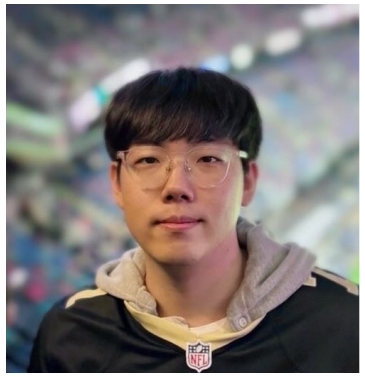
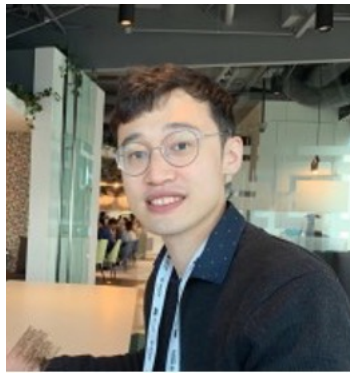
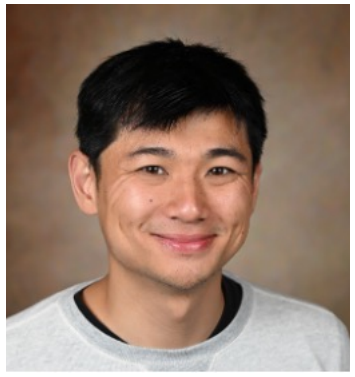
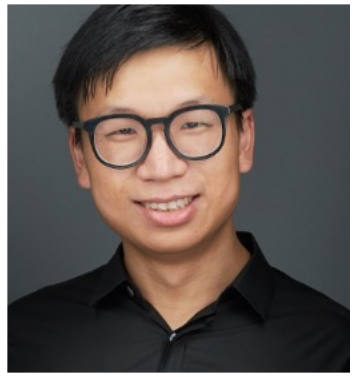
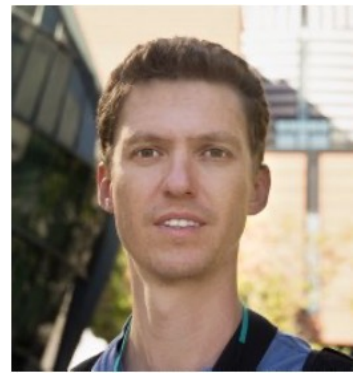
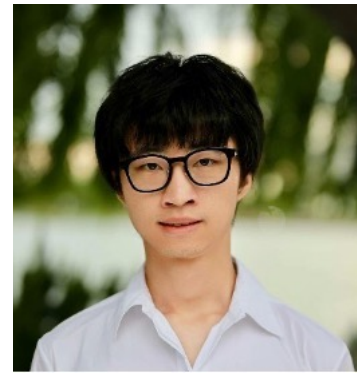
MotionStream: Real-Time Video Generation With Interactive Motion Controls

[1]Seoul National University [2]Adobe Research [3]Carnegie Mellon University [4]Morpheus AI

Joonghyuk Shin[1,2] Zhengqi Li[2] Richard Zhang[2] Jun-Yan Zhu[3] Jaesik Park[1] Eli Shechtman[2] Xun Huang[2,4]

Text-to-Video Models

SORA

VEO

Adding Motion Controls

(It's not real-time; 12 min. for 5s)

Geng et al, "Motion Prompting", CVPR 2025

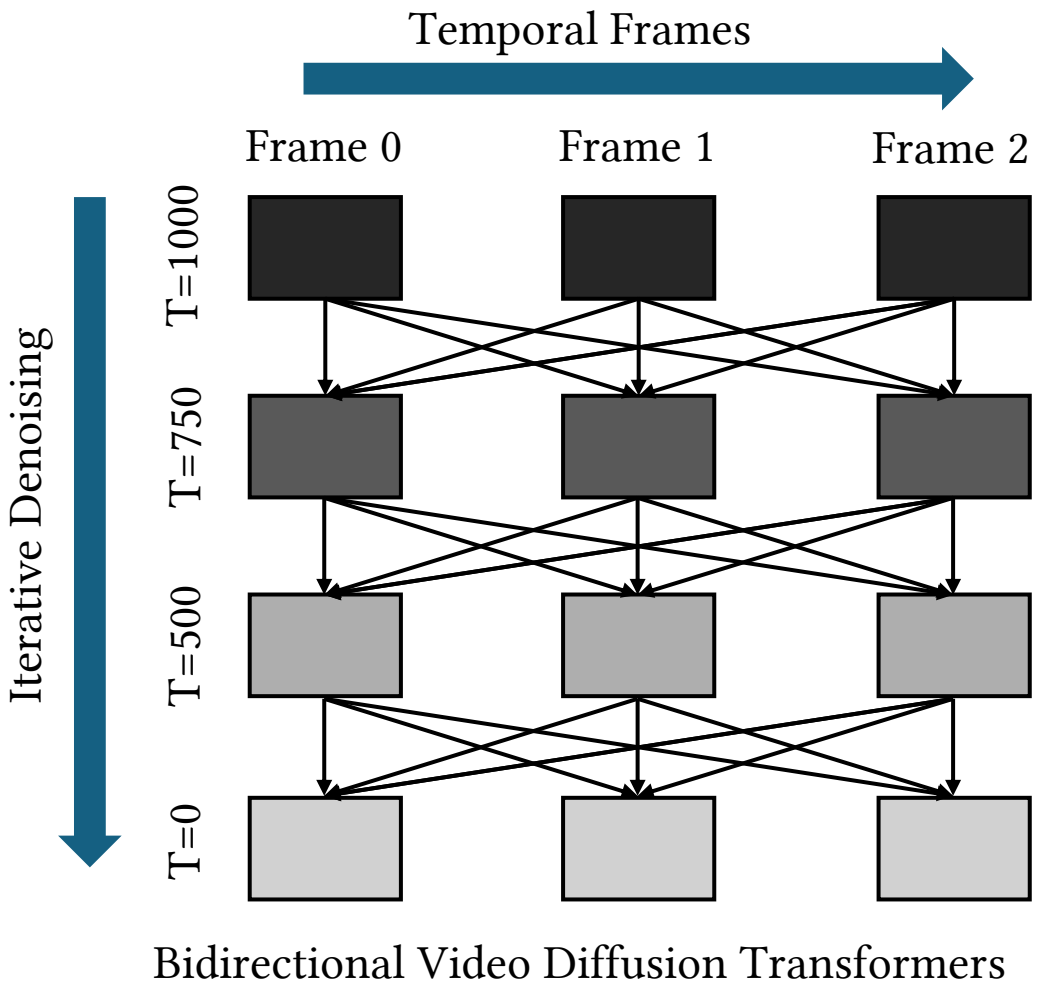
Coarse annotation

Generated Video

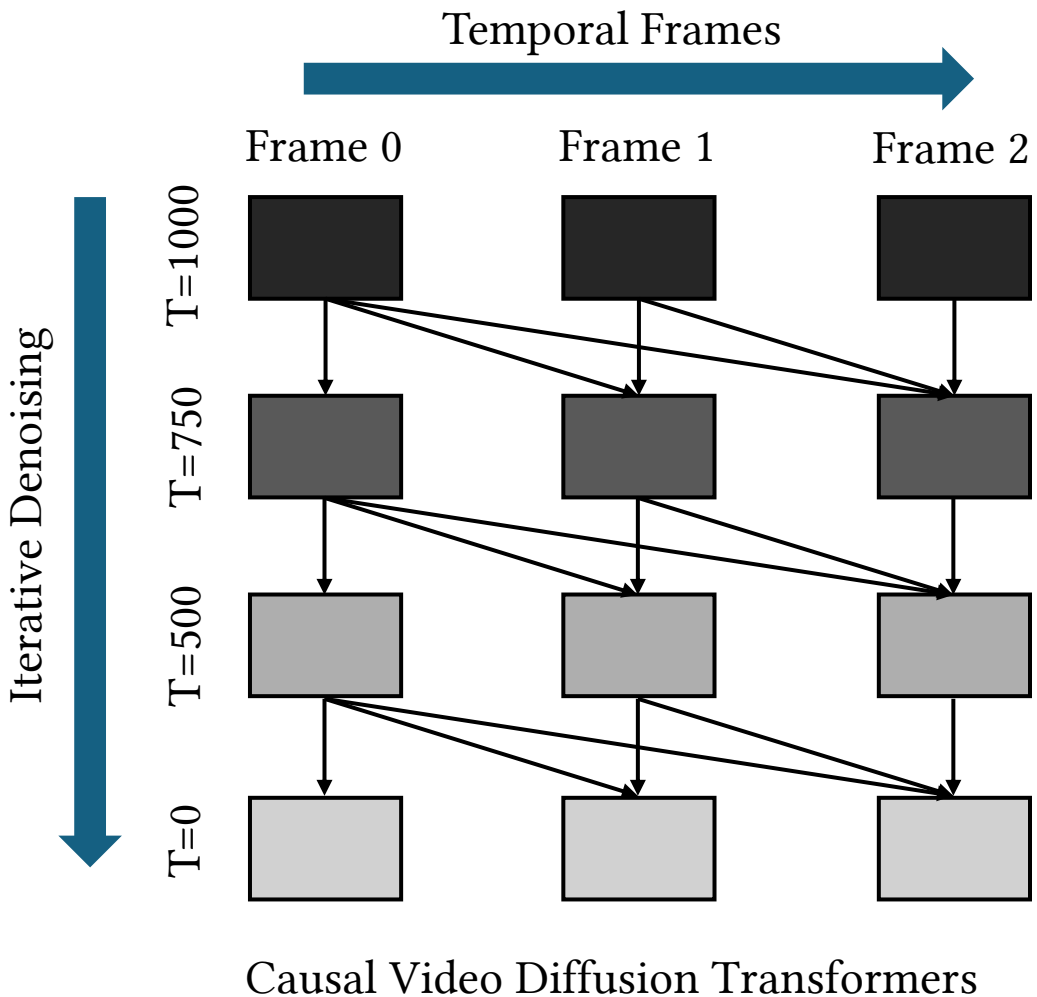
Burgert et al, "Go-with-the-Flow", CVPR 2025

