AttnForge
FlashAttention-2
FlashAttention-3
FlexAttention
FlashInfer
PyTorch Inductor
FlashSigmoid
Flash-Linear-Attention

Normalized latency Vs. AttnForge
(lower is better)

(a) Softmax Attention (DeepSeek-V2-Lite)

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 | BS1S1 KV2048 | BS1S1 KV4096 | BS1S1 KV8192 | BS8S1 KV2048 | BS8S1 KV4096 | BS8S1 KV8192 — FWD

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — BWD

(b) Softmax Attention (LLAMA-3.1-8B)

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — FWD

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — BWD

(c) ReLU Attention (ViT-s/16-style)

BS32 S512 | BS32 S1024 | BS32 S2048 | BS64 S512 | BS64 S1024 | BS64 S2048 — FWD

BS32 S512 | BS32 S1024 | BS32 S2048 | BS64 S512 | BS64 S1024 | BS64 S2048 — BWD

(d) Softmax Attention (Diff-Transformer-3B)

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — FWD

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — BWD

(e) Retention Parallel (RetNet-6.7B)

BS1 S2048 | BS1 S4096 | BS8 S2048 | BS8 S4096 — FWD

BS1 S2048 | BS1 S4096 | BS8 S2048 | BS8 S4096 — BWD

(f) Sigmoid Attention (LLAMA-3.1-8B)

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — FWD (PyTorch Inductor Failed)

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — BWD (PyTorch Inductor Failed)

(g) Mamba2 SSM (Mamba2-2.7B)

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — FWD

BS1 S2048 | BS1 S4096 | BS1 S8192 | BS8 S2048 | BS8 S4096 | BS8 S8192 — BWD

(h) Gated Retention (RFA-Big)

BS64 S1024 | BS64 S2048 | BS64 S4096 — FWD

BS64 S1024 | BS64 S2048 | BS64 S4096 — BWD (FLA Failed)

(i) Gated Retention (YOCO-13B)

BS8 S1024 | BS8 S2048 | BS8 S4096 — FWD

BS8 S1024 | BS8 S2048 | BS8 S4096 — BWD

(j) RetNet Recurrent (RetNet-6.7B)

BS1 S2048 | BS1 S4096 | BS8 S2048 | BS8 S4096 — FWD

BS1 S2048 | BS1 S4096 | BS8 S2048 | BS8 S4096 — BWD