Motion Consistency Model: Accelerating Video Diffusion with Disentangled Motion-Appearance Distillation

Yuanhao Zhai[1], Kevin Lin[2], Zhengyuan Yang[2], Linjie Li[2], Jianfeng Wang[2], Chung-Ching Lin[2], David Doermann[1], Junsong Yuan[1], and Lijuan Wang[2]

[1]State University of New York at Buffalo, [2]Microsoft

University at Buffalo The State University of New York

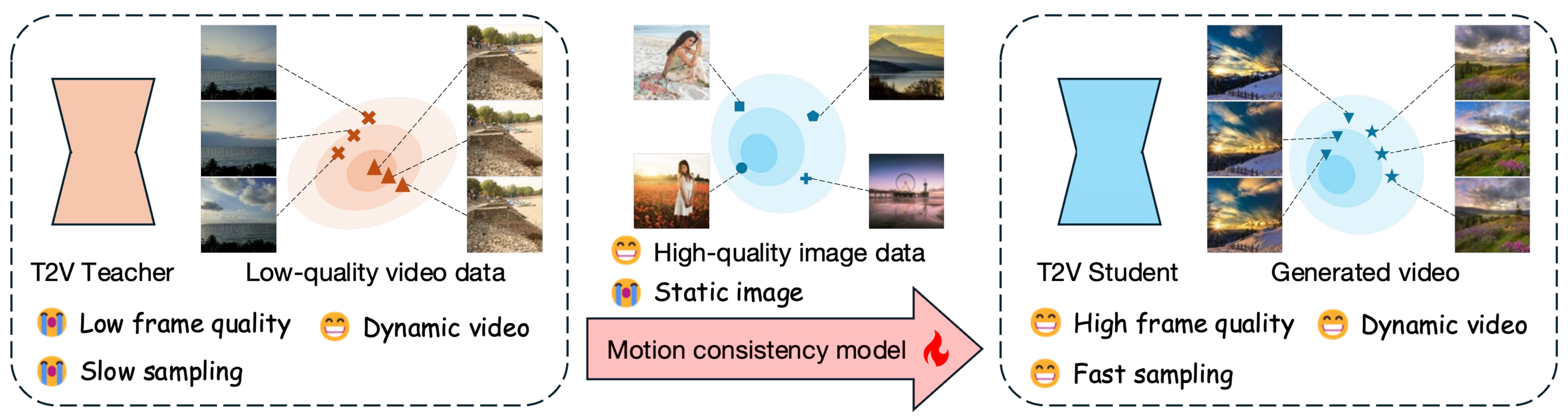
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- Video diffusion models suffer from **slow sampling**
- Video datasets often have **low-quality appearance**
- **High-quality image datasets** are underutilized in video model training