(1) Slow (2) Offline Processing (3) Short

Causal Video Diffusion Models



Bidirectional Video Diffusion Transformers

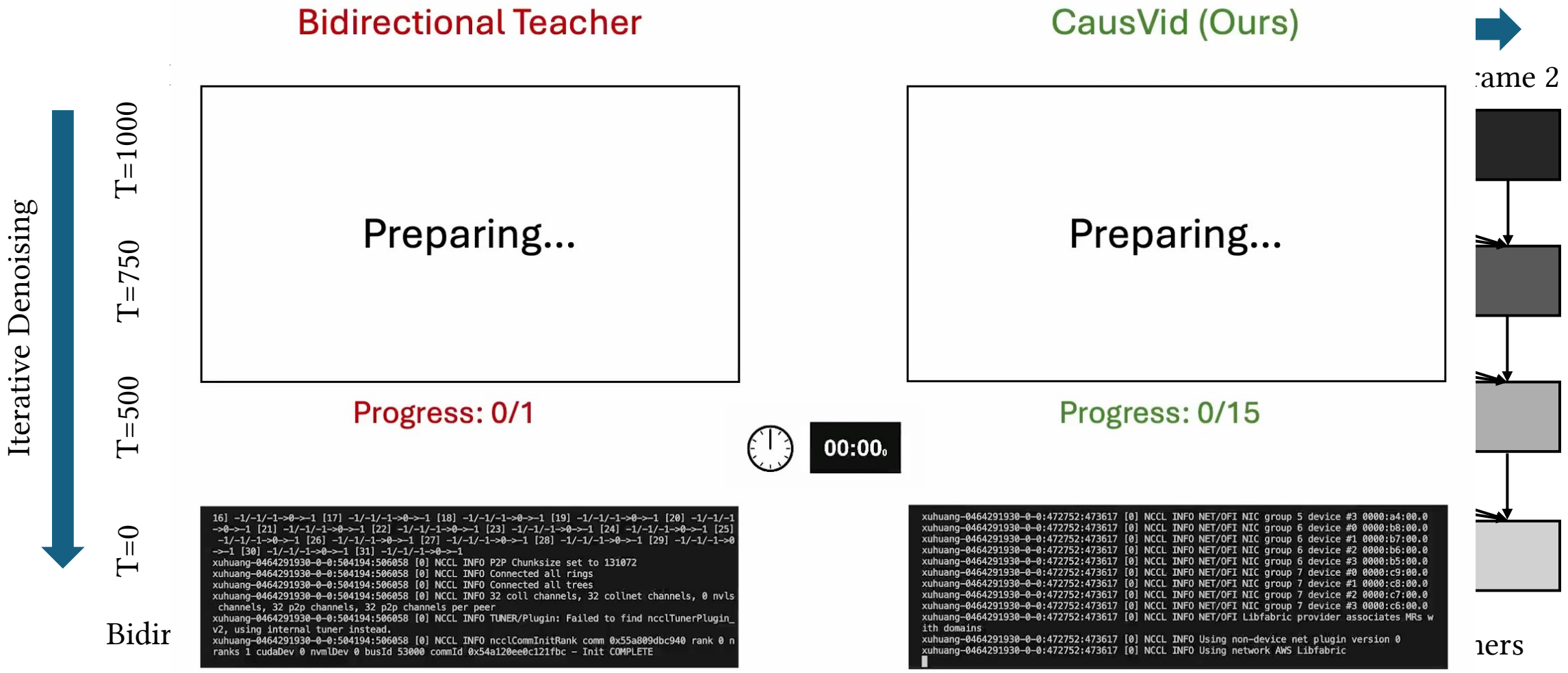


Causal Video Diffusion Transformers

Yin et al., “CausVid”, CVPR 2025

Huang et al., “Self Forcing”, NeurIPS 2025

Causal Video Diffusion Models



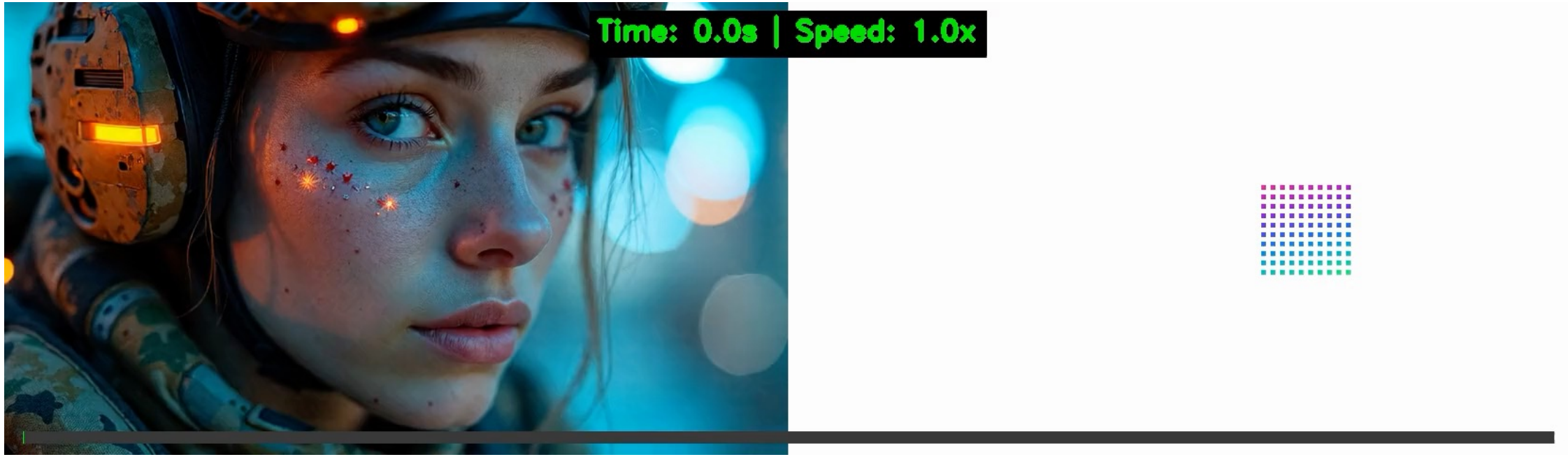
Yin et al., “CausVid”, CVPR 2025

Huang et al., “Self Forcing”, NeurIPS 2025

Causal Video Diffusion Models

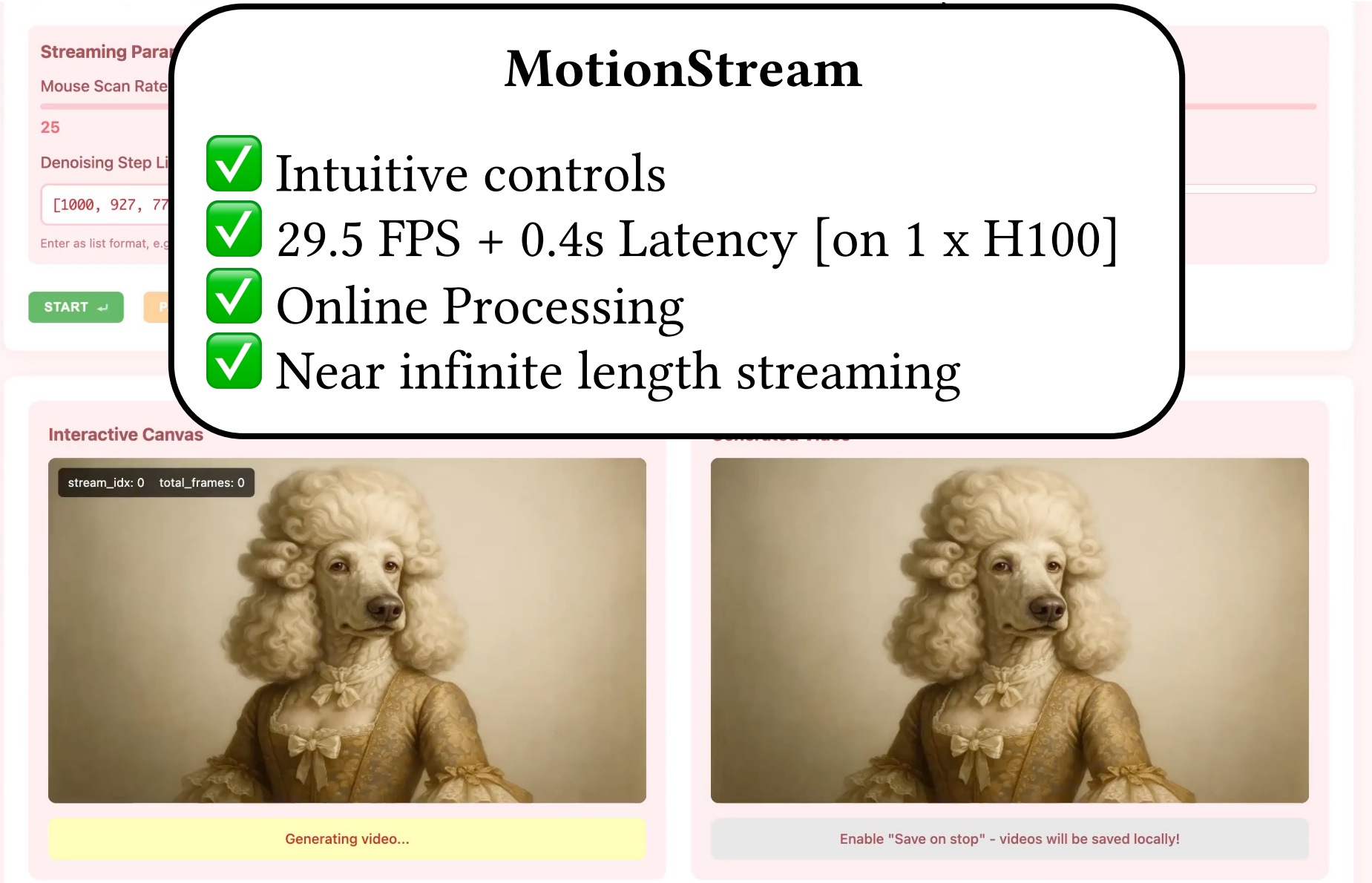
Video Generated with
Self Forcing Distillation (Output)

User Mouse Drags (Input)



