SeerAttention-R

Sparse Attention Adaptation for Long Reasoning

Yizhao Gao[1,2,*] Shuming Guo[3,*] **Shijie Cao**[2,†] Yuqing Xia[2] Yu Cheng[4] Lei Wang[4] Lingxiao Ma[2] Yutao Sun[5] Tianzhu Ye[2] Li Dong[2] Hayden Kwok-Hay So[1] Yu Hua[3] Ting Cao[5] Fan Yang[2] Mao Yang[2]

[1]HKU [2]Microsoft Research [3]HUST [4]Peking University [5]Tsinghua University (*Equal contribution, †Corresponding author)

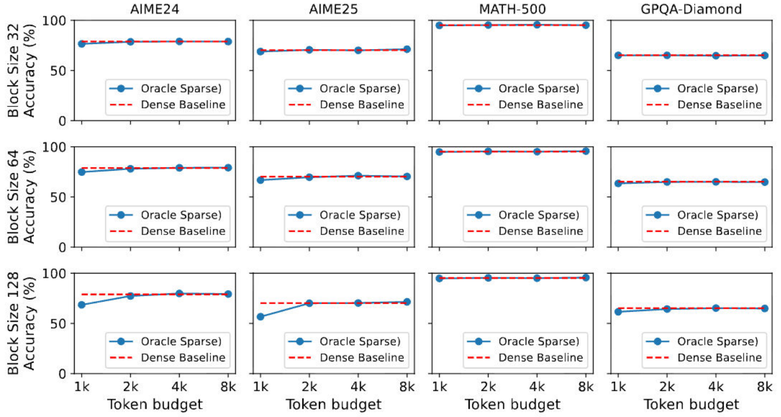
TL;DR

SeerAttention-R is a **lightweight, plug-in sparse-attention framework** for the *long decoding* phase of reasoning LLMs. With only **0.4B** training tokens for a tiny gate (base model frozen), it keeps AIME accuracy **near-lossless at a 4K token budget** and delivers up to **9× kernel speedup** over FlashAttention-3 on H100. The learned gate predicts *which* KV blocks matter at each step, so the base model stays frozen and the approach drops into existing reasoning LLMs without retraining.

Problem: Reasoning = Long Decoding

- Reasoning models (o1, DeepSeek-R1, Qwen3) scale performance by generating **longer** chains of thought.
- Auto-regressive decoding is **memory-bound**: per-token cost grows linearly, total cost grows **quadratically** with length.
- Sparse attention is promising, but existing methods target *prefilling* or rely on training-free heuristics that break at large block sizes.

Key Insight: Decoding Attention Is Sparse

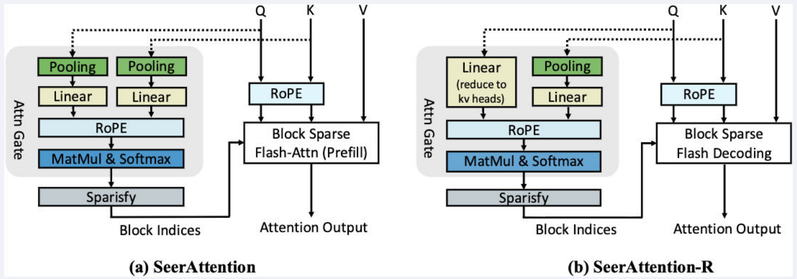


Oracle sparsity on Qwen3-14B. At each decoding step we use the *true* column-wise max-pooled attention map to pick the top-k KV blocks — an upper bound on what any learned gate could achieve. Accuracy stays *lossless* from a ∼2K-token budget across AIME24/25, MATH-500, and GPQA-Diamond, even at coarse block sizes of 64 and 128. **Reasoning attention is inherently block-sparse** — the remaining challenge is *predicting* which blocks to keep without peeking at the true map.

Contributions

- **AttnGate for sparse decoding:** drop query pooling; add GQA-aware query aggregation & 3-way (Max/Min/Avg) key pooling.
- **Lightweight self-distillation:** 1D column-wise max-pool as ground truth; *only* the gate is trained (∼0.4B tokens, base model frozen).
- **Block-sparse flash-decoding kernel** (TileLang + Triton) — up to 9× over FA3, 1.7× over Triton on H100.
- **Consistently beats Quest** across Qwen3-4B/8B/14B and R1-Distill-14B on AIME24/25, MATH-500, and GPQA-Diamond.