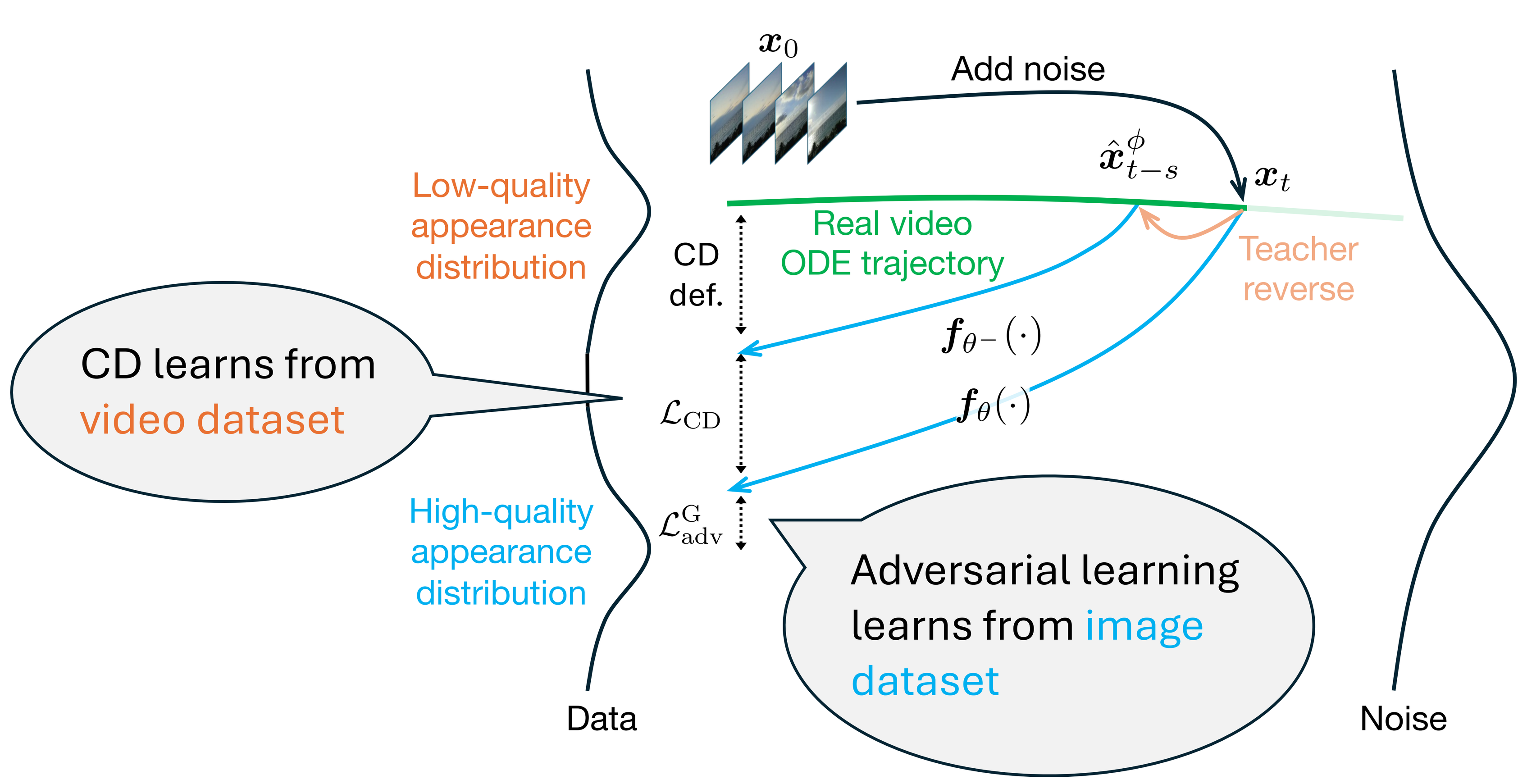
How can we speed up video diffusion and improve video appearance quality?

We present **motion consistency model**, a video diffusion distillation method that
- **Accelerate sampling**
- **Enhance frame appearance** by leveraging image datasets



Baseline

- Video latent consistency distillation (CD)
  - Learns to generate from the video dataset
- Frame-wise adversarial learning
  - Learns the appearance from the image dataset

