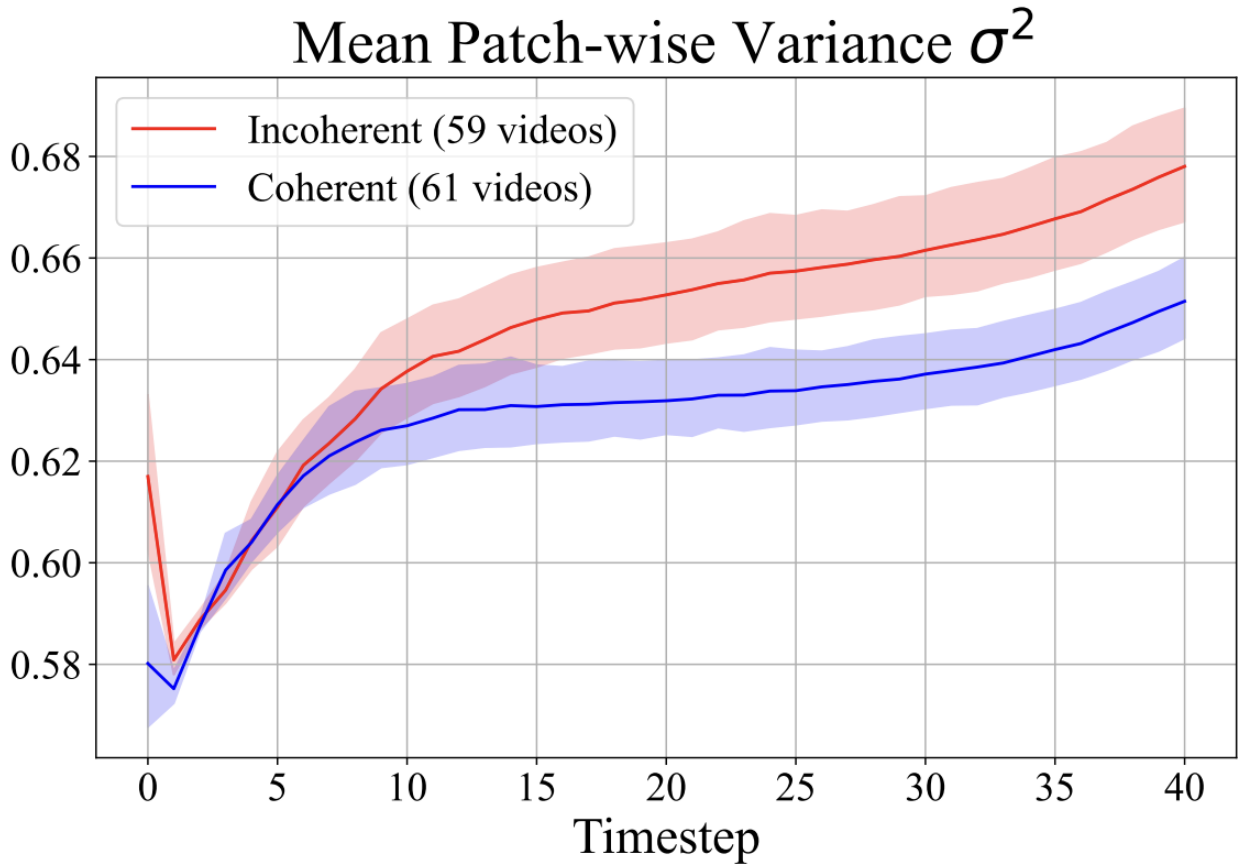
$$\sigma^2_{w,h,c} = \mathbb{V}_{f \sim [F-1]} \left[(\Delta u_{\theta,t})_{f,w,h,c} \right]$$

where $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Quantitative motivation

$$\sigma^2_{w,h,c} = \mathbb{V}_{f\sim[F-1]}\left[(\Delta u_{\theta,t})_{f,w,h,c}\right]$$

where $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.



