ApertureAttn: Native 4K Video Generation with Image-Only Supervision

Ruonan Yu[1], Zigeng Chen[1], Zhenxiong Tan[1], Songhua Liu[2]*, Xinchao Wang[1]*

National University of Singapore, xML Lab[1] Shanghai Jiao Tong University, Go-There Lab[2]

ruonan@u.nus.edu, liusonghua@sjtu.edu.cn, xinchao@nus.edu.sg

*Equal Corresponding Authors

NUS National University of Singapore xML Lab

NEURAL INFORMATION PROCESSING SYSTEMS

Native 4K video generation is expensive to train: Requires large-scale 4K video data

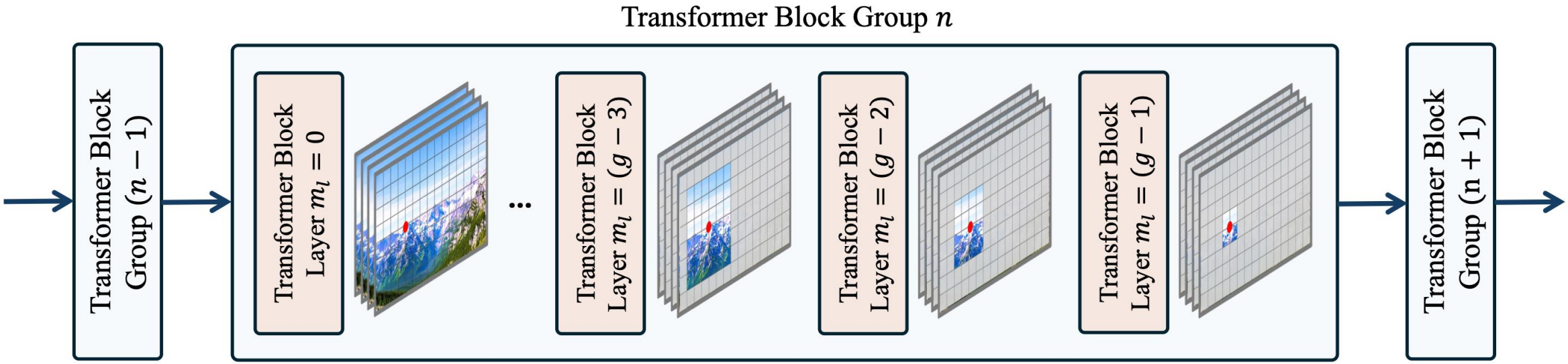
	Video data	Training data	Frames	Training cost
UltraWan-4K	Yes	42,000 4K videos	29	128 GPUs
UltraGen	Yes	42,000 4K videos	29	128 GPUs
T3-Video	Yes	42,000 4K videos	81	64 GPUs
Ours	**No**	**3,000 2K images**	**81**	**4 GPUs, < 8 h**

Thousands of GPU Hours!

How can we do this with images only?

