PLASH: Provably Linear-Time Attention with Selective Higher-Order Feature Sketching

Yuwen Huang
The Hong Kong University of Science and Technology (Guangzhou)
yuwenhuang@hkust-gz.edu.cn

Xiang Pan*
Lingnan University
xxiangpan@ln.edu.hk *Corresponding author

Hiring PhD students at HKUST(GZ) · ICML 2026 · Seoul

Scan for the paper (PDF)

Softmax attention compares every query with every key.

Its time and memory grow quadratically with the context.
PLASH runs in linear time and bounds its error on every input.

- Linear time and memory, $O\big((N_q + N_k) \cdot M\big)$
- Exact softmax readout over M prototypes
- Per-input error bound, no labels needed

2.4× faster at $N = 32$K
than FlashAttention-2 [1]
4,358 → 1,809 ms per layer

14.3× faster at $N = 8{,}192$
attention layer, $d = 512$
615.2 → 43.0 ms

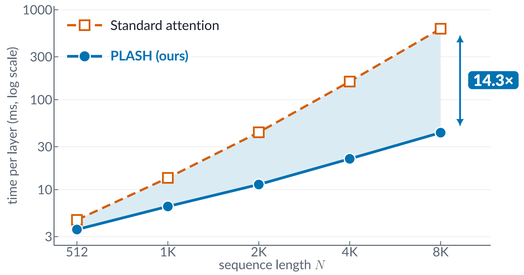
100% of error bounds hold
35,058 real test inputs, zero violations

4/4 forecasting datasets
lower MSE than Standard attention at $L = 336$

1 The problem: quadratic attention

Attention-layer time: quadratic vs. linear

Forward time per layer; A800-80GB, $d = 512$, batch 16



The idea: route N keys through $M \ll N$ prototypes



all pairs: $O(N^2)$ — M prototypes: $O(N \cdot M)$

Keys and values are compressed to M learned prototypes and enriched once; every query then reads out the prototypes with exact softmax.

Where PLASH helps

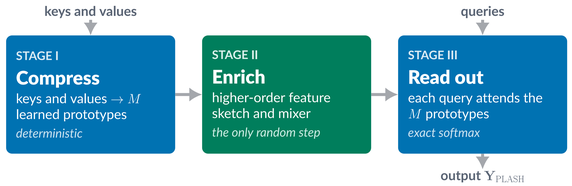
Long-context LLMs
up to 32K tokens on Qwen3-4B [2]

Time-series forecasting
ETT [7], ECL, and Weather

Long-document QA and RAG
long inputs with exact readout

Trustworthy inference
flags inputs whose bound is loose

2 PLASH: compress, enrich, read out



Stage II: degree-k features without the d^k tensor

Degree-k features describe how k tokens interact. TensorSketch [6] builds them from CountSketch [5] and the fast Fourier transform (FFT) in $O(D_k \log D_k)$ time.



Sizing rule: the sketch error stays under control

$$D_k \geq \frac{2 \cdot C_k \cdot M}{\eta^2 \cdot \delta_k} \quad \text{for all } k \in \mathcal{K}, \qquad \sum_{k \in \mathcal{K}} \delta_k \leq \delta$$

With these sketch sizes, one bound covers the sketched features of all M prototypes and all degrees, with probability at least $1 - \delta$ over the sketch randomness.

Linear time and memory

Fixed head dimensions and sketch sizes; self- and cross-attention

$O\big((N_q + N_k) \cdot M\big)$
PLASH: time and peak memory

$O(N_q \cdot N_k)$
softmax attention: time

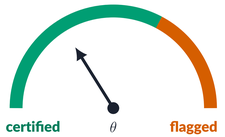
Randomness in one stage only

PLASH: one random stage, exact readout, per-input certificate

Random-feature kernels
randomness everywhere

Deterministic compression
no per-input guarantee

3 A certificate on every input



Each forward pass computes a bound θ on its own error. An input with $\theta \leq \epsilon_{\mathrm{out}}$ is certified, and any other input is flagged. No labels are needed.

Per-input certificate, computed in the forward pass

$$\Pr\Big(\big\|\mathbf{Y}_{\mathrm{soft}} - \mathbf{Y}_{\mathrm{PLASH}}\big\|_F \leq \theta\Big) \geq 1 - \delta, \qquad \theta \triangleq \epsilon_{\mathrm{I}} + \epsilon_{\mathrm{det}} + \epsilon_{\mathrm{II}}$$

$\mathbf{Y}_{\mathrm{soft}}$: exact softmax attention; θ: the sum of three error terms from the forward pass; δ: the probability budget of the sizing rule.

Guarantees

- **Per input, without labels.** The bound θ uses only quantities from the forward pass.
- **With high probability.** It holds with probability at least $1 - \delta$ over the sketch randomness.
- **For causal models.** A block-causal variant keeps the certificate; across 40 leakage tests, the output change at protected positions stays below 3.24×10^{-5}, the bf16 noise level.
- **For self- and cross-attention.** The construction applies to both.

100% of 35,058 real test inputs stay within their bound
zero violations on four forecasting datasets

The sketch makes the certificate possible

ETTm1 forecasting MSE; Stage II replaced by an equal-capacity MLP

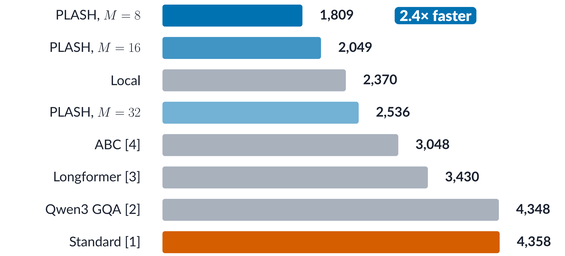
0.760
PLASH with TensorSketch
certificate kept

0.893
TensorSketch → MLP
+17.5% MSE, certificate lost

4 Results

Fastest of eight methods at 32K tokens

Forward time per layer (ms), $N = 32$K; Qwen3-4B [2], A800-80GB, batch 1, bf16; Standard uses FlashAttention-2 [1]



Language modeling at long context

EMA loss, lower is better

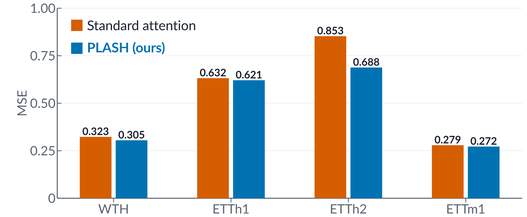
23.5
lowest of eight at 16K
PLASH, $M = 16$

40.1
ties Standard at 32K
PLASH, $M = 32$, 1.7× faster

+0.13
gap to Standard, 1.5B
+0.31 at 124M

Forecasting ($L = 336$): lower MSE on all four datasets

Mean squared error (MSE), lower is better



Largest gain on ETTh2: 0.853 → 0.688 (−19%). Next steps: higher-degree sketches, fused GPU kernels, and adaptive per-input budgets.

References. [1] T. Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. ICLR 2024. [2] A. Yang et al. Qwen3 technical report. arXiv, 2025. [3] I. Beltagy, M. E. Peters, A. Cohan. Longformer: The long-document transformer. arXiv, 2020. [4] H. Peng et al. ABC: Attention with bounded-memory control. ACL 2022. [5] M. Charikar, K. Chen, M. Farach-Colton. Finding frequent items in data streams. ICALP 2002. [6] N. Pham, R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. KDD 2013. [7] H. Zhou et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. AAAI 2021.

Acknowledgments. Supported in part by the Hong Kong University of Science and Technology (Guangzhou) (G0101000351) and Lingnan University (SDS24A16).