Qualitative motivation

- Denoised latent estimation:

$$\bar{z}_1 = z_t - \sigma_t \cdot u_{\theta,t}$$

- Grayscale video frames are visualized using:

$$V_{\bar{z}_{1,c0}} = 255 \cdot \frac{\bar{z}_{1,c0} - \min_{F,H,W}(\bar{z}_{1,c0})}{\max_{F,H,W}(\bar{z}_{1,c0}) - \min_{F,H,W}(\bar{z}_{1,c0})}$$

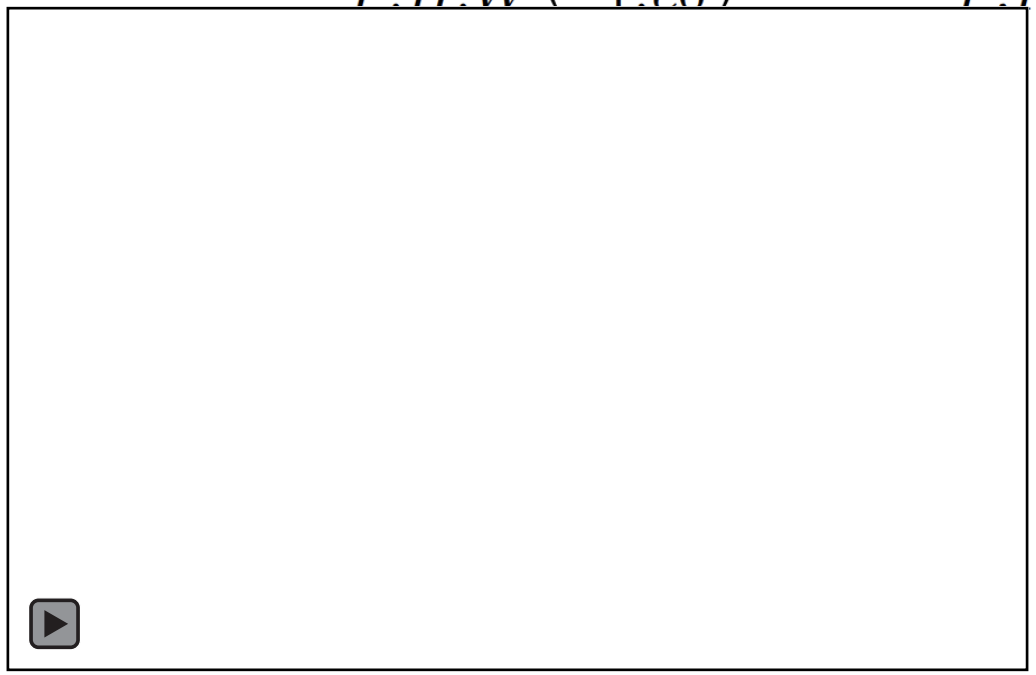
Qualitative motivation

- Denoised latent estimation:

$$\bar{z}_1 = z_t - \sigma_t \cdot u_{\theta,t}$$

- Grayscale video frames are visualized using:

$$V_{\bar{z}_{1,c0}} = 255 \cdot \frac{\bar{z}_{1,c0} - \min_{F,H,W}(\bar{z}_{1,c0})}{\max_{F,H,W}(\bar{z}_{1,c0}) - \min_{F,H,W}(\bar{z}_{1,c0})}$$



Qualitative motivation

A standard FM step

- Input: A text prompt $\mathcal{P}$, a timestep t_i, and a trained Flow Matching model FM.

$$u_{\theta,t_i|\mathcal{P}} \leftarrow FM(z_{t_i}, t_i, \mathcal{P})$$
$$u_{\theta,t_i|\emptyset} \leftarrow FM(z_{t_i}, t_i, \emptyset)$$
$$u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right) \quad \text{CFG}$$

- Output: A noised latent $z_{t_{i+1}}$ for the next timestep t_{i+1}.

$$z_{t_{i+1}} \leftarrow (1 - \sigma_{t_i}) \cdot z_{t_i} - \sigma_{t_i} \cdot u_{\theta,t_i}$$
Return $z_{t_{i+1}}$

A standard FM step

- Input: A text prompt $\mathcal{P}$, a timestep t_i, and a trained Flow Matching model FM.

$$u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right)$$

- Output: A noised latent $z_{t_{i+1}}$ for the next timestep t_{i+1}.

$$z_{t_{i+1}} \leftarrow (1 - \sigma_{t_i}) \cdot z_{t_i} - \sigma_{t_i} \cdot u_{\theta,t_i}$$

Return $z_{t_{i+1}}$

A standard FM step

- Input: A text prompt $\mathcal{P}$, a timestep t_i, and a trained Flow Matching model FM.

$$u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right)$$

- Output: A noised latent $z_{t_{i+1}}$ for the next timestep t_{i+1}.

$$z_{t_{i+1}} \leftarrow (1 - \sigma_{t_i}) \cdot z_{t_i} - \sigma_{t_i} \cdot u_{\theta,t_i}$$