`Streaming = Real Time + Long Duration + Interactive Controls`

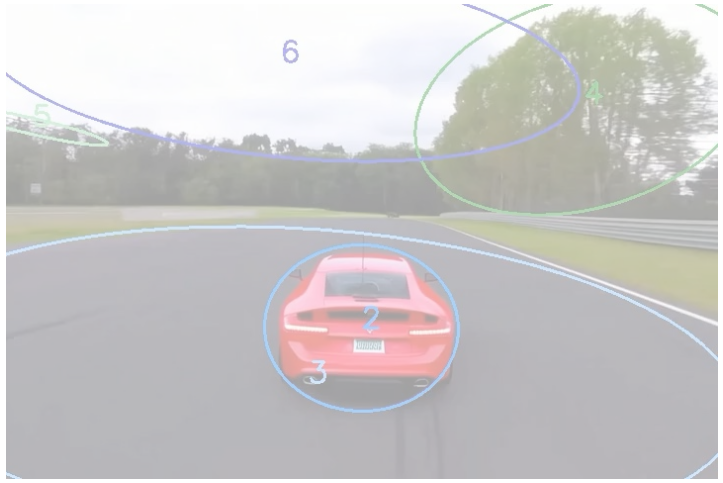
MotionStream (Preview)



Played at 2X speed

Choosing a Good Motion Representation

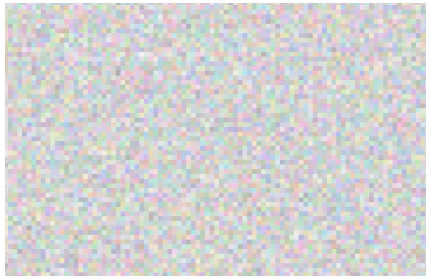
Feng et al., “BlobGen-Vid”, 2025



Burgert et al., “Go-with-the-Flow”, 2025

User Control

Noise Warping



Generated Video

Wang et al., “ATI”, 2025

Bounding Box / Blob

Ma et al., “TrailBlazer”, 2023
Luo et al., “Ctrl-V”, 2024
Feng et al., “BlobGen-Vid”, 2025

Optical Flow

Shin et al., “InstantDrag”, 2024
Chefer et al., “VideoJAM”, 2025
Burgert et al., “Go-with-the-Flow”, 2025

Sparse Trajectories

Wang et al., “MotionCtrl”, 2024
Zhang et al., “Tora”, 2025
Wang et al., “ATI”, 2025

Choosing a Good Motion Representation

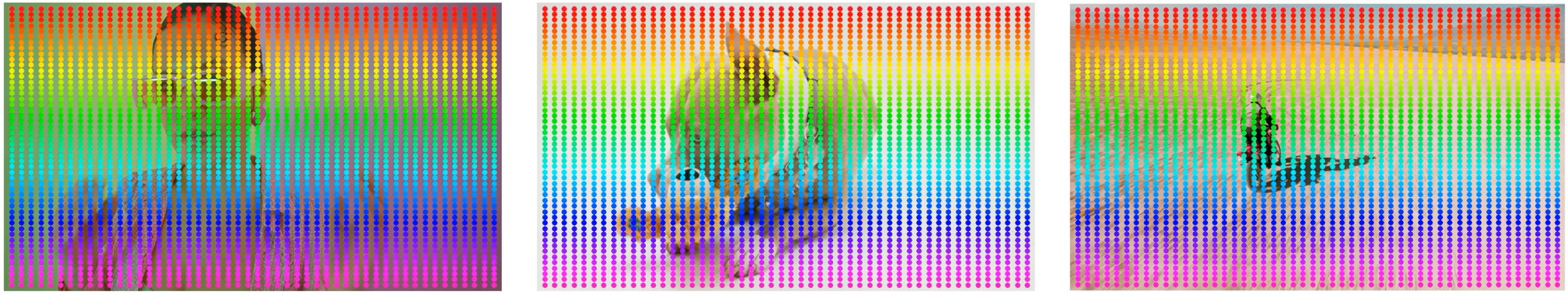
Geng et al., “Motion Prompting”, 2025

(Dense) Point Trajectories

Geng et al., “Motion Prompting”, 2025

Gu et al., “Diffusion as Shader”, 2025

Preparing Data

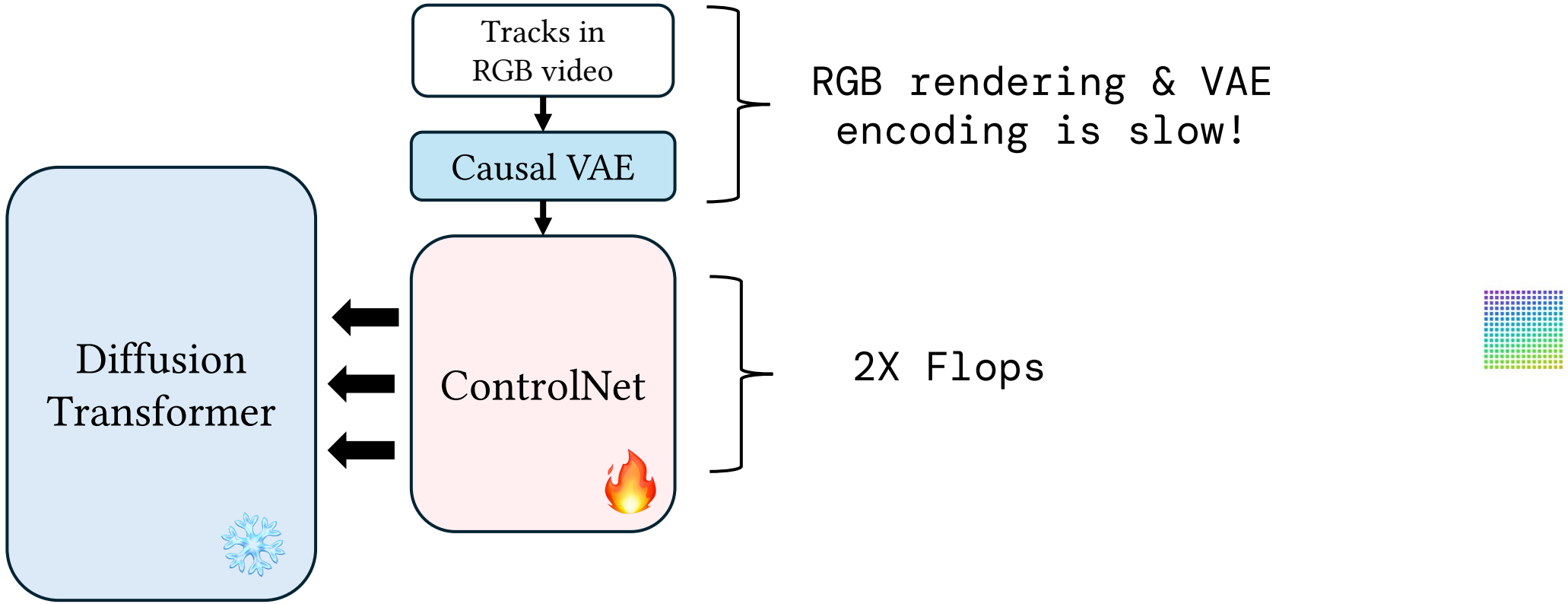


OpenVid-1M (filtered 0.6M) + Generated Wan Videos (30~70K)
Point Tracking with **CoTracker3** (50x50 points)

