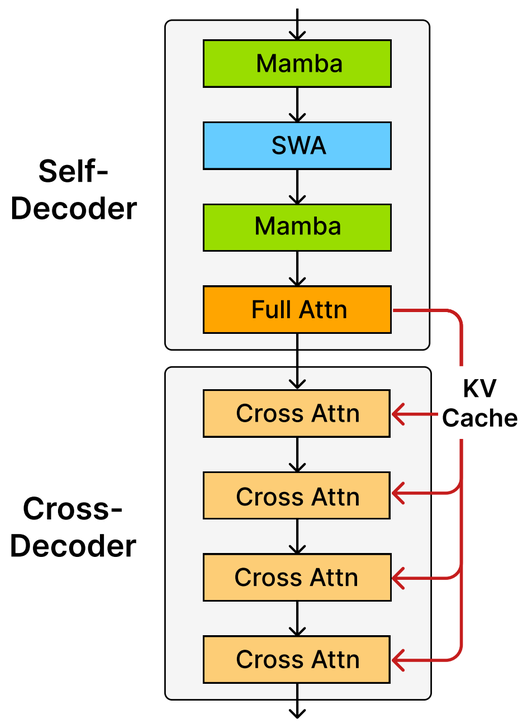
Decoder-Hybrid-Decoder Architecture for Efficient Reasoning with Long Generation

You Only Cache Once (YOCO)

- **Linear** prefill complexity with decoder-decoder architecture (Samba+YOCO).



- For cross-decoder, Memory I/O cost = Full attention
  Slow for Long CoT generation!

How to Share Memory between SSMs?

- Just share the output state!
- **Gated Memory Unit (GMU):**

$$\mathbf{y}_l = (\mathbf{m}_{l'} \odot \sigma(W_1 \mathbf{x}_l)) W_2$$

- It reweights the previous layer's token mixing with current layer input:

$$G^{(l')} = \sigma(W_1^{(l')} \mathbf{x}^{(l')})$$

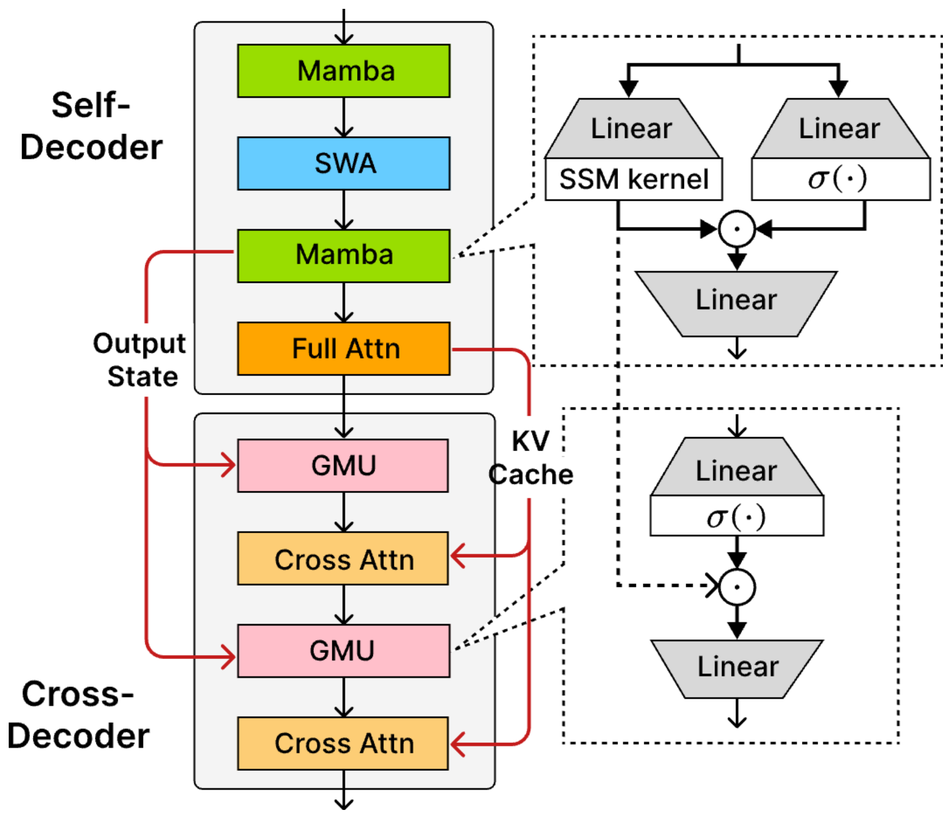
$$h_{ik} = G_{ik} \sum_j A_{ij}^{(l')} \, v_{jk}^{(l')} = \sum_j G_{ik} \, A_{ij}^{(l')} \, v_{jk}^{(l')}$$

$$= \sum_j \underbrace{A_{ij}^{(l')} \, G_{ik}}_{\tilde{A}_{ijk}} \, v_{jk}^{(l')},$$

