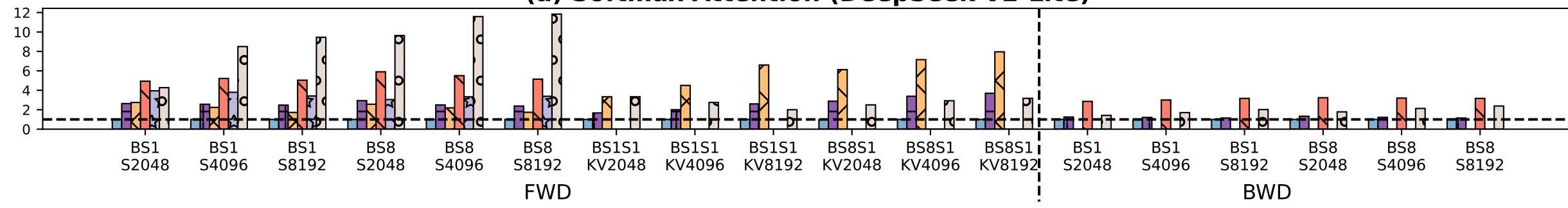
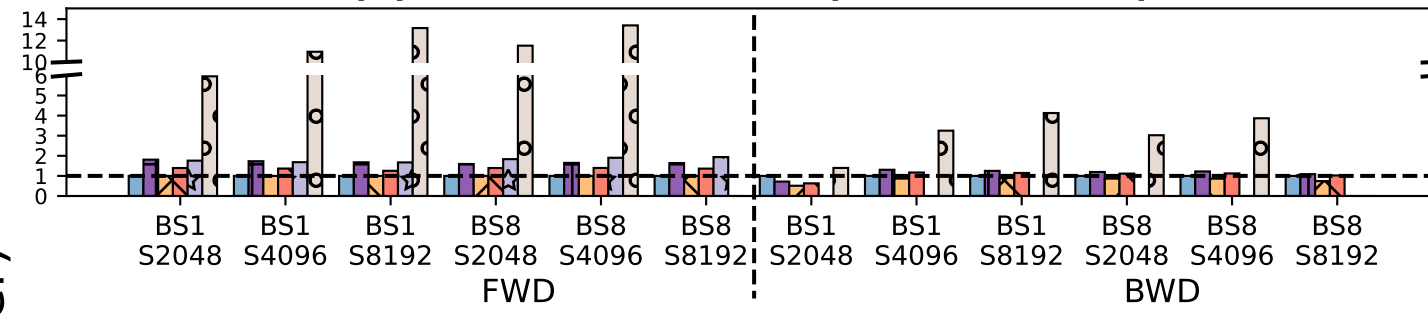


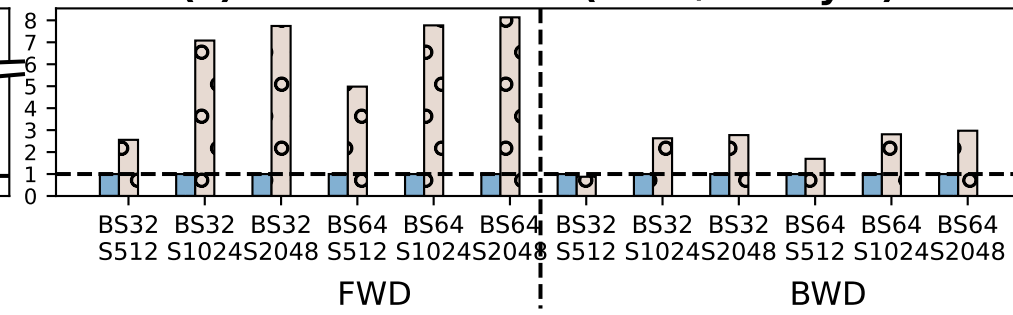
(a) Softmax Attention (DeepSeek-V2-Lite)