Karaev et al., “CoTracker3”, ICCV 2025

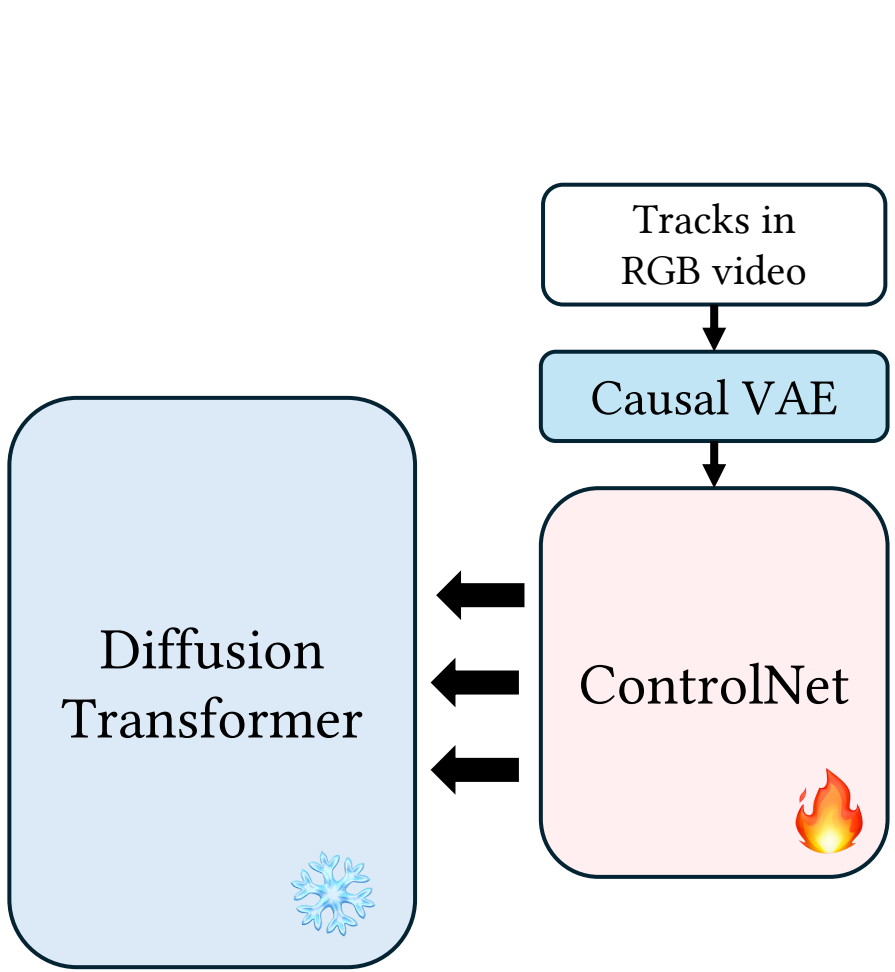
Designing a Motion-Control Module

ControlNet-based Approaches

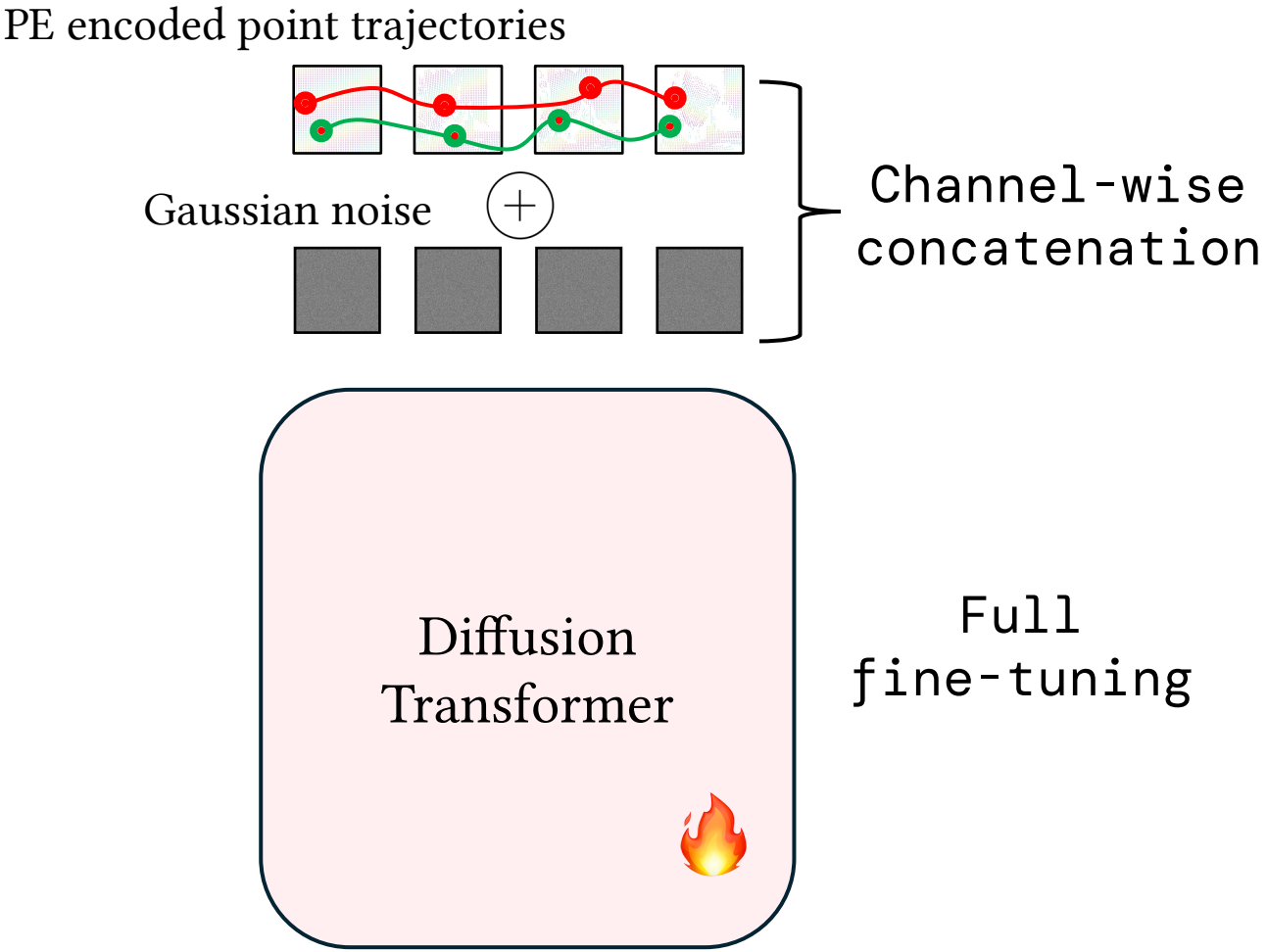


Gu et al., “Diffusion As Shader”, 2025

Designing a Motion-Control Module

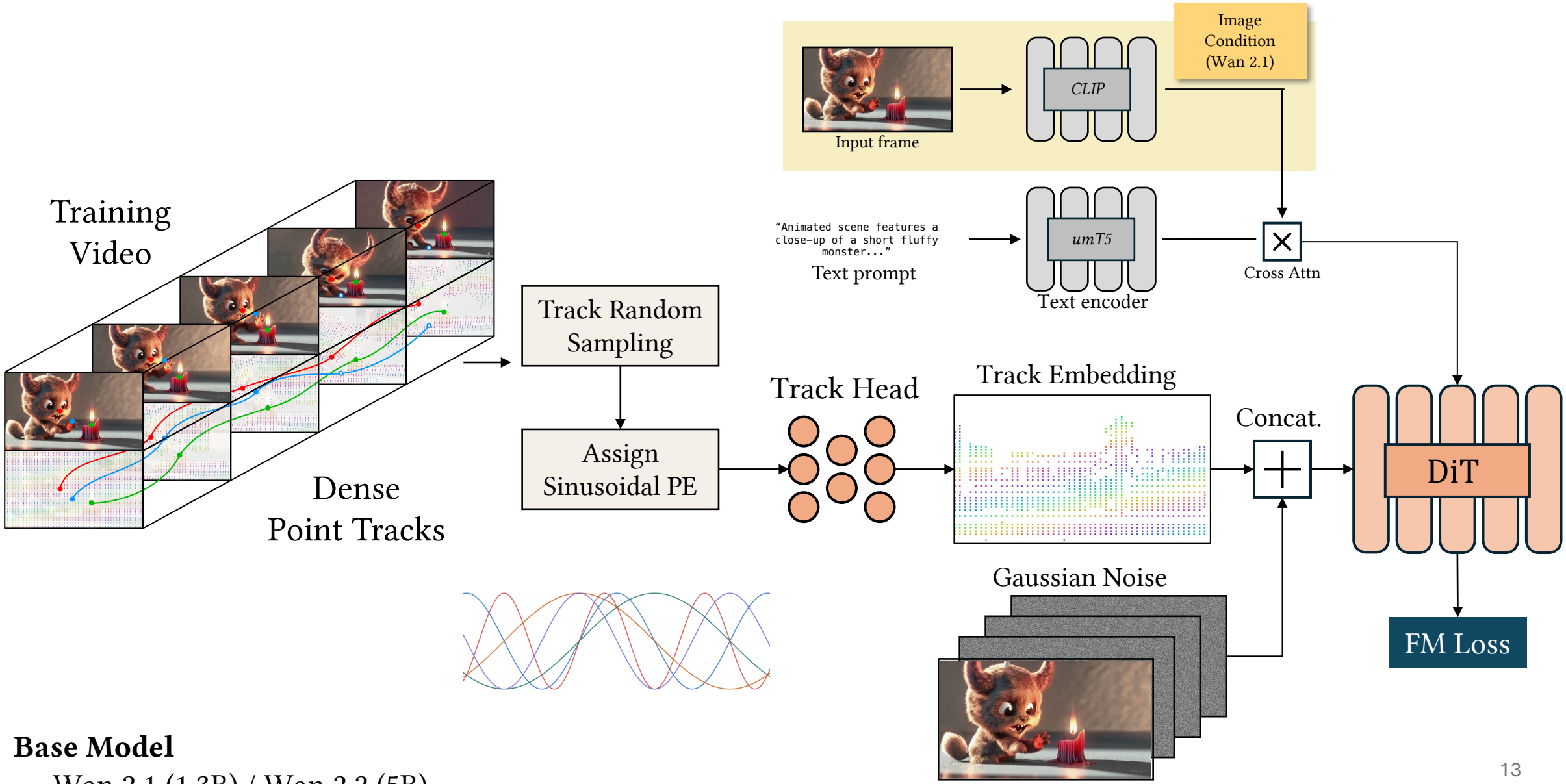


Gu et al., “Diffusion As Shader”, 2025



Ours, Teacher

Training a Motion-Controlled Teacher Model

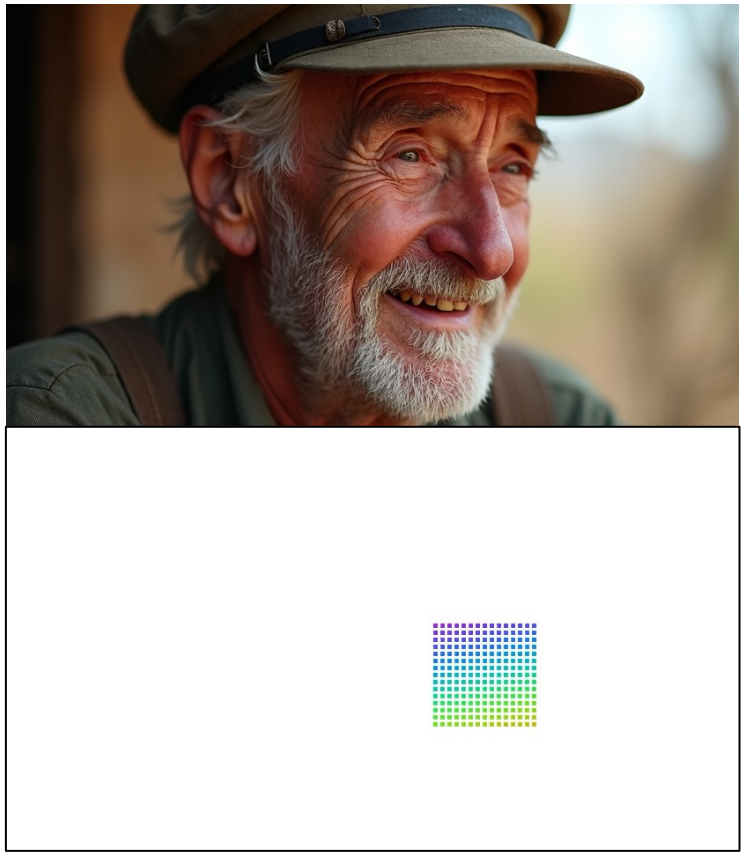


Base Model

- Wan 2.1 (1.3B) / Wan 2.2 (5B)

Joint Guidance with Text and Motion Conditions

Inputs



CFG: $\hat{v} = v(\mathbf{x}_t, \varnothing) + w \cdot \big(v(\mathbf{x}_t, c) - v(\mathbf{x}_t, \varnothing)\big)$

(or ance

Joint Guidance

$$\hat{v} = v_{\text{base}} + w_t \cdot \Big(v(c_t, c_m) - v(\varnothing, c_m)\Big) + w_m \cdot \Big(v(c_t, c_m) - v(c_t, \varnothing)\Big)$$

Joint Guidance with Text and Motion Conditions

Inputs

"**rainbow appears** in background"

CFG: $\hat{v} = v(\mathbf{x}_t, \varnothing) + w \cdot \Big(v(\mathbf{x}_t, c) - v(\mathbf{x}_t, \varnothing)\Big)$

(or ance

Joint Guidance

$$\hat{v} = v_{\text{base}} + w_t \cdot \Big(v(c_t, c_m) - v(\varnothing, c_m)\Big) + w_m \cdot \Big(v(c_t, c_m) - v(c_t, \varnothing)\Big)$$