ApertureAttn

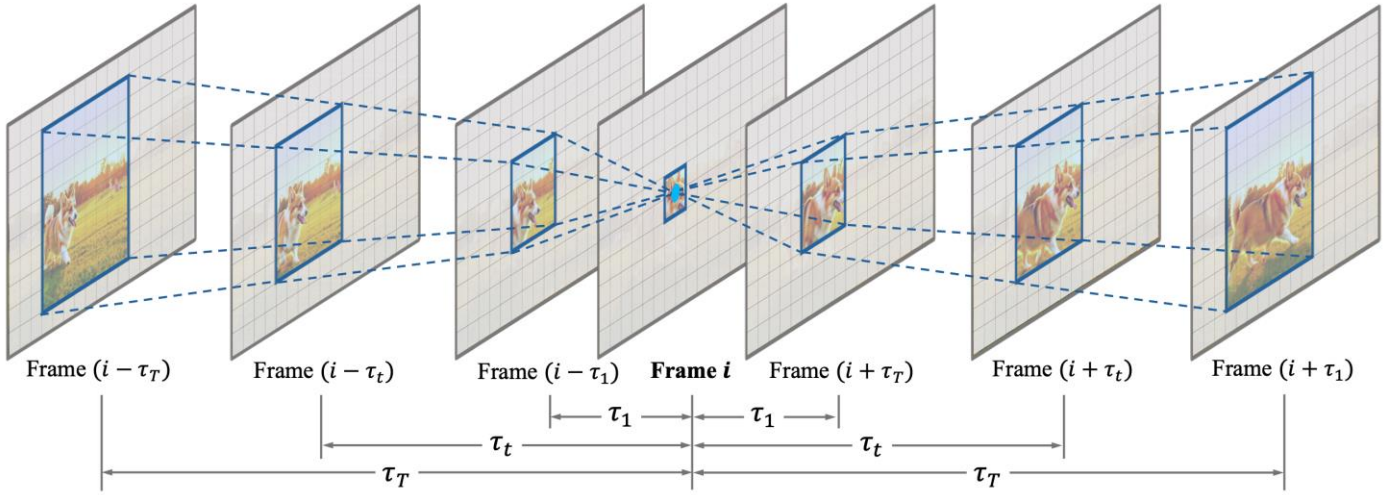
Train on images as single-frame videos



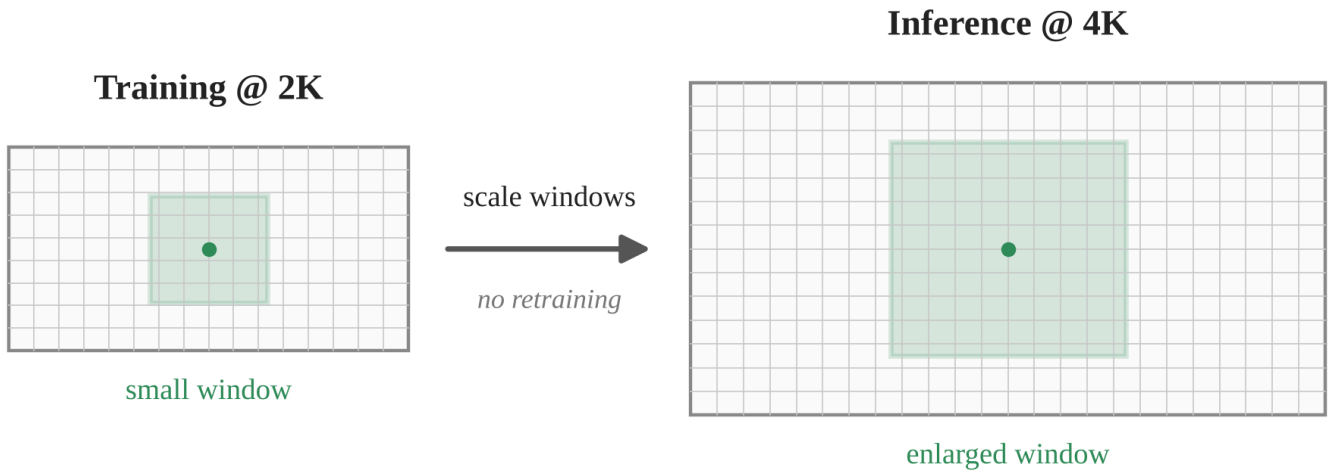
Overview: ApertureAttn

ApertureAttn

Train on images as single-frame videos



PGTAttn



Window-Scaled Extrapolation

Main Results

Visualization



Main Results

Quantitative

Table 1: The evaluation results of baselines and our proposed method on VBench. Here, $^{+}$ means applying NTK RoPE scaling, proportional attention, and a larger timestep shift to the original Wan2.2.

Method/Setting	Aesthetic Quality	Background Consistency	Imaging Quality	Motion Smoothness	Overall Consistency	Subject Consistency
Original Wan2.2-TI2V-5B	42.02%	98.72%	39.79%	99.42%	17.07%	97.33%
Wan2.2-TI2V-5B^{+}	57.45%	98.94%	61.78%	98.81%	23.58%	96.80%
SDEdit	60.16%	98.53%	57.87%	98.17%	26.77%	96.26%
RealViformer	61.55%	98.48%	62.69%	97.80%	26.89%	96.02%
FlashVSR	62.08%	98.37%	65.60%	97.73%	26.70%	95.72%
Direct Tune	55.77%	98.70%	56.25%	99.09%	22.70%	98.09%
UltraWan	57.32%	98.73%	56.63%	99.23%	24.46%	98.09%
T3-Video	59.63%	98.39%	65.62%	98.54%	25.66%	96.60%
ApertureAttn (Ours)	**66.62%**	**99.24%**	**66.94%**	**98.84%**	**26.82%**	**97.72%**