- We can also add RMSNorm after gating => **nGMU.**

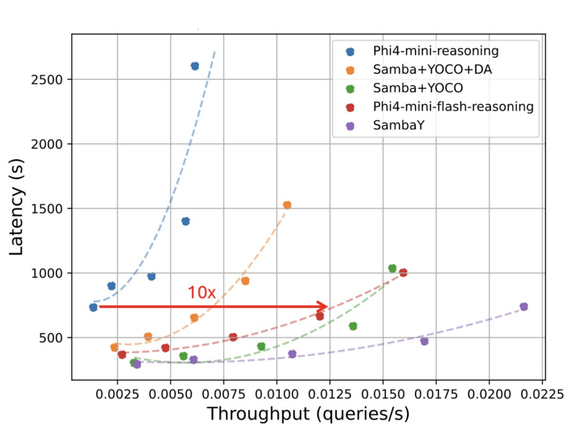
SambaY

- **Linear** prefill complexity with half cross-attention layers replaced with GMUs. => **Decoder-Hybrid-Decoder Architecture**

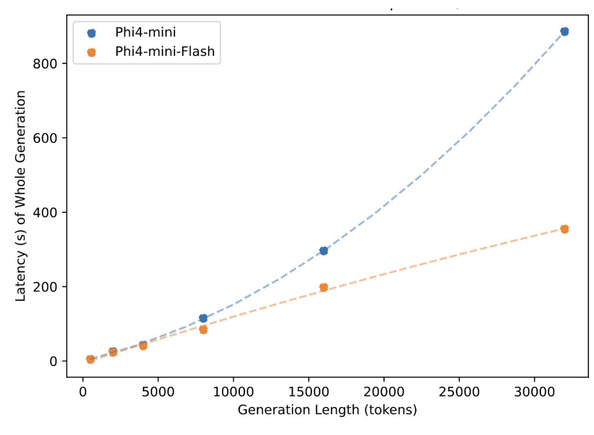


Long Generation Efficiency

- 10x Throughput **even with slow** Differential Attention (DA) and **sub-optimal** vLLM implementation.



(b) Prompt: 2000, Generation: 32000



Better Reasoning

- Pass@1 avg. over 64 runs for AIME 24/25, over 8 for MATH 500 and GPQA-Diamond.

Model	AIME24	AIME25	Math500	GPQA Diamond
DeepSeek-R1-Distill-Qwen-1.5B	29.58	20.78	84.50	37.69
DeepSeek-R1-Distill-Qwen-7B	53.70	35.94	93.03	47.85
DeepSeek-R1-Distill-Llama-8B	43.96	27.34	87.48	45.83
Bespoke-Stratos-7B	21.51	18.28	80.73	38.51
OpenThinker-7B	29.69	24.32	87.25	41.60
Phi4-mini-Reasoning (3.8B)	48.13	31.77	91.20	44.51
Phi4-mini-Flash-Reasoning (3.8B)	**52.29**	**33.59**	**92.45**	**45.08**

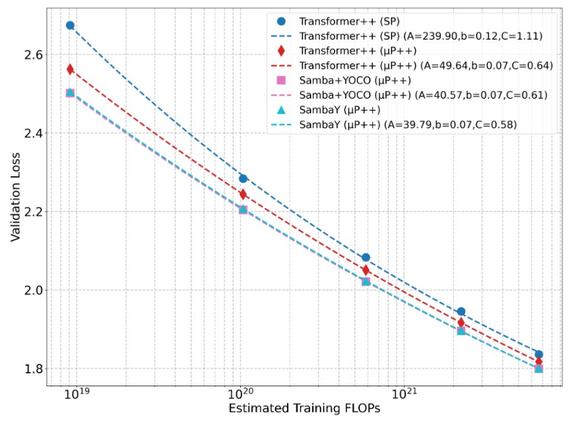
- Better long context performance on RULER with 1B/40B ablation study:

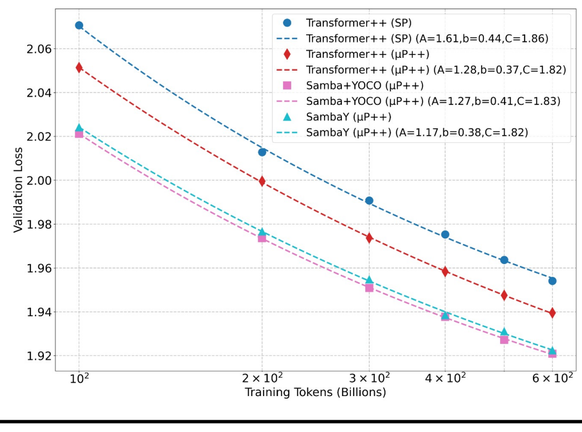
Model	SWA	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	Avg.
Transformer++	-	36.4	3.8	0.0	27.9	24.1	94.8	66.0	31.0	35.5
TransformerLS	256	42.8	6.0	0.0	**29.8**	**27.5**	91.8	49.6	23.4	33.9
Samba+YOCO	1024	49.0	**28.0**	**2.6**	12.8	18.3	**100.0**	63.2	23.6	37.2
SambaY	256	54.6	27.8	0.4	12.7	19.4	83.2	81.2	63.8	42.9
SambaY+DA	512	**64.6**	27.6	0.2	12.8	19.9	99.8	**86.4**	**69.6**	**47.6**

- Why? Hybrid models converge faster and long context data is limited.

Better Scaling

- μP++ = μP + Depth-μP + zero weight decay on vector-like parameters.





Fin.