Distilling into a causal, few-step student

We have a great teacher now, but it's …

`0.79 FPS (~1 min. for 5s vid.)`

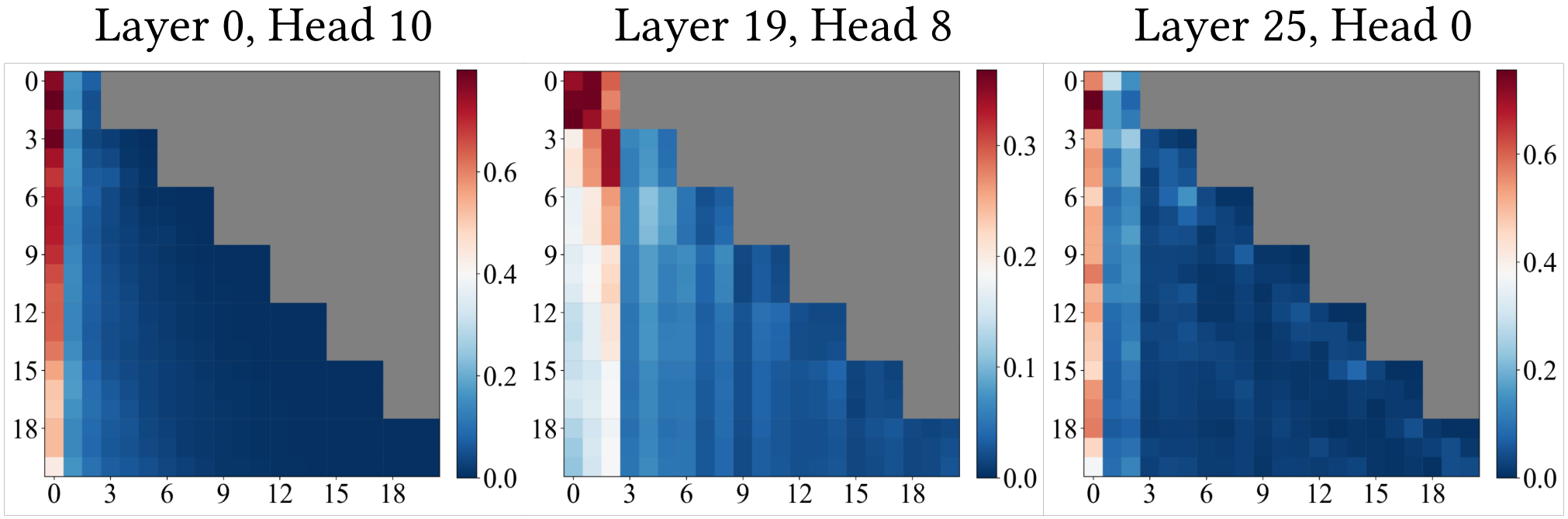
`Offline Processing`

`Maximum 5s (81f) duration`

Yin et al., "CausVid", CVPR 2025

Huang et al., "Self Forcing", NeurIPS 2025

Distilling into a causal, few-step student

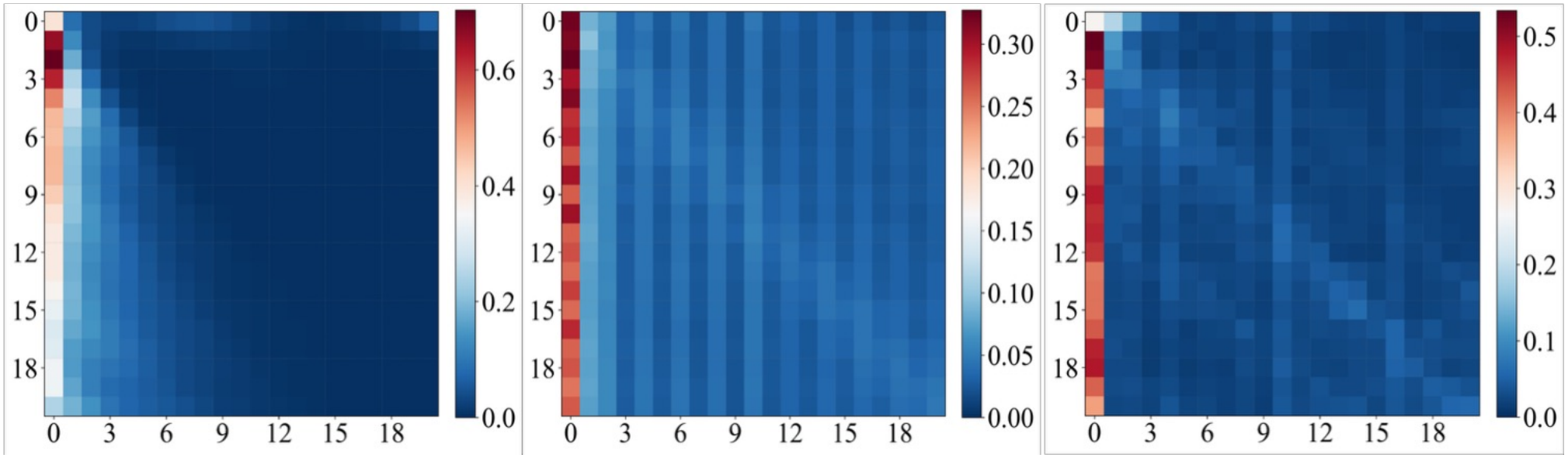


(Chunk-wise) Autoregressive Diffusion Model with Full Context

Xiao et al., “StreamingLLM”, ICLR 2024

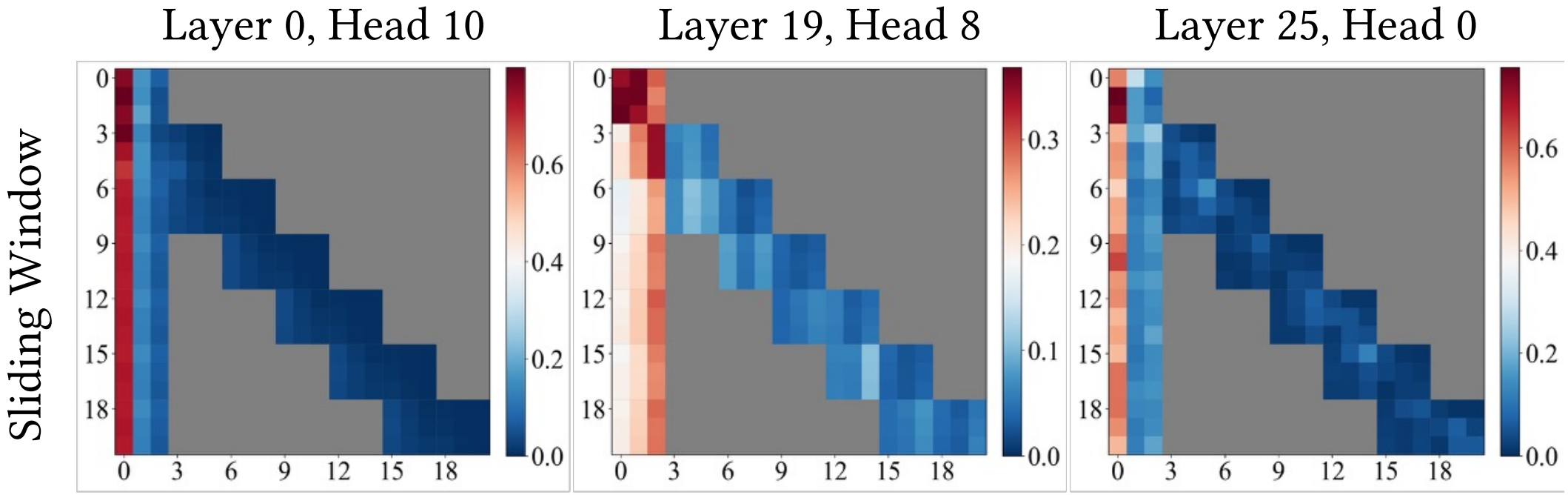
Distilling into a causal, few-step student