Return $z_{t_{i+1}}$

A standard FM step

- Input: A text prompt $\mathcal{P}$ a timestep t_i, and a trained Flow Matching model FM.

$$u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right)$$

Recall

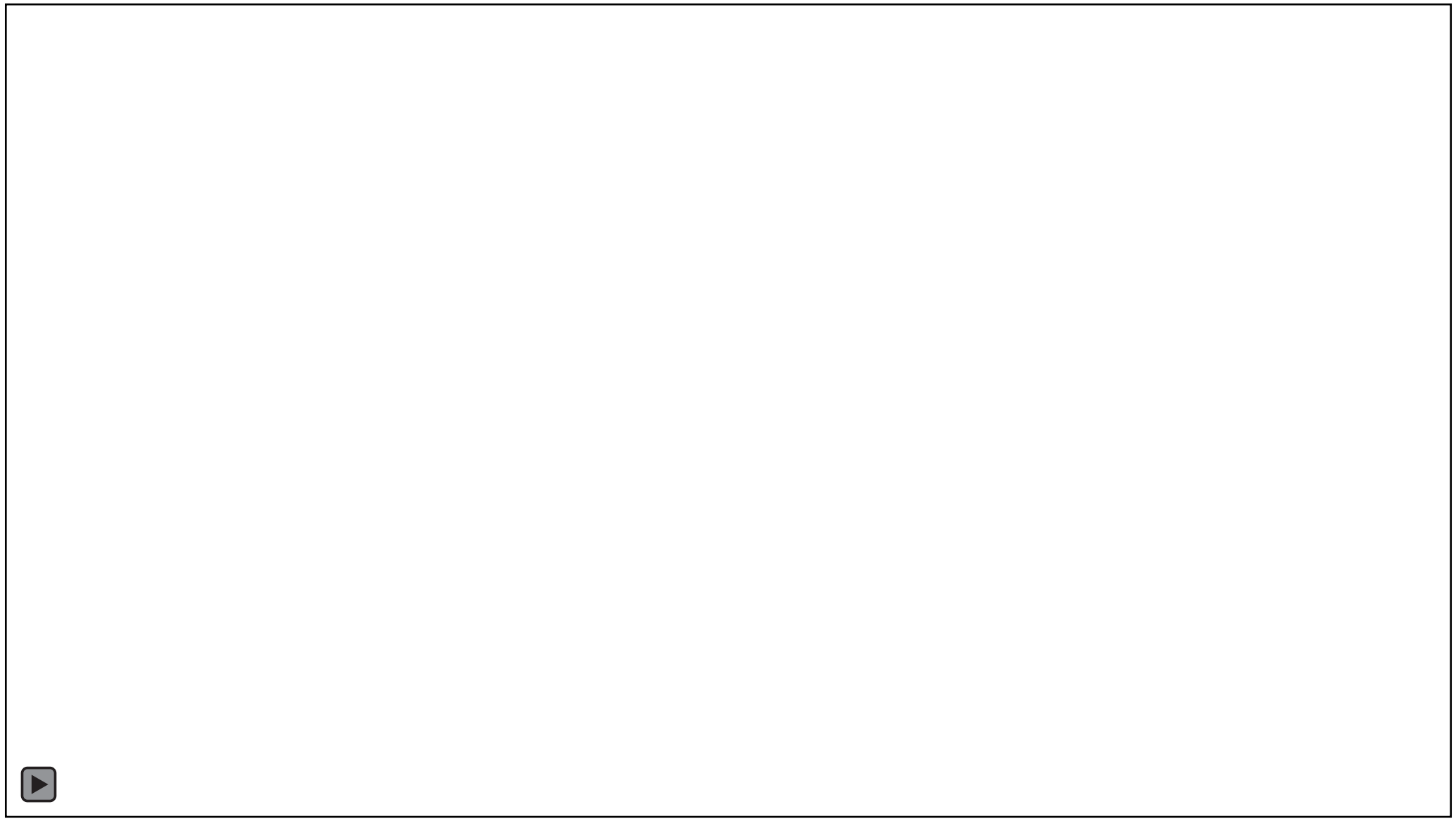
- Debiasing operator:

