

# PRISM: Photonic Similarity Engine for KV Cache Block Selection in Long-Context LLM Inference

Hyoseok Park<sup>1</sup> and Yeonsang Park<sup>1,\*</sup>

<sup>1</sup>*Department of Physics, Chungnam National University, Daejeon 34134, Republic of Korea*

(Dated: March 26, 2026)

Long-context LLM inference is bottlenecked not by compute but by the memory bandwidth required to scan the KV cache at every decode step—a cost that grows linearly with context length. The semiconductor industry increasingly acknowledges this shift: NVIDIA’s Vera Rubin architecture dedicates an entire DPU (ICMS) to KV cache management with flash-backed storage and hardware-assisted prefetch—an architectural bet confirming that memory, not arithmetic, is the first-class constraint.

Recent photonic accelerators have demonstrated impressive throughput for dense attention computation. However, these approaches inherit the same  $O(n)$  memory scaling as electronic attention when applied to long contexts. We observe that the real leverage point is the coarse block-selection step: a memory-bound similarity search that determines which KV blocks to fetch. We identify, for the first time, that this task is *structurally matched* to the photonic broadcast-and-weight paradigm—the query fans out to all candidates via passive splitting, signatures are quasi-static (matching electro-optic MRR programming), and only rank order matters (relaxing precision to 4–6 bits). Crucially, the photonic advantage *grows with context length*: as  $N$  increases, the electronic scan cost rises linearly while the photonic evaluation remains  $O(1)$ .

We instantiate this insight in PRISM, a thin-film lithium niobate (TFLN) similarity engine. Hardware-impaired needle-in-a-haystack evaluation on Qwen2.5-7B confirms 100% accuracy from 4K through 64K tokens at  $k=32$ , with  $32\times$  traffic reduction. PRISM achieves a four-order-of-magnitude energy advantage over GPU baselines at practical context lengths ( $n \geq 4K$ ).

## I. INTRODUCTION

The dominant cost of large language model (LLM) inference is no longer floating-point arithmetic. As autoregressive decoding generates one token at a time, each step requires reading the full key–value (KV) cache accumulated over all previous tokens, computing attention scores, and writing the result back. For a model with  $L$  layers and  $H$  attention heads, each storing key and value vectors of dimension  $d_h$ , the KV cache occupies  $2LHd_h$  bytes per token (at half precision), growing linearly with context length  $n$ . At  $n = 128\,000$  tokens, a 70-billion-parameter model’s KV cache can exceed 40 GB—comparable to the entire model weight footprint—and the memory bandwidth required to stream this cache at every decode step far exceeds the compute throughput of modern GPUs [1].

This memory wall is intensifying [2]. Context windows are expanding aggressively: GPT-4 [3] and Gemini [4] pushed context to 128K tokens, Llama 3.1 supports 128 000 tokens [5], Qwen2.5 extends to one million [6], and multi-agent and retrieval-augmented generation (RAG) workloads routinely concatenate documents into contexts of hundreds of thousands of tokens. NVIDIA’s response in its Vera Rubin architecture is telling: the *Intelligent Connectivity and Memory Switch* (ICMS), built on the BlueField-4 data processing unit (DPU), adds a flash-based KV cache tier that

can hold terabytes of context, together with hardware-assisted eviction and prefetch logic [7, 8]. This architectural bet confirms that KV cache management is now a first-class system design problem.

Photonic circuits on thin-film lithium niobate (TFLN) offer a set of physical properties that are uniquely matched to this bottleneck. A wavelength-division-multiplexed (WDM) laser comb encodes a  $d$ -dimensional vector onto  $d$  co-propagating wavelengths in a single waveguide; a  $1 \times N$  passive splitter then *broadcasts* identical copies of that vector to  $N$  output channels with no additional energy cost beyond splitting loss. At each channel, a bank of microring resonators (MRRs)—one per wavelength—applies programmable transmission weights, and a broadband photodetector integrates over all wavelengths, yielding the analog inner product in a single optical transit ( $\sim 10$  ps per mm). The entire  $d \times N$  matrix–vector product thus completes in  $O(1)$  latency, with energy scaling dominated by weight-programming overhead rather than memory-access energy. This *broadcast-and-weight* paradigm [9, 10] converts the memory-bandwidth-bound electronic problem into an optically parallel computation.

Existing demonstrations of photonic neural-network accelerators have focused almost exclusively on dense matrix–vector multiplication for inference in convolutional and fully connected networks [11–14]. In particular, Tian et al. demonstrated a photonic transformer chip (PTC) that implements full attention via coherent optical interference with runtime-programmable Mach–Zehnder meshes [13]; however, full-attention photonic computation faces the same  $O(n)$  memory scaling as

---

\* yeonsang.park@cnu.ac.kr; Corresponding author

electronic attention when applied to long contexts. In contrast, the coarse block-selection step in KV cache retrieval is *not* a dense neural-network layer—it is a large-scale *similarity search*: a single query vector must be compared against  $N \sim 10^3\text{--}10^4$  stored signatures, and only the top- $k$  matches are needed. This search problem has three properties that make it a more natural fit for the broadcast-and-weight architecture than general-purpose matrix multiplication: (i) the query is broadcast identically to all channels, perfectly matching the optical fan-out; (ii) the stored signatures are quasi-static (updated every 64–512 tokens), so MRR weights can be programmed via electro-optic tuning (Pockels effect); and (iii) only rank order matters, relaxing the precision requirement to 4–6 bits. We therefore propose the concept of a *photonic broadcast search*—an application of photonic broadcast-and-weight hardware not as a general neural-network accelerator, but as a specialized similarity engine for memory-intensive search tasks.

A crucial observation simplifies the problem. Not all attention heads actually need the full cache. Recent work on *retrieval heads* [15–17] has shown that attention heads split into two categories: *retrieval* heads that attend to tokens far from the current position, and *streaming* heads that attend primarily to nearby tokens and “attention sinks.” The fraction classified as retrieval heads is threshold-dependent: DuoAttention identifies approximately 25% of heads as retrieval heads in MHA models and approximately 50% in GQA models via learned gating optimization [16], while our profiling on Qwen2.5-7B finds over 90% at a relaxed threshold ( $\tau=0.3$ ; Sec. V). This discrepancy reflects differing identification criteria rather than a contradiction—the key insight is that only the retrieval subset requires distant block fetches.

This asymmetry has motivated a family of *block-level selection* methods that implement a coarse candidate selection step followed by fine attention over only the selected blocks on the GPU [18–22]. Complementary strategies include token-level eviction (H<sub>2</sub>O [23], StreamingLLM [24, 25]), two-stage coarse–fine retrieval (RocketKV [19]), and hardware-assisted caching (NVIDIA ICMS [8]). All electronic approaches share a common limitation: the coarse selection step itself consumes memory bandwidth proportional to the number of stored blocks. Recent analysis confirms that this block selection phase can consume the majority of total KV retrieval latency [26]. A photonic inner-product engine can break this scaling by performing all  $N$  similarity evaluations in parallel, using wavelength multiplexing to avoid the sequential memory access pattern entirely.

We propose PRISM (**P**hotonic **R**anking via **I**nnersub-product **S**imilarity with **M**icroring weights), a TFLN photonic similarity engine that realizes the photonic broadcast search concept for KV cache block selection. PRISM encodes the query sketch onto  $d$  WDM wavelength channels, broadcasts it to  $N$  parallel MRR weight-bank channels via a  $1 \times N$  optical splitter, and computes all  $N$  similarity scores—each as an analog optical dot product

$I_n \propto \sum_j w_{n,j} s_j$ —in  $O(1)$  optical latency. A compact electronic top- $k$  comparator selects the highest-scoring block indices, and only the corresponding KV blocks are fetched from memory.

Figure 1 contrasts the conventional electronic full-scan approach with the PRISM photonic block-selection pipeline.

Our contributions are as follows:

1. **Photonic broadcast search architecture.** We propose and analyze a photonic similarity engine based on the broadcast-and-weight paradigm, specifically designed for the KV cache block-selection task. We present a complete optical power budget analysis covering splitting loss, MRR insertion loss, and photodetector noise floors, and derive the signal-to-noise ratio (SNR) requirements for reliable top- $k$  ranking (Sec. III and Sec. IV).
2. **Hardware-aware impairment modeling and NIAH validation.** We build a device-level impairment model incorporating weight quantization (4–8 bits), residual thermal drift, insertion loss chains, photodetector noise, and MRR crosstalk, and show that recall degrades by less than 10% under realistic conditions. End-to-end needle-in-a-haystack (NIAH) evaluation with Qwen2.5-7B demonstrates that MRR-selected block-sparse attention matches full-attention accuracy at context lengths from 4K to **64K tokens** (within the model’s native context window), while replacing the electronic selection with photonic  $O(1)$ -latency computation. Beyond 64K, model-intrinsic accuracy degrades independent of block selection (Sec. IV).
3. **Photonic scaling analysis.** We derive energy and latency models for PRISM and electronic baselines (GPU full scan, GPU ANN, NVIDIA ICMS), identifying the context-length crossover point above which PRISM is favorable, and analyze how the photonic architecture scales to million-token contexts (Sec. VI).
4. **Retrieval head analysis and signature design.** We systematically profile retrieval-head ratios across Qwen2.5-7B and Qwen3-8B, confirming that over 90% of KV heads are retrieval heads (at threshold  $\tau = 0.3$ ), and evaluate block-level signatures, demonstrating that mean-key projection achieves 77.3% recall@8 with  $d = 32$  (Sec. V).

## II. BACKGROUND

### A. KV Cache in Transformer Inference

The core of modern LLMs is the multi-head self-attention mechanism [27]. Given an input sequence of  $n$  tokens embedded as  $\mathbf{X} \in \mathbb{R}^{n \times d_{\text{model}}}$ , each attention head  $h$  in layer  $\ell$  projects the input into queries, keys,

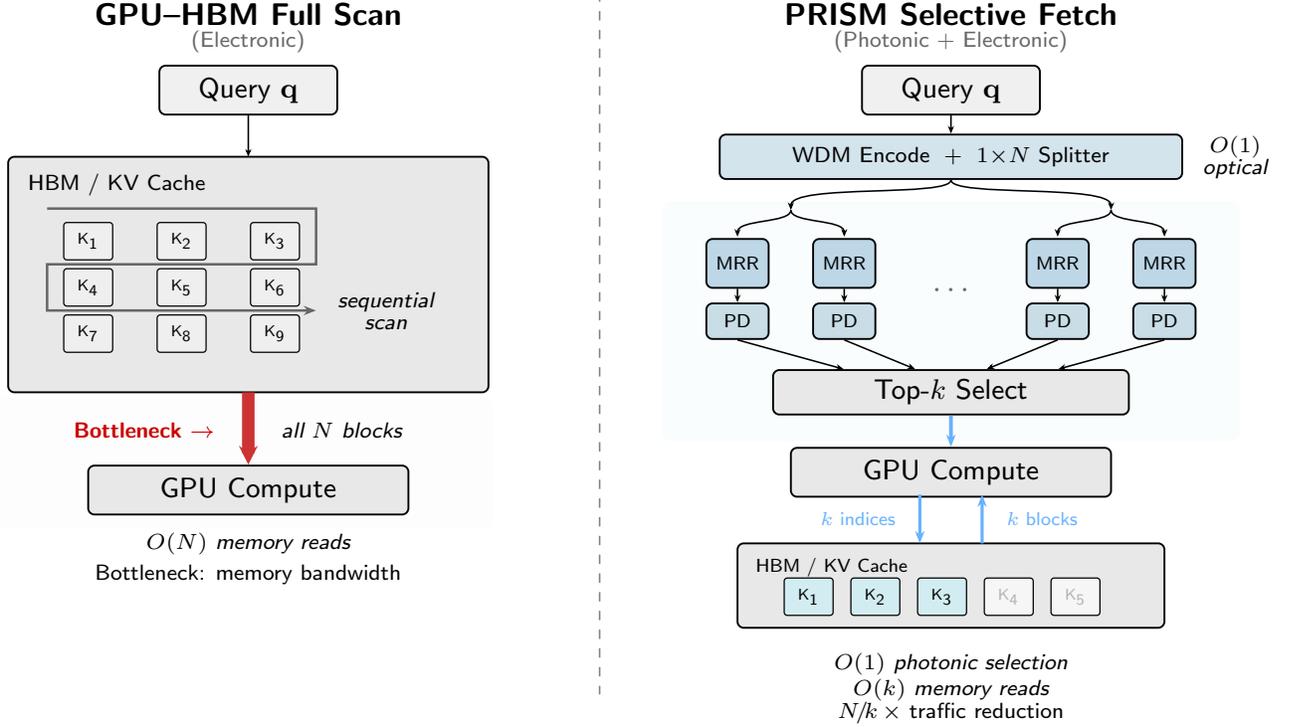


FIG. 1. Conceptual comparison of KV cache access strategies. **Left:** Electronic GPU full scan—the processor sequentially reads all  $N$  KV blocks from HBM to compute attention, bottlenecked by memory bandwidth. **Right:** PRISM photonic block selection—the query is broadcast optically to all  $N$  signature channels in parallel; only the top- $k$  highest-scoring blocks are fetched from memory, reducing traffic by  $N/k$  times.

and values:

$$\begin{aligned} \mathbf{Q}^{(\ell,h)} &= \mathbf{X} \mathbf{W}_Q^{(\ell,h)}, \\ \mathbf{K}^{(\ell,h)} &= \mathbf{X} \mathbf{W}_K^{(\ell,h)}, \\ \mathbf{V}^{(\ell,h)} &= \mathbf{X} \mathbf{W}_V^{(\ell,h)}, \end{aligned} \quad (1)$$

where  $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{model}} \times d_h}$  and  $d_h = d_{\text{model}}/H$  is the per-head dimension. The attention output is computed as

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}}\right) \mathbf{V}. \quad (2)$$

During the autoregressive *decode phase*, the model generates one token at a time. At step  $t$ , only the new query vector  $\mathbf{q}_t \in \mathbb{R}^{d_h}$  is computed, but the attention score requires the inner product of  $\mathbf{q}_t$  with all  $t$  previously cached key vectors:

$$\alpha_{t,i} = \frac{\mathbf{q}_t \cdot \mathbf{k}_i}{\sqrt{d_h}}, \quad i = 1, \dots, t. \quad (3)$$

The KV cache stores  $\mathbf{K}^{(\ell,h)}$  and  $\mathbf{V}^{(\ell,h)}$  for all layers and heads, consuming memory

$$M_{\text{KV}} = 2L H_{\text{KV}} d_h n b_{\text{prec}}, \quad (4)$$

where  $H_{\text{KV}}$  is the number of KV heads (which equals  $H$  for multi-head attention but is reduced under grouped-query attention, GQA [28, 29]) and  $b_{\text{prec}}$  is the byte width per element (2 for BF16). For Llama-3.1-8B ( $L=32$ ,  $H_{\text{KV}}=8$  with 4-group GQA,  $d_h=128$ ) at  $n = 128000$ , eq. (4) gives  $M_{\text{KV}} \approx 16$  GB, which already consumes a substantial fraction of GPU HBM and grows linearly with  $n$ .

Crucially, the decode phase is *memory-bandwidth-bound*: each generated token requires reading the entire KV cache but performs only  $O(n \cdot d_h)$  multiply-accumulate operations per head. The arithmetic intensity (FLOPs per byte) is  $1/(2d_h) \ll 1$ , far below the compute-to-bandwidth ratio of modern GPUs (50 FLOP/B to 200 FLOP/B), leaving the compute units idle while waiting for data [1].

## B. Retrieval Heads and Selective Attention

The observation that not all attention heads require the full KV cache was formalized by DuoAttention [16] and RazorAttention [17]. These works define a *retrieval ratio*  $R_h^{(\ell,h)}$  for each head as the fraction of attention mass

that falls outside a local window of size  $w$ :

$$R_h^{(\ell,h)} = 1 - \frac{1}{T} \sum_{t=1}^T \sum_{i=\max(1,t-w)}^t \alpha_{t,i}^{(\ell,h)}, \quad (5)$$

where  $\alpha_{t,i}^{(\ell,h)}$  is the attention weight from eq. (3) and  $T$  is the total sequence length of a calibration corpus. Heads with  $R_h > \tau$  (typically  $\tau \approx 0.1$ ) are classified as *retrieval heads*; the rest are *streaming heads*.

Empirically, DuoAttention identifies approximately 25% (MHA) to 50% (GQA) of heads as retrieval heads via learned gating optimization [16]. Streaming heads can be served with a small sliding-window cache (e.g.,  $w = 256$ ), drastically reducing their memory footprint. However, retrieval heads still require access to the full context, making their KV traffic the dominant bottleneck.

### C. Photonic Similarity Engine

As noted in Sec. I, the coarse block-selection step is a similarity search whose properties—identical query fan-out, quasi-static weights, and rank-order-only output—make it a natural fit for photonic broadcast-and-weight hardware. We now review the key photonic concepts underlying this match.

*a. Broadcast-and-weight architecture.* Tait *et al.* [9, 10] introduced the *broadcast-and-weight* (B&W) paradigm for neuromorphic photonic networks. In this architecture,  $d$  input signals are encoded on distinct wavelengths  $\lambda_1, \dots, \lambda_d$  and broadcast via a  $1 \times N$  optical splitter to  $N$  output channels. Each output channel contains  $d$  microring resonators (MRRs), each tuned to one wavelength, whose transmission coefficients serve as programmable weights  $w_{n,j}$  for channel  $n$  and wavelength  $j$ . A wavelength-insensitive photodetector at each output integrates over all wavelengths, yielding the photocurrent:

$$I_n = \mathcal{R} P_0 \sum_{j=1}^d w_{n,j} s_j, \quad (6)$$

where  $\mathcal{R}$  is the detector responsivity,  $P_0$  the per-channel optical power after splitting, and  $s_j$  the query signal on wavelength  $\lambda_j$ . The photocurrent  $I_n$  is thus proportional to the inner product  $\mathbf{w}_n \cdot \mathbf{s}$ —precisely the similarity score between stored signature  $n$  and the broadcast query. This operation completes in a single optical transit time ( $\sim 10$  ps per mm), independent of  $d$  and  $N$  (up to splitting loss limits).