Layer 0, Head 10 | Layer 19, Head 8 | Layer 25, Head 0



Bidirectional Diffusion Model

Distilling into a causal, few-step student

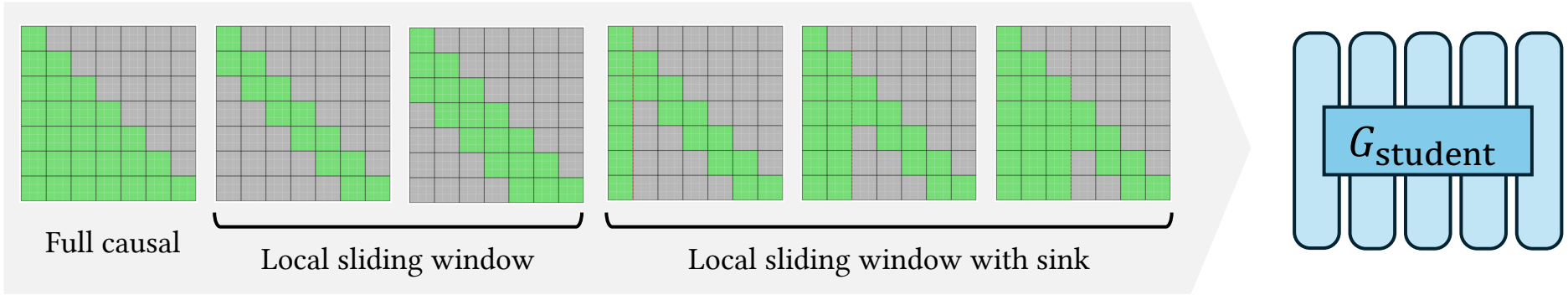


`Attention Sink + Small Sliding Window Attention`

Distilling into a causal, few-step student

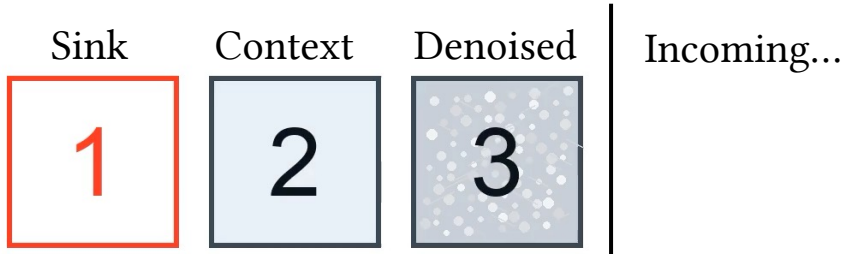
Stage 2: Causal Distillation

Phase 1: ODE initialization

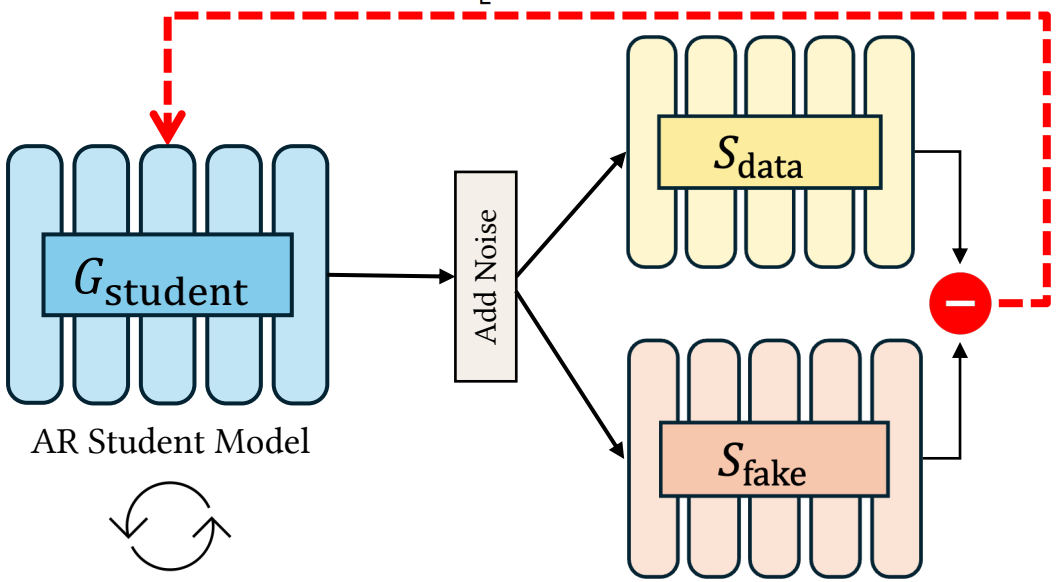


Phase 2: Self-Forcing distillation with rollouts & DMD

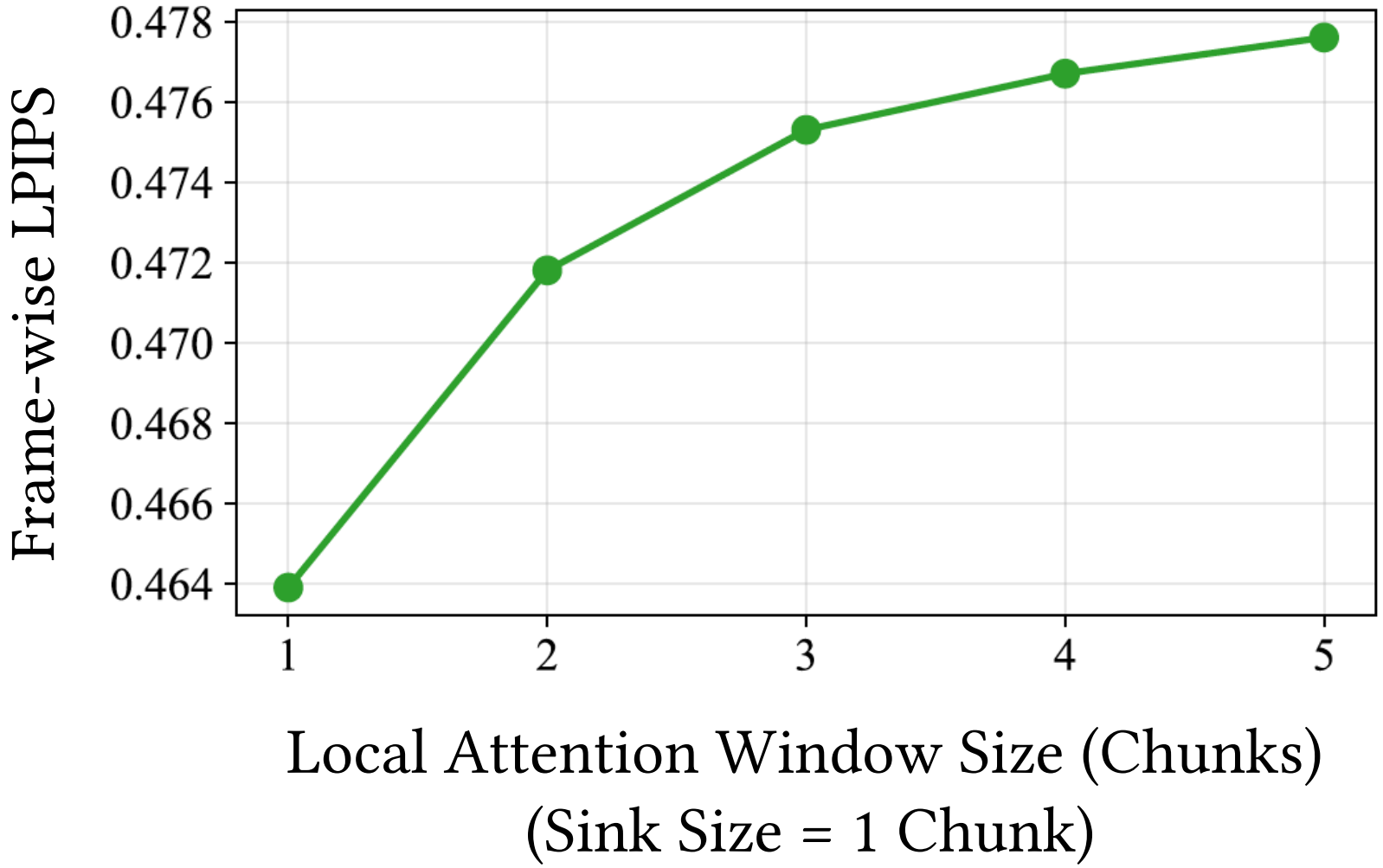
$$\nabla_\theta \mathcal{L}_{\text{DMD}} \approx -\mathbb{E}_{t,\hat{\mathbf{z}}_0}\left[\left(s_{\text{real}}(\Psi(\hat{\mathbf{z}}_0,t),t) - s_{\text{fake}}(\Psi(\hat{\mathbf{z}}_0,t),t)\right)\cdot\frac{\partial\hat{\mathbf{z}}_0}{\partial\theta}\right],$$



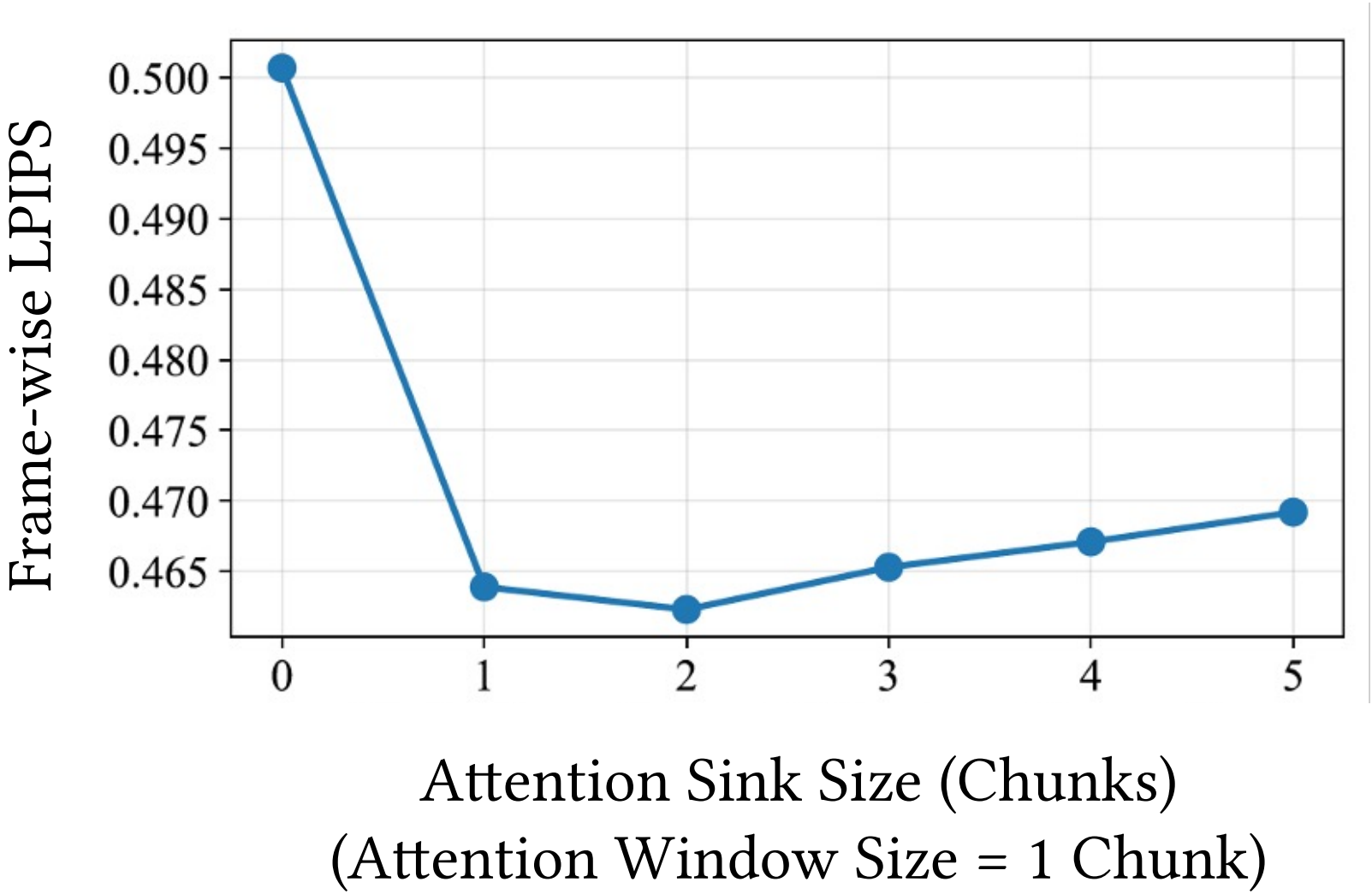
Simulated Rolling KV Cache **during Training**!