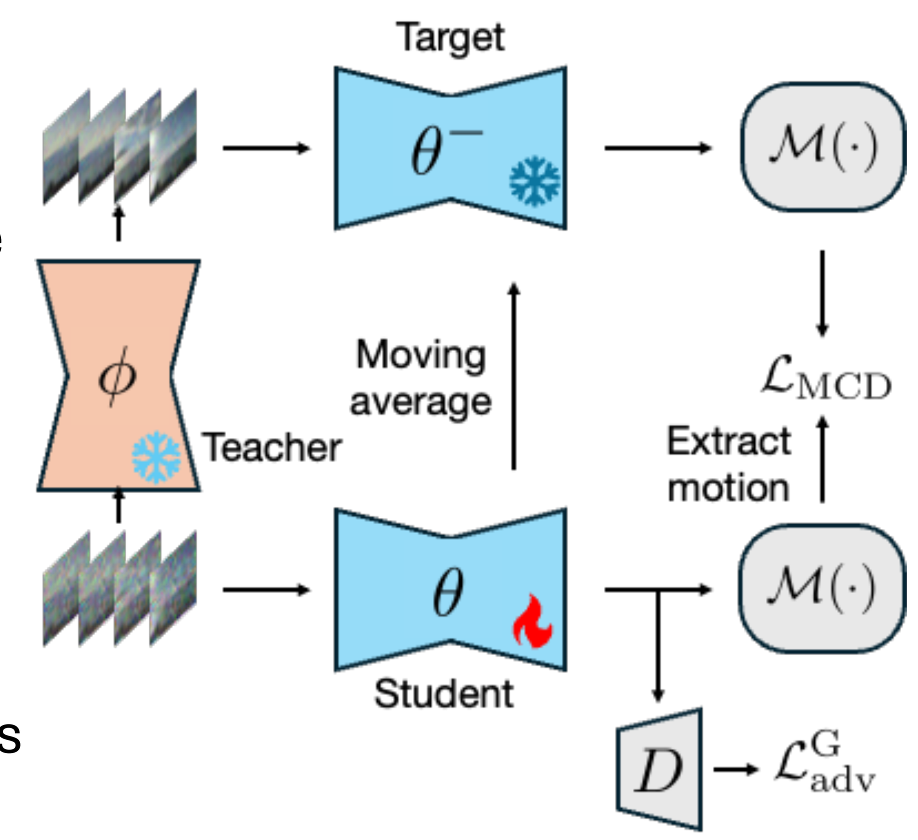


Disentangled motion consistency distillation

- Conflict objectives in baseline
  - CD also learns low-quality frame from the video dataset
  - Adversarial learning learns high-quality frame from images