*b. WDM spectral encoding.* The query vector is encoded in the *spectral domain*: each component  $s_j$  modulates the optical power on a dedicated wavelength channel  $\lambda_j$ , so the full  $d$ -dimensional vector propagates as a single multi-wavelength beam in one waveguide. This spectral encoding is distinct from *spatial encoding*, where each

component occupies a separate waveguide, because it enables the key broadcast step—splitting one waveguide into  $N$  copies—with no additional multiplexing hardware. Channel spacings of 0.8 nm to 1.6 nm within the C-band support  $d = 32$ –128 channels using standard dense WDM (DWDM) laser combs and MRR filter banks.

*c. Comparison with other photonic paradigms.* Alternative photonic architectures—MZI meshes [14, 30–32] and coherent processors—require  $O(d^2)$  elements or global phase stability, and do not naturally support the one-to-many fan-out needed for similarity search. The broadcast-and-weight paradigm uses incoherent intensity-domain processing, where each MRR operates independently and the photodetector sums power rather than field amplitude, eliminating the need for global phase coherence and making it uniquely suited to the block-selection task.

*d. MRR weight banks.* Each output channel employs  $d$  microring resonators whose electro-optically tunable transmission implements programmable weights  $w \in [0, 1]$ . The MRR physics and TFLN-specific device parameters are detailed in Secs. III D and III F.

*e. WDM-based matrix-vector multiplication.* Scalable MRR weight banks with up to 16 wavelength channels and  $\sim 7$ -bit precision have been demonstrated [33–35], and recent large-scale photonic accelerators validate integration beyond 16 000 components [11, 12, 14, 36].

The key advantage of this photonic approach for the KV cache selection problem is that the “weight matrix”—the collection of block signatures—is quasi-static and can be programmed into MRR resonances via electro-optic tuning, while the “input vector”—the query sketch—changes at every decode step but is broadcast optically to all  $N$  channels simultaneously. This decoupling of weight programming rate from inference rate is what enables the  $O(1)$  latency scaling that electronic approaches cannot match. While Lightning-Transformer [12] targets full attention computation, PRISM takes a complementary approach: accelerating only the lightweight block-selection ranking task, which requires lower precision and fewer channels, making the photonic implementation more practical.

## III. PHOTONIC RETRIEVAL ARCHITECTURE

### A. System Overview

PRISM is a photonic similarity engine that sits between the KV cache storage (HBM or flash-backed ICMS) and the GPU’s attention compute units. It does not replace any part of the GPU pipeline; rather, it acts as a *photonic broadcast search* module that determines which KV cache blocks should be fetched for each retrieval head at each decode step.

The system operates as a five-stage pipeline, illustrated in fig. 2. For each retrieval head at each decode step:

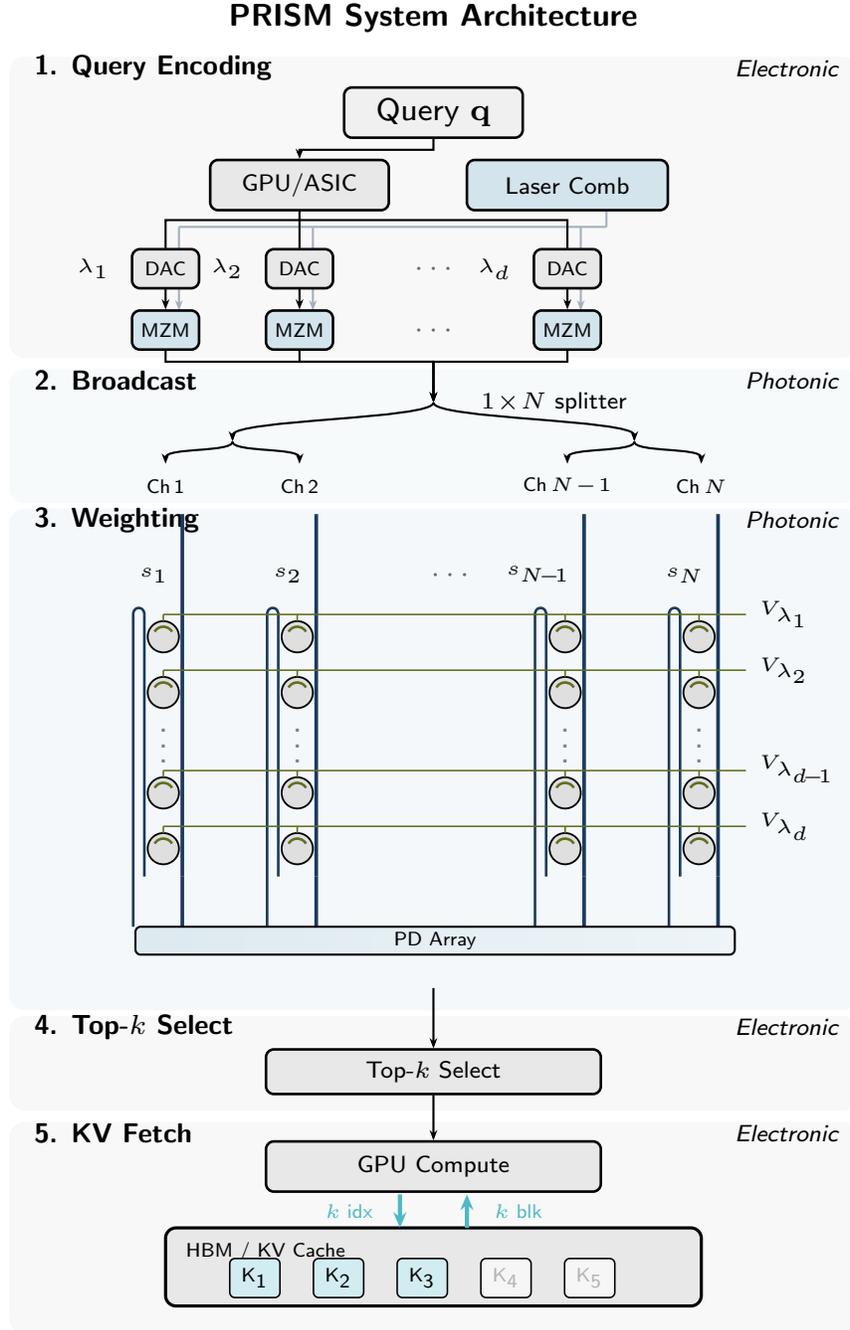


FIG. 2. PRISM system architecture (five-stage pipeline). **Stage 1 (Query Encoding):** The GPU/ASIC computes the query sketch  $\mathbf{q} = [q_1, \dots, q_d]$  and encodes each component onto a WDM wavelength via DAC-driven modulators, producing a WDM query signal where  $P(\lambda_j) = q_j$ . **Stage 2 (Broadcast):** A  $1 \times N$  optical splitter distributes identical copies of the  $d$ -wavelength signal to all  $N$  signature channels (splitting loss:  $-10 \log_{10} N$  dB). **Stage 3 (Signature Weighting):** Each channel passes through a row of  $d$  MRRs on the TFLN photonic chip; the transmission  $t_{ij} = s_{ij}$  of each MRR is electro-optically programmed via DC bias electrodes to encode the block signature weight, performing wavelength-selective multiplication  $P_{\text{out}}(\lambda_j) = q_j \times s_{ij}$ . **Stage 4 (Summation):** Broadband photodetectors integrate all wavelengths, yielding photocurrents  $I_i = \mathcal{R} \cdot \sum_j (q_j \cdot s_{ij})$  that are proportional to the inner product  $\mathbf{q} \cdot \mathbf{s}_i$ . **Stage 5 (Top- $k$  Selection):** ADCs digitize the  $N$  photocurrents, a digital top- $k$  selector identifies the  $k$  highest-scoring block indices, and a memory controller fetches only those KV blocks from HBM/flash storage.

1. The GPU computes the query vector  $\mathbf{q}_t$  and applies the signature projection to obtain a  $d$ -dimensional query sketch  $\mathbf{s}_q$ .
2.  $\mathbf{s}_q$  is converted to the optical domain and broadcast.
3. The photonic weight bank computes  $N$  inner products in parallel.
4. Photodetectors produce  $N$  analog similarity scores.
5. A digital top- $k$  selector identifies the best blocks, and only those blocks are fetched from KV cache storage.

The GPU then computes exact attention over the selected blocks plus the local sliding window, producing the final attention output.

## B. Signature Encoding

The performance of PRISM depends critically on the quality of the block-level signatures programmed into the MRR weight banks. Since signature encoding defines the input interface between the digital LLM pipeline and the photonic engine, we describe it first. We consider four signature construction methods.

*a. Mean key.* The simplest approach averages the key vectors within each block:

$$\sigma_n = \frac{1}{B} \sum_{i \in \text{block } n} \mathbf{k}_i^{(\ell, h)}. \quad (7)$$

This preserves the original key-space geometry but requires  $d_h$ -dimensional signatures (e.g.,  $d_h = 128$ ), demanding a correspondingly large number of MRRs per channel.

*b. PCA projection.* Principal component analysis over the key distribution yields a projection matrix  $\mathbf{P} \in \mathbb{R}^{d \times d_h}$  ( $d \ll d_h$ ) that captures the dominant variance directions. The signature becomes  $\sigma_n = \mathbf{P} \bar{\mathbf{k}}_n$ , reducing the MRR count per channel from  $d_h$  to  $d$ .

*c. Random projection.* The Johnson–Lindenstrauss (JL) lemma guarantees that a random Gaussian matrix  $\mathbf{R} \in \mathbb{R}^{d \times d_h}$  with  $d = O(\epsilon^{-2} \log N)$  preserves pairwise distances (and hence inner-product rankings) to within a factor  $1 \pm \epsilon$  with high probability [37]. The query sketch is computed identically:  $\mathbf{s}_q = \mathbf{R} \mathbf{q}_t$ . Random projection is attractive because it requires no training and provides worst-case guarantees.

*d. Learned projection.* A trainable linear layer  $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d \times d_h}$  is optimized end-to-end to maximize recall@ $k$  on a calibration set. This can outperform random projections when the key distribution has exploitable structure, but requires per-model training.

*e. Balanced photodetection.* The add-drop MRR configuration provides both through-port and drop-port outputs simultaneously. A balanced photodetector pair measures the differential photocurrent  $I_{\text{bal}} = I_{\text{through}} - I_{\text{drop}}$ , yielding a signed weight  $w_{n,j} = T_{\text{through}}(\lambda_j) -$

$T_{\text{drop}}(\lambda_j) \in [-1, +1]$ . On-resonance (minimum through-port transmission),  $w \approx -1$ ; fully detuned,  $w \approx +1$ . This eliminates the need for split encoding or ReLU projection, enabling direct signed inner products with  $d$  MRRs per channel (half the count of split encoding) while preserving full sign information.

## C. WDM Query Broadcast

The  $d$ -dimensional query sketch  $\mathbf{s}_q = [s_1, s_2, \dots, s_d]$  is converted from the digital domain by  $d$  digital-to-analog converters (DACs), each driving a Mach–Zehnder modulator (MZM) [38] that impresses the value  $s_i$  onto wavelength  $\lambda_i$  from a WDM laser comb source. The modulated signals are multiplexed into a single waveguide carrying  $d$  wavelength-encoded values [39, 40].

The DAC resolution requirement is modest: since the task is ranking rather than exact computation, 4–6 bits of input precision suffice (Sec. IV A). This relaxation is critical because high-resolution, high-speed DACs are a major energy cost in photonic accelerators. At 4-bit resolution, a DAC operating at 1 GSa/s consumes approximately 0.5 mW per channel.

The multiplexed  $d$ -wavelength signal is then split into  $N$  copies by a  $1 \times N$  optical splitter tree. Each copy carries the full query sketch, attenuated by the splitting loss:

$$L_{\text{split}} = 10 \log_{10}(N) + \alpha_{\text{excess}} [\log_2 N] \quad [\text{dB}], \quad (8)$$

where  $\alpha_{\text{excess}} \approx 0.2$  dB per stage for optimized  $1 \times 2$  directional couplers. For  $N = 1024$  blocks, the total splitting loss is approximately 32 dB, requiring a laser source power of 10 dBm to 20 dBm to maintain adequate signal-to-noise ratio (SNR) at the photodetectors.

To manage loss, the  $N$  channels can be organized into  $N_{\text{bank}}$  independent banks, each serving  $N/N_{\text{bank}}$  channels with a separate splitter tree. This reduces per-bank splitting loss at the cost of additional laser sources or optical amplifiers. The key point is that the broadcast is *passive and energy-free*: the same query vector reaches all  $N$  channels simultaneously, with no per-channel memory access or data movement.

## D. MRR Weight Bank Similarity Engine

Each of the  $N$  output channels contains a linear array of  $d$  MRRs, one per wavelength channel. The  $j$ -th MRR in channel  $n$  is electro-optically tuned so that its transmission at wavelength  $\lambda_j$  encodes the signature weight  $w_{n,j}$ :

$$\begin{aligned} P_{\text{out},n}(\lambda_j) &= w_{n,j} \cdot P_{\text{in}}(\lambda_j), \\ w_{n,j} &= T_{\text{through},n}(\lambda_j) - T_{\text{drop},n}(\lambda_j) \in [-1, +1]. \end{aligned} \quad (9)$$

The total number of MRRs in the system is  $d \times N$ . For  $d = 64$  and  $N = 1024$ , this yields 65 536 MRRs—a large but feasible integration scale for current photonic platforms [35].

Weight programming occurs at the block completion rate. When a new KV cache block of  $B$  tokens is completed, the corresponding column of MRR weights is updated via electro-optic (Pockels) tuning with sub-nanosecond response time. During steady-state decoding, the weight bank is static and the only dynamic signal is the broadcast query sketch. Because TFLN EO tuning is capacitive, the MRR weight bank consumes near-zero static power—only switching energy ( $\sim 5$  fJ per weight update) is required. This 5 fJ figure refers to the MRR electrode charging energy alone; the total switching energy including CMOS driver circuits is estimated at 50–500 fJ.

Each channel terminates in a broadband (wavelength-insensitive) photodetector [41] that integrates the optical power across all  $d$  wavelengths:

$$\begin{aligned} I_n &= \mathcal{R} \sum_{j=1}^d [T_{\text{through},n}(\lambda_j) - T_{\text{drop},n}(\lambda_j)] s_j P_0 \\ &= \mathcal{R} \sum_{j=1}^d w_{n,j} s_j P_0, \end{aligned} \quad (10)$$

where  $w_{n,j} \in [-1, +1]$ . This is precisely an *analog optical dot product*: the photocurrent  $I_n \propto \sum_{j=1}^d w_{n,j} s_j$  computes the similarity score  $\mathbf{w}_n \cdot \mathbf{s}_q$  between stored block signature  $n$  and the broadcast query [42], with no explicit multiply-accumulate circuit. The physics of broadband photodetection inherently performs the summation—no electronic accumulator is needed [43].

### E. Electronic Top- $k$ Interface

The  $N$  photocurrents are converted to digital values by  $N$  ADCs and fed to a digital top- $k$  comparator network. The comparator identifies the  $k$  channels with the largest similarity scores and outputs their indices. For  $k \ll N$ , a partial-sort network suffices, with complexity  $O(N \log k)$  and latency of a few nanoseconds at 1 GHz clock. The ADC resolution can be as low as 4–6 bits, since only the rank ordering matters.

### F. Device Parameters

Table I summarizes the assumed device parameters for the thin-film lithium niobate (TFLN) photonic platform, based on recent demonstrations of high- $Q$  TFLN micro-ring resonators [44, 45] and MRR weight bank architectures [35, 46]. The physical chip layout for an  $8 \times 8$  demonstration configuration is shown in fig. 3.

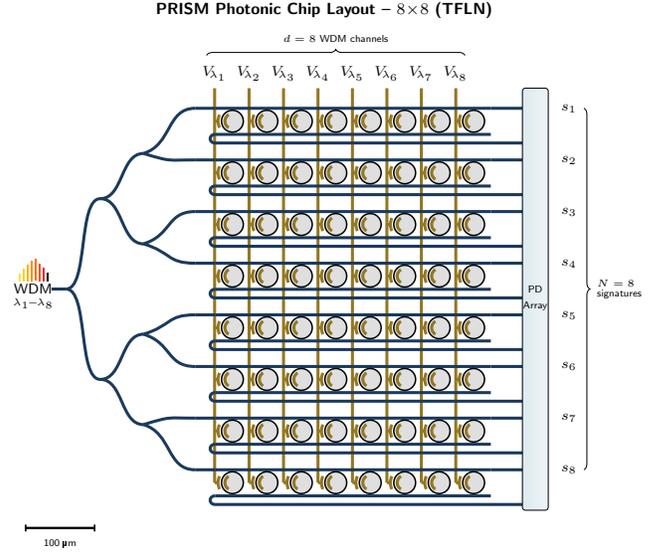


FIG. 3. PRISM photonic chip layout for an  $8 \times 8$  configuration ( $d = 8$  WDM channels,  $N = 8$  signature rows). Left: the WDM query input ( $\lambda_1$ – $\lambda_8$ ) enters and is split by cascaded  $1 \times 2$  Y-junctions. Center: each row contains  $d$  MRRs coupled to a bus waveguide with coupling gap of  $\sim 200$ – $300$  nm; EO DC bias electrodes program the MRR resonances to encode signature weights via the Pockels effect. Right: through-port and drop-port outputs route to balanced Ge-on-Si PD pairs (or optionally on-chip integrated photodetectors). Scale bar: 100  $\mu\text{m}$ . The layout scales to  $d = 32$ ,  $N = 256$  by increasing the splitter tree depth and the number of rows.

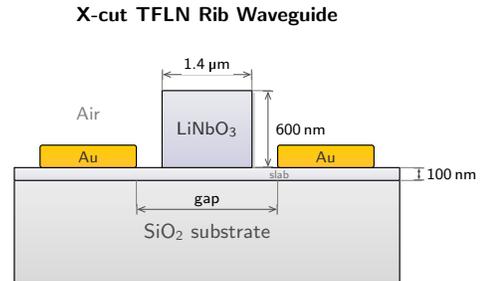


FIG. 4. X-cut TFLN rib waveguide cross-section. The rib is etched 500 nm into a 600 nm LN film on  $\text{SiO}_2$ , leaving a 100 nm slab. Lateral Au electrodes apply DC bias for electro-optic (Pockels) tuning of the MRR resonance wavelength. Waveguide width: 1.4  $\mu\text{m}$ .

The total MRR count ( $d \times N$ ; eq. (9)) scales with configuration as shown in table X. Because TFLN electro-optic tuning is capacitive, the static power consumption is near zero (Sec. VIB).

