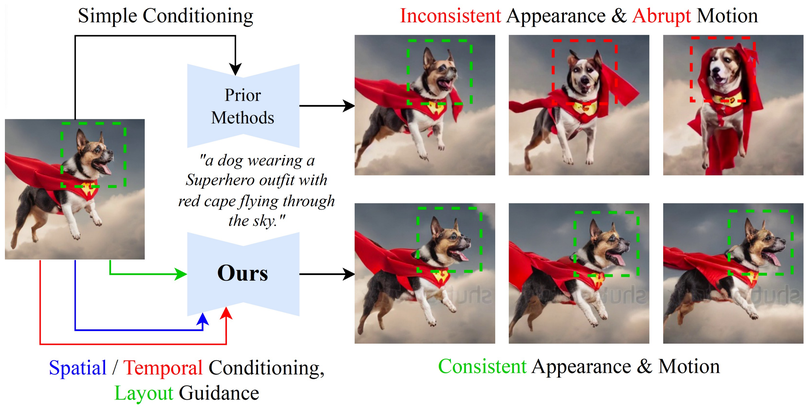
Towards Practical Video Generation and Editing

Weiming Ren

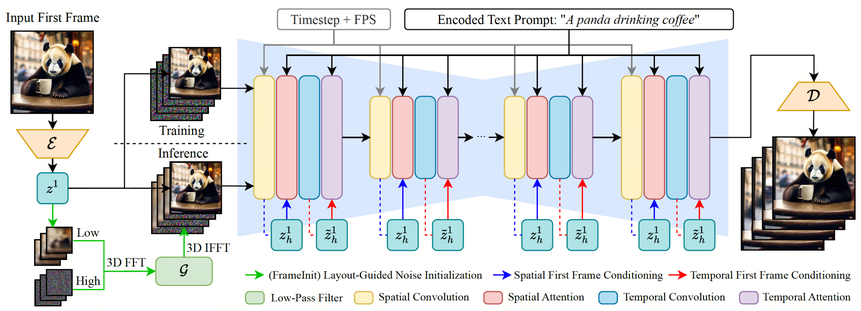
David R. Cheriton School of Computer Science, University of Waterloo

ConsistI2V AnyV2V

ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation



- Problem with text-to-video (T2V) generation: lack of **precise control** of video contents (e.g. object appearance).
- Solution: adding additional **image conditioning** in videogen
- Existing methods: often result in inconsistent appearance and motion due to **weak image conditioning.**
- We propose a novel diffusion-based I2V generation framework to enhance visual consistency, which contains:
  1. **Spatiotemporal attention** over the first frame to maintain spatial and motion consistency.
  2. Inference-time noise initialization strategy that uses the **low-frequency band from the first frame** to stabilize video generation (FrameInit).



- Spatiotemporal feature conditioning in attention layers

Spatial Attention $Q_s = W_s^Q z^i, K_s = W_s^K z^i, V_s = W_s^V z^i,$

Temporal Attention $Q_s = W_s^Q z^i, K_s' = W_s^K [z^i, z^1], V_s' = W_s^V [z^i, z^1]$

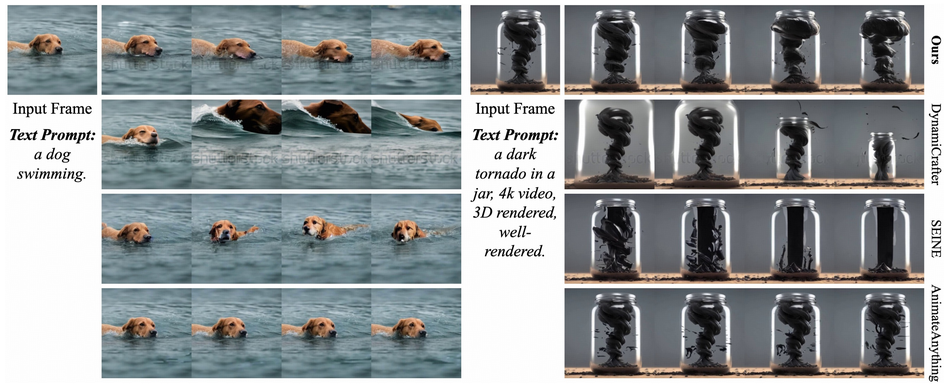
- Low-frequency component for noise initialization (FrameInit)

$$\mathcal{F}^{low}_{\mathbf{z}_\tau} = \texttt{FFT\_3D}(\mathbf{z}_\tau) \odot \mathcal{G}(D_0),$$
$$\mathcal{F}^{high}_{\epsilon} = \texttt{FFT\_3D}(\epsilon) \odot (1 - \mathcal{G}(D_0)),$$
$$\epsilon' = \texttt{IFFT\_3D}(\mathcal{F}^{low}_{\mathbf{z}_\tau} + \mathcal{F}^{high}_{\epsilon}),$$