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

- $\sigma^2_{w,h,c} = \mathbb{V}_{f\sim[F-1]}\left[(\Delta u_{\theta,t})_{f,w,h,c}\right]$
  where $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.
- Averaging across the channel dimension:

$$s_{w,h} = \mathbb{E}_{c\sim[C]}\left[\sigma^2_{w,h,c}\right] = \frac{1}{C}\sum_{c=1}^{C}\sigma^2_{w,h,c}$$

Recall



A standard FM step

- Input: A text prompt $\mathcal{P}$ a timestep t_i, and a trained Flow Matching model FM.

$$u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right)$$

A Single FlowMo Denoising Step

- Input: A text prompt P, a timestep t_i , a set of iterations for refinement $\{\tau 1, \ldots, \tau \ell\}$, and a trained Flow Matching model FM.

$$u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right)$$

if $t_i \in \{\tau_1, \ldots, \tau_\ell\}$ **then**

Compute $(\Delta u_{\theta,t_i})$ as in Eq. (2)

Compute σ^2 as in Eq. (3)

$$s_{w,h} \leftarrow \mathbb{E}_{c\sim[C]}\left[\sigma^2_{w,h,c}\right] \quad \forall w \forall h$$

$$\mathcal{L} \leftarrow \max_{w\sim[W],h\sim[H]} s_{w,h}$$

$$z_{t_i} \leftarrow z_{t_i} - \eta \cdot \nabla_{z_{t_i}} \mathcal{L}$$

$$u_{\theta,t_i|\mathcal{P}} \leftarrow FM(z_{t_i}, t_i, \mathcal{P})$$

$$u_{\theta,t_i|\emptyset} \leftarrow FM(z_{t_i}, t_i, \emptyset)$$

$$u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right)$$

end if

A Single FlowMo Denoising Step

$u_{\theta,t_i|\mathcal{P}} \leftarrow FM(z_{t_i}, t_i, \mathcal{P})$

$u_{\theta,t_i|\emptyset} \leftarrow FM(z_{t_i}, t_i, \emptyset)$

$u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right)$

if $t_i \in \{\tau_1, \ldots, \tau_\ell\}$ **then**

  Compute $(\Delta u_{\theta,t_i})$ as in Eq. (2)

  Compute σ^2 as in Eq. (3)

  $s_{w,h} \leftarrow \mathbb{E}_{c\sim[C]}\left[\sigma^2_{w,h,c}\right] \quad \forall w \forall h$

  $\mathcal{L} \leftarrow \max_{w\sim[W],h\sim[H]} s_{w,h}$

  $z_{t_i} \leftarrow z_{t_i} - \eta \cdot \nabla_{z_{t_i}} \mathcal{L}$

  $u_{\theta,t_i|\mathcal{P}} \leftarrow FM(z_{t_i}, t_i, \mathcal{P})$

  $u_{\theta,t_i|\emptyset} \leftarrow FM(z_{t_i}, t_i, \emptyset)$

  $u_{\theta,t_i} \leftarrow u_{\theta,t_i|\mathcal{P}} + \rho \cdot \left(u_{\theta,t_i|\mathcal{P}} - u_{\theta,t_i|\emptyset}\right)$