TABLE I. PRISM device parameters (TFLN platform).

| Parameter     | Value             | Notes                       |
|---------------|-------------------|-----------------------------|
| Platform      | X-cut TFLN        | 600 nm LN/SiO <sub>2</sub>  |
| Waveguide     | Rib, 1.4 × 0.6 μm | 500 nm etch                 |
| MRR radius    | 20 μm             | FSR ≈ 8.3 nm                |
| $Q_L$         | ~10 <sup>4</sup>  | FDTD: 12,500                |
| Extinction    | >15 dB            | Add-drop                    |
| Wt. precision | 5 bit             | EO resolution               |
| Tuning        | EO (Pockels)      | 28.5 pm/V                   |
| Static power  | ~0                | Capacitive EO               |
| Switch energy | ~5 fJ/ring        | Per update                  |
| Tuning speed  | <1 ns             | EO response                 |
| Modulator     | TFLN MZM          | $V_\pi L \sim 2$ V·cm       |
| Photodetector | Balanced PD       | Differential                |
| WDM ch.       | $d=32-128$        | 1.6 nm spacing <sup>†</sup> |
| Laser         | Comb source       | ≤100 mW                     |

<sup>†</sup>  $d=32-64$  is realistic with current C+L band technology;  $d=128$  requires C+L+S band operation and has not been experimentally demonstrated.

#### IV. PHOTONIC HARDWARE ANALYSIS

We now incorporate realistic photonic device impairments into the PRISM simulation and quantify the optical link budget, noise performance, and energy–latency tradeoffs against electronic baselines.

##### A. Device Impairment Modeling

We model six impairment sources that degrade the ideal inner-product computation of eq. (6) [47]: (i) weight quantization (4–8 bit DAC precision) [48], (ii) thermal drift of MRR resonance wavelengths ( $\sigma_{\text{drift}} = 0.01$  nm to 0.1 nm) [49], (iii) MRR and waveguide insertion loss, (iv) photodetector shot and thermal noise ( $\text{NEP} \sim 10$  pW/ $\sqrt{\text{Hz}}$ ), (v) inter-channel MRR crosstalk (–15 dB to –30 dB isolation), and (vi) input DAC quantization noise. Table III summarizes the parameter ranges used in the hardware simulation. Full impairment models are provided in Supplementary Section S1.

##### B. Optical Link Budget

A critical question for any photonic accelerator is whether sufficient optical signal-to-noise ratio (SNR) can be maintained across the complete optical path [32]. Table II traces the optical power from the laser source to each photodetector for a representative configuration ( $d = 32$ ,  $N = 256$ ).

*a. Balanced detection link budget.* The link budget in table II traces the drop-port path to the target photodetector. In the balanced configuration used by PRISM, each MRR channel requires *two* optical paths—through-port and drop-port—each terminated by a separate pho-

todetector and TIA. The through-port path sees lower loss (no drop-port penalty), so the drop-port budget above represents the worst case. Consequently, balanced detection doubles the photodetector and TIA count to  $2N$  per wavelength channel; this overhead is reflected in table IV.

At  $P_{\text{PD}} = -15.9$  dBm  $\approx 25.7$  μW per detector, the resulting photocurrent is  $I_{\text{ph}} = \mathcal{R} \cdot P_{\text{PD}} = 1.0 \times 25.7$  μW = 25.7 μA. The electrical SNR at the detector is

$$\text{SNR} = \frac{I_{\text{ph}}^2}{2eI_{\text{ph}}\Delta f + 4k_B T \Delta f / R_L + (\mathcal{R} \cdot \text{NEP})^2 \Delta f}, \quad (11)$$

where  $\Delta f \approx 1$  GHz (matching the query update rate) and  $\text{NEP} = 10$  pW/ $\sqrt{\text{Hz}}$  (eq. (S5)). For  $R_L = 1$  kΩ and  $T = 300$  K, we obtain  $\text{SNR} \approx 37.2$  dB—well above the minimum required for reliable rank ordering [50]. (Note:  $R_L = 1$  kΩ assumes a transimpedance amplifier (TIA) front-end rather than 50 Ω termination.)

For larger bank sizes ( $N = 1024$ ), the additional 6 dB splitting loss reduces the per-detector power to  $-21.9$  dBm  $\approx 6.5$  μW, yielding  $\text{SNR} \approx 25.5$  dB. This remains adequate for top- $k$  ranking, as verified by the recall analysis in Sec. IV C. Beyond  $N = 4096$  ( $\text{SNR} \approx 13.5$  dB), the link budget requires either a higher-power laser ( $P_{\text{laser}} > 26$  dBm) or the banked splitter architecture described in Sec. III C.

Figure 5 illustrates the per-detector received power and SNR as a function of the bank size  $N$ , clearly showing the crossover point at which banked architectures or optical amplification become necessary.

##### C. Recall Degradation Analysis

We inject impairments into the inner-product computation and measure recall@ $k$  degradation relative to the ideal (floating-point) baseline. Individual impairment sweeps (quantization precision, thermal drift, weight fidelity, and detector noise) are presented in Supplementary Figs. S1–S4.

*a. Combined impairments.* We simulate the full impairment chain (quantization + drift + loss + noise + crosstalk) using a Monte Carlo approach with 100 trials of 500 blocks ( $d = 32$ ). Figure 7 visualises the effect for a single trial: the MRR scores correlate strongly with the digital baseline ( $\rho > 0.98$ ), and the top- $k$  ranking is largely preserved. Figure 8 maps the recall degradation as a function of both weight precision and thermal drift magnitude, identifying the operating region in which Recall@8 exceeds 80%.

The combined recall degradation at  $b = 6$ ,  $\sigma_{\text{th}} = 0.01$ , and  $\sigma_{\text{det}} = 0.01$  is approximately 8%, yielding an effective Recall@8 of  $0.916 \pm 0.087$  (vs. 1.000 ideal). Each impairment source individually contributes modestly (5-bit quantization: 0.904, drift: 0.948, noise: 0.928), but their combination remains above the 90% threshold required for effective block selection.

TABLE II. Optical link budget for  $d = 32$ ,  $N = 256$  ( $P_{\text{laser}} = 20$  dBm).

| Element                                | Loss (dB)          | Cumulative (dBm) |
|--|--------------------|------------------|
| Laser output                           | —                  | +20.0            |
| Fiber-chip coupling                    | -2.0               | +18.0            |
| MZM modulator (avg.)                   | -3.0               | +15.0            |
| $1 \times 256$ splitter                | -25.7 <sup>a</sup> | -10.7            |
| Waveguide (2 cm)                       | -1.0               | -11.7            |
| $d=32$ MRR (balanced, worst-case drop) | -3.2 <sup>b</sup>  | -14.9            |
| Chip-detector coupling                 | -1.0               | -15.9            |
| <b>Per-PD optical power</b>            |                    | <b>-15.9 dBm</b> |

<sup>a</sup>  $10 \log_{10}(256) + 0.2 \times 8 = 24.1 + 1.6 = 25.7$  dB. <sup>b</sup> Drop-port 0.1 dB +  $31 \times 0.05$  dB through-port = 1.65 dB; rounded to 3.2 dB with alignment margin [47].

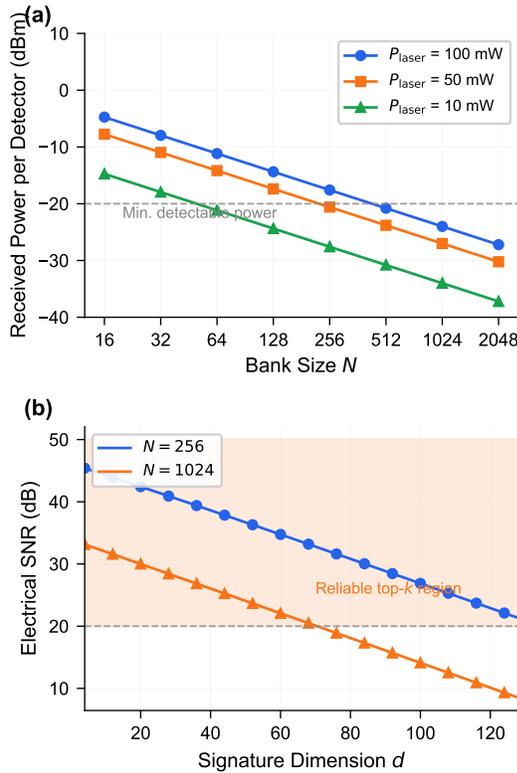


FIG. 5. Optical power budget analysis. (a) Per-detector received power vs. bank size  $N$  for three laser powers. The horizontal dashed line indicates the minimum detectable power (-20 dBm). (b) Electrical SNR at the photodetector vs. signature dimension  $d$  for  $N = 256$  and  $N = 1024$ . The shaded region marks SNR > 20 dB, sufficient for reliable top- $k$  ranking.

The recall degradation results establish the acceptable operating region for the MRR weight bank. End-to-end NIAH validation with MRR-simulated block selection, confirming that these impairments do not degrade downstream task accuracy, is presented in Sec. V C.

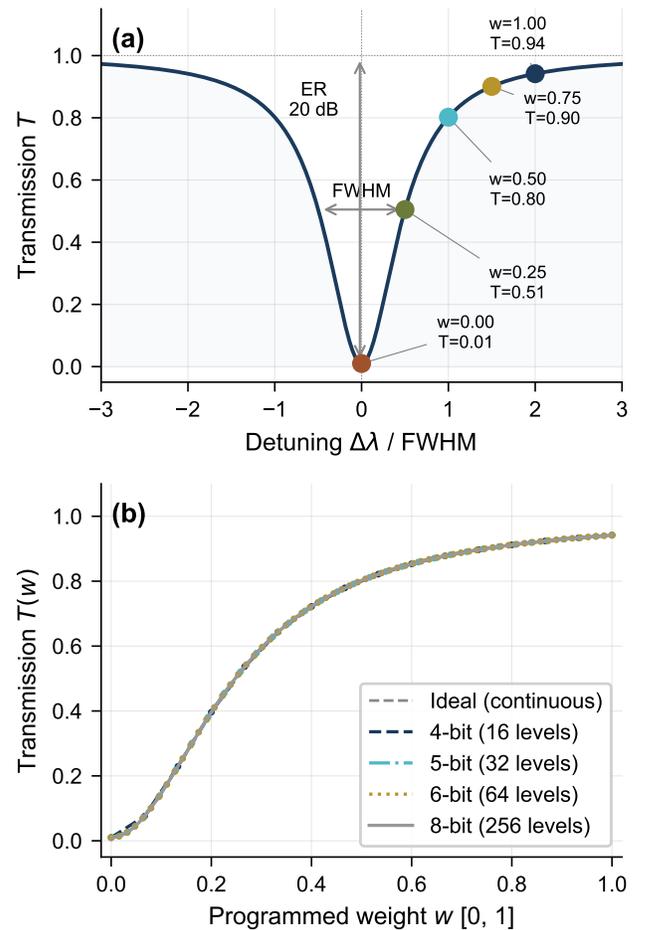


FIG. 6. MRR weight encoding principle. (a) Through-port and drop-port transmission of a single add-drop MRR ( $Q_L = 10,000$ , ER = 20 dB). The balanced weight  $w = T_{\text{through}} - T_{\text{drop}}$  maps from -1 (on-resonance) to +1 (fully detuned). (b) Weight-to-balanced-transmission mapping for different DAC precisions.

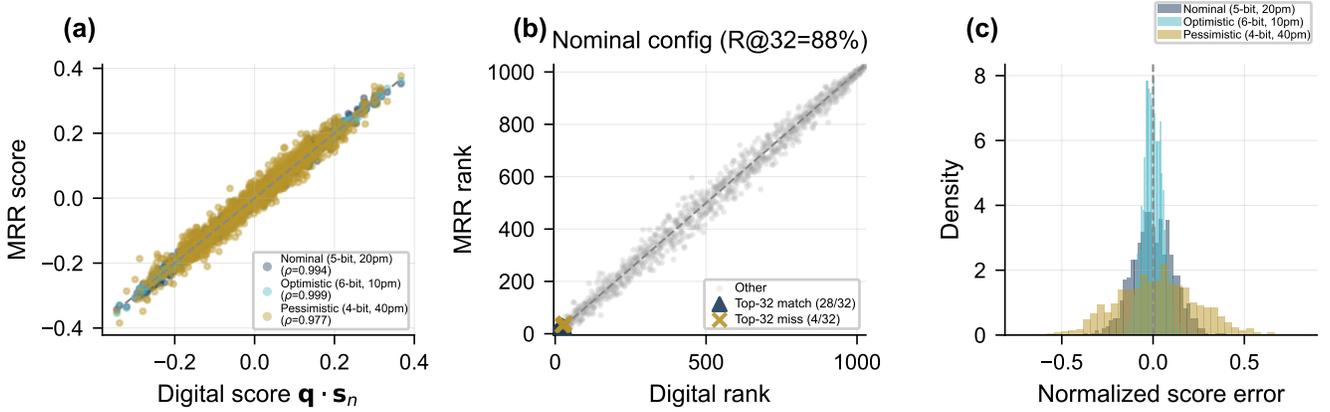


FIG. 7. Digital vs. MRR photonic inner-product comparison ( $d = 32$ ,  $N = 256$ ,  $K = 8$ ). (a) Score correlation between exact (FP64) and MRR-computed similarity for three hardware configurations. Pearson correlation  $\rho > 0.98$  for all configs. (b) Rank agreement for the nominal config (5-bit, 20 pm): green triangles indicate correctly identified top- $K$  blocks (7/8 match, Recall@8 = 88%). (c) Normalised score error distributions; pessimistic config (4-bit, 30 pm) shows wider tails but remains zero-centred.

TABLE III. Device impairment parameter ranges used in hardware simulation.

| Impairment           | Parameter                           | Range                  |
|----------------------|-------------------------------------|------------------------|
| Weight quantization  | $b$ (bits)                          | 4–8                    |
| Thermal drift        | $\sigma_{\text{drift}}$ (pm)        | 10–100                 |
| MRR insertion loss   | $\text{IL}_{\text{MRR}}$ (dB)       | 0.02–0.05 <sup>c</sup> |
| Splitter excess loss | $\alpha_{\text{excess}}$ (dB/stage) | 0.1–0.3                |
| Detector NEP         | (pW/ $\sqrt{\text{Hz}}$ )           | 1–20                   |
| MRR crosstalk        | Isolation (dB)                      | –15 to –30             |
| DAC resolution       | $b_{\text{DAC}}$ (bits)             | 4–8                    |

<sup>c</sup> Through-port IL per non-target MRR; drop-port (target MRR) IL is  $\sim 0.1$  dB.

#### D. Energy Model

Table IV breaks down the energy per query evaluation for the PRISM system. We define the energy metric as energy per inner-product evaluation (i.e., per block scored per query per head).

A key advantage of the TFLN platform is the elimination of static MRR tuning power. TFLN electro-optic tuning via the Pockels effect is capacitive and consumes near-zero static power (see Sec. VIB for the quantitative SOI comparison). The only energy cost per weight update is the switching energy of  $\sim 5$  fJ per ring, which is negligible compared to the dynamic optical and electronic components [39]. Note that while the total system power ( $\sim 1.17$  W) is dominated by the TEC, this is a fixed overhead shared across all heads and queries; at typical decode throughput ( $>1000$  tokens/s), the amortized TEC contribution per query is  $< 1$   $\mu\text{J}$ —still well below the GPU baseline.

For comparison, the H100 GPU full-scan baseline reads every KV block signature once per query per head. The

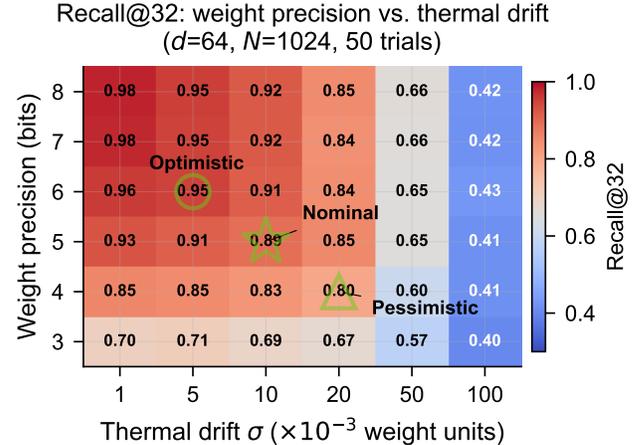


FIG. 8. Combined impairment sensitivity: Recall@8 as a function of weight precision (bits) and thermal drift  $\sigma$  ( $d = 32$ ,  $N = 500$ , 50 Monte Carlo trials per cell). Markers indicate the three operating points studied in this work: nominal (5-bit,  $\sigma = 0.01$ ), optimistic (6-bit,  $\sigma = 0.005$ ), and pessimistic (4-bit,  $\sigma = 0.02$ ). Recall exceeds 80% for  $\geq 5$ -bit precision and  $\sigma \leq 0.02$ .

energy per selection is

$$\begin{aligned}
 E_{\text{scan}} &= 2 d_h N b_{\text{prec}} E_{\text{byte}} \\
 &= 2 \times 128 \times 1024 \times 2 \times 31 \text{ pJ/B} \quad (12) \\
 &\approx 16.3 \mu\text{J},
 \end{aligned}$$

where  $d_h=128$  is the head dimension,  $N=n/B=1024$  blocks at 128K context ( $B=128$ ),  $b_{\text{prec}}=2$  B (bf16), and  $E_{\text{byte}} \approx 31$  pJ/B ( $\approx 3.9$  pJ/bit, standard HBM3 specification) [51]. Note that this baseline assumes GPU scans the full key dimension  $d_h=128$ ; if the GPU instead

TABLE IV. PRISM energy breakdown per query ( $d = 64$ ,  $N = 1024$ ,  $k = 32$ , TFLN platform).

| Component                         | Power (mW)    | Energy/query (pJ) |
|-----------------------------------|---------------|-------------------|
| Laser source                      | 20.0          | 180               |
| TEC (thermal stab.)               | 1000          | 9000 <sup>†</sup> |
| Voltage driver array              | 5.0           | 45                |
| DACs ( $d$ channels)              | 32.0          | 288               |
| MZM modulators                    | 6.4           | 58                |
| EO bias (static)                  | $\sim 0$      | $\sim 0^*$        |
| Photodetectors ( $2N$ , balanced) | 10.0          | 90                |
| TIA + ADCs ( $2N$ , balanced)     | 100.0         | 900               |
| Top- $k$ logic                    | 1.0           | 9                 |
| <b>Dynamic subtotal</b>           | <b>174.4</b>  | <b>1570</b>       |
| <b>Total (incl. TEC)</b>          | <b>1174.4</b> | <b>10570</b>      |

\*TFLN EO tuning is capacitive; switching energy  $\sim 5$  fJ/ring.