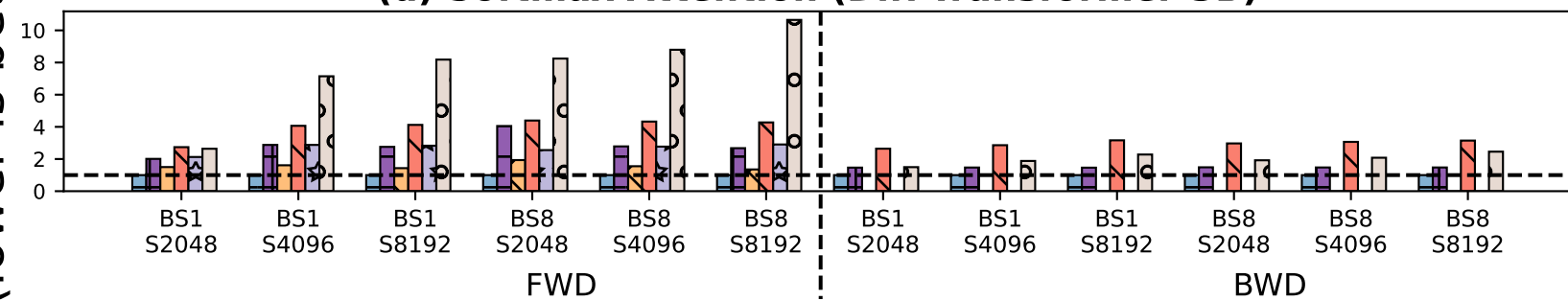
(b) Softmax Attention (LLAMA-3.1-8B)



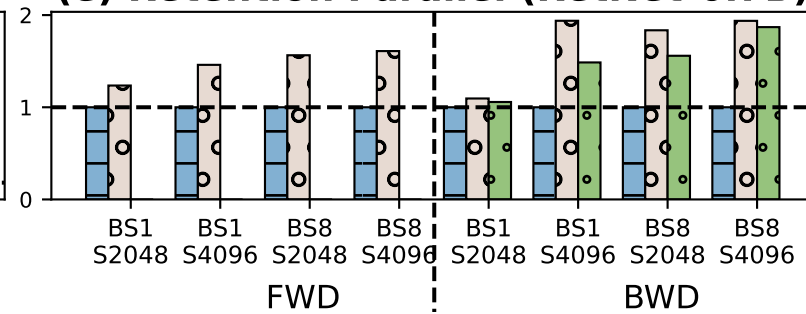
(c) ReLU Attention (ViT-s/16-style)



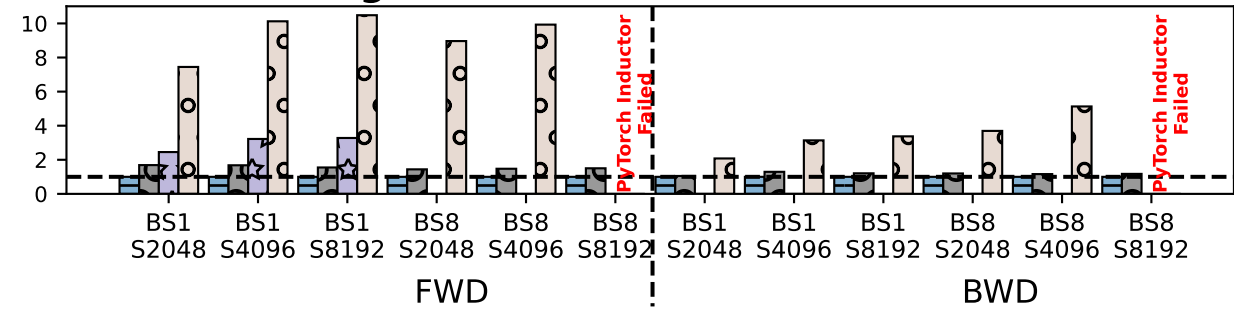
(d) Softmax Attention (Diff-Transformer-3B)