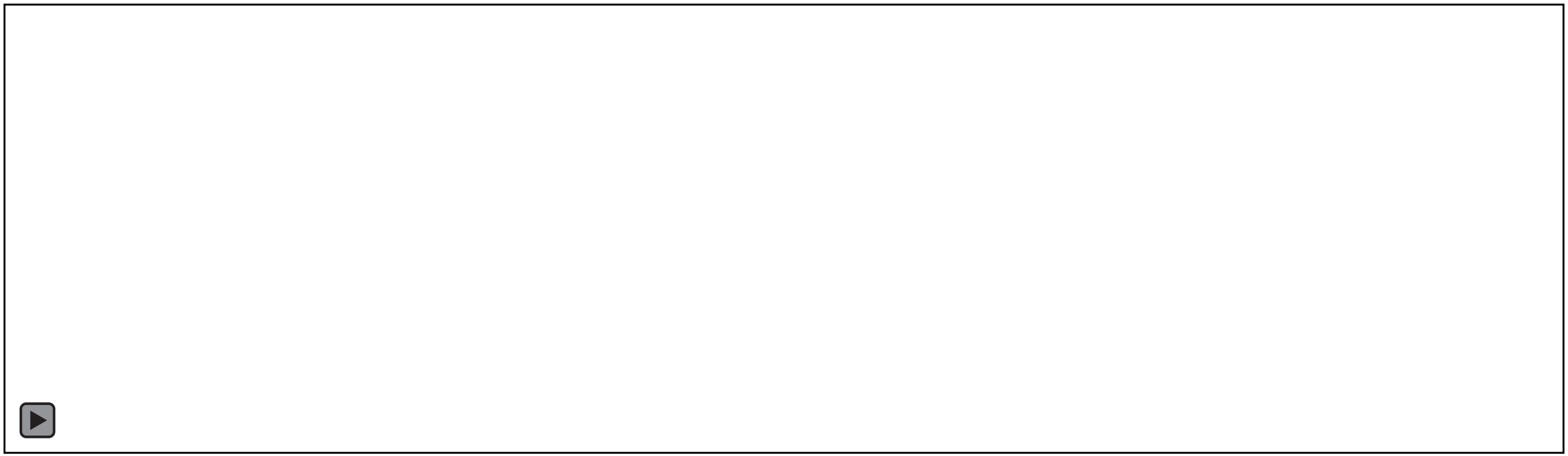
end if

$z_{t_{i+1}} \leftarrow (1 - \sigma_{t_i}) \cdot z_{t_i} - \sigma_{t_i} \cdot u_{\theta,t_i}$

Return $z_{t_{i+1}}$

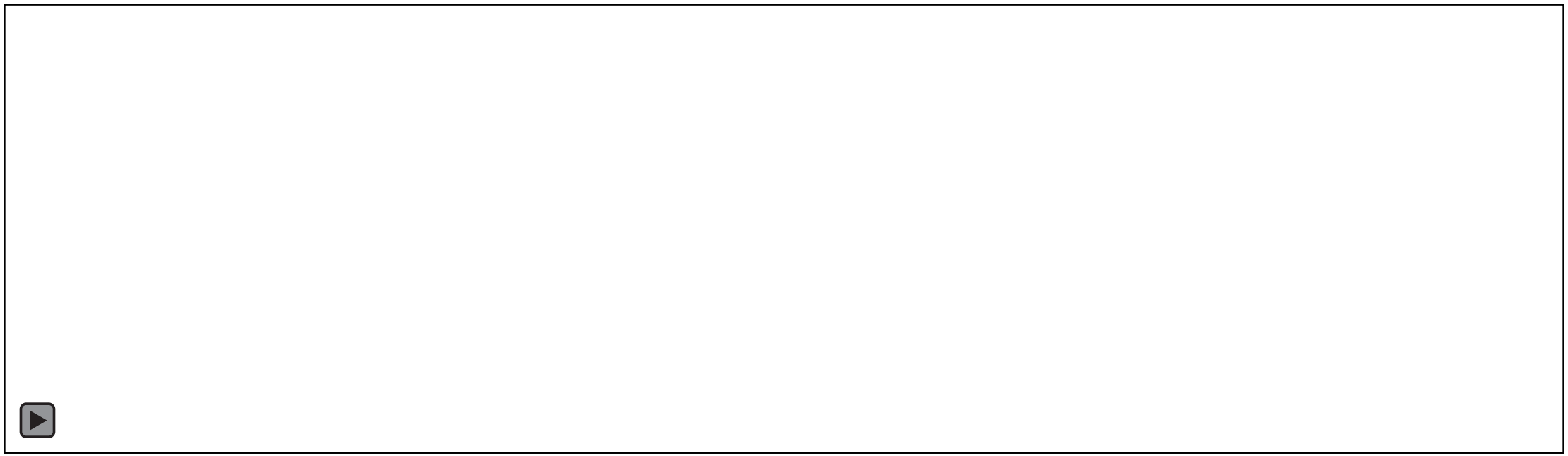
Qualitative Comparison: FlowMo vs. Base Models