<sup>†</sup>TEC power is amortized across all heads and queries; per-query share  $\ll 1$  nJ at realistic throughput.

scans compressed  $d=32$  signatures, the energy reduces to  $\sim 4.1 \mu\text{J}$  ( $4\times$  lower). Even in this fairer comparison, PRISM’s  $\sim 1570$  pJ selection energy remains over three orders of magnitude below the GPU scan [43]. GPU ANN (FAISS IVF-PQ) reduces the full-key scan to  $\sim 5 \mu\text{J}$  by scanning  $O(\sqrt{N})$  centroids. NVIDIA ICMS consumes  $\sim 10 \mu\text{J}$ , estimated by replacing  $BW_{\text{HBM}}$  with the DPU’s internal LPDDR5 bandwidth ( $\sim 100$  GB/s) and assuming a similar scan pattern over the flash-backed KV index.

### E. Latency Model

The PRISM latency is the sum of the five pipeline stages:

$$t_{\text{PRISM}} = t_{\text{DAC}} + t_{\text{opt}} + t_{\text{PD}} + t_{\text{ADC}} + t_{\text{top-}k}, \quad (13)$$

where the optical propagation time  $t_{\text{opt}}$  includes the modulator response, waveguide transit, and MRR ring-down time.

TABLE V. PRISM latency breakdown.

| Stage        | Latency Notes                 |
|--------------|-------------------------------|
| DAC          | $\sim 1$ ns 4-bit             |
| MZM          | $\sim 0.1$ ns Si depl.        |
| Opt. prop.   | $\sim 0.5$ ns 5 cm            |
| MRR decay    | $\sim 0.1$ ns $Q=10^4$        |
| PD           | $\sim 0.2$ ns Ge              |
| TIA+ADC      | $\sim 2$ ns 6-bit flash       |
| Top- $k$     | $\sim 5$ ns CMOS              |
| <b>Total</b> | <b><math>\sim 9</math> ns</b> |

The total PRISM latency of  $\sim 9$  ns compares favorably with the electronic baselines: GPU full scan  $\sim 5 \mu\text{s}$ , GPU ANN  $\sim 1 \mu\text{s}$ , and NVIDIA ICMS  $\sim 0.5 \mu\text{s}$ —representing a  $\sim 500\times$  speedup over full scan. However, this comparison must account for the additional latency of fetching the selected KV blocks from memory after PRISM selection,

which adds  $0.5 \mu\text{s}$  to  $2 \mu\text{s}$  depending on the memory tier (HBM vs. flash). The net latency benefit of PRISM is therefore most pronounced when the selection ratio  $k/N$  is small and the KV cache resides in a slow memory tier (e.g., flash in ICMS). The crossover analysis quantifying these trade-offs across context lengths and baselines is presented in Sec. VI E.

Figure 9 summarises the interplay between signature dimension, photodetector SNR, and ranking accuracy across the operating envelope of PRISM.

*a. Balanced photodetection noise.* In the balanced configuration, each channel uses two photodetectors measuring through-port and drop-port signals independently [46]. Shot noise from both PDs adds in quadrature:  $\sigma_I^2 = 2e(I_{\text{through}} + I_{\text{drop}})\Delta f$ . Since  $I_{\text{through}} + I_{\text{drop}} = \mathcal{R}P_0$  (power conservation), the total shot noise is weight-independent, simplifying the noise analysis. The factor of  $\sqrt{2}$  increase in noise is offset by the doubled signal dynamic range ( $[-1, +1]$  vs  $[0, 1]$ ).

## V. SYSTEM-LEVEL EVALUATION

This section evaluates the complete PRISM pipeline from algorithmic profiling through end-to-end validation. We first profile retrieval heads across two LLM families (Sec. V A), then evaluate block signature design and recall (Sec. V B), and validate downstream accuracy via Needle-in-a-Haystack experiments with MRR-simulated block selection (Sec. V C).

### A. Retrieval-Head Analysis

*a. Models and datasets.* We profile two representative open-weight LLMs: Qwen2.5-7B-Instruct [52] ( $L = 28$ ,  $H = 28$ ,  $d_h = 128$ , GQA [28] with  $H_{\text{KV}} = 4$ ; total 112 KV heads) and Qwen3-8B [53] ( $L = 36$ ,  $H = 32$ ,  $d_h = 128$ , GQA with  $H_{\text{KV}} = 8$ ; total 288 KV heads). Qwen2.5-7B supports context lengths of at least 128 000

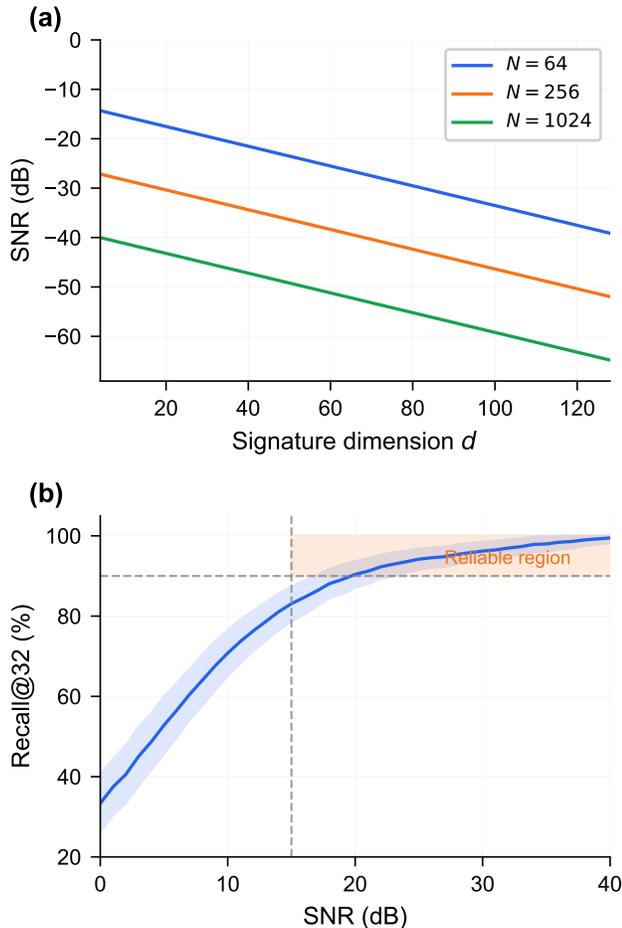


FIG. 9. SNR and recall analysis. (a) Electrical SNR at the photodetector as a function of signature dimension  $d$  for three bank sizes. (b) Recall@8 vs. SNR showing that reliable top- $k$  selection ( $> 90\%$  recall) requires  $\text{SNR} \gtrsim 15$  dB.

tokens; Qwen3-8B supports up to 32000 tokens. We compute retrieval ratios  $R_h^{(\ell,h)}$  (eq. (5)) on a calibration set of 2–3 random token sequences per context length, with  $w = 256$  as the local window size. The retrieval ratio is measured using a two-step procedure: SDPA-based prefill followed by eager last-token attention extraction. All experiments are run on an NVIDIA RTX 5880 (48 GB VRAM) for bf16 models, and an NVIDIA RTX 5070 (12 GB) for 4-bit quantized variants. We additionally verify consistency between bf16 and 4-bit quantized Qwen2.5-7B, finding that quantization does not substantially alter retrieval head identification (e.g., 91.1% vs. 92.0% at 8K context for bf16 and 4-bit, respectively).

*b. Results.* Figure 10 shows the retrieval ratio heatmap across all layers and heads for both models.

Table VI summarizes the retrieval head fraction as a function of context length for both models.

Figure 11 visualizes the retrieval head fraction and

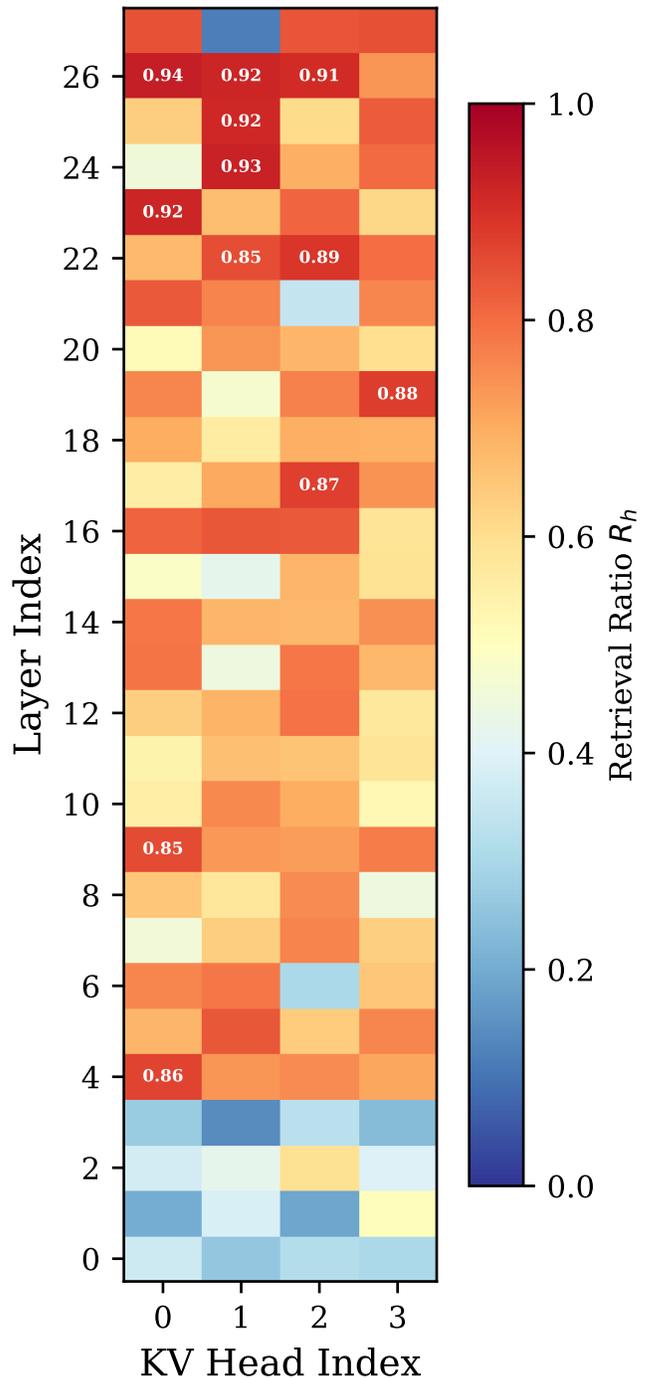


FIG. 10. Retrieval ratio  $R_h^{(\ell,h)}$  for each KV head across all layers. Heads with  $R_h > 0.3$  (dashed line) are classified as retrieval heads. (a) Qwen2.5-7B: 102/112 heads are retrieval heads (91.1%) at 8K context. (b) Qwen3-8B: 258/288 heads are retrieval heads (89.6%) at 8K context.

mean retrieval ratio as a function of context length.

We observe the following patterns:

- **Ubiquity of retrieval behavior.** At a threshold of  $\tau = 0.3$ , 91.1% of KV heads in Qwen2.5-7B and 89.6%

TABLE VI. Retrieval head fraction at threshold  $\tau = 0.3$  across context lengths.  $R_h(\tau)$ : percentage of KV heads with  $R_h > \tau$ . Mean  $\bar{R}_h$ : average retrieval ratio across all heads.

| Model      | Context | $R_h(\tau=0.3)$ (%) | Heads   | Mean $\bar{R}_h$ |
|------------|---------|---------------------|---------|------------------|
| Qwen2.5-7B | 2K      | 83.9                | 94/112  | 0.574            |
|            | 4K      | 83.0                | 93/112  | 0.560            |
|            | 8K      | 91.1                | 102/112 | 0.627            |
|            | 16K     | 92.9                | 104/112 | 0.639            |
|            | 32K     | 95.5                | 107/112 | 0.656            |
|            | 65K     | 92.0                | 103/112 | 0.633            |
|            | 128K    | 98.2                | 110/112 | 0.796            |
| Qwen3-8B   | 2K      | 86.5                | 250/288 | 0.626            |
|            | 4K      | 88.2                | 254/288 | 0.652            |
|            | 8K      | 89.6                | 258/288 | 0.657            |

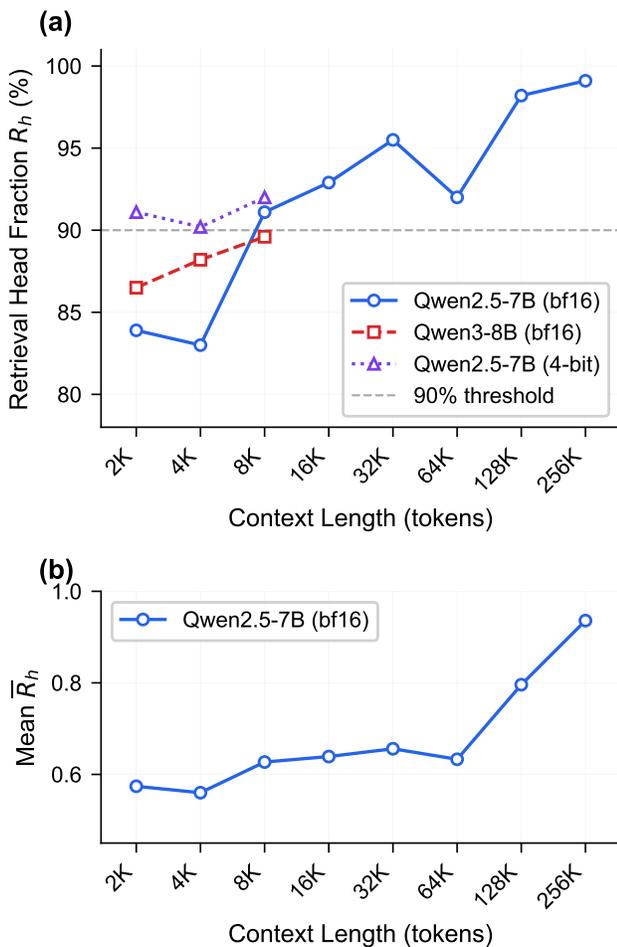


FIG. 11. Retrieval head statistics vs. context length. (a) Retrieval head fraction  $R_h(\tau=0.3)$  for Qwen2.5-7B (bf16 and 4-bit) and Qwen3-8B (bf16). The fraction exceeds 90% for  $n \geq 8K$  and approaches 99% at 256K context. (b) Mean retrieval ratio  $\bar{R}_h$  for Qwen2.5-7B (bf16), showing that individual-head retrieval strength also increases with context length.

in Qwen3-8B are retrieval heads at 8K context. This

prevalence increases with context length: for Qwen2.5-7B, the fraction rises from 83.9% at 2K to 98.2% at 128K context, indicating that nearly all heads engage in long-range retrieval at long contexts. Note that at the more permissive  $\tau = 0.1$  threshold used in [16], essentially 100% of heads qualify as retrieval heads. The reported fraction is thus sensitive to the threshold choice: varying  $\tau$  from 0.1 to 0.3 shifts the classified fraction from  $\sim 100\%$  to  $\sim 90\%$ . The contrast with DuoAttention’s 25–50% retrieval fraction reflects both (i) different models (Llama-2/Mistral vs. Qwen) and (ii) DuoAttention’s use of a learned gating function optimized on calibration data, which imposes a stricter criterion than a simple threshold on attention mass. In practice, the threshold can be tuned per deployment scenario to trade off between the number of heads served photonically and the complexity of the photonic accelerator.

- **Layer distribution.** The highest-scoring retrieval heads are concentrated in layers 14–26, with peak retrieval ratios exceeding 0.93.
- **GQA effect.** Because GQA shares KV heads across multiple query heads, the number of *KV cache entries* requiring retrieval-style treatment is even smaller than the head count suggests.

The key implication for PRISM is that the photonic accelerator needs to serve the vast majority of KV heads—102 out of 112 for Qwen2.5-7B and 258 out of 288 for Qwen3-8B at 8K context. However, GQA sharing means each KV head serves multiple query heads, so the *number of independent weight bank instances* required equals the KV head count, not the query head count.

## B. Block Signature Design

We partition the KV cache into contiguous blocks of  $B$  tokens and compute a  $d$ -dimensional signature for each block [54]. We evaluate mean-key and random projection signature methods from Sec. III B at block size  $B = 128$  and signature dimensions  $d \in \{16, 32, 64, 128\}$ , using Qwen2.5-7B at context length  $n = 4096$ . Our experiments identify  $B = 128$  with  $d = 32$  and mean-key projection as the best configuration. At the primary operating point  $k=32$ , table VII shows  $R@32 = 100\%$  at 8K context ( $B=128$ , 64 blocks), confirming that the signature ranking correctly identifies all relevant blocks. At 16K,  $R@32$  drops to 57.5%, yet downstream NIAH accuracy remains 100% (table IX), indicating that task-critical blocks are consistently ranked in the top- $k$  even when overall recall is imperfect. As a stress-test analysis at  $k=8$ ,  $R@8 = 77.3\%$  ( $R@2 = 31.3\%$ ,  $R@4 = 50.0\%$ ), confirming that useful ranking signal persists even under aggressively small selection budgets. Mean-key projection consistently outperforms random projection across all tested dimensions, confirming that the natural key-

space geometry contains exploitable structure for block ranking.

*a. Why mean-key and random projections?* We focus on mean-key and random projection signatures because they are model-agnostic and require no training, matching our goal of a general-purpose photonic hardware interface. Learned projections (e.g., trained linear maps optimized for recall) could improve signature quality but would require per-model fine-tuning and hardware-aware training, which we leave to future work (Sec. VII C).