How about disentangling motion and appearance learning
- Extract motion from the video latent
- Apply CD only on the motion (MCD)

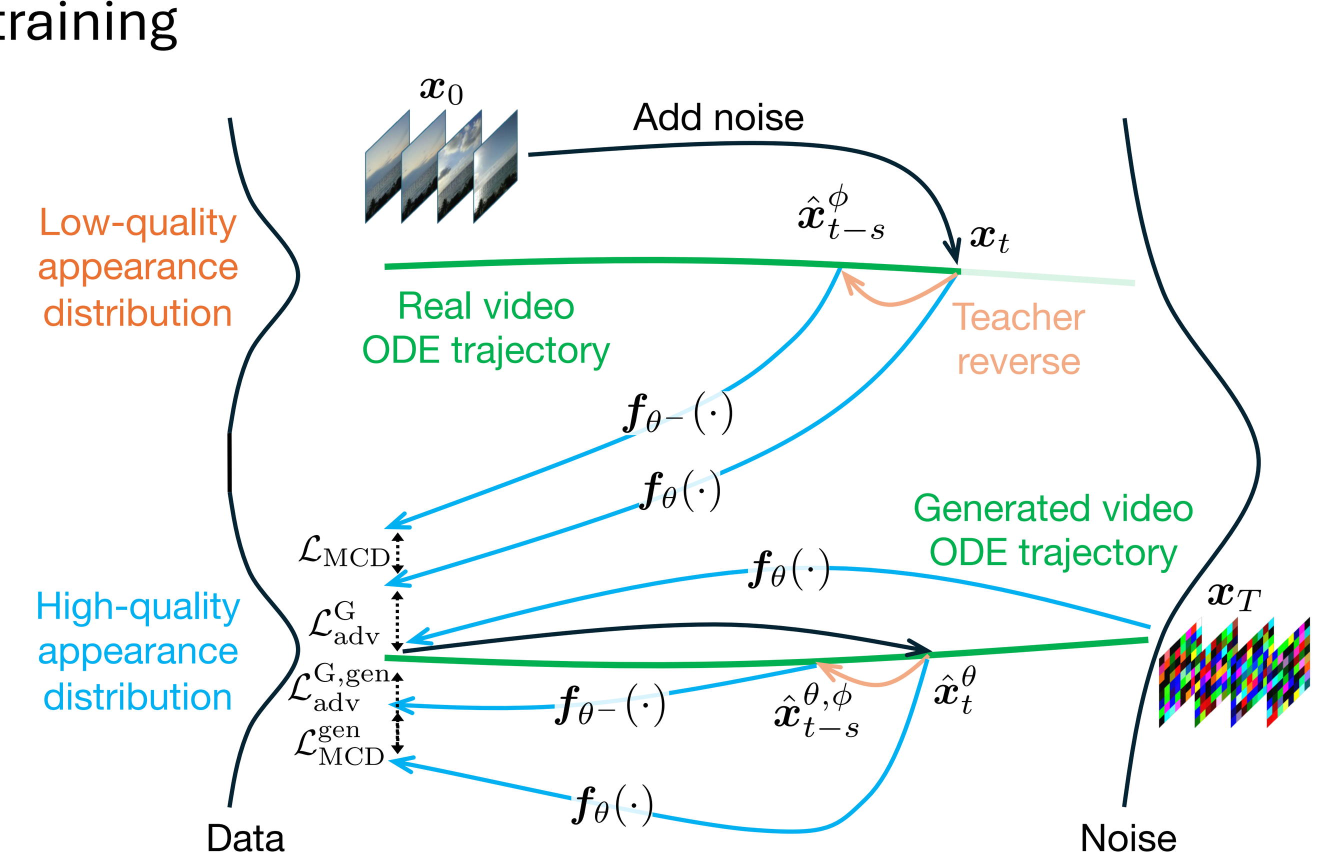
- MCD only learns the motion
- Adversarial learning learns the appearance