Distilling into a causal, few-step student

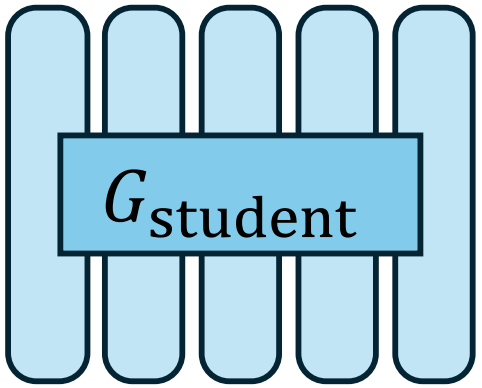


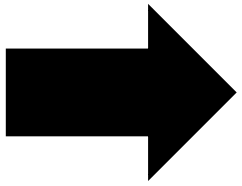
Distilling into a causal, few-step student

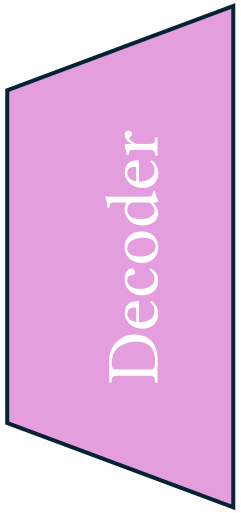


Faster streaming with Tiny VAE

Latent Space







Pixel (RGB) Space

Few-step AR DiT

53% Wall Time

Causal Decoder

47% Wall Time

→ Faster & Smaller VAE[1]!

16.7 → **29.5 FPS (1.3B, 480p)**
10.4 → **23.9 FPS (5B, 720p)**

[1]**@madebyolin:** https://github.com/madebyollin/taehv

Quantitative Evaluations

Table 1: **Benchmark on Motion Transfer (Reconstruction).**

Method	Backbone & Resolution	FPS	DAVIS Validation Set				Sora Demo Subset			
			PSNR	SSIM	LPIPS	EPE	PSNR	SSIM	LPIPS	EPE
Image Conductor (Li et al., 2025d)	AnimateDiff (256P)	2.98	11.30	0.214	0.664	91.64	10.29	0.192	0.644	31.22
Go-With-The-Flow Burgert et al. (2025)	CogVideoX-5B (480P)	0.60	15.62	0.392	0.490	41.99	14.59	0.410	0.425	10.27
Diffusion-As-Shader (Gu et al., 2025b)	CogVideoX-5B (480P)	0.29	15.80	0.372	0.483	40.23	14.51	0.382	0.437	18.76
ATI (Wang et al., 2025b)	Wan 2.1-14B (480P)	0.23	15.33	0.374	0.473	17.41	16.04	0.502	0.366	6.12
Ours Teacher (Joint CFG)	Wan 2.1-1.3B (480P)	0.79	**16.61**	**0.477**	**0.427**	**5.35**	**17.82**	**0.586**	0.333	**2.71**
Ours Causal (Distilled)	Wan 2.1-1.3B (480P)	**16.7**	16.20	0.447	0.443	7.80	16.67	0.531	0.360	4.21
Ours Teacher (Joint CFG)	Wan 2.2-5B (720P)	0.74	16.10	0.466	**0.427**	7.86	17.18	0.571	**0.331**	3.16
Ours Causal (Distilled)	Wan 2.2-5B (720P)	10.4	16.30	0.456	0.438	11.18	16.62	0.545	0.343	4.30

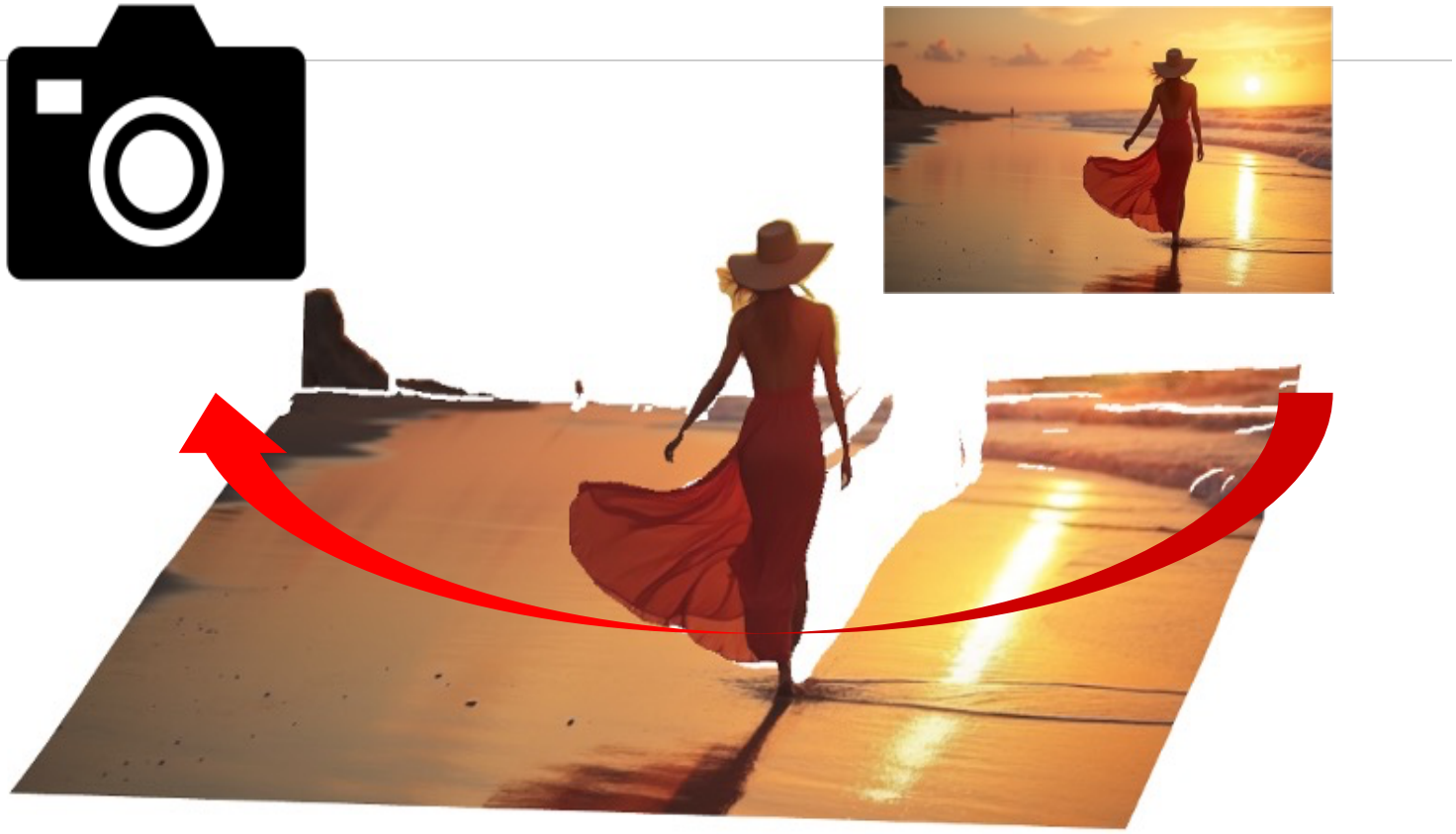
Table 2: **Evaluation on Novel View Synthesis.**

Method	Resolution	FPS	LLFF		
			PSNR	SSIM	LPIPS
DepthSplat (Xu et al., 2025)	576P	1.40	13.9	0.28	0.30
ViewCrafter (Yu et al., 2024)	576P	0.26	14.0	0.30	0.30
SEVA (Yu et al., 2024)	576P	0.20	14.1	0.30	0.29
Ours Teacher (1.3B)	480P	0.79	**16.0**	**0.42**	**0.21**
Ours Causal (1.3B)	480P	**16.7**	15.7	0.38	0.23
Ours Teacher (5B)	720P	0.74	14.0	0.40	0.22
Ours Causal (5B)	720P	10.4	15.0	0.39	0.23

Table 3: **Comparing track representation methods.** Our sinusoidal PE with learnable track head outperforms RGB-VAE in both quality and efficiency, achieving 40× faster encoding critical for real-time streaming.

Method	Time (ms)	DAVIS / Sora			
		PSNR	SSIM	LPIPS	EPE
RGB-VAE	1053	16.03 / 16.99	0.433 / 0.544	0.463 / 0.363	8.57 / 3.96
PE-Head	**24.8**	**16.29 / 17.15**	**0.452 / 0.559**	**0.456 / 0.359**	**6.54 / 3.13**

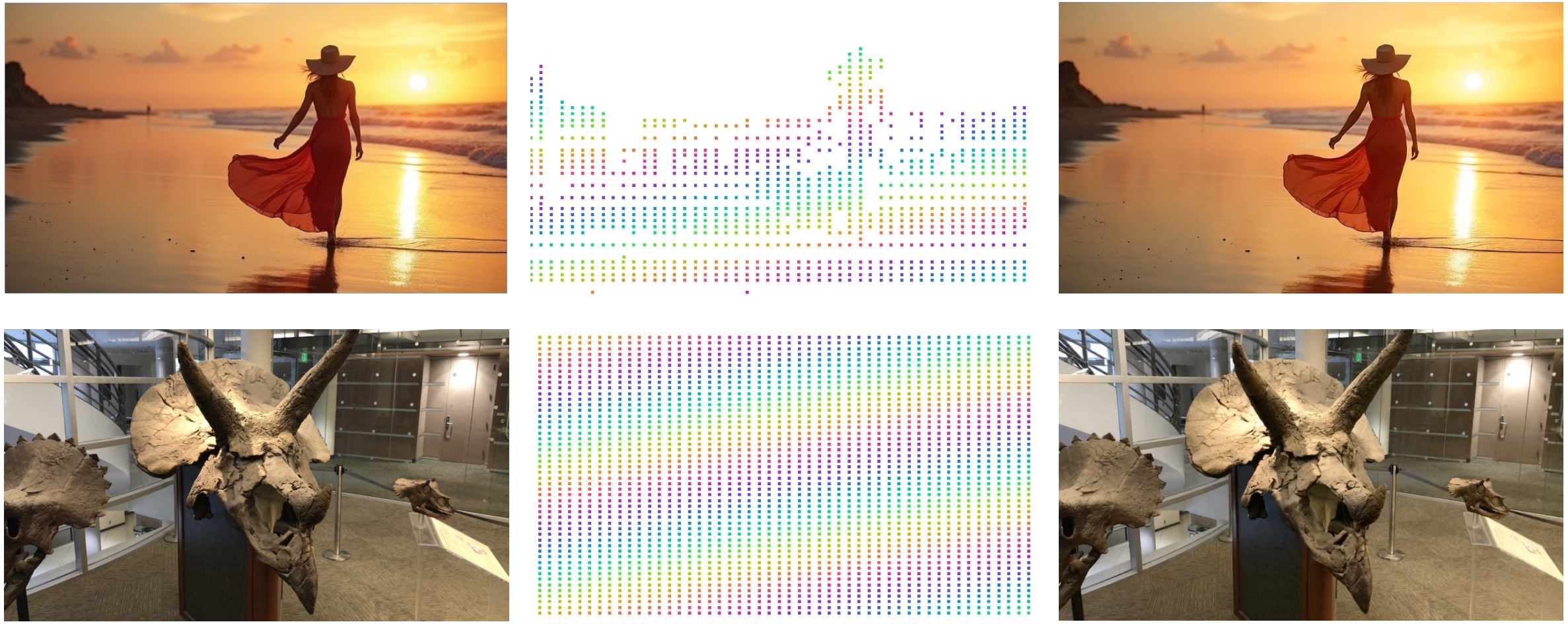
Downstream Applications (Camera Control)