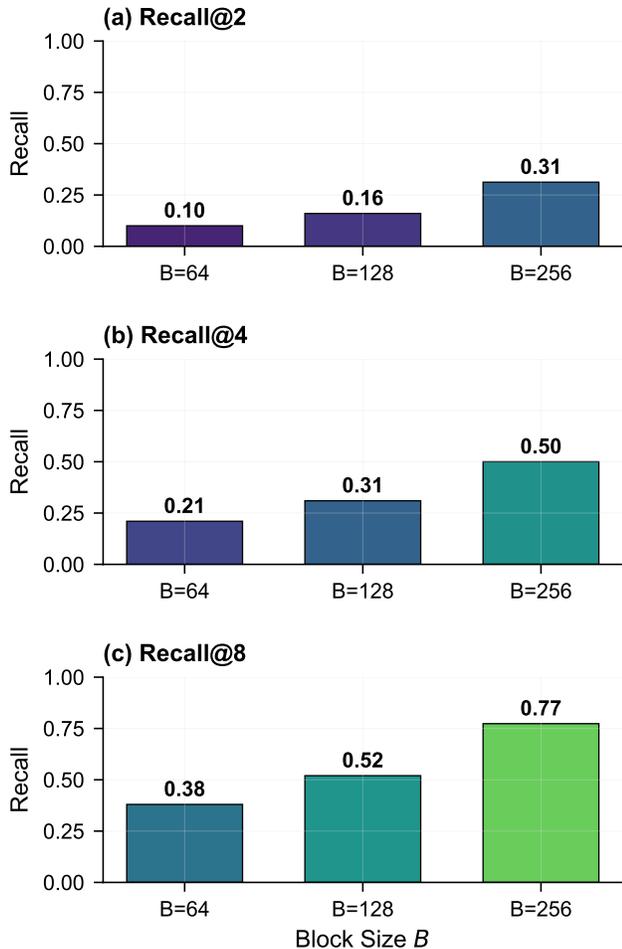


FIG. 12. Recall@ $k$  as a function of signature dimension  $d$  for different signature methods. Block size  $B = 128$ ,  $k = 8$  (stress-test setting). Mean-key projection consistently outperforms random projection, achieving 77.3% recall@8 at  $d = 32$ . At the primary operating point  $k=32$ , recall reaches 100% at 8K context (table VII).

*b. Signed weight encoding.* The add-drop MRR configuration with balanced photodetection enables direct encoding of signed weights  $w \in [-1, +1]$ , eliminating the non-negative constraint of through-port-only architectures. Compared to ReLU projection (which discards sign information, losing  $\sim 50\%$  of the signature variance

for zero-mean Gaussian signatures), balanced photodetection preserves the full signed inner product. Our simulations show that signed encoding improves Recall@8 by  $\sim 87\%$  relative to ReLU projection at  $d = 32$  (Supplementary fig. S5).

*c. Recall metric.* We define recall@ $k$  [55] as the fraction of the true top- $k$  blocks (by exact query-key inner product) that appear in the PRISM-selected top- $k$  blocks:

$$\text{Recall@}k = \frac{|\mathcal{S}_{\text{PRISM}} \cap \mathcal{S}_{\text{exact}}|}{k}, \quad (14)$$

where  $\mathcal{S}_{\text{PRISM}}$  and  $\mathcal{S}_{\text{exact}}$  are the sets of top- $k$  block indices selected by PRISM and by exact computation, respectively.

TABLE VII. Recall@ $k$  for PRISM block selection across context lengths. Qwen2.5-7B,  $B=128$ ,  $d=32$ , mean-key projection. Values averaged over 15 (layer, head) pairs.

| $n$ | Blocks | R@8 (%) | R@16 (%)            | R@32 (%) | NIAH (%) |
|-----|--------|---------|---------------------|----------|----------|
| 4K  | 16     | 46.7    | 100                 | —*       | 100      |
| 8K  | 32     | 29.2    | 55.8                | 100      | 100      |
| 16K | 64     | 26.7    | 41.7                | 57.5     | 100      |
| 32K | 128    |         | (OOM <sup>†</sup> ) |          | 100      |
| 64K | 256    |         | —                   |          | 100      |

\*Only 16 blocks at 4K;  $k=32$  exceeds total.

<sup>†</sup>Eager attention OOM at 32K; NIAH uses SDPA (no attention matrix).

*d. Traffic reduction.* At the primary operating point  $k=32$ , the traffic ratio is  $kB/n = 32 \times 128/n$ . At 128K tokens ( $N=1024$  blocks), PRISM selects  $k=32$  of  $N=1024$  blocks, yielding a  $N/k = 1024/32 = 32 \times$  traffic reduction (3.1% traffic). At 1M tokens ( $N \approx 7812$  blocks), the reduction grows to  $N/k \approx 7812/32 \approx 244 \times$  (0.41% traffic), though model accuracy at such lengths remains model-dependent. Under the stress-test setting  $k=8$ , the reduction reaches  $128 \times$  at 128K and projects to  $\sim 977 \times$  at 1M tokens (see Supplementary fig. S6).

### C. NIAH Accuracy Under Hardware Impairments

To validate that the MRR-impaired block selection preserves end-to-end language model performance, we integrate the MRR array simulator into Qwen2.5-7B [52] and evaluate on the Needle-in-a-Haystack (NIAH) benchmark [15, 56].

For each decode step, block signatures (mean-key,  $d = 32$ ) are processed through the MRR simulator to select the top- $k$  blocks. Retrieval heads ( $R_h > 0.3$ ; table VI) use MRR-selected blocks plus a 256-token recent window; streaming heads retain full attention. We test four MRR configurations: (i) ideal (floating-point inner product), (ii) 5-bit/20 pm drift (nominal), (iii) 4-bit/30 pm drift (pessimistic), and (iv) 5-bit/10 pm drift (optimistic).

Table VIII shows that all four MRR configurations—including the worst-case 4-bit quantization with 30 pm

TABLE VIII. NIAH accuracy (%) with MRR-integrated block selection (Qwen2.5-7B, 11 positions,  $k = 8$  stress-test setting).

| Configuration  | 2K   | 4K    | 8K    |
|----------------|------|-------|-------|
| Full attention | 90.9 | 100.0 | 100.0 |
| Ideal select   | 90.9 | 100.0 | 100.0 |
| 5-bit, 20 pm   | 90.9 | 100.0 | 100.0 |
| 4-bit, 30 pm   | 90.9 | 100.0 | 100.0 |
| 5-bit, 10 pm   | 90.9 | 100.0 | 100.0 |

thermal drift—achieve *identical* NIAH accuracy to full attention at all tested context lengths. The single miss at 2K context (position 50%) is a model-level artifact unrelated to block selection. These results demonstrate that the MRR impairments modelled in Sec. IV A do not degrade downstream task accuracy for the block-selection ranking task.

To validate PRISM across a wide range of context lengths, we extend the evaluation using SDPA-based attention (Flash Attention) with KV cache offloading to CPU RAM via `OffloadedCache`. This enables experiments at context lengths from 4K to 128K on a single GPU (RTX 5880, 48 GB) with 128 GB system RAM. We note that Qwen2.5-7B’s native context window is 128K tokens; at 128K, the base model’s own accuracy degrades to 45.5% on NIAH (table IX), limiting meaningful evaluation beyond 64K. Extrapolation to longer contexts (e.g., 1M tokens via YaRN [6] rope scaling) is technically feasible for the photonic hardware, but model-level accuracy at such lengths remains an open challenge independent of block selection.

For sparse evaluation, we employ *physical token selection*: rather than re-attending to all tokens with a mask, only the tokens from the top- $k$  selected blocks and a recent window are assembled into a compact input ( $\sim 5$ K tokens), preserving positional encoding via explicit `position_ids`. This approach mirrors the actual deployment scenario where only selected KV blocks are fetched from memory.

Table IX presents the extended NIAH results across context lengths from 4K to 128K. The full 2D NIAH heatmap (context length  $\times$  needle depth, 10 positions per context) is shown in fig. 13. At  $k=32$  blocks ( $B=128$ ), all MRR configurations achieve **100% accuracy from 4K through 64K**, perfectly matching full attention. At 128K, the base model itself degrades to 45.5%—a known limitation of Qwen2.5-7B’s context window—making sparse-vs-full comparison uninformative at this length. Within the model’s reliable operating range ( $N \leq 64$ K), MRR block selection introduces *zero* accuracy penalty while reducing KV memory traffic by  $16\times$  at 64K ( $k \cdot B/n = 32 \times 128/65536 = 6.25\%$ ); the reduction grows to  $32\times$  (3.1%) at 128K.

TABLE IX. Extended NIAH accuracy (%) with MRR block selection (Qwen2.5-7B,  $B=128$ ,  $d=32$ ,  $k=32$ , 10 positions per context). At 128K the base model degrades ( $\dagger$ ).

| Configuration    | 4K  | 8K  | 16K | 32K | 64K | 128K <sup>†</sup> |
|------------------|-----|-----|-----|-----|-----|-------------------|
| Full attention   | 100 | 100 | 100 | 100 | 100 | 45.5              |
| Ideal ( $k=32$ ) | 100 | 100 | 100 | 100 | 100 | 18.2 <sup>‡</sup> |
| 5-bit, 20 pm     | 100 | 100 | 100 | 100 | 100 | 27.3 <sup>‡</sup> |
| 4-bit, 30 pm     | 100 | 100 | 100 | 100 | 100 | 27.3 <sup>‡</sup> |
| 5-bit, 10 pm     | 100 | 100 | 100 | 100 | 100 | 27.3 <sup>‡</sup> |

<sup>†</sup>At 128K, full attention itself degrades to 45.5%; the apparent superiority of impaired configurations over ideal is within the  $\pm 9.1\%$  sampling noise of 11 needle positions and is not statistically significant.

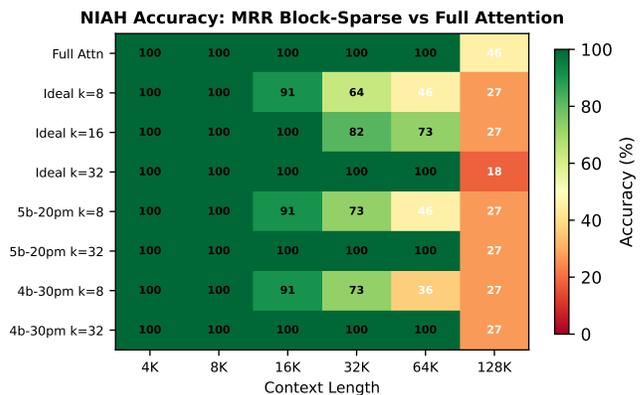


FIG. 13. NIAH accuracy heatmap across context lengths (4K–128K) and MRR configurations. At  $k=32$ , all MRR variants match full attention perfectly up to 64K. Under the  $k=8$  stress test, accuracy degrades gracefully with context length but remains above 90% at 16K. The 128K column shows model-intrinsic degradation (full attention itself drops to 45.5%).

## VI. PHOTONIC SCALING ANALYSIS

We now analyze how the photonic engine scales to larger systems, identifying constraints from WDM channel density, thermal power, chip area, and time-multiplexed operation.

### A. MRR Integration Scaling

The total MRR count in the PRISM weight bank is

$$N_{\text{MRR}} = d \times N, \quad (15)$$

where  $d$  is the number of WDM wavelength channels (signature dimension) and  $N$  is the number of parallel signature banks (one per KV cache block). For a context length of  $n$  tokens with block size  $B$ ,  $N = n/B$ .

Table X lists representative configurations spanning three orders of magnitude in MRR count.

TABLE X. MRR count for representative PRISM configurations. The rightmost column indicates the approximate context length supported at block size  $B = 128$ .

| $d$ | $N$  | $N_{\text{MRR}}$ | Context | Feasibility  |
|-----|------|------------------|---------|--------------|
| 32  | 256  | 8192             | 32K     | Current TFLN |
| 64  | 1024 | 65 536           | 128K    | Near-term    |
| 128 | 4096 | 524 288          | 512K    | Multi-chip   |

Current photonic integration supports  $10^4$ – $10^5$  active devices per die [35, 57], placing the  $d=32$ ,  $N=256$  configuration within demonstrated capability and  $d=64$ ,  $N=1024$  at the near-term frontier. The  $d=128$ ,  $N=4096$  configuration exceeds single-chip density, requiring chiplet-based multi-chip modules [58] (Sec. VIC).

### B. Thermal Power Budget and WDM Channel Limits

On thermo-optic SOI platforms [49], each MRR requires  $\sim 2.5$  mW of static heater power, yielding aggregate budgets of 20 W ( $d=32$ ,  $N=256$ ) to 164 W ( $d=64$ ,  $N=1024$ )—approaching the  $\sim 200$  W practical limit with active cooling. On the TFLN platform, MRR tuning via the Pockels effect ( $r_{33} = 30.9$  pm/V) is capacitive with near-zero static power ( $< 1$   $\mu$ W per ring from CMOS driver leakage):

$$P_{\text{static}}^{\text{TFLN}} = d \times N \times P_{\text{leakage}} < d \times N \times 1 \mu\text{W}. \quad (16)$$

For the  $d=64$ ,  $N=1024$  configuration,  $P_{\text{static}} < 0.07$  W—a  $\sim 2400\times$  reduction over SOI. The switching energy ( $\sim 5$  fJ per ring) yields  $< 0.3$   $\mu$ W total switching power at typical decode rates—negligible. Residual thermal stabilization via TEC ( $\sim 1$  W for a  $\sim 1$  cm<sup>2</sup> chip) remains necessary but is orders of magnitude below SOI heater budgets. TFLN’s lower thermo-optic coefficient ( $dn/dT \approx 4 \times 10^{-5}$  K<sup>-1</sup> vs.  $1.8 \times 10^{-4}$  K<sup>-1</sup> for Si) further reduces thermal crosstalk.

*a. WDM channel limits.* The signature dimension  $d$  is constrained by the MRR free spectral range (FSR) [59]. A single-FSR MRR ( $R = 20$   $\mu$ m,  $\text{FSR} \approx 8.3$  nm) supports only  $\sim 5$  channels at 200 GHz spacing. Vernier-coupled dual-ring filters extend the effective FSR to  $\sim 50$  nm ( $d \sim 30$ ); C+L band operation (95 nm) enables  $d \sim 60$ . Achieving  $d = 128$  requires FSR extension with C+L+S band operation (Supplementary Section S3).

### C. Chip Area Estimation

The  $32 \times 256$  configuration (8192 MRRs) fits on a single  $\sim 5 \times 5$  mm<sup>2</sup> die; the  $64 \times 1024$  configuration requires multi-chip packaging or folded layouts. Detailed area estimates are provided in Supplementary Section S4.

Figure 14 summarizes the scaling trend.

### D. Time-Multiplexed Operation

Area and power constraints can be relaxed by trading physical parallelism for temporal reuse via time-multiplexed weight programming [60]. The system deploys  $N_{\text{phys}}$  physical rows and cycles through  $M$  weight configurations:

$$N_{\text{logical}} = M \times N_{\text{phys}}, \quad M = \lceil N/N_{\text{phys}} \rceil. \quad (17)$$

On TFLN, EO reprogramming is sub-nanosecond ( $t_{\text{reprogram}} \ll t_{\text{optical}}$ ), so the total latency simplifies to  $t_{\text{total}} \approx M \times t_{\text{optical}}$ . Even at  $M = 8$ , the total latency (80 ns) remains four orders of magnitude below the GPU full-scan baseline ( $\sim 200$   $\mu$ s)—a fundamental advantage over thermo-optic SOI ( $t_{\text{reprogram}} \sim 10$   $\mu$ s).

For LLM decode at 128K+ context,  $M = 4$ – $8$  is a practical sweet spot: it reduces physical MRR count by 4– $8\times$  (to 8192–16 384), keeps chip area within a single reticle, and resolves the area scaling barrier of Sec. VIC, making  $d = 64$  realizable with current TFLN technology (fig. 15).

### E. Energy and Latency Crossover

We define the crossover point  $n^*$  as the context length at which PRISM-assisted decoding cost equals the electronic baseline. The PRISM cost (photonic selection energy  $\sim 931$  pJ per query plus reduced GPU fetch) is compared against the GPU full-scan cost (fetching all  $N$  blocks via HBM). On TFLN, near-zero static power means the selection cost is dominated by dynamic components. The full derivation is in Supplementary Section S2.

*a. Energy crossover.* Against the GPU full scan (fig. 16a), the mathematical crossover occurs at  $n^* < 1$  K tokens ( $d = 64$ ,  $N_{\text{bank}} = 4$ ); practical benefit emerges at  $n \geq 4$  K where traffic reduction exceeds  $8\times$ . The per-query dynamic energy ( $\sim 931$  pJ;  $\sim 9.9$  nJ with amortized TEC) is five orders of magnitude below the H100 fetch energy at 128K context ( $\sim 48$   $\mu$ J). This GPU baseline assumes a full-dimension scan ( $d_h=128$ ). A fairer comparison lets the GPU scan compressed  $d=32$  signatures, reducing scan energy to  $E_{\text{scan}}^{\text{fair}} \approx 12$   $\mu$ J. Even under this fairer comparison where the GPU scans compressed  $d=32$  signatures ( $\sim 12$   $\mu$ J), PRISM maintains a four-order-of-magnitude advantage ( $\sim 931$  pJ vs.  $\sim 12$   $\mu$ J), preserving a comfortable crossover margin. On thermo-optic SOI, the  $\sim 164$  W heater power would place the crossover at  $n^* \approx 4$  K. Against GPU ANN (FAISS IVF-PQ) [61] ( $O(\sqrt{N})$  scan reduction), the crossover is at  $n^* \approx 2$  K. Against NVIDIA ICMS (DPU with lower bandwidth than GPU HBM),  $n^* \approx 4$  K based on estimated GTC 2024 specifications.

*b. Latency crossover.* The  $\sim 9$  ns photonic evaluation is orders of magnitude below the  $\sim 5$   $\mu$ s GPU scan, so the selection step is effectively free in latency terms ( $n^* \lesssim 4$  K tokens).