Wan2.1-1.3B



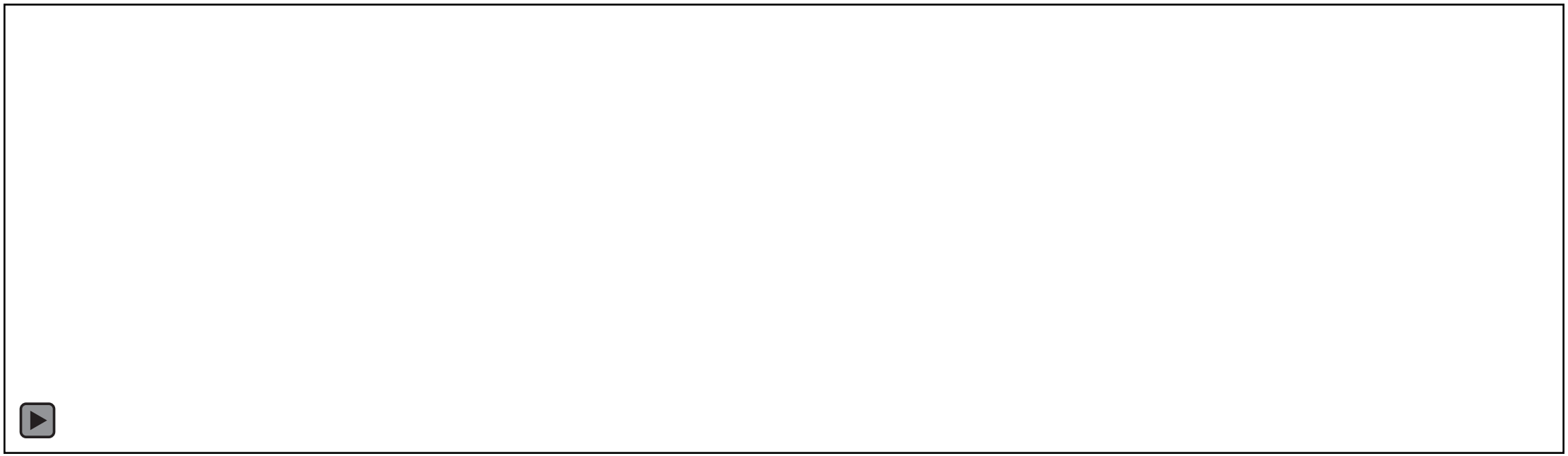
A boy with glasses flying a kite in a grassy field.

Wan2.1-1.3B



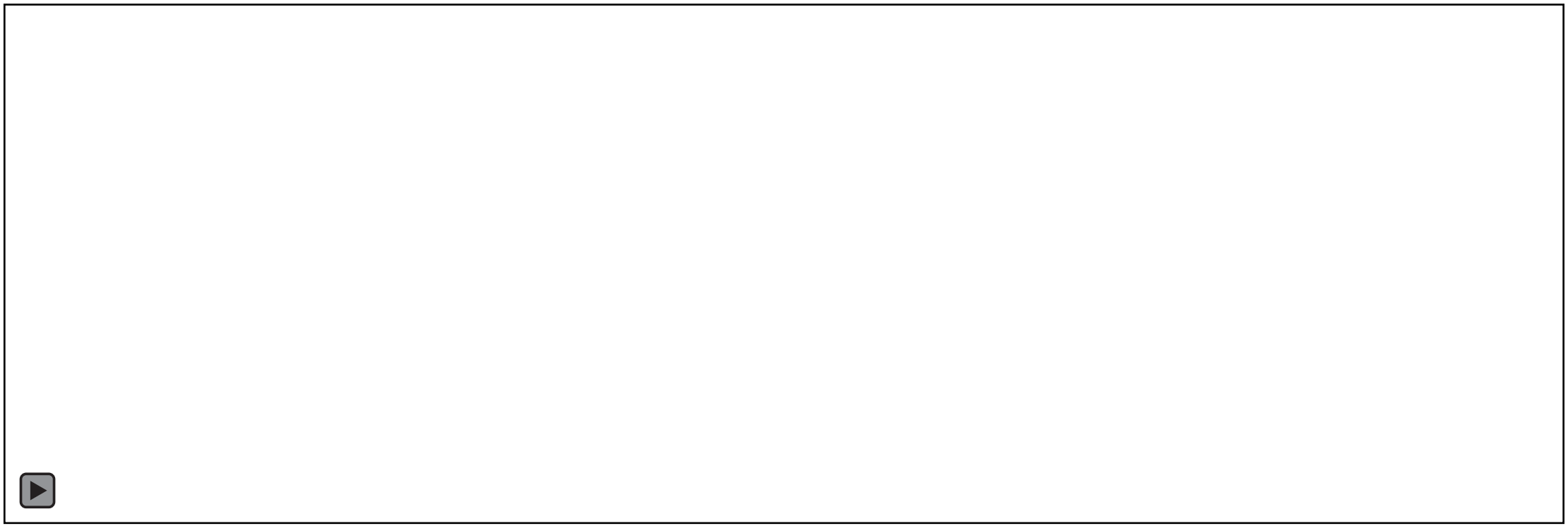
A ninja flipping through a bamboo fores.

