Breaking the Exactness Barrier: Interleaved DeepSeek Sparse Attention for Efficient Long-Context Reasoning

Interleaved partitioning replaces synchronized global refinement with independent local selection.

Yifan Guo[1]
Shanghai Jiao Tong University
yfguo26@cse.cuhk.edu.hk

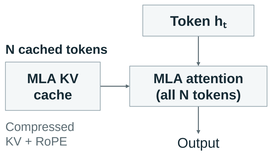
Wei Cui[2]
Microsoft Research Asia
weicu@microsoft.com

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2026

1 Motivation and local selection

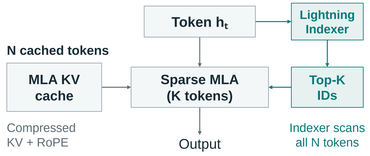
Sparse attention architecture

DeepSeek-V3
MLA over the full causal context



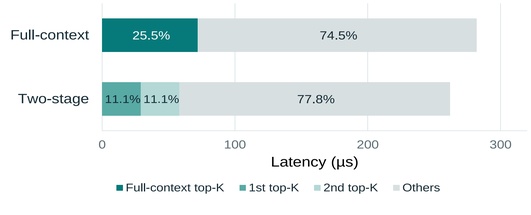
Smaller KV storage per token

DeepSeek-V3.2
DSA instantiated on MLA



Fewer tokens in attention

Exact global Top-K latency



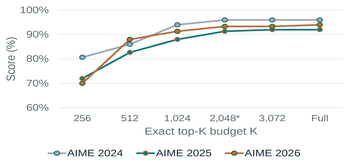
22.2%

of two-stage latency comes from top-K

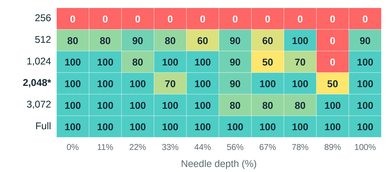
11.1% local selection
11.1% global refinement

Quality versus inference budget

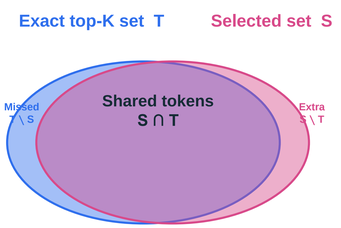
Mathematical reasoning
AIME 2024 / 2025 / 2026



Needle in a Haystack
128K context, accuracy (%)



Local-selection objective



Goal: cover T while keeping the selected set S small

Minimize local budget

$$\min_{m \in \mathbb{N}} \; m$$

$$\text{s.t.} \quad \Pr[T \subseteq S(m)] \geq \gamma$$

$$m \geq \frac{K}{G}$$

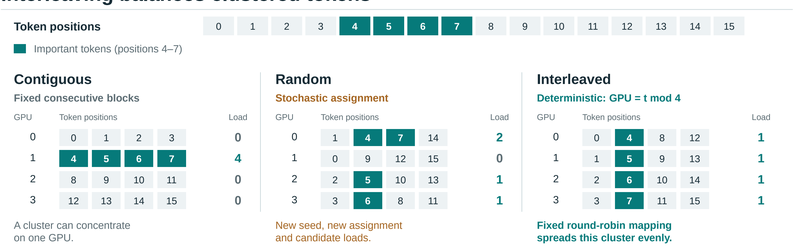
$$T = \operatorname*{Top}_{G-1}(C, K)$$

$$S(m) = \bigcup_{g=0} \operatorname{Top}(C_g, m)$$

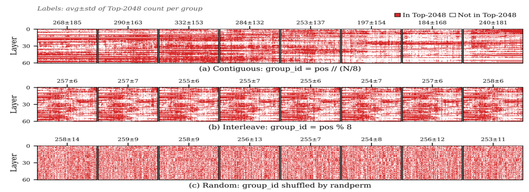
γ: target probability of full top-K recall

2 Interleaving and budget

Interleaving balances clustered tokens



Measured candidate-load imbalance



128K context	Std.	Max load
Contiguous	318.3	1,766
Random	14.9	320
Interleaved	**11.7**	**311**

Full recall as a tail event

Load model
$$X_i = |T \cap C_i| \sim \text{Binomial}(K, 1/G)$$

Full recall
$$T \subseteq S(m) \iff \max_i X_i \leq m$$

Union bound
$$\Pr[T \subseteq S(m)] = 1 - \Pr\Big(\bigcup_i \{X_i > m\}\Big) \geq 1 - \sum_i \Pr(X_i > m) = 1 - G\Pr(X_1 > m)$$