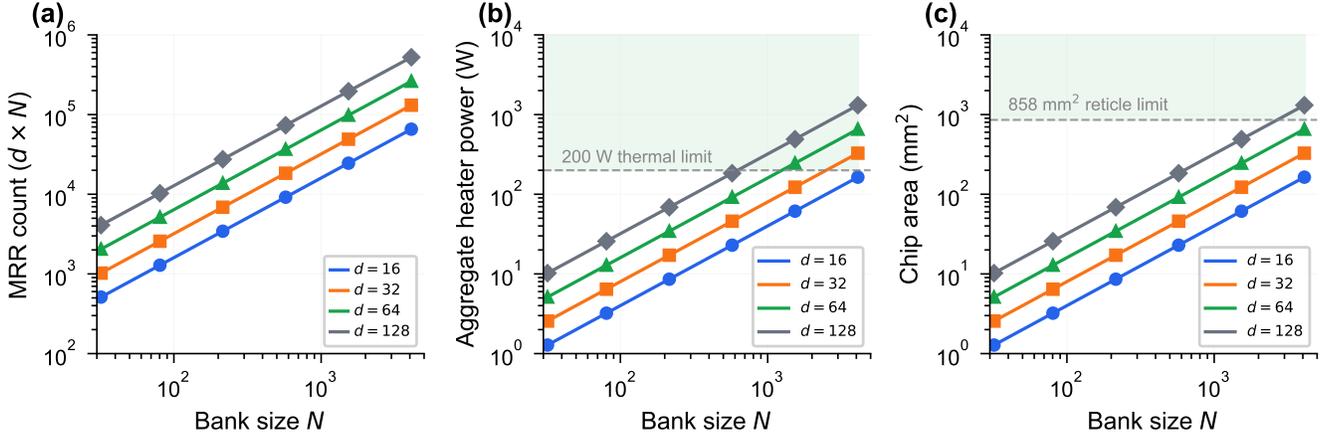


FIG. 14. PRISM photonic scaling projections. MRR count, aggregate heater power (SOI), and estimated chip area as functions of the configuration parameters  $d$  and  $N$ . The dashed horizontal lines indicate practical limits: 200 W thermal dissipation (active cooling) and 858 mm<sup>2</sup> single-reticle area. Configurations below both limits (shaded region) are realizable on a single photonic chip.

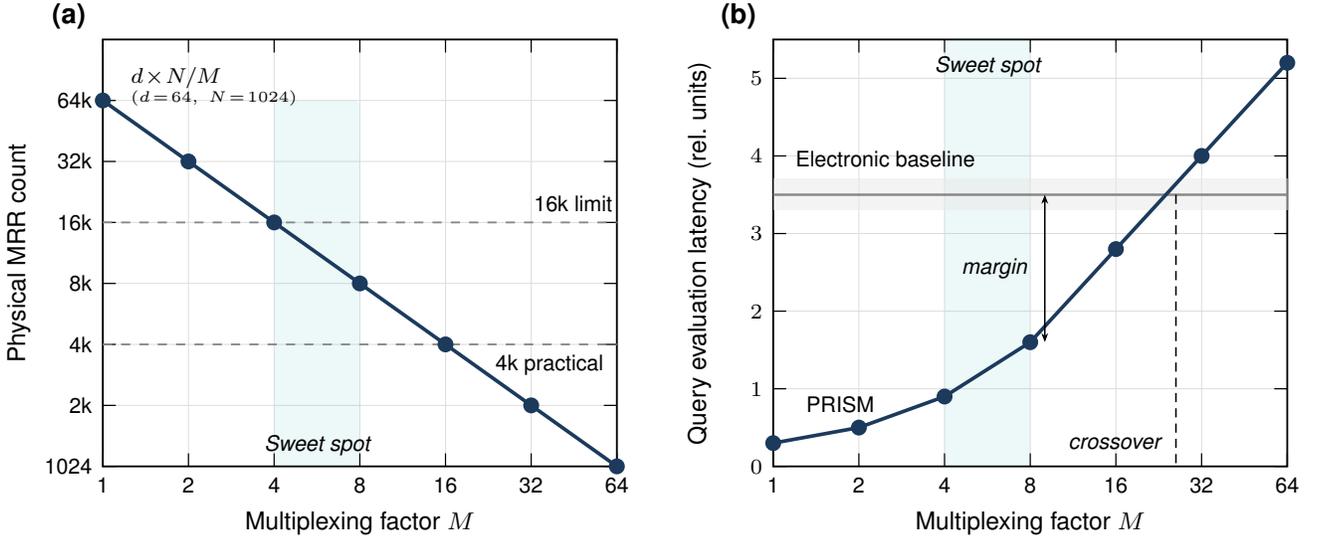


FIG. 15. Time-multiplexed PRISM operation. (a) Physical MRR count vs. multiplexing factor  $M$  ( $d = 64$ ,  $N = 1024$ ). (b) Query evaluation latency vs.  $M$ , compared to electronic baselines. The shaded region indicates the sweet spot ( $M = 4$ – $8$ ).

*c. Sensitivity.* The dominant factors are: (i) dynamic power ( $\sim 88$  mW, dominated by TIAs/ADCs/DACs, negligible vs. electronic baselines); (ii) signature dimension  $d$  (controls MRR count and area, traded against recall quality, Sec. VB); (iii) bank count  $N_{\text{bank}}$  (splitting loss vs. parallelism); and (iv) HBM bandwidth (HBM4 improvements shift the crossover to longer contexts).

*d. Scaling projections.* The energy ratio  $C_{\text{PRISM}}/C_{\text{GPU}}$  decreases as  $\sim 1/n$  because electronic scan cost grows linearly while PRISM selection cost is fixed. At  $n = 1000000$  ( $N \approx 8000$  blocks), the GPU

reads  $\sim 1$  MB of signatures per head per query; PRISM accommodates this with  $N_{\text{bank}} = 8$  banks (512 000 MRRs total). In multi-agent scenarios, a single weight bank serves  $A$  agents simultaneously (only the query sketch changes), amortizing dynamic power by  $1/A$ .

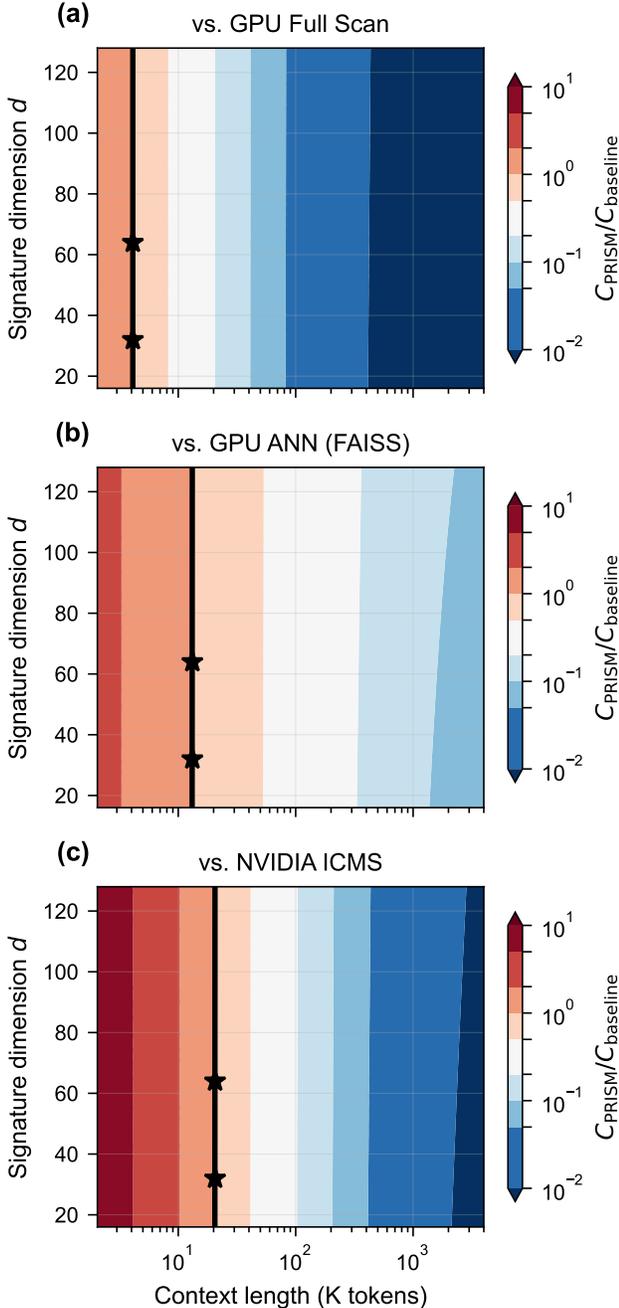


FIG. 16. Energy crossover map for PRISM vs. electronic baselines. The crossover contour ( $C_{\text{PRISM}}/C_{\text{baseline}} = 1$ ) shifts to shorter context lengths as  $d$  decreases. (a) vs. GPU full scan: practical benefit at  $n^* \approx 4\text{K}$ . (b) vs. GPU ANN:  $n^* \approx 2\text{K}$ . (c) vs. NVIDIA ICMS:  $n^* \approx 4\text{K}$ .

## VII. DISCUSSION

### A. Limitations and Practical Considerations

All hardware results in this work are based on device-level simulations with parameters extracted from FDTD

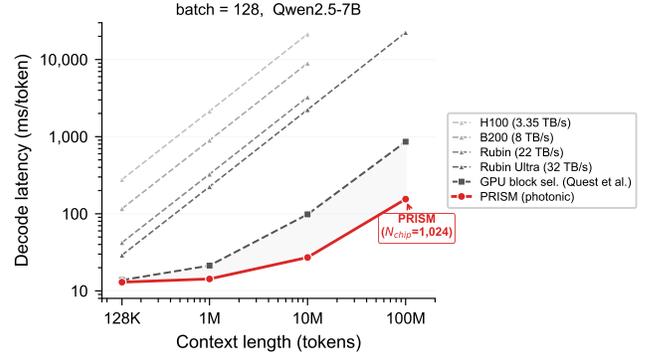


FIG. 17. Energy advantage vs. context length. The ratio decreases as  $\sim 1/n$ ; at 1 000 000 tokens ( $B = 128$ ,  $k=32$ ), the traffic reduction reaches  $244\times$ .

and supplemented by literature values; no physical prototype has been fabricated or measured. The impairment models, while grounded in FDTD simulation and published device data, may not capture all fabrication-dependent effects such as waveguide roughness variations, EO electrode non-uniformity across a large array, and packaging-induced stress. At  $d = 64$  and  $N = 1024$ , the system requires 65 536 MRRs; systematic characterization of  $>10\,000$ -MRR arrays on TFLN has not been reported, though recent progress in TFLN foundry processes suggests that large-scale integration is feasible [44]. Fabrication non-uniformity causes resonance wavelength variations of  $\sigma_\lambda \sim 0.5\text{ nm}$  to  $2\text{ nm}$  across a wafer [10], but on the TFLN platform EO tuning can compensate via DC bias adjustment without static power penalty. Residual thermal drift, while mitigated by lithium niobate's  $\sim 4\times$  lower thermo-optic coefficient compared to silicon, still requires chip-level thermal stabilization ( $\sim 1\text{ W}$  TEC budget). Including the TEC thermal stabilization overhead (amortized at  $>1000$  queries/s throughput), the per-query energy rises from the  $\sim 931\text{ pJ}$  dynamic-only figure to  $\sim 9.9\text{ nJ}$ ; the  $931\text{ pJ}$  value cited elsewhere in this paper refers to the dynamic photonic pipeline alone.

*a. Interface latency.* The  $\sim 9\text{ ns}$  latency reported for PRISM reflects the photonic pipeline alone (DAC through top- $k$  selection) and does not include the host interface overhead. A PCIe 5.0 round-trip (DMA setup and transfer) adds  $\sim 1\text{ }\mu\text{s}$  to  $2\text{ }\mu\text{s}$ ; CXL-attached memory semantics reduce this to  $\sim 200\text{ ns}$  to  $500\text{ ns}$ ; direct interposer or co-packaged integration would add only  $\sim 10\text{ ns}$  to  $50\text{ ns}$ . Even with PCIe overhead, the total system latency of  $\sim 2\text{ }\mu\text{s}$  remains below the GPU full-scan latency ( $\sim 5\text{ }\mu\text{s}$ ), yielding a system-level 2–3 $\times$  speedup. Co-packaging—the long-term integration target—would preserve the  $\sim 100\times$  raw photonic advantage. Speedup claims should therefore be interpreted as system-level 2–3 $\times$  with PCIe, potentially  $100\times$  with co-packaging.

*b. Demonstrated vs. projected scale.* To clarify the maturity of the MRR integration scales assumed in this work: demonstrated TFLN arrays have reached  $\sim 10$ –

100 MRRs [44], while SOI platforms have demonstrated  $\sim 1000$ – $10\,000$  MRRs [35]. PRISM’s “current” configuration (8192 MRRs at  $d=32$ ,  $N=256$ ) is a *projected* design point that extrapolates from these demonstrations; the flagship configuration (65 536 MRRs at  $d=64$ ,  $N=1024$ ) is also projected and would likely require multi-chip or wafer-scale integration.

At  $d = 64$  and  $N = 1024$ , the system requires 65 536 individually addressable voltage bias lines for fabrication—offset compensation of each MRR, presenting a significant packaging and routing challenge that will require advanced fan-out or interposer-based solutions.

The add-drop MRR configuration with balanced photodetection resolves the sign limitation of through-port-only architectures. The balanced differential photocurrent  $I_{\text{through}} - I_{\text{drop}}$  naturally encodes signed weights in  $[-1, +1]$ , enabling true signed inner products without ReLU projection or split encoding. The trade-off is a doubling of the photodetector count (two PDs per channel), but since PDs are orders of magnitude smaller than MRRs, the area penalty is negligible.

The retrieval head classification threshold  $\tau = 0.3$  used throughout this work becomes less discriminative at longer contexts, where most heads tend to exhibit high retrieval scores; DuoAttention’s learned gating identifies only 25–50% of heads as retrieval heads. The 90%+ fraction reported here should therefore be interpreted as an upper bound estimate at the evaluated context lengths.

For multi-head serving, GQA [28] reduces the number of independent weight bank instances from the retrieval head count (102 for Qwen2.5-7B) to the KV head count ( $H_{\text{KV}} = 4$ ), since block signatures are derived from key vectors at KV-head granularity. These 4 heads can be served by time-multiplexed reprogramming ( $\sim 4$  ns on TFLN, negligible vs. the  $\sim 5$   $\mu$ s KV fetch) or by parallel replication of 4 weight banks. When the layer dimension is included, the full configuration space is  $H_{\text{KV}} \times L = 4 \times 28 = 112$  weight bank instances per decode step. Under time-multiplexing, this amounts to  $112 \times \sim 1$  ns  $\approx 112$  ns total reprogramming overhead—still  $\sim 45\times$  smaller than a single KV block fetch ( $\sim 5$   $\mu$ s) and therefore negligible in the decode-step budget. Alternatively, a layer-parallel deployment with 28 PRISM chips (one per layer, each serving 4 KV heads) would eliminate the layer serialization entirely at the cost of additional chip area.

## B. Comparison with Related Approaches

The block-level top- $k$  selection mechanism at the core of PRISM builds on a strategy independently validated by several works: Quest [18] preserves over 99% of full-attention accuracy on long-context benchmarks including NIAH up to 1M tokens, DuoAttention [16] maintains LongBench performance within 1–2% of full attention, and InfLLM [20] and RocketKV [19] provide additional evidence for block-level selection at long con-

text. PRISM’s contribution is orthogonal: the key question is not whether block selection preserves quality (answered affirmatively above) but whether MRR-based analog computation introduces sufficient error to degrade the selection. Our NIAH results (Sec. V C) confirm that it does not, even under pessimistic hardware impairments.

Tian et al.’s photonic transformer chip (PTC) [13] demonstrates that coherent optical interference can implement full transformer attention with high throughput ( $>200$  POPS); however, it targets dense attention computation rather than the coarse block-selection task addressed by PRISM, and its  $O(n)$  memory access scaling remains for long-context KV caches.

InfLLM is the most directly comparable system, as it offloads the full KV cache to CPU RAM and retrieves blocks via electronic inner products. The key distinction is selection latency scaling: InfLLM’s selection time grows as  $O(N)$  with the number of cached blocks, while PRISM’s photonic engine evaluates all  $N$  similarities in  $O(1)$  optical transit time. This advantage grows with context length—precisely the regime where the KV cache bottleneck is most severe.

Relative to Quest [18] and RocketKV [19], which perform block selection *digitally* on the GPU, PRISM targets a different bottleneck: these methods reduce *compute* by pruning low-scoring KV blocks but still require the GPU to read all block signatures from HBM (costing  $O(N)$  memory traffic per decode step). PRISM eliminates this signature scan entirely by offloading it to a photonic coprocessor with  $O(1)$  latency and near-zero energy, making it complementary—Quest- or RocketKV-style scoring policies could be used to *define* which blocks are selected, while PRISM accelerates the *execution* of that selection. The GPU ANN baseline used in our crossover analysis (FAISS IVF-PQ [61]) represents a well-established but not state-of-the-art GPU search library; more recent GPU-accelerated ANN libraries (e.g., CAGRA, cuVS) may further reduce the electronic baseline latency and energy, narrowing the crossover window.

NVIDIA’s ICMS [8] addresses the complementary *capacity* problem (terabyte-scale flash-backed KV storage with DPU-managed prefetch), while PRISM solves the *selection* problem via photonic parallel inner products. Note that the ICMS energy and bandwidth specifications used in our comparisons are estimated from public announcements; no published measurements are available, and actual performance may differ. A natural integration would place PRISM within or adjacent to the ICMS, combining storage capacity with photonic selection speed. The recently announced NVIDIA Rubin platform [62] further underscores industry momentum toward dedicated KV cache acceleration hardware, complementary to PRISM’s photonic approach.

### C. Outlook

The immediate next step is fabrication of a small-scale TFLN MRR prototype ( $8 \times 8$  weight bank) to validate inner-product accuracy under real device impairments and provide measured values for parameters currently extracted from simulation. Scaling to a full module ( $d = 64$ ,  $N = 256$ ) integrated with GPU-based LLM inference would validate the crossover predictions of Sec. VI E. Integrating non-volatile weight storage (e.g., phase-change trimming [63, 64]) could further reduce switching energy for quasi-static block signatures [36, 65]. More challenging benchmarks such as SCBench [66] and query-focused retrieval analysis [67] would strengthen confidence in the robustness of photonicallly selected blocks beyond the NIAH validation presented here.

*a. Practical integration.* A deployable PRISM module would package the photonic chip, laser source, and TEC onto a single substrate, offered in one of three form factors: a PCIe add-in card for drop-in datacenter use, a CXL-attached device for lower-latency memory-semantic access, or a co-packaged chiplet on an interposer for maximum performance. Integration with existing LLM serving stacks (e.g., vLLM, TensorRT-LLM) would proceed via a block-index API: the host submits a query sketch and receives ranked block indices, transparently replacing the software signature-scan kernel.

*b. Benchmark scope.* NIAH is a retrieval-oriented benchmark that tests single-needle recall; it does not exercise multi-hop reasoning, summarization, or other long-context capabilities. Our results therefore validate retrieval fidelity but not general long-context quality. We note, however, that the block selection mechanism is inherited from Quest [18] and InfLLM [20], which have been validated on broader benchmarks (LongBench,  $\infty$ Bench); PRISM’s contribution is the photonic hardware mapping of this selection, not the selection algorithm itself.