Wang et al., “MoGE-2”, CVPR 2025

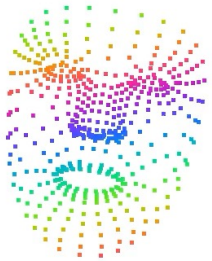
Yang et al., “Depth Anything V2”, NeurIPS 2024

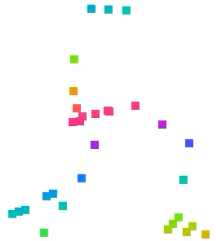
Downstream Applications (Camera Control)

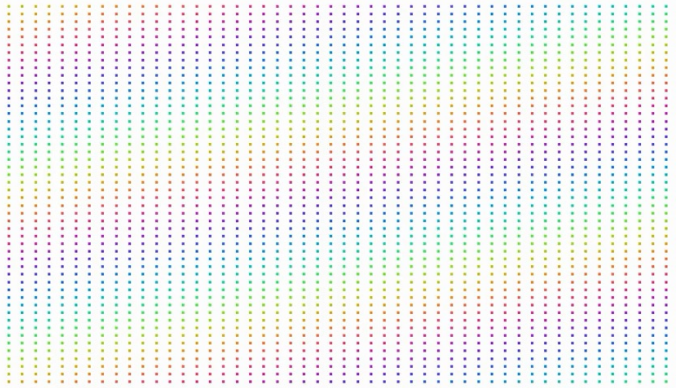


Real Time AR Diffusion as a Renderer!

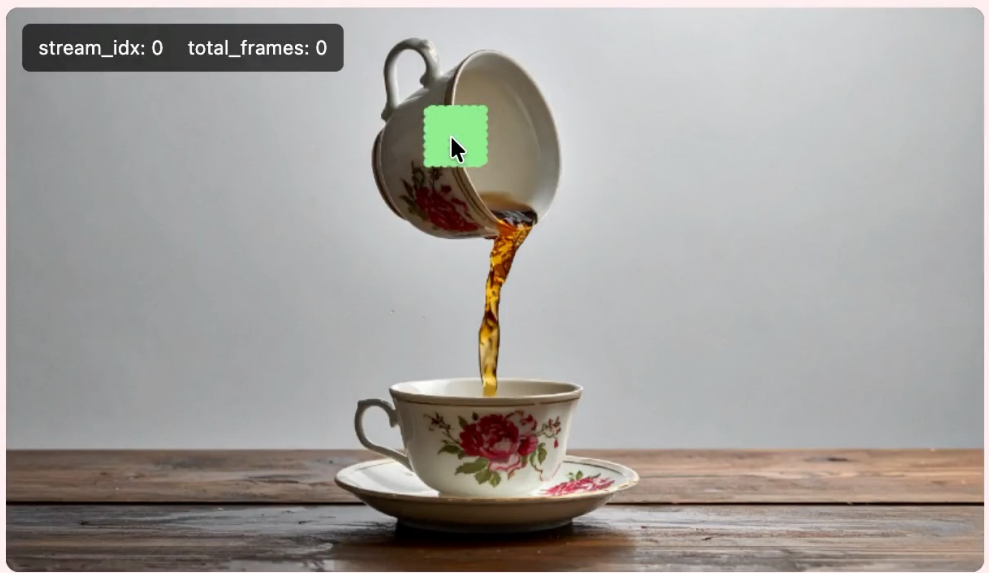
Downstream Applications (RT Motion Transfer)

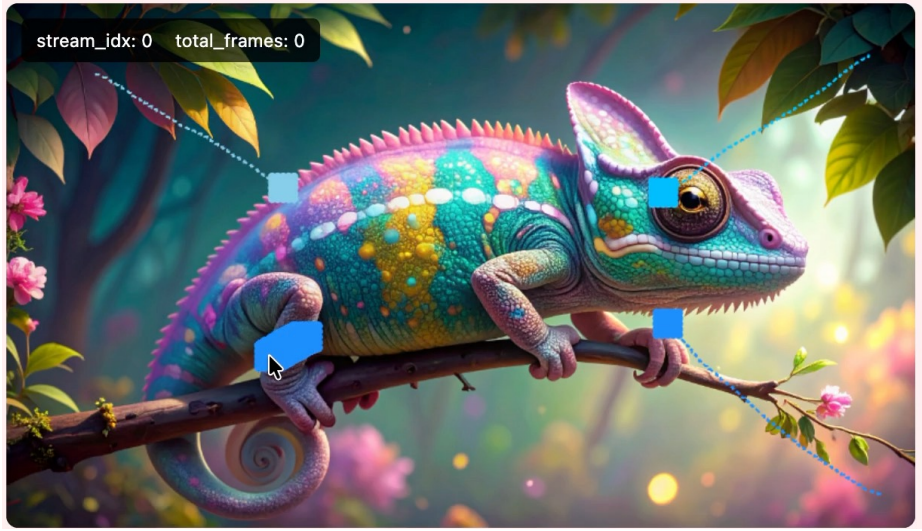




“Motion” Streaming Demos



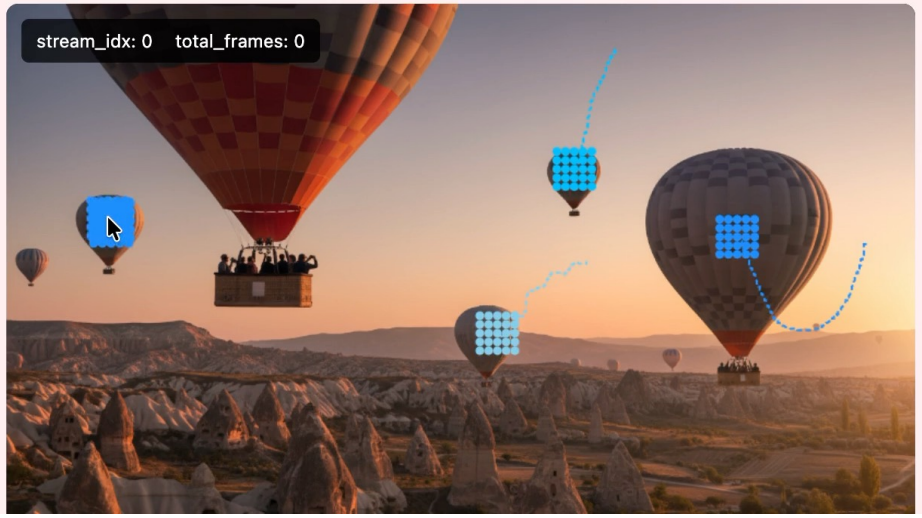
“Motion” Streaming Demos

Interactive Canvas



Generated Video

Interactive Canvas



Generated Video

“Motion” Streaming Demos

Interactive Canvas

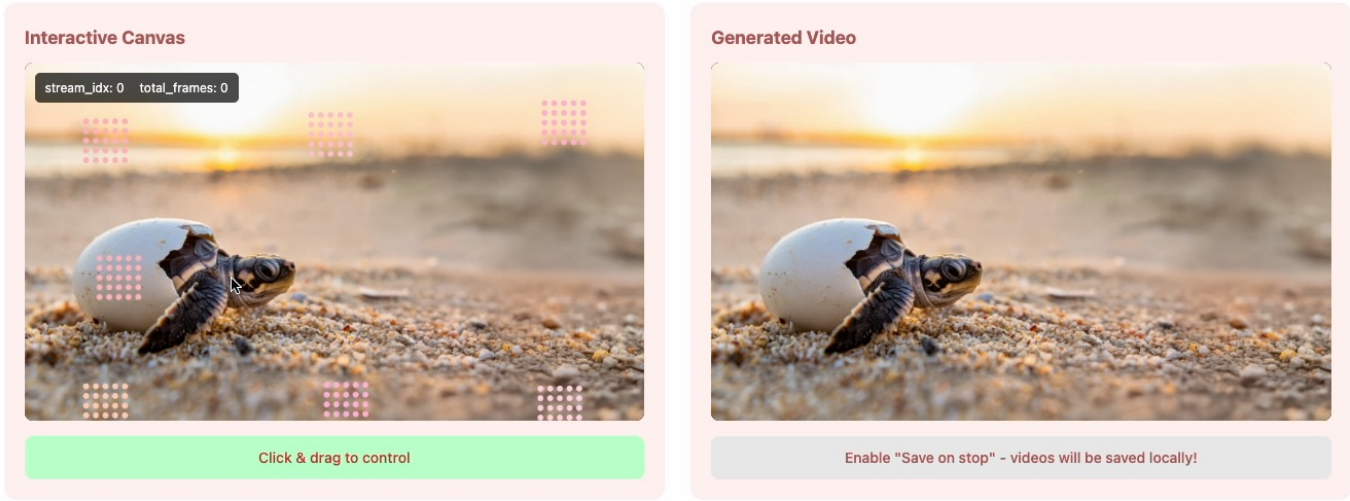
Click & drag to control

Generated Video

Enable "Save on stop" - videos will be saved locally!

Limitations

Limited Memory



Complex Motions (backbone)

Thank You (Questions!)

Concurrent Works (Acknowledgements)
LongLive / Rolling-Forcing / Reward-Forcing / Context-Forcing /
Self-Forcing++ / Matrix-Game / RELIC / …