Target probability
$\Pr[T \subseteq S(m)] \geq \gamma$ is ensured by
$\Pr(X_1 > m) \leq (1-\gamma)/G \iff F_X(m) \geq p := 1 - (1-\gamma)/G$

Exact and approximate local budgets

Exact threshold
$$m_{ex} = \min\{m \in \mathbb{N} : F_X(m) \geq p\}$$
$$F_X(m) = \sum_{j=0}^{m} \binom{K}{j} (1/G)^j (1-1/G)^{K-j}$$

Normal model
$$\mu = K/G, \quad \sigma = \sqrt{[K(G-1)]/G}$$
$$F_X(m) \approx \Phi((m + ½ - \mu)/\sigma) \geq p \quad \Rightarrow \; m \gtrsim \mu + z_p\sigma - ½$$

One-token margin
$m_{ap} = \lceil \mu + z_p\sigma - ½ + 1 \rceil = \lceil \mu + z_p\sigma + ½ \rceil$

32-token rounding
$m = 32\lceil m_{ap}/32 \rceil$

K = 2,048, G = 8, γ = 99%
$m_{ex} = m_{ap} = 302, \quad m = 320$ tokens/GPU

Closed-form 8-GPU budget

$$m_{ap} = \left\lceil \frac{K}{G} + z_p \frac{\sqrt{[K(G-1)]}}{G} + ½ \right\rceil$$

$z_p = \Phi^{-1}(p)$
$p = 1 - (1-\gamma)/G$

$$m_{ap} = \lceil 256 + 3.0233 \times 14.9666 + 0.5 \rceil = 302$$

Before alignment: m_{ap} = 302 tokens per GPU

Local budget m (tokens / GPU)
320
$m = 32\lceil 302/32 \rceil$

Equivalent counts for m = 320
ε = m − K/G = 64 extra tokens / GPU
Gm = 2,560 tokens in total (+25% over K)

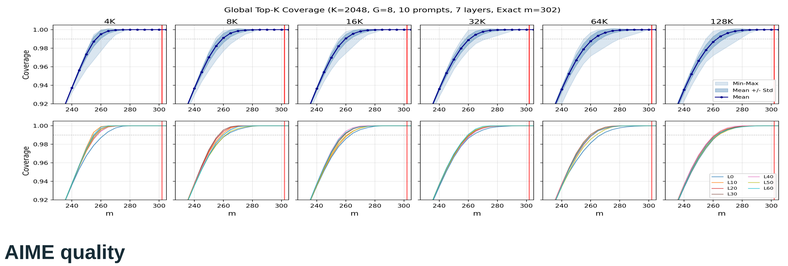
3 Communication and evidence

Communication and attention budgets

Per decoding query	Two-stage exact selection	IDSA (m = 320)
Selected per GPU	K = 2,048	m = 320
Candidates gathered (all GPUs)	G·K = 16,384	Gm = 2,560
Second global top-K	Required	Removed
Tokens used by attention	K = 2,048	Gm = 2,560

6.4× fewer candidate entries to gather
25% more attention candidates

Coverage across context lengths



AIME quality

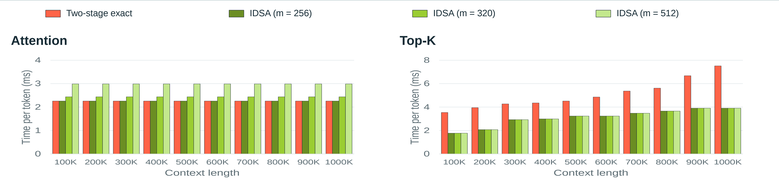
DeepSeek-V3.2 (%)

Configuration	2024	2025	2026
Exact K = 2,048	96.0%	91.3%	93.3%
IDSA m = 256	96.0%	90.0%	92.0%
IDSA m = 320	**96.0%**	**91.3%**	**93.3%**
IDSA m = 512	96.0%	92.0%	94.0%

GLM-5 (%)

Configuration	2024	2025	2026
Exact K = 2,048	94.0%	93.3%	94.7%
IDSA m = 256	93.3%	92.7%	94.0%
IDSA m = 320	**94.0%**	**93.3%**	**95.3%**
IDSA m = 512	96.7%	96.0%	95.3%

Attention and selection latency



Attention cost stays nearly flat.
IDSA removes global refinement.

End-to-end decoding latency



< 20 ms / token
at the 1M-token systems test