SageAttention3: Microscaling FP4 Attention for Inference and An Exploration of 8-bit Training

Jintao Zhang*, Jia Wei*, Haoxu Wang, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Kai Jiang, Jun Zhu#, Jianfei Chen# | *Tsinghua University*

Spotlight Paper

GitHub

Goal

- **SageAttn3-FP4**:
  Propose an attention method that utilizes the FP4 Tensor Core for inference acceleration in a plug-and-play way.
- **SageBwd**:
  Design a trainable 8-bit attention for training acceleration.

Method

SageAttention3-FP4 :

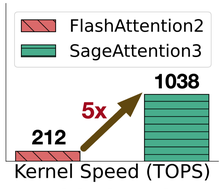
- FP4 microscaling quantization.
- Two-level quantization for P.

SageBwd:

- Maintain one MatMul among the five in backpropagation in FP16.

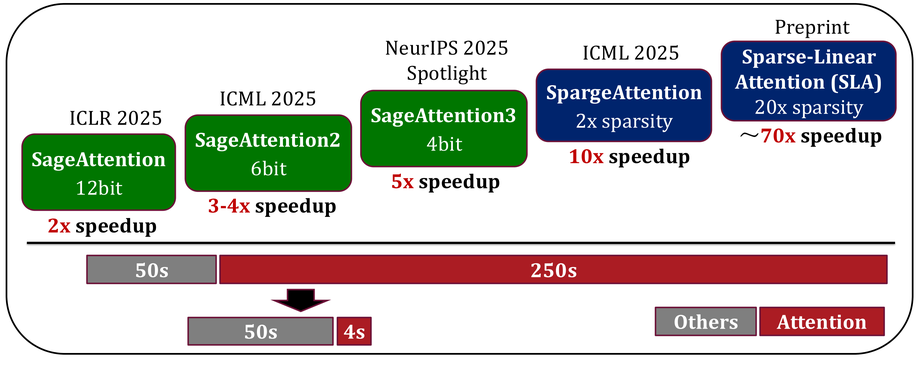
Contribution

- First FP4 attention for inference, 1000 TOPS on RTX5090.



- First trainable low-bit attention, enabling accelerated training with lossless fine-tuning.

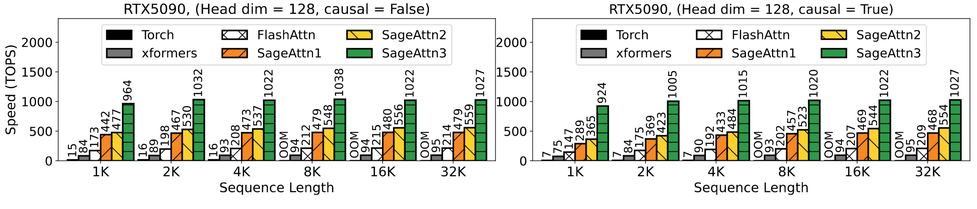
Our Progress on Attention Acceleration



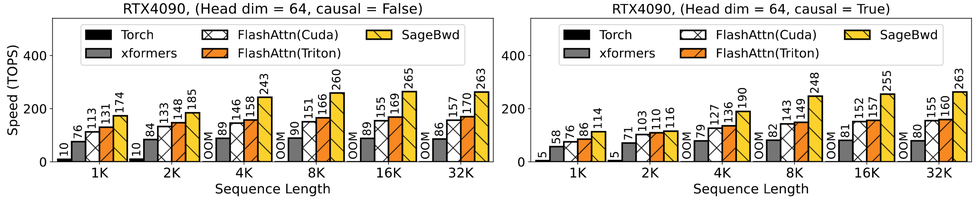
The **70x speedup** is already used in a commercial video generation product.

GPU Kernel Speedup

- **SageAttn3-FP4 kernel speed: 4 - 5x** speedup than FlashAttention.



- **SageBwd kernel speed: 1.65x** speedup than FlashAttention.

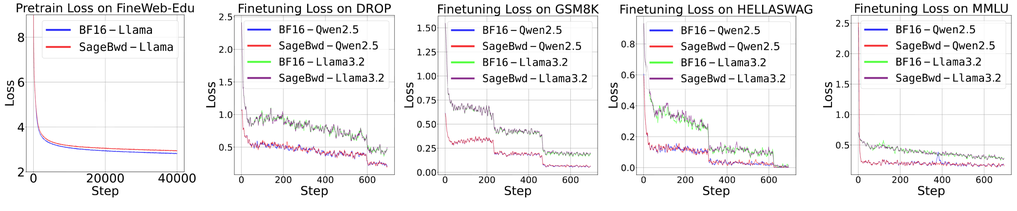


End-to-End Evaluation

- **SageAttention3-FP4**: No loss in end-to-end quality on diffusion models.

HunyuanVideo *Stable-Diffusion3.5*

- **SageBwd**: No gap with FP/BF16 attention in fine-tuning tasks.



- End-to-end speedup:

Model	Original	Sage1	Sage2	Sage3
CogvideoX (2B)	64 s	55 s	46 s	**27 s**
HunyuanVideo	489 s	257 s	240 s	**164 s**

Model	Original	SageBwd
Llama (8K)	2.1 s	**1.9 s**
Llama (16K)	6.0 s	**5.2 s**

- SageBwd : INT8 vs. FP8:

(b) Qwen2.5-3B

Method	GSM8K ↑	MMLU ↑
INT8 Fine-tuning	**0.5945**	**0.6032**
FP8 Fine-tuning	0.5868	0.5907