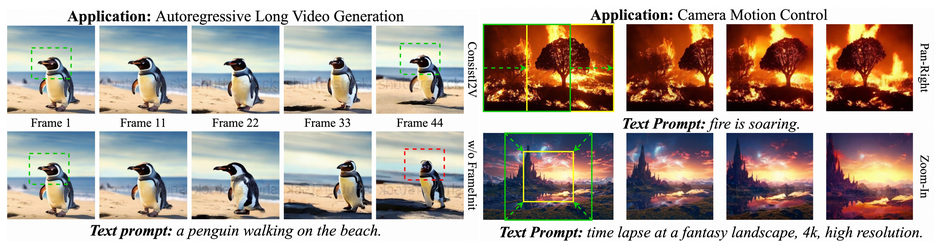
ConsistI2V achieves better results with less training data

Method	#Data	UCF-101			MSR-VTT		Human Eval: Consistency	
		FVD ↓	IS ↑	FID ↓	FVD ↓	CLIPSIM ↑	Appearance ↑	Motion ↑
AnimateAnything	10M+20K†	642.64	**63.87**	**10.00**	218.10	0.2661	43.07%	20.26%
I2VGen-XL	35M	597.42	18.20	42.39	270.78	0.2541	1.79%	9.43%
DynamiCrafter	10M+10M†	404.50	41.97	32.35	219.31	0.2659	44.49%	31.10%
SEINE	25M+10M†	306.49	54.02	26.00	152.63	**0.2774**	48.16%	36.76%
ConsistI2V	10M	**177.66**	56.22	15.74	**104.58**	0.2674	**53.62%**	**37.04%**

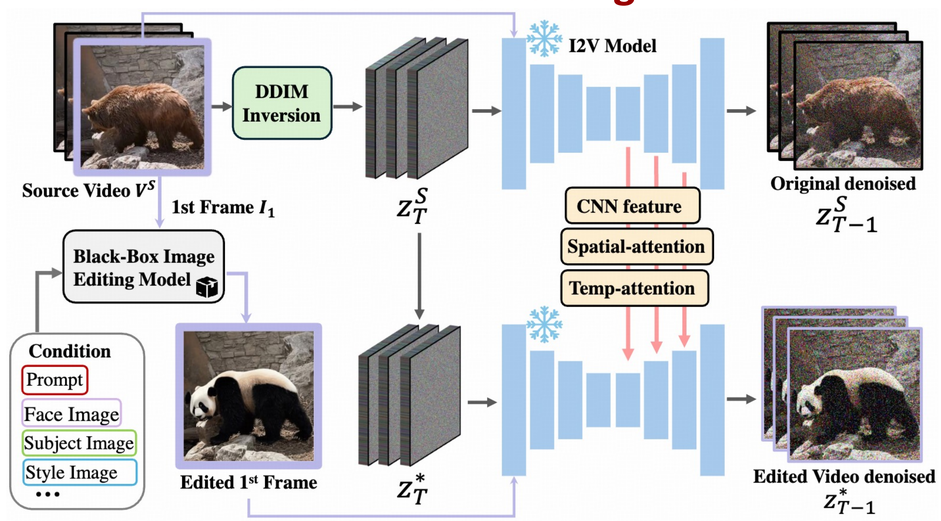
Visual Comparisons of ConsistI2V vs. Baseline Methods



Applications of ConsistI2V



AnyV2V: A Tuning-Free Framework For Any Video-to-Video Editing Tasks



- Key idea: pretrained I2V generation models are **zero-shot** video editing models.
- Video editing pipeline: edit the **first frame** -> get video latents using **DDIM inversion** -> animate the first frame to get the edited video using **I2V models**
- How to preserve source video content?
  - Structural guidance: **DDIM Inversion**
  - Appearance guidance: spatial feature injection
  - Motion guidance: temporal feature injection

(Appearance and motion guidance are adapted from the **PnP framework**)

- Highlight: AnyV2V works with **any** image editing methods and **any** I2V generation models.

AnyV2V can perform a wide range of editing tasks