CVPR Nashville JUNE 11-15, 2025

SnapGen-V: Generating a Five-Second Video within Five Seconds on a Mobile Device

Yushu Wu, Zhixing Zhang, Yanyu Li, Yanwu Xu, Anil Kag, Yang Sui, Huseyin Coskun, Ke Ma,
Aleksei Lebedev, Ju Hu, Dimitris Metaxas, Yanzhi Wang, Sergey Tulyakov, Jian Ren

Northeastern University

RUTGERS

Video Generation is Computationally Intensive

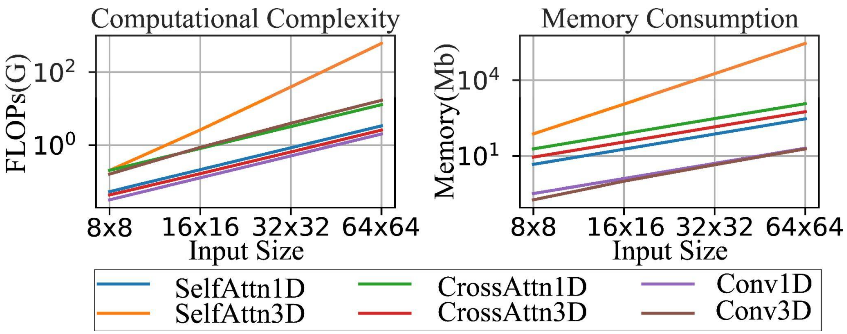
- **Diffusion Models are Powerful:** Recent success in generating cinematic, high-resolution videos.
- **High Computational Cost:** Video generation is significantly more demanding than image generation.
- **Cloud-Reliant:** Most advanced models run on powerful cloud servers (e.g., CogVideoX-5B takes 5 mins on an NVIDIA A100 GPU).
- **Accessibility Barrier:** This limits broader adoption, especially for mobile users.
- **The Gap:** Little focus on accelerating video models for mobile devices.

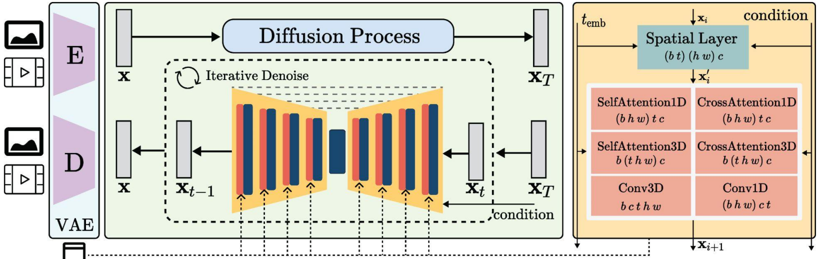
SnapGen-V: Real-Time Video on Your Phone

- **What is SnapGen-V?** A comprehensive acceleration framework to bring large-scale video diffusion to mobile devices.
- **Key Achievement:** Generates a 5-second, high-quality, motion-consistent video on an iPhone 16 Pro Max within 5 seconds.
- **Compact & Efficient:** Achieved with only 0.6B parameters.
- **Impact:** Shifts video generation from minutes on GPUs to seconds on mobile.
- **A First:** Represents the first successful mobile deployment of this kind of video diffusion model, showcasing real-time potential.

Efficient Architecture

- ❖ Efficient Spatial Backbone:
  - ➢ Started with a pre-trained text-to-image model (Stable Diffusion v1.5).
  - ➢ Pruned it to achieve 2.5x size compression and >10x speedup for the image backbone on mobile.
- ❖ Hardware-Efficient Temporal Layer Design:
  - ➢ Systematically investigated various temporal layers (attention, convolutions).
  - ➢ Conducted a latency-memory joint architecture search to find the optimal design specifically for mobile constraints.