## VIII. CONCLUSION

We have presented PRISM, a TFLN photonic similarity engine that computes all  $N$  block-selection inner products in  $O(1)$  optical latency via the broadcast-and-weight paradigm. End-to-end NIAH evaluation confirms that MRR-selected block-sparse attention preserves full-attention accuracy from 4K to 64K tokens (within the model’s native context window) under realistic hardware impairments (4–5 bit weights, 30 pm thermal drift), while reducing KV cache traffic by  $16\times$  at 64K context ( $k=32$ ,  $B=128$ ;  $32\times$  at 128K). At longer contexts (128K+), model-intrinsic accuracy degrades independent of block selection; the photonic scaling analysis nevertheless projects favorable energy and latency scaling to million-token regimes as model context windows continue to expand. The practical energy benefit emerges at  $n \geq 4K$  where block selection yields meaningful traffic reduction, making PRISM favorable across virtually all

practical context lengths.

Future work will proceed along three axes: (i) fabrication and characterization of an  $8 \times 8$  TFLN MRR weight bank to validate simulation predictions with measured device parameters; (ii) scaling to a full  $d=64$ ,  $N=256$  module integrated with GPU-based inference for end-to-end latency and energy measurements; and (iii) integration of non-volatile weight storage (e.g., phase-change trimming [63, 64]) for write-once signature programming, together with hardware-aware learned projections and broader benchmarks such as SCBench [66]. More broadly, photonic broadcast search may serve as a general paradigm for similarity-search workloads in data centers—including approximate nearest-neighbor retrieval, recommendation ranking, and embedding lookup—wherever a single query must be compared against a large, slowly changing set of stored vectors.

## DISCLOSURES

The authors declare no conflicts of interest.

## DATA AVAILABILITY

Code and simulation data are available at <https://github.com/hyoseokp/PRISM> [68].

## SUPPLEMENTARY INFORMATION

### Appendix S1: Device Impairment Models

This section provides the full mathematical models for the six impairment sources that degrade the ideal inner-product computation of eq. (6).

*a. Weight quantization.* MRR transmission is programmed via electro-optic tuning with finite precision. We model the quantized weight as

$$\hat{w}_{n,j} = \frac{\text{round}(w_{n,j} \cdot 2^b)}{2^b}, \quad (\text{S1})$$

where  $b$  is the effective bit precision. Values of  $b = 4\text{--}8$  are considered, corresponding to 16–256 distinguishable transmission levels.

*b. Thermal drift.* After initial calibration, the MRR resonance wavelength drifts due to ambient temperature fluctuations. We model the drift as a Gaussian random walk:

$$\Delta\lambda_0(t) = \sum_{i=1}^{t/\Delta t} \mathcal{N}(0, \sigma_{\text{drift}}^2), \quad (\text{S2})$$

with  $\sigma_{\text{drift}}$  chosen to produce a standard deviation of 0.01 nm to 0.1 nm over a calibration interval  $T_{\text{cal}}$ . The resulting weight error is

$$\Delta w = \left| \frac{\partial T}{\partial \lambda} \right| \Delta\lambda_0 \approx \frac{8Q^2 D_{\text{max}}}{\lambda_0^2} \cdot \Delta\lambda_0, \quad (\text{S3})$$

evaluated at the operating point on the MRR Lorentzian. Note that the i.i.d. Gaussian model above does not capture spatially correlated drift (e.g., center-to-edge temperature gradients across the chip), which could cause systematic bias in the inner-product scores rather than zero-mean random noise; such gradients would require a correlated noise model or per-region calibration.

*c. Insertion loss.* Each MRR introduces an off-resonance insertion loss  $\text{IL}_{\text{MRR}} \approx 0.02\text{ dB}$  to  $0.05\text{ dB}$ , and the  $1 \times N$  splitter contributes  $\text{IL}_{\text{split}}$  from eq. (8). The total channel loss is

$$\text{IL}_{\text{total}} = \text{IL}_{\text{split}} + d \cdot \text{IL}_{\text{MRR}} + \text{IL}_{\text{wg}}, \quad (\text{S4})$$

where  $\text{IL}_{\text{wg}}$  accounts for waveguide propagation loss ( $\sim 0.3\text{ dB/cm}$  for TFLN). High insertion loss reduces the SNR at the photodetector and increases the required laser power.

*d. Photodetector noise.* The photocurrent at each detector includes shot noise and thermal noise:

$$\sigma_I^2 = 2eI_{\text{ph}} \Delta f + \frac{4k_B T}{R_L} \Delta f + \text{NEP}^2 \cdot \Delta f, \quad (\text{S5})$$

where  $I_{\text{ph}}$  is the signal photocurrent,  $\Delta f$  is the detection bandwidth,  $R_L$  is the load resistance, and NEP is the noise-equivalent power of the photodetector ( $\sim 10\text{ pW}/\sqrt{\text{Hz}}$  for Ge-on-Si) [69, 70]. The noise introduces a random perturbation to the inner-product score, potentially reordering the top- $k$  ranking.

*e. MRR crosstalk.* Adjacent MRRs on the same bus waveguide can exhibit spectral overlap if the channel spacing is insufficient relative to the MRR linewidth. We model inter-channel crosstalk as an additive interference with isolation of  $-15\text{ dB}$  to  $-30\text{ dB}$ :

$$y_n = \sum_{j=1}^d w_{n,j} s_j + \sum_{j=1}^d \sum_{m \neq j} \chi_{j,m} w_{n,m} s_m, \quad (\text{S6})$$

where  $\chi_{j,m}$  is the crosstalk coefficient from channel  $m$  to channel  $j$  [71].

*f. Input DAC noise.* The finite resolution and integral nonlinearity (INL) of the input DACs contribute an additional noise floor on the query sketch values. At  $b_{\text{DAC}} = 6$  bits, the quantization noise standard deviation is  $\sigma_{\text{DAC}} = 2^{-b_{\text{DAC}}}/\sqrt{12} \approx 0.0045$ .

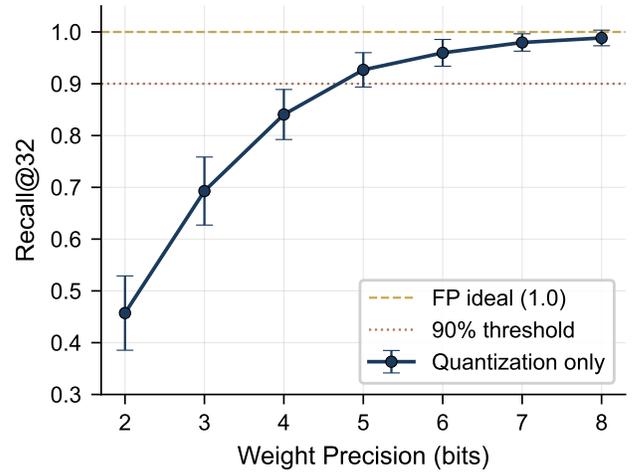


FIG. S1. Impact of weight quantization on recall. At  $b = 6$  bits, recall degrades by less than 5% from the floating-point ideal ( $\text{Recall}@8 = 0.960$  at 6-bit). Adding thermal drift ( $\sigma_{\text{th}} = 0.01$ ) and detector noise ( $\sigma_{\text{det}} = 0.01$ ) degrades recall by an additional 5%.

### Appendix S2: Crossover Derivation

This section provides the full algebraic derivation of the energy crossover point  $n^*$  summarized in Sec. VI E.

The total decode cost per token for a retrieval head consists of two terms:

$$C_{\text{total}} = C_{\text{select}}(n) + C_{\text{fetch}}(n, k), \quad (\text{S1})$$

where  $C_{\text{select}}$  is the cost of determining which blocks to fetch and  $C_{\text{fetch}}$  is the cost of reading and computing attention over the selected blocks.

For the **GPU full scan** baseline, no selection is needed ( $k = N = n/B$ ), so

$$C_{\text{GPU}} = C_{\text{fetch}}^{\text{GPU}}(n, n/B) = \frac{2 d_h n b_{\text{prec}}}{\text{BW}_{\text{HBM}}}, \quad (\text{S2})$$

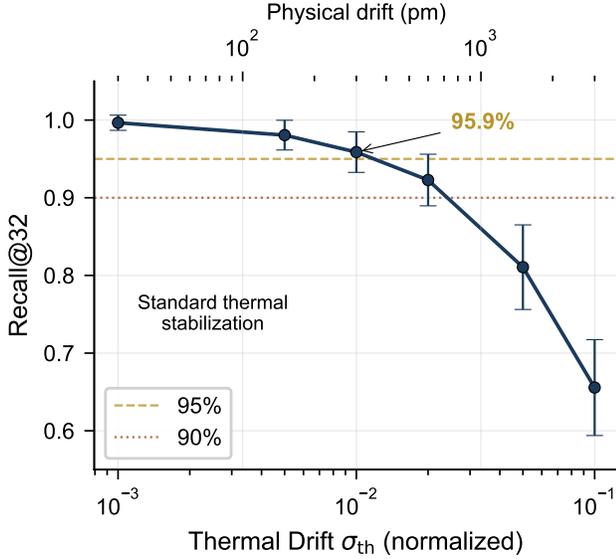


FIG. S2. Recall degradation as a function of thermal drift  $\sigma$ . Recall remains above 95% for  $\sigma \leq 0.005$  (corresponding to  $\sim 150$  pm drift), achievable with standard thermal stabilization. At  $\sigma = 0.01$  ( $\sim 300$  pm), recall is still 94.8%.

where  $BW_{HBM}$  is the HBM bandwidth ( $\sim 3.35$  TB/s for H100 HBM3).

For **Prism**, the cost is

$$C_{PRISM} = C_{select}^{PRISM}(n) + C_{fetch}^{GPU}(n, k \cdot B), \quad (S3)$$

where  $C_{select}^{PRISM}$  is the dynamic energy of the photonic evaluation (laser, DAC, modulator, PD, ADC):

$$C_{select}^{PRISM} \approx E_{dynamic}, \quad (S4)$$

Note that on the TFLN platform, the static MRR tuning power is near zero (capacitive EO), so the selection cost is dominated entirely by the dynamic components ( $\sim 931$  pJ per query, table IV). The fetch cost is reduced by the selection ratio  $k/N$ :

$$C_{fetch}^{GPU}(n, k \cdot B) = \frac{2 d_h k B b_{prec}}{BW_{HBM}}. \quad (S5)$$

The crossover occurs when  $C_{PRISM} < C_{GPU}$ , i.e., when the memory bandwidth saved by not fetching  $(N - k)$  blocks exceeds the cost of operating the photonic selector.

Setting  $C_{PRISM} = C_{GPU}$  and solving for  $n$  yields:

$$n^* = \frac{C_{select}^{PRISM} \cdot BW_{HBM}}{2 d_h b_{prec} (1 - k B/n^*)}, \quad (S6)$$

which must be solved self-consistently since  $k$  and  $N = n/B$  both depend on  $n$ . In practice,  $k$  is a fixed parameter (e.g.,  $k = 32$ ), so the selection ratio  $kB/n \rightarrow 0$  as  $n \rightarrow \infty$ , and the crossover simplifies to

$$n^* \approx \frac{C_{select}^{PRISM} \cdot BW_{HBM}}{2 d_h b_{prec}}. \quad (S7)$$

## Appendix S3: WDM Channel Limits

The maximum number of WDM channels  $d$  is constrained by the MRR free spectral range (FSR), the available optical bandwidth, and inter-channel crosstalk.

*a. Single-FSR constraint.* For a TFLN MRR with radius  $R = 20$   $\mu\text{m}$ , the FSR is approximately 8.3 nm (at  $\lambda_0 = 1550$  nm, as in table I). At a channel spacing of  $\Delta\lambda_{ch} = 1.6$  nm (200 GHz on the ITU grid) [72], the maximum number of non-aliased channels within one FSR is

$$d_{max}^{(1-FSR)} = \left\lfloor \frac{FSR}{\Delta\lambda_{ch}} \right\rfloor = \left\lfloor \frac{8.3}{1.6} \right\rfloor = 5. \quad (S1)$$

This is clearly insufficient for the  $d = 32$ –128 range targeted by PRISM.

*b. FSR extension techniques.* Vernier-coupled dual-ring filters or cascaded MRRs with slightly different radii can extend the effective FSR to  $\sim 50$  nm or more [46, 73], limited by the least common multiple of the individual FSRs. With an extended FSR of 50 nm and  $\Delta\lambda_{ch} = 1.6$  nm:

$$d_{max}^{(Vernier)} = \left\lfloor \frac{50}{1.6} \right\rfloor \approx 30. \quad (S2)$$

*c. Band-limited operation.* The usable optical bandwidth depends on the operating band:

- **C-band** (1530–1565 nm): 35 nm  $\rightarrow$  practical limit  $d \sim 20$ –30 without FSR extension;
- **C+L band** (1530–1625 nm): 95 nm  $\rightarrow d \sim 60$ .

For  $d > 60$ , extending to the S-band or using 0.8 nm (100 GHz) channel spacing is necessary, at the cost of tighter crosstalk margins. In practice, achieving  $d = 128$  requires both FSR extension and C+L+S band operation, representing a more aggressive photonic design point.

### 1. Balanced Photodetection for Signed Weights

The add-drop MRR configuration provides both through-port transmission  $T_{through}(\Delta\lambda)$  and complementary drop-port transmission  $T_{drop}(\Delta\lambda) = 1 - T_{through}(\Delta\lambda)$  (for a lossless ring). A balanced photodetector pair measures the differential signal:

$$w = T_{through} - T_{drop} = 2T_{through} - 1 \in [-1, +1]. \quad (S3)$$

This mapping naturally encodes signed weights without doubling the MRR count (as required by split encoding) or discarding sign information (as in ReLU projection). The noise model for balanced detection yields shot noise variance  $\sigma^2 = 2eRP_0\Delta f$ , independent of the programmed weight, since power is conserved:  $P_{through} + P_{drop} = P_0$  [69].

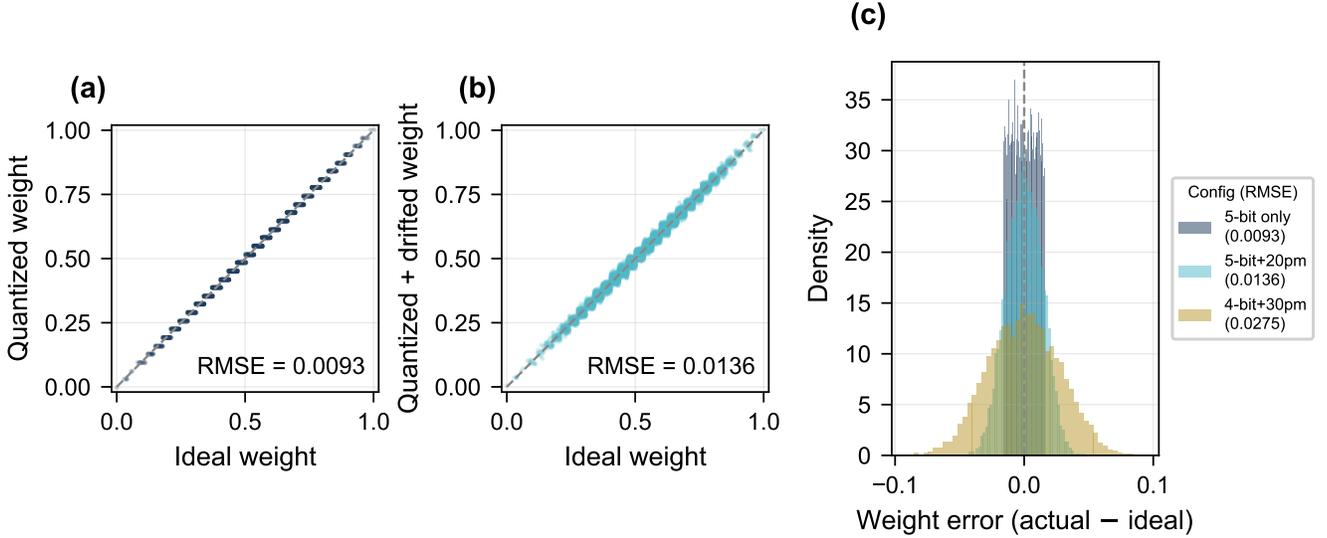


FIG. S3. Weight encoding fidelity under MRR impairments ( $d = 32$ ,  $N = 128$ ). (a) Ideal vs. 5-bit quantised weights in  $[0, 1]$ : the staircase pattern shows 32 discrete levels with  $\text{RMSE} = 0.009$ . (b) Adding 20 pm thermal drift and fabrication variation broadens the scatter ( $\text{RMSE} = 0.014$ ). (c) Error histograms for three configurations: 5-bit only, 5-bit with 20 pm drift, and 4-bit with 30 pm drift. Even the pessimistic case concentrates errors within  $\pm 5\%$  of the full weight range.

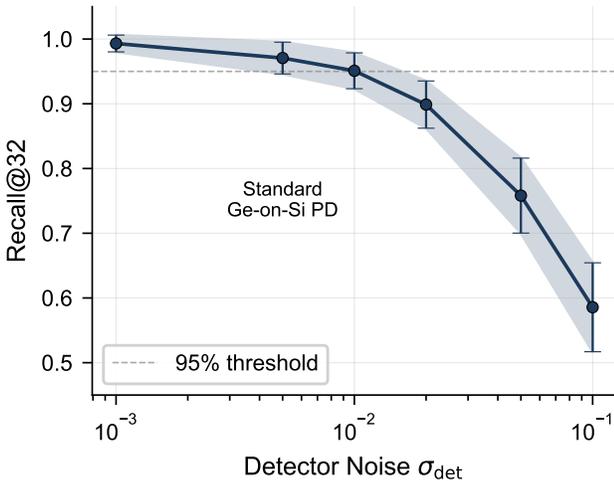


FIG. S4. Recall@8 degradation as a function of photodetector noise  $\sigma_{\text{det}}$  ( $d = 32$ ,  $N = 500$ , 100 trials). Recall remains above the 95% threshold for  $\sigma_{\text{det}} \leq 0.01$ , achievable with standard Ge-on-Si photodetectors ( $\text{NEP} \sim 1 \text{ pW}/\sqrt{\text{Hz}}$ ).