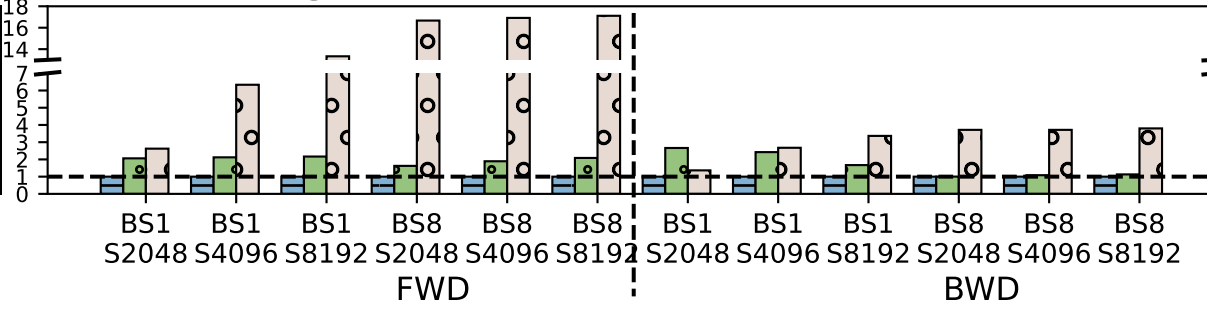
(e) Retention Parallel (RetNet-6.7B)



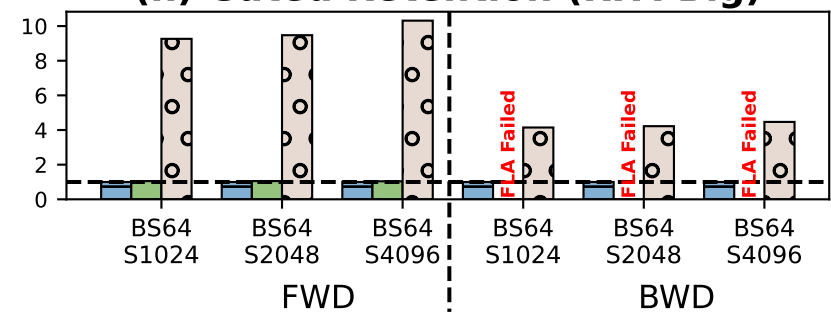
(f) Sigmoid Attention (LLAMA-3.1-8B)



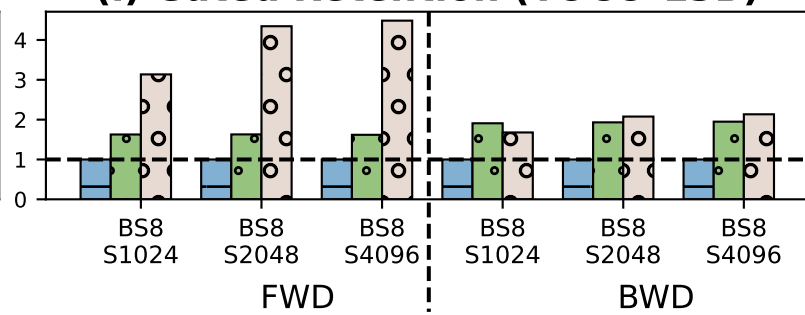
(g) Mamba2 SSM (Mamba2-2.7B)



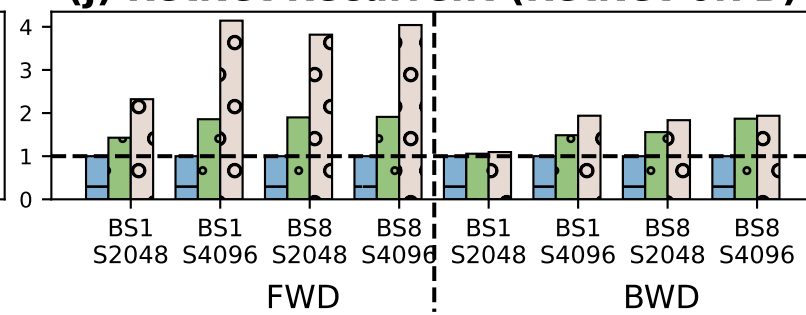
(h) Gated Retention (RFA-Big)



(i) Gated Retention (YOCO-13B)



(j) RetNet Recurrent (RetNet-6.7B)



Normalized latency Vs. AttnForge
(lower is better)