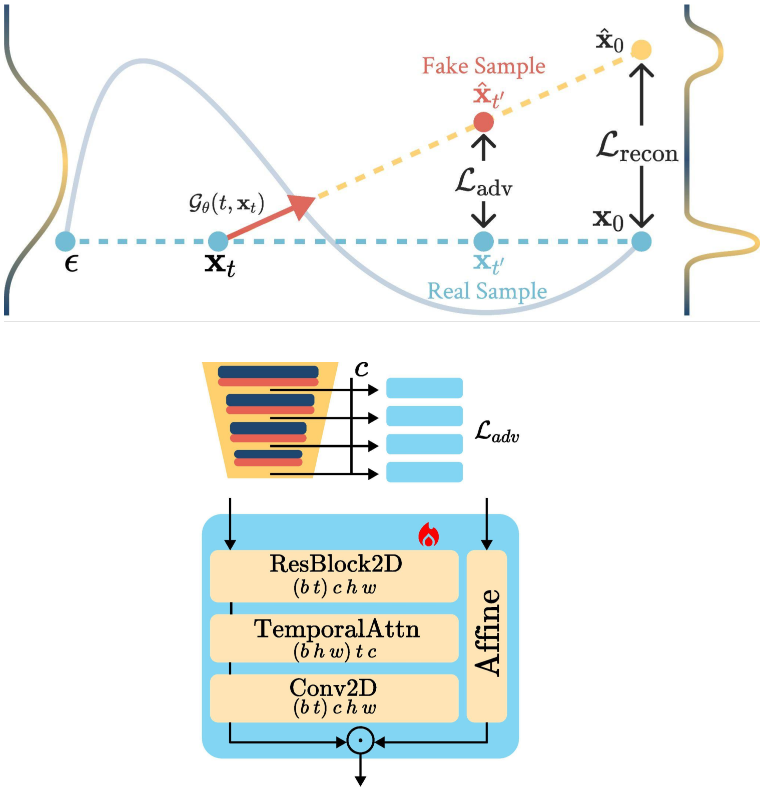




Latency and Memory Guided Architecture Search

Latent Adversarial Fine-tuning:

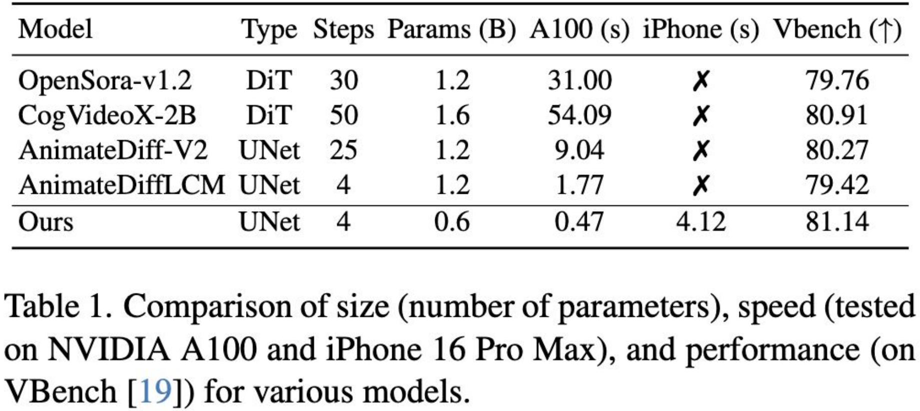
- ❖ Developed a tailored adversarial fine-tuning method.
- ❖ Reduced denoising steps from 25 to just 4 steps (and eliminated classifier-free guidance), leading to >12x speedup without sacrificing quality.
- ❖ Incorporating image-video discriminator heads for image-video joint-training.