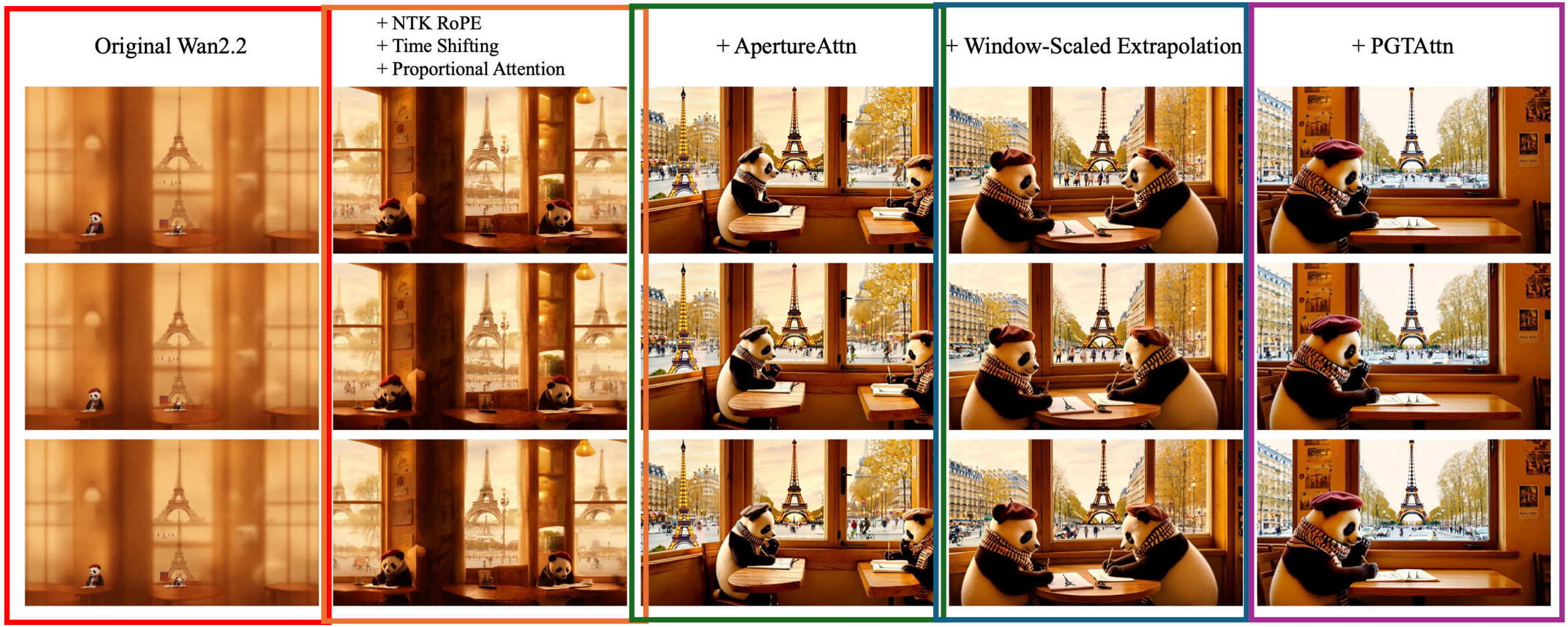
NUS National University of Singapore

xML Lab

Ablation

Each Component

NEURAL INFORMATION PROCESSING SYSTEMS



NUS National University of Singapore xML Lab

NEURAL INFORMATION PROCESSING SYSTEMS

🚀 Native 4K, 81 Frames, Single Stage
🚀 Image-Only Training - No Video Data
🚀 3,000 Images, 4 GPUs, Under 8 Hours

🌟 Rich Detail, Coherent Structure, No Duplicated Subjects
🌟 Lightweight LoRA, Plug-in Attention Design
🌟 Outperforms 4K-Video-Trained Methods





Thanks for Your Listening!

If you have any questions, please contact to us via email ruonan@u.nus.edu