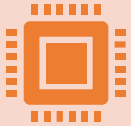
Fast and Memory-Efficient Video Diffusion Using Streamlined Inference

Northeastern University

Zheng Zhan, **Yushu Wu**, Yifan Gong, Zichong Meng, Zhenglun Kong, Changdi Yang, Geng Yuan, Wei Niu, Yanzhi Wang

Introduction



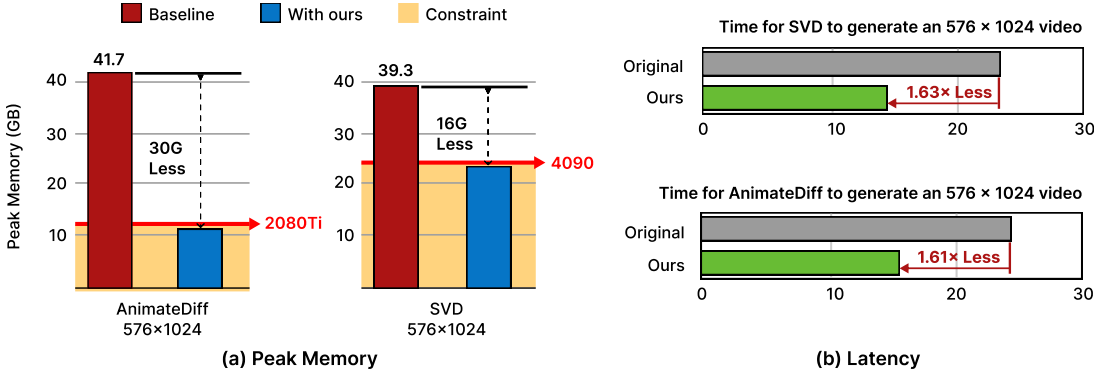
Diffusion models allow people to create visual content with text prompts for video generation.

Current SOTA Video Generation Model is not user friendly

- High-memory consumption
- Slow inference speed

Video Diffusion models are Memory and Computation Intensive

- Video Diffusion models need to inference multiple frames in a forward computation, leading ultra-high memory usage.
- High resolution video generation are slow in inference and usually get **out-of-memory** error.
- Unfeasible for consumer GPUs.