Wan2.1-1.3B



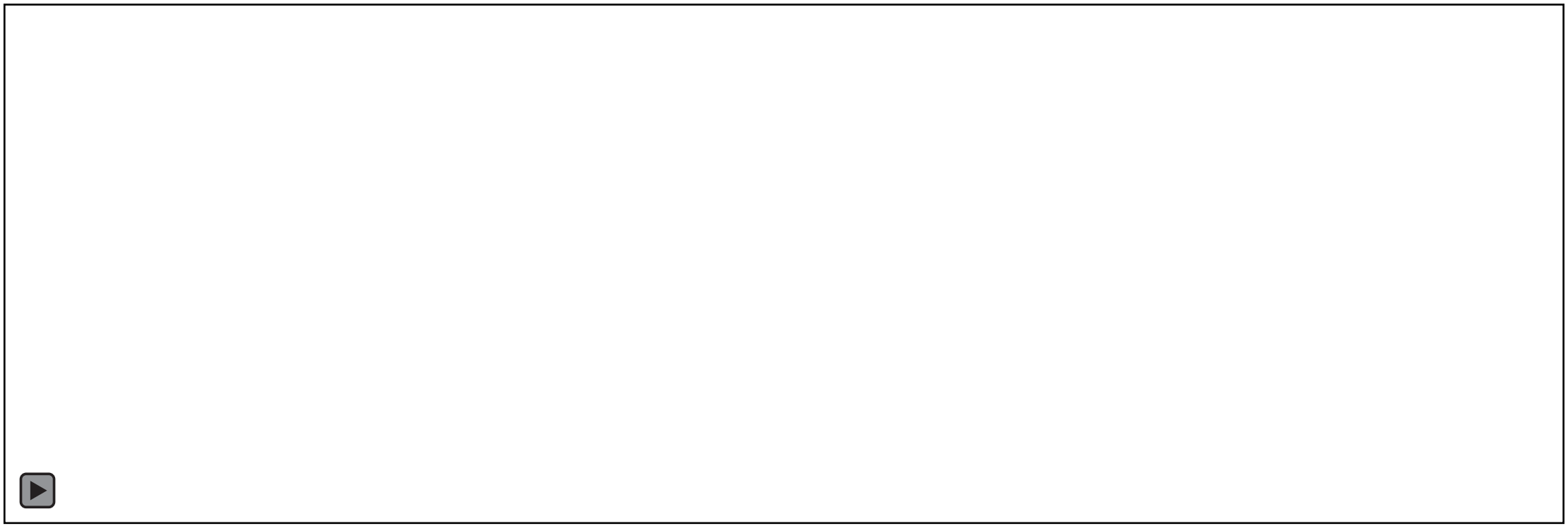
A painter creating a landscape on canvas.

CogVideoX-5B



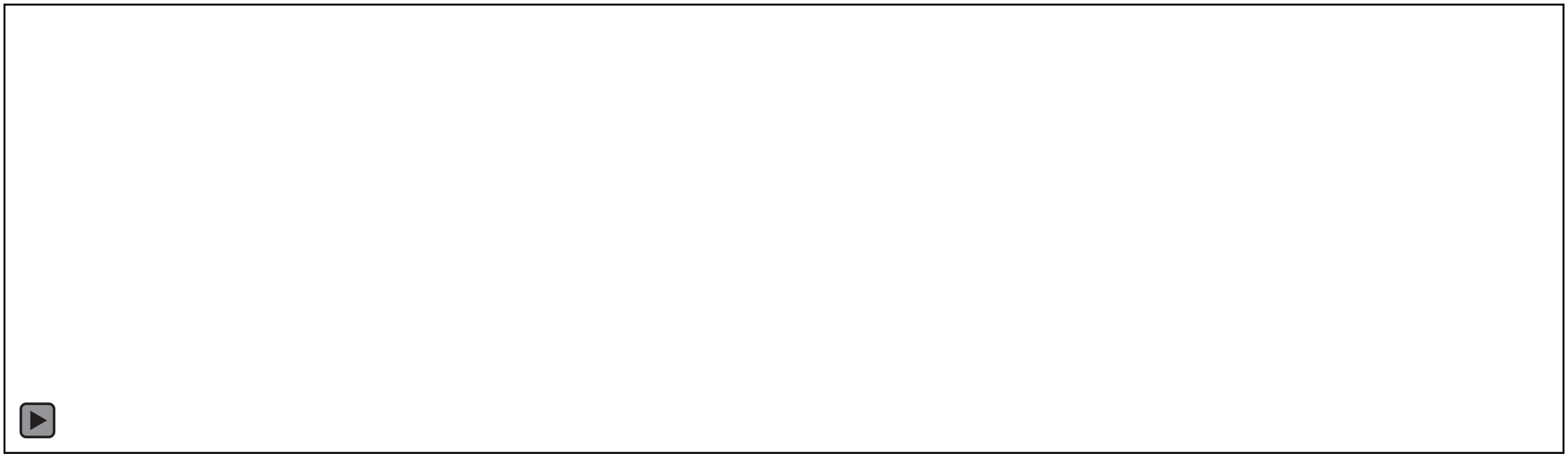
A pair of flamingos wading through shallow water.

