

OF COMPUTER SCIENCE



VECTOR INSTITUTE

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



Spatial / Temporal Conditioning, Layout Guidance

Consistent Appearance & Motion

- 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).

Input First Frame	Timestep + FPS Encoded Text Prompt: "A panda drinking coffee
$ \begin{array}{c} $	\rightarrow
Low 3D IFFT	$egin{array}{cccccccccccccccccccccccccccccccccccc$
High 3D FF I G	\rightarrow (FrameInit) Layout-Guided Noise Initialization \rightarrow Spatial First Frame Conditioning \rightarrow T Low-Pass Filter Spatial Convolution Spatial Attention Temporal Conv

- 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)

 $\mathcal{F}^{low}_{\mathbf{z}_{\tau}} = \mathtt{FFT}_{3D}(\mathbf{z}_{\tau}) \odot \mathcal{G}(D_0),$ $\mathcal{F}^{high}_{\epsilon} = \mathrm{FFT}_{3D}(\epsilon) \odot (1 - \mathcal{G}(D_0)),$ $\epsilon' = \text{IFFT}_{3D}(\mathcal{F}_{\mathbf{z}_{\tau}}^{low} + \mathcal{F}_{\epsilon}^{high}),$

Towards Practical Video Generation and Editing Weiming Ren

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





emporal First Frame Conditioning olution 🛛 Temporal Attentior



...

ConsistI2V achieves better results with less training data

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

Visual Comparisons of ConsistI2V vs. Baseline Methods



tornado in d jar, 4k video 3D rendered, wellrendered.

Applications of ConsistI2V



Text prompt: a penguin walking on the beach



 Z_T^*

Edited 1st Frame



Text Prompt: time lapse at a fantasy landscape, 4k, high resolution

Edited Video denoised Z_{T-1}^{*}

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 Prompt-based Editing







- Key idea: pretrained I2V generation models are