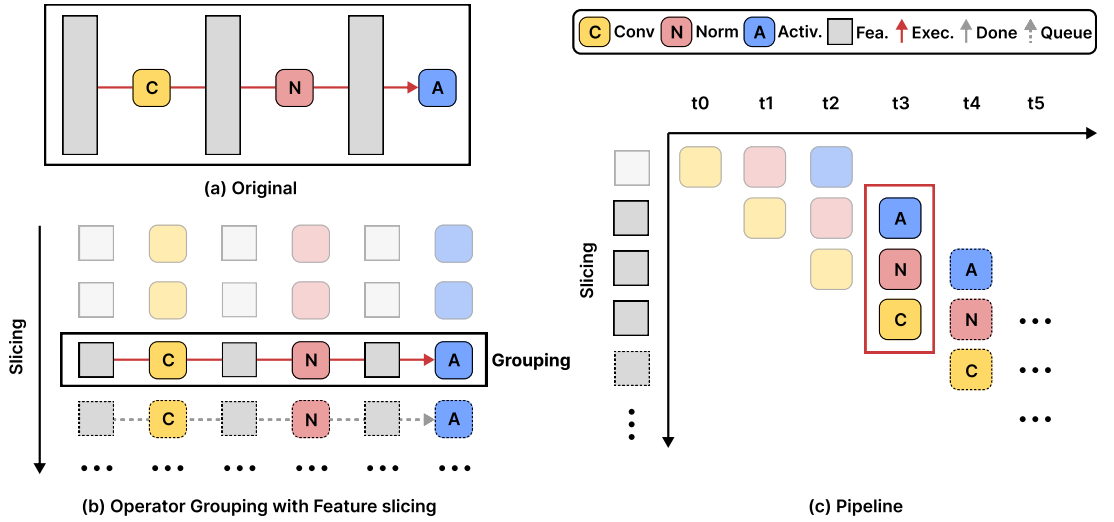


Streamlined Inference Framework

- Training Free
- Reduce Denoising Steps
- Preserving models performance

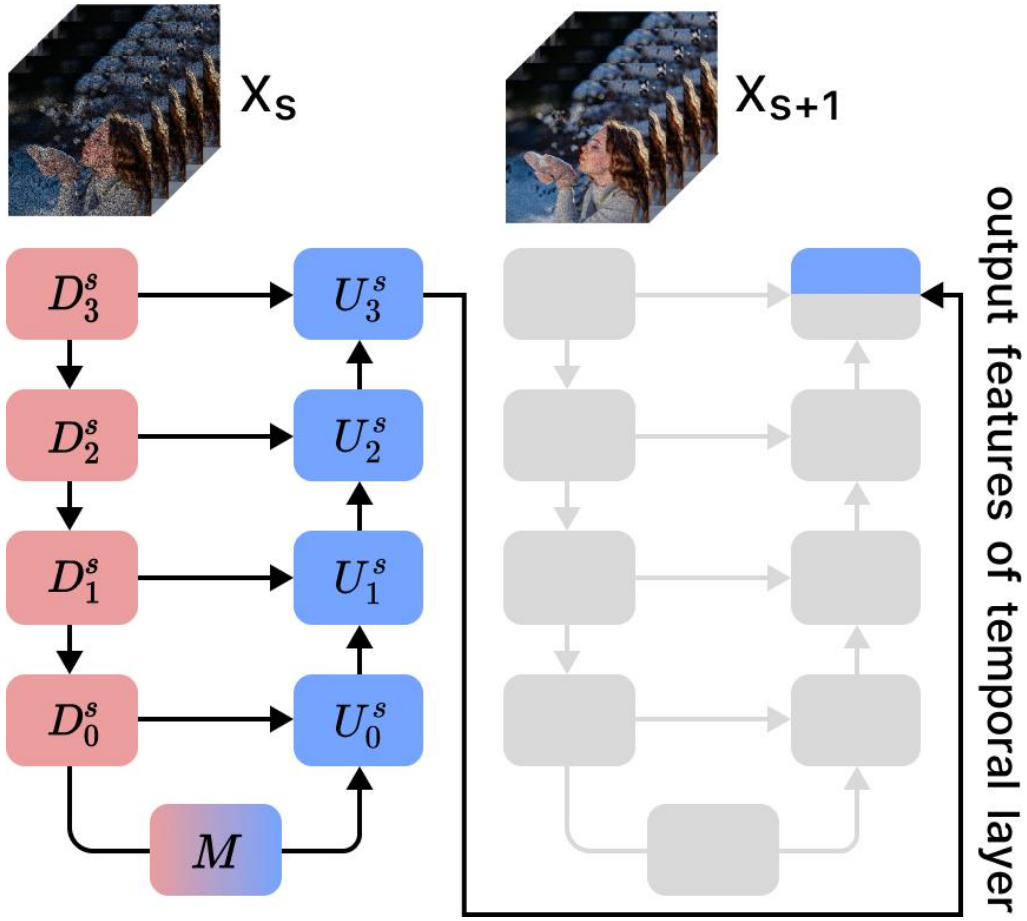
Feature Slicer and Operator Grouping with Pipeline

- Video Diffusion Models commonly adopt spatial-temporal architecture.
- Feature Slicer:
  - Spatial layer – slicing temporal dim
  - Temporal layer – slicing spatial dim
- Operator Grouping and pipelining:
  - Group as much as out-of-place operation



Step Rehash

- The high similarity of features after temporal layers.
- By reuse the high similarity features, we can reduce inference time.
- where-to-skip → when-to-skip



Experiments Results

Table 1: Comparison of our Streamlined Inference with baseline methods in video visual quality (on UCF101), PM (Peak Memory), and latency (measured with 50 runs with the average value).

Model	Method	FVD↓	CLIP-Score↑	512 × 512		576 × 1024	
				PM	Latency	PM	Latency
SVD #F=14	Original	307.7	29.25	20.91G	10.23s	39.49G	23.29s
	Naïve Slicing	1127.5	26.32	8.12G	31.85s	10.72G	65.56s
	Ours	340.6	28.98	13.67G	7.36s	23.42G	14.24s
SVD-XT #F=25	Original	387.9	28.18	31.97G	17.05s	61.17G	40.77s
	Naïve Slicing	2180.0	24.42	8.12G	59.86s	10.72G	121.82s
	Ours	424.7	27.94	19.37G	12.10s	36.32G	25.47s
AnimateDiff #F=16	Original	758.7	28.89	21.83G	9.65s	41.71G	24.38s
	Naïve Slicing	2403.9	26.63	7.22G	19.98s	9.92G	38.69s
	Ours	784.5	28.71	7.51G	7.08s	11.07G	15.15s

Visual Quality

Original

Ours

AnimateDiff

Stable Video Diffusion

More Results







NEURAL INFORMATION
PROCESSING SYSTEMS


NORTHEASTERN UNIVERSITY
LVX
VERITAS
VIRTVS
FOUNDED
BOSTON, MASSACHUSETTS

Thank You!