Adversarial Fine-tuning

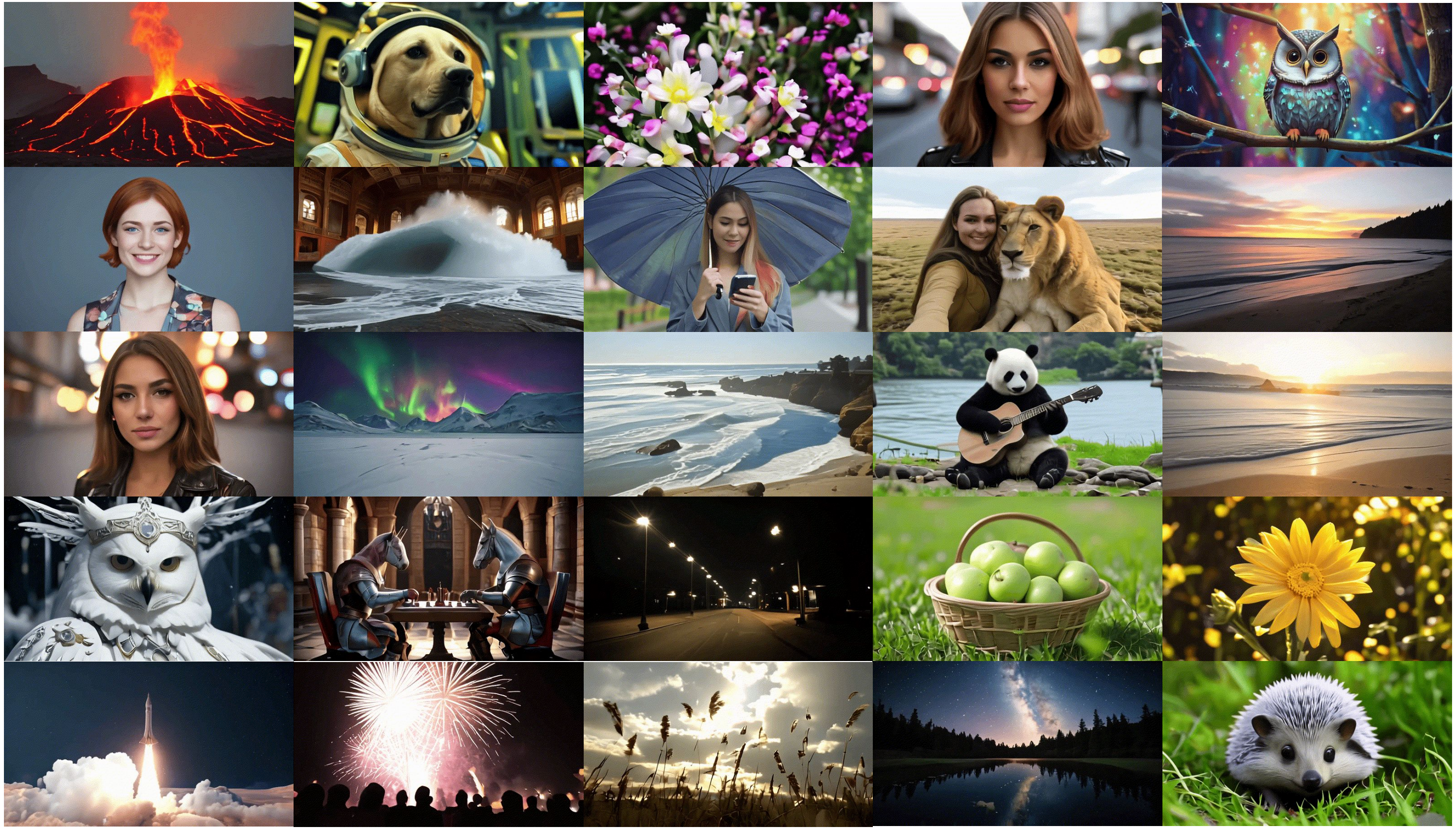
Results

- **Speed:** Generate 5-second video on iPhone 16 Pro Max in ~4.12 seconds.
- **Size:** Only 0.6 Billion parameters – significantly smaller than server-side models.
- **Quality:** Achieves a competitive VBench score of 81.14, outperforming several larger and slower models.



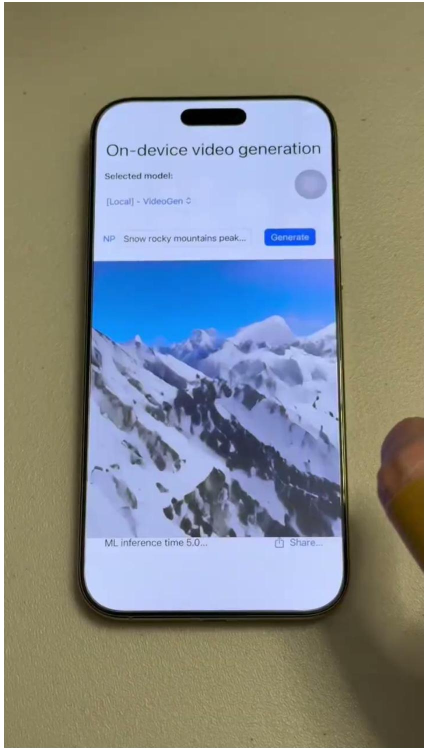
Model	Type	Steps	Params (B)	A100 (s)	iPhone (s)	Vbench (↑)
OpenSora-v1.2	DiT	30	1.2	31.00	✗	79.76
CogVideoX-2B	DiT	50	1.6	54.09	✗	80.91
AnimateDiff-V2	UNet	25	1.2	9.04	✗	80.27
AnimateDiffLCM	UNet	4	1.2	1.77	✗	79.42
Ours	UNet	4	0.6	0.47	4.12	81.14

Table 1. Comparison of size (number of parameters), speed (tested on NVIDIA A100 and iPhone 16 Pro Max), and performance (on VBench [19]) for various models.



Quality Results

Quality Results



On-device video generation

Selected model:

[Local] - VideoGen ◊

NP Snow rocky mountains peak...

ML inference time 5.0...

Demo

Thank You

Project Page