CogVideoX-5B



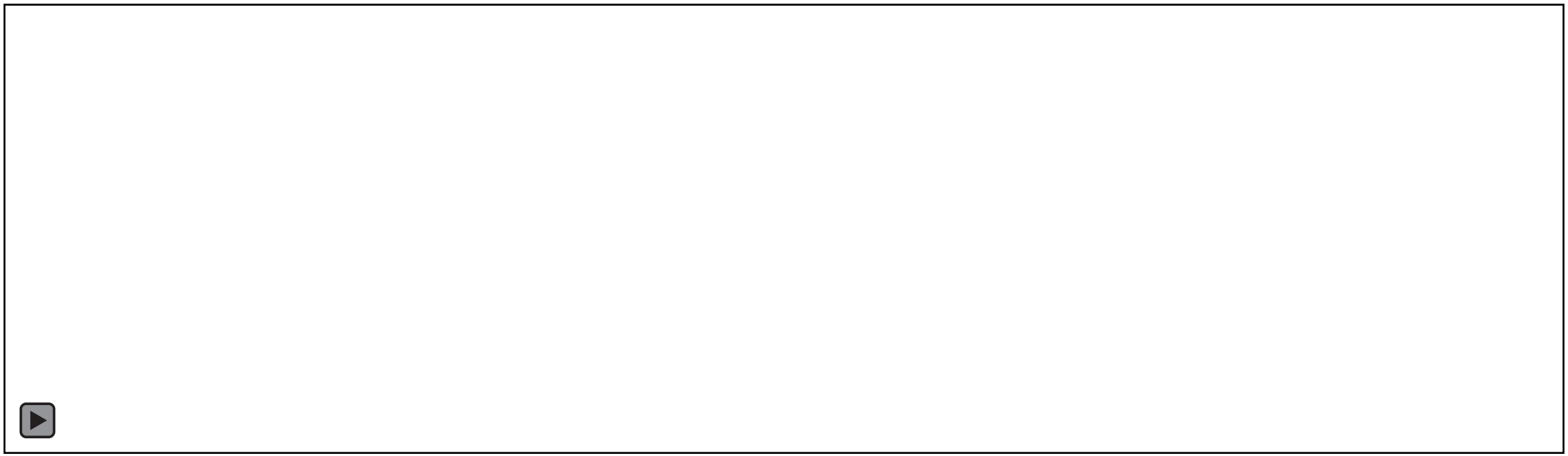
A ballerina leaping through the air.

FreeInit



Athletic man doing gymnastics on a horizontal bar.

FreeInit



A female kayaker paddling through white water rapids.

VBench evaluation results

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%)

Similar idea to Debiasing operator Δ

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

Attention contribution:

$$\text{cont}_{i,j} = \left\| \sum_{h=1}^{H} \text{attn}_{i,j}^{h} \, \mathbf{v}_j^h \, W_o^h \right\|_2,$$

$$\text{attn}_{i,j}^{h} = \textsc{Softmax}\left(\left\{ \frac{\langle \mathbf{q}_i^h, \mathbf{k}_r^h \rangle}{\sqrt{d_h}} \right\}_{r=1}^{T} \right)_j$$

Attention contribution

1. Analyzing transformers in embedding space, G. Dar et al-2022.
2. Understanding and mitigating compositional issues in text-to-image generative models, A. Zarei-2024.
3. Localizing Knowledge in Diffusion Transformers, A. Zarei-2025.

Thank you!

- Thank you for your attention!
- I appreciate your time and interest.
- If you have any questions, please feel free to ask.
- Contact information: alimohammadiamirhossein@gmail.com