Method: AttnGate for Sparse Decoding



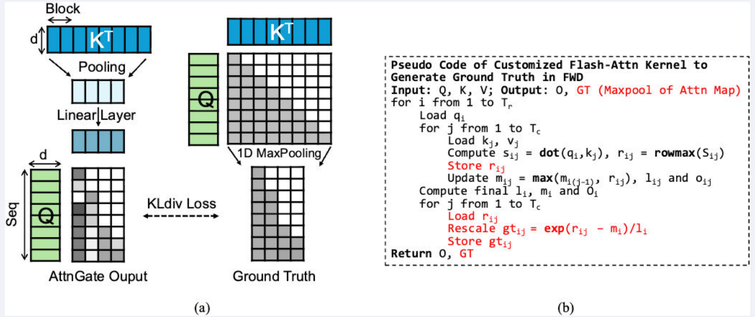
Sparse prefill (SeerAttention) vs. sparse decode (SeerAttention-R). The decode gate keeps the full query sequence (no sequence-dim pooling), and a linear layer collapses each GQA group of query heads to a single gate head for *shared* sparsity within the group.

$$\mathbf{Q}_{\text{gate}} = \mathrm{RoPE}\big(\mathbf{W}^{q}_{\text{gate}}\,\mathrm{reshape}(\mathbf{Q}, [\ldots, g\cdot d])\big)$$
$$\mathbf{K}_{\text{gate}} = \mathrm{RoPE}\big(\mathbf{W}^{k}_{\text{gate}}\,[\mathrm{P}_{\max}(\mathbf{K}) \,\|\, \mathrm{P}_{\min}(\mathbf{K}) \,\|\, \mathrm{P}_{\text{avg}}(\mathbf{K})]\big)$$
$$\mathbf{S} = \mathrm{softmax}\big(\mathbf{Q}_{\text{gate}}\mathbf{K}_{\text{gate}}{}^{\top}/\sqrt{d_{gate}}\big)$$

Design points.

- **GQA-shared sparsity:** one block-level decision per KV head ⇒ hardware-friendly.
- **Composite K pooling** (Max/Min/Avg) preserves outliers *and* distribution.
- **Pre-RoPE gate inputs**, RoPE re-applied inside the gate with per-block position.
- **Base model frozen** — AttnGate is the only trainable component.

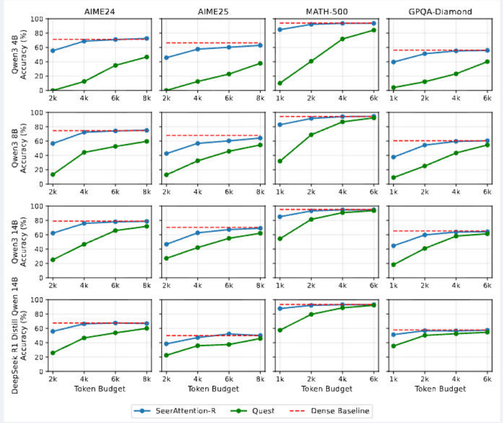
Training: Self-Distillation




```
Pseudo Code of Customized Flash-Attn Kernel to
Generate Ground Truth in FWD
Input: Q, K, V; Output: O, GT (Maxpool of Attn Map)
for i from 1 to T_r
    Load q_i
    for j from 1 to T_c
        Load k_j, v_j
        Compute s_ij = dot(q_i,k_j), r_ij = rowmax(S_ij)
        Store r_ij
        Update m_ij = max(m_i(j-1), r_ij), l_ij and o_ij
    Compute final l_i, m_i and O_i
    for j from 1 to T_c
        Load r_ij
        Rescale gt_ij = exp(r_ij - m_i)/l_i
        Store gt_ij
Return O, GT
```

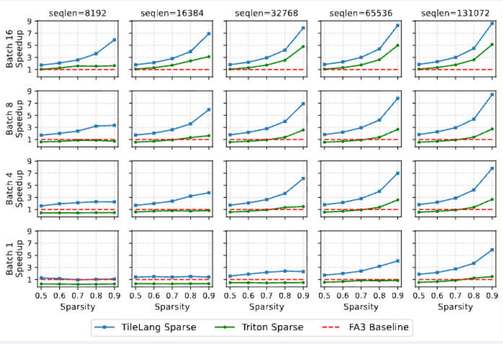
- **Ground truth:** *column-wise 1D max-pool* of the true attention map, then max-pool within each GQA group, then row-normalized.
- **Loss:** KL divergence between gate softmax and ground truth.
- **FlashAttention-2 fused kernel** emits ground truth alongside the attention output — avoids materializing the full $\mathbf{QK}^{\top}$.
- **Setup:** OpenR1-MATH-220K, seq-len 32K, 800 steps, batch 16, LR 1e-3, MI300x, DeepSpeed ZeRO-2 — only gate weights updated, ∼**0.4B tokens**.

Accuracy vs. Token Budget



- Block size 64, sparse attention in *all* layers (same config for Quest).
- **SeerAttention-R** matches full attention on **AIME24/25 at 4K**; Quest fails even at 8K.
- On MATH-500 & GPQA-Diamond: **SeerAttention-R** matches dense at **2K**; Quest needs ∼8K.
- Larger models are *more* robust to sparsity — gap closes fastest on 14B.

Kernel Speedup on H100



- TileLang beats both Triton and FlashAttention-3. 1.7× over Triton; gains scale with seq-len & batch size.
- Up to 9× **over FA3** at 90% sparsity (bs=16, seq≥32K) — near I/O bound.