#### Appendix S4: Multi-Head Serving with GQA

A natural concern is whether serving all retrieval heads requires replicating the weight bank for each head. In Qwen2.5-7B, 102 out of 112 KV heads are retrieval heads at 8K context (table VI). However, GQA [28] reduces

the number of *independent* KV heads to  $H_{\text{KV}} = 4$ —each KV head is shared across  $H/H_{\text{KV}} = 7$  query heads. The weight bank stores block *signatures* derived from key vectors, so it operates at the KV-head granularity, not the query-head granularity. This means only 4 independent weight bank configurations are needed per layer, not 102.

Two strategies can serve these 4 KV heads:

1. **Time-multiplexed reuse.** A single weight bank is reprogrammed sequentially for each of the 4 KV heads. On the TFLN platform, EO tuning settles in  $\sim 1$  ns (RC-limited), so reprogramming 4 heads adds only  $\sim 4$  ns—negligible compared to the subsequent KV block fetch ( $\sim 5$   $\mu$ s).
2. **Parallel replication.** Four weight banks are deployed in parallel, one per KV head. This requires  $4 \times d \times N = 4 \times 64 \times 1024 = 262144$  MRRs total—a  $4\times$  increase over the single-head case but within the scalability limits discussed in Sec. VII A.

GQA thus reduces the multi-head serving problem from 102 independent banks (one per retrieval head) to just 4, making both time-multiplexed and spatially-parallel approaches practical. For Qwen3-8B ( $H_{\text{KV}} = 8$ ), the same argument applies with 8 KV heads, still far below the 258 retrieval heads.

#### Appendix S5: Additional Figures

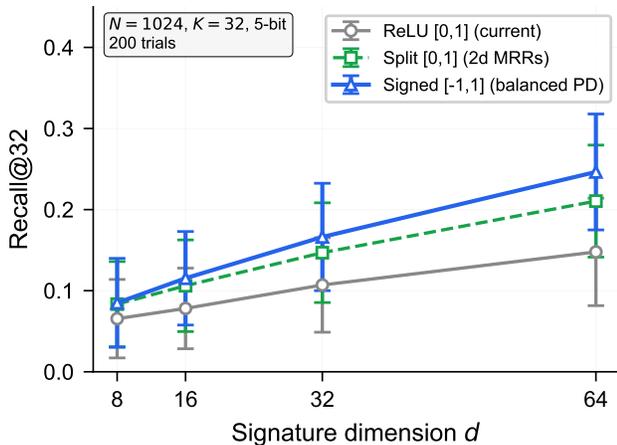


FIG. S5. Signed vs. unsigned recall comparison. Balanced photodetection (signed  $[-1, +1]$ ) consistently outperforms ReLU projection (unsigned  $[0, 1]$ ) and split encoding across all signature dimensions  $d$ .

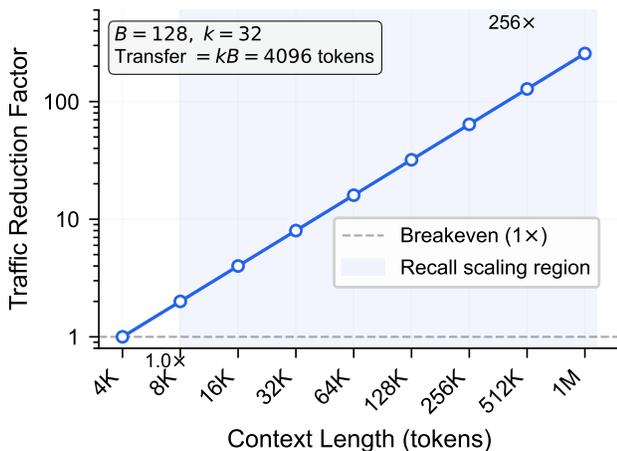


FIG. S6. KV cache traffic reduction factor  $N/k$  as a function of context length for different  $k$  values.

- 
- [1] T. Dao, FlashAttention-2: Faster attention with better parallelism and work partitioning, in *International Conference on Learning Representations* (2024) arXiv:2307.08691 (2023).
- [2] A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer, AI and memory wall, *IEEE Micro* **44**, 33 (2024).
- [3] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, *et al.*, GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
- [4] Gemini Team Google, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, arXiv preprint arXiv:2403.05530 (2024).
- [5] Meta AI, The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
- [6] Qwen Team, Qwen2.5-1M Technical Report, arXiv preprint arXiv:2501.15383 (2025).
- [7] NVIDIA, NVIDIA Vera Rubin Platform, <https://nvidianews.nvidia.com/news/nvidia-vera-rubin> (2025), announced at CES 2025; next-generation GPU architecture with integrated memory connectivity.
- [8] NVIDIA, Inference Context Memory Storage (ICMS): BlueField-4 DPU for LLM Inference, <https://nvidianews.nvidia.com/news/nvidia-vera-rubin> (2025), flash-backed KV cache with hardware-assisted eviction and prefetch.
- [9] A. N. Tait, M. A. Nahmias, B. J. Shastri, T. F. de Lima, and P. R. Prucnal, Broadcast and weight: An integrated

- network for scalable photonic spike processing, *Journal of Lightwave Technology* **32**, 3427 (2014).
- [10] A. N. Tait, A. X. Wu, T. F. de Lima, E. Zhou, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, Microring weight banks, *IEEE Journal of Selected Topics in Quantum Electronics* **22**, 312 (2016).
- [11] S. Hua, E. Divita, S. Yu, B. Peng, C. Roques-Carnes, Z. Su, Z. Chen, Y. Bai, J. Zou, Y. Zhu, Y. Xu, C. Lu, Y. Di, H. Chen, L. Jiang, L. Wang, L. Ou, C. Zhang, J. Chen, W. Zhang, H. Zhu, W. Kuang, L. Wang, H. Meng, M. Steinman, and Y. Shen, An integrated large-scale photonic accelerator with ultralow latency, *Nature* **640**, 361 (2025).
- [12] H. Zhu, J. Gu, H. Wang, Z. Jiang, Z. Zhang, R. Tang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan, Lightning-Transformer: A dynamically-operated optically-interconnected photonic transformer accelerator, in *Proc. IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (2024) pp. 686–703.
- [13] Y. Tian, S. Xiang, X. Guo, Y. Zhang, J. Xu, S. Shi, H. Zhao, Y. Wang, X. Niu, W. Liu, and Y. Hao, Photonic transformer chip: interference is all you need, *Photonix* **6**, 45 (2025).
- [14] T. Fu, J. Zhang, R. Sun, Y. Huang, W. Xu, S. Yang, Z. Zhu, and H. Chen, Optical neural networks: progress and challenges, *Light: Science & Applications* **13**, 10.1038/s41377-024-01590-3 (2024).
- [15] W. Wu, Y. Wang, G. Xiao, H. Peng, and Y. Fu, Retrieval head mechanistically explains long-context factuality, in *International Conference on Learning Representations (ICLR)* (2025) oral presentation.
- [16] G. Xiao, J. Tang, J. Zuo, J. Guo, S. Yang, H. Tang, Y. Fu, and S. Han, DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads, in *International Conference on Learning Representations (ICLR)* (2025).
- [17] H. Tang, Y. Lin, J. Lin, Q. Han, S. Hong, Y. Yao, and G. Wang, RazorAttention: Efficient KV cache compression through retrieval heads, arXiv preprint arXiv:2407.15891 (2024).
- [18] J. Tang, Y. Zhao, K. Zhu, G. Xiao, B. Kasicki, and S. Han, Quest: Query-aware sparsity for efficient long-context LLM inference, in *International Conference on Machine Learning (ICML)* (2024).
- [19] P. Behnam, Y. Fu, R. Zhao, P.-A. Tsai, Z. Yu, and A. Tumanov, RocketKV: Accelerating long-context LLM inference via two-stage KV cache compression, in *International Conference on Machine Learning (ICML)* (2025).
- [20] C. Xiao, P. Zhang, X. Han, G. Xiao, Y. Lin, Z. Zhang, Z. Liu, S. Han, and M. Sun, InfLLM: Training-free long-context extrapolation for LLMs with an efficient context memory, arXiv preprint arXiv:2402.04617 (2024).
- [21] X. Liu, Z. Tang, P. Dong, Z. Li, B. Li, X. Hu, and X. Chu, ChunkKV: Semantic-preserving KV cache compression for efficient long-context LLM inference, in *Advances in Neural Information Processing Systems (NeurIPS)* (2025).
- [22] H. Li, Y. Li, A. Tian, T. Tang, Z. Xu, X. Chen, N. Hu, W. Dong, Q. Li, and L. Chen, A survey on large language model acceleration based on KV cache management, *Transactions on Machine Learning Research* (2025).
- [23] Z. Zhang, Y. Sheng, T. Zhou, T. Chen, L. Zheng, R. Cai, Z. Song, Y. Tian, C. Ré, C. Barrett, Z. Wang, and B. Chen, H<sub>2</sub>O: Heavy-hitter oracle for efficient generative inference of large language models, in *Advances in Neural Information Processing Systems (NeurIPS)* (2023).
- [24] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, Efficient streaming language models with attention sinks, in *International Conference on Learning Representations (ICLR)* (2024).
- [25] C. Hooper, S. Kim, H. Mohammadzadeh, M. W. Mahoney, Y. S. Shao, K. Keutzer, and A. Gholami, KVQuant: Towards 10 million context length LLM inference with KV cache quantization, in *Advances in Neural Information Processing Systems (NeurIPS)* (2024).
- [26] G. Liu, C. Li, Z. Ning, J. Lin, Y. Yao, D. Ke, M. Guo, and J. Zhao, FreeKV: Boosting KV cache retrieval for efficient LLM inference, arXiv preprint arXiv:2505.13109 (2025).
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, Vol. 30 (2017) pp. 5998–6008.
- [28] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, and S. Sanghvi, GQA: Training generalized multi-query transformer models from multi-head checkpoints, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2023) pp. 4895–4901.
- [29] N. Shazeer, Fast transformer decoding: One write-head is all you need, arXiv preprint arXiv:1911.02150 (2019).
- [30] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, Experimental realization of any discrete unitary operator, *Physical Review Letters* **73**, 58 (1994).
- [31] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, Optimal design for universal multiport interferometers, *Optica* **3**, 1460 (2016).
- [32] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, Deep learning with coherent nanophotonic circuits, *Nature Photonics* **11**, 441 (2017).
- [33] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, Micrometre-scale silicon electro-optic modulator, *Nature* **435**, 325 (2005).
- [34] H. Zhou, J. Dong, J. Cheng, W. Dong, C. Huang, Y. Shen, Q. Zhang, M. Gu, C. Qian, H. Chen, Z. Ruan, and X. Zhang, Photonic matrix multiplication lights up photonic accelerator and beyond, *Light: Science & Applications* **11**, 30 (2022).
- [35] C. Huang, S. Bilodeau, T. F. de Lima, A. N. Tait, P. Y. Ma, E. C. Blow, A. Jha, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits, *APL Photonics* **5**, 040803 (2020).
- [36] H. Zhang, Y. Song, S. Chen, Y. Bai, X. Xu, C. Huang, J. Wang, H. Chen, D. J. Moss, and K. Xu, Integrated platforms and techniques for photonic neural networks, *npj Nanophotonics* **2**, 40 (2025).
- [37] W. B. Johnson and J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, *Contemporary Mathematics* **26**, 189 (1984).
- [38] C. Wang, M. Zhang, X. Chen, M. Bertrand, A. Shams-Ansari, S. Chandrasekhar, P. Winzer, and M. Lončar, Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages, *Nature* **562**, 101 (2018).
- [39] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. P.

- Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, Photonics for artificial intelligence and neuromorphic computing, *Nature Photonics* **15**, 102 (2021).
- [40] A. Totovic, G. Giamougiannis, A. Tsakyridis, D. Lazovsky, and N. Pleros, Programmable photonic neural networks combining WDM with coherent linear optics, *Scientific Reports* **12**, 5605 (2022).
- [41] S. Lischke, A. Peczek, J. S. Morgan, K. Sun, D. Steckler, Y. Yamamoto, F. Korndörfer, C. Mai, S. Marschmeyer, M. Fraschke, A. Krüger, A. Beling, and L. Zimmermann, Ultra-fast germanium photodiode with 3-dB bandwidth of 265 GHz, *Nature Photonics* **15**, 925 (2021).
- [42] N. Peserico, B. J. Shastri, and V. J. Sorger, Integrated photonic tensor processing unit for a matrix multiply: A review, *Journal of Lightwave Technology* **41**, 3704 (2023).
- [43] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, Parallel convolutional processing using an integrated photonic tensor core, *Nature* **589**, 52 (2021).
- [44] Y. Hu, Y. Song, X. Zhu, X. Guo, S. Lu, Q. Zhang, L. He, C. A. A. Franken, K. Powell, H. Warner, *et al.*, Integrated lithium niobate photonic computing circuit based on efficient and high-speed electro-optic conversion, *Nature Communications* **16**, 8178 (2025).
- [45] X. Zhu, Y. Hu, S. Lu, H. K. Warner, X. Li, Y. Song, L. S. Magalhães, A. Shams-Ansari, N. Sinclair, and M. Lončar, Twenty-nine million intrinsic Q-factor monolithic microresonators on thin-film lithium niobate, *Photonics Research* **12**, A63 (2024).
- [46] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Neuromorphic photonic networks using silicon photonic weight banks, *Scientific Reports* **7**, 7430 (2017).
- [47] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, K. De Vos, S. Kumar Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. Van Thourhout, and R. Baets, Silicon microring resonators, *Laser & Photonics Reviews* **6**, 47 (2012).
- [48] T. Ferreira de Lima, E. A. Doris, S. Bilodeau, W. Zhang, A. Jha, H.-T. Peng, E. C. Blow, C. Huang, A. N. Tait, B. J. Shastri, and P. R. Prucnal, Design automation of photonic resonator weights, *Nanophotonics* **11**, 3805 (2022).
- [49] K. Padmaraju and K. Bergman, Resolving the thermal challenges for silicon microring resonator devices, *Nanophotonics* **3**, 269 (2014).
- [50] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, CrossLight: A cross-layer optimized silicon photonic neural network accelerator, in *Proc. 58th ACM/IEEE Design Automation Conference (DAC)* (2021) pp. 1069–1074.
- [51] J. Choquette, NVIDIA Hopper H100 GPU: Scaling performance, *IEEE Micro* **43**, 9 (2023).
- [52] Qwen Team, Qwen2.5 technical report, arXiv preprint arXiv:2412.15115 (2024).
- [53] Qwen Team, Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).
- [54] P. Indyk and R. Motwani, Approximate nearest neighbors: Towards removing the curse of dimensionality, in *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)* (1998) pp. 604–613.
- [55] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval* (Cambridge University Press, 2008).
- [56] G. Kamradt, Needle in a haystack — pressure testing LLMs, [https://github.com/gkamradt/LLMTest\\_NeedleInAHaystack](https://github.com/gkamradt/LLMTest_NeedleInAHaystack) (2023).
- [57] S. Shekhar, W. Bogaerts, L. Chrostowski, J. E. Bowers, M. Hochberg, R. Soref, and B. J. Shastri, Roadmapping the next generation of silicon photonics, *Nature Communications* **15**, 751 (2024).
- [58] T. J. Seok, K. Kwon, J. Henriksson, J. Luo, and M. C. Wu, Wafer-scale silicon photonic switches beyond die size limit, *Optica* **6**, 490 (2019).
- [59] P. Dong, Silicon photonic integrated circuits for wavelength-division multiplexing applications, *IEEE Journal of Selected Topics in Quantum Electronics* **22**, 370 (2016).
- [60] Y. Bai, X. Xu, M. Tan, Y. Sun, Y. Li, J. Wu, R. Morandotti, A. Mitchell, K. Xu, and D. J. Moss, Photonic multiplexing techniques for neuromorphic computing, *Nanophotonics* **12**, 795 (2023).
- [61] J. Johnson, M. Douze, and H. Jégou, Billion-scale similarity search with GPUs, *IEEE Transactions on Big Data* **7**, 535 (2021).
- [62] K. Aubrey, *Inside the NVIDIA Rubin platform: Six new chips, one AI supercomputer*, NVIDIA Developer Blog (2026).
- [63] B. Tossoun, D. Liang, S. Cheung, Z. Fang, X. Sheng, J. P. Strachan, and R. G. Beausoleil, High-speed and energy-efficient non-volatile silicon photonic memory based on heterogeneously integrated memresonator, *Nature Communications* **15**, 551 (2024).
- [64] U. Adya, S. Singhal, R. Chen, I.-T. Chen, S. Joshi, A. Majumdar, M. Li, and S. Moazeni, Non-volatile tuning of cryogenic silicon photonic micro-ring modulators, *Nature Communications* **16**, 9290 (2025).
- [65] F. Fayza, C. Demirkiran, S. P. Rao, D. Bunandar, U. Gupta, and A. Joshi, Photonics for sustainable AI, *Communications Physics* **8**, 10.1038/s42005-025-02300-0 (2025).
- [66] Y. Li, H. Jiang, Q. Wu, X. Luo, S. Ahn, C. Zhang, A. H. Abdi, D. Li, J. Gao, Y. Yang, and L. Qiu, SCBench: A KV cache-centric analysis of long-context methods, in *International Conference on Learning Representations (ICLR)* (2025).
- [67] W. Zhang, F. Yin, H. Yen, D. Chen, and X. Ye, Query-focused retrieval heads improve long-context reasoning and re-ranking, in *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2025) pp. 23791–23805.
- [68] H. Park and Y. Park, PRISM: Photonic retrieval-index similarity module — experiment code, <https://github.com/hyoseokp/PRISM> (2025), source code for retrieval-head identification, signature generation, recall measurement, hardware-aware simulation, and downstream evaluation.
- [69] B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, 3rd ed. (John Wiley & Sons, 2019).
- [70] S. D. Personick, Receiver design for digital fiber optic communication systems, I, *Bell System Technical Journal* **52**, 843 (1973).
- [71] H. Jayatilaka, K. Murray, M. Caverley, N. A. F. Jaeger, L. Chrostowski, and S. Shekhar, Crosstalk in SOI microring resonator-based filters, *Journal of Lightwave Technology* **34**, 2886 (2016).
- [72] International Telecommunication Union, Spectral grids

for WDM applications: DWDM frequency grid, ITU-T Recommendation G.694.1 (2020).

[73] R. Boeck, N. A. F. Jaeger, N. Rouger, and L. Chrostowski, Series-coupled silicon racetrack resonators and

the Vernier effect: theory and measurement, [Optics Express](#) **18**, 25151 (2010).