Mixed trajectory distillation

- Training-inference discrepancy
  - Training: ODE trajectories sampled from low-quality video
  - Inference: sample in the high-quality video space

Simulate inference-time ODE trajectories using multi-step sampling
- Represent high-quality appearance
- Apply MCD and adversarial learning on latents sampled from these trajectories
- Mixing the real- and generated-video ODE trajectories for training



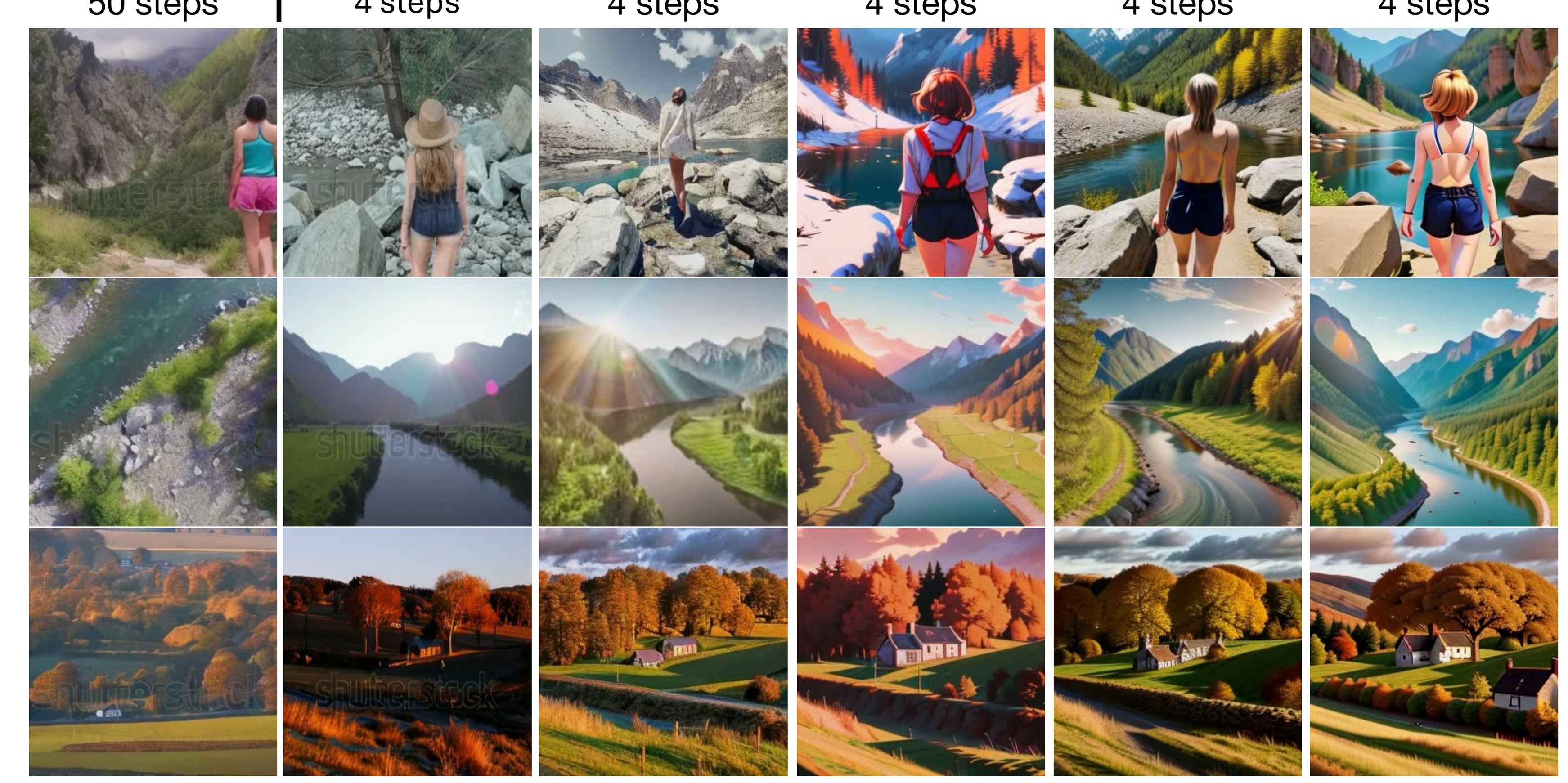
Results

High-resolution video generation (4 steps)

Pose-conditioned video generation (4 steps)

Frame quality improvement

Teacher 50 steps | Ours w/ WebVid 4 steps | LAION-aes 4 steps | Anime 4 steps | Realistic 4 steps | 3D cartoon 4 steps



Quantitative results

Teacher	Method	FVD@Step ↓ 1	2	4	8	CLIPSIM@Step ↑ 1	2	4	8
AnimateDiff [19] 512 × 512 × 16	DDIM [54]	4782	4350	2774	933	20.90	20.94	22.87	27.36
	DPM++ [37]	2004	1447	876	794	22.93	24.5	27.62	29.10
	LCM [38] (our impl.)	1276	1180	956	830	25.75	27.33	28.37	28.65
	AnimateLCM [61]	1578	1278	824	740	27.56	28.52	29.58	27.67
	AnimateDiff-Lightning [35]	1260	1259	892	932	27.38	28.77	29.12	28.77
	MCM (ours)	**1197**	**1036**	**801**	**675**	**28.95**	**29.40**	**29.64**	**29.13**
ModelScopeT2V [62] 256 × 256 × 16	DDIM [54]	6459	2305	1445	841	21.49	20.33	22.57	26.76
	DPM++ [37]	2039	1336	467	552	23.48	24.85	28.51	**29.70**
	LCM [38] (our impl.)	1094	820	713	717	26.78	28.01	28.45	29.01
	MCM (ours)	**501**	**434**	**414**	**482**	**28.37**	**29.02**	**28.86**	28.28