# Dynamic High-Speed Integrated Photonic Tensor Core for Full-Range Transformer Self-Attention

**Shupeng Ning,[1] Hanqing Zhu,[1] Chenghao Feng,[1] Zhenxiang Xu,[1] Jiaqi Gu,[1,2] David Z. Pan,[1] and Ray T. Chen[1,*]**

[1] *Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78758, USA*
[2] *School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA*

[*] *chenrt@austin.utexas.edu*

**Abstract:** We demonstrate a coherent photonic tensor core for the Transformer self-attention mechanism using wavelength-division multiplexed interference to realize dynamic full-range matrix multiplications with compact dot-product arrays. Experiments validate scalable, hardware-efficient optical acceleration of attention workloads.

## 1. Introduction

Machine learning (ML) based on deep neural networks (DNNs) has transformed a wide range of scientific and technological fields. In particular, the widespread deployment and immense computational demands of attention-based Transformer architectures, such as large language models, have intensified the need for efficient hardware accelerators [1]. Unlike conventional fixed-weight matrix-vector multiplications, the multi-head attention mechanism in Transformer models relies on dynamic matrix multiplications, where both operands are activations generated on the fly rather than pre-programmed static weights. Although integrated photonics has emerged as a promising platform for high-performance computing [2]. these dynamic characteristics pose significant challenges for existing photonic tensor cores (PTCs), including the requirement for high-speed operand mapping and device reprogramming, as well as limited opportunities for operand reuse.

In this work, we propose a PTC tailored for the implementation of the self-attention mechanism. The PTC employs a compact crossbar array of interference-based photonic dot-product engines capable of high-speed, dynamic, full-range matrix multiplications. On-chip experimental results verify the effectiveness of the architecture and demonstrate its potential to support advanced ML models and large-scale computational workloads.

## 2. Design and Working Mechanism

### 2.1. Dynamically Operated Dot-Product (DDot) Unit

The architecture is built around coherent DDot units and a crossbar-style dynamically operated photonic tensor core (DoPTC). It directly targets the core bottleneck of Transformer workloads of dynamic, full-range matrix multiplications among query/key/value activations. As shown in Fig. 1(a), elements of the input vectors $\mathbf{x}$ and $\mathbf{y}$ are encoded onto optical fields at distinct wavelengths, and a $-90^\circ$ phase shift is applied to one input branch. Using broadband modulators such as Mach-Zehnder interferometers (MZIs), we can maintain an approximately identical transfer function over the wavelength range of interest. After a 50:50 directional coupler or a $2\times 2$ multimode interferometer (MMI), the two orthogonal signals recombine to form $(x_i \pm y_i)$, which are then detected by balanced photodetectors (PDs). This process can be expressed as:

$$\begin{pmatrix} z_i^+ \\ z_i^- \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e^{-j\pi/2} \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} x_i + y_i \\ j(x_i - y_i) \end{pmatrix}. \tag{1}$$

$$\begin{pmatrix} I^+ \\ I^- \end{pmatrix} \propto \frac{1}{2} \begin{pmatrix} \sum_i |x_i + y_i|^2 \\ \sum_i |j(x_i - y_i)|^2 \end{pmatrix}, \qquad I = I^+ - I^- \propto \sum_i x_i y_i \equiv \mathbf{x}\cdot\mathbf{y}. \tag{2}$$

This "compute-in-interference" mechanism performs signed arithmetic in a single shot by encoding each operand's sign and magnitude in the complex optical field (phase and amplitude), while wavelength-division multiplexing (WDM) enables multiple $(x_i, y_i)$ pairs to be processed concurrently at distinct wavelengths. Because the interference unit and phase shifter are passive and remain fixed after calibration, the DDot unit itself incurs no thermal control overheads.

### 2.2. Crossbar-Style DoPTC

To scale from single dot products to large-scale matrix multiplications, the DoPTC arranges multiple DDot units in a compact crossbar topology. Each bus waveguide carries a WDM stack, and an optical broadcast network fans these WDM channels across rows and columns within the crossbar array, enabling both operands to be shared among multiple DDot units (Fig. 1(b)). The intra-core optical broadcasting significantly reduces signal modulation complexity, thereby providing a hardware-efficient and scalable solution for optical Transformer accelerators.

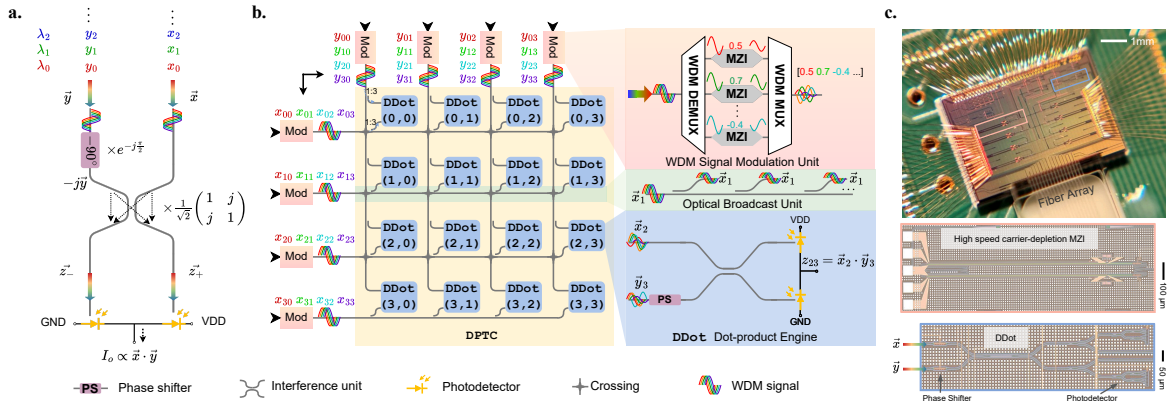


Fig. 1: (a) & (b) Schematic of DDot unit and DoPTC, adapted from our previous work [3] with modifications for clarity. (c) Micrographs of the taped-out DoPTC with integrated optical and electrical packaging.

## 3. Results and Discussion

In this work, we taped out a DoPTC with a $2 \times 2$ DDot array and high-speed carrier-depletion MZIs (Fig. 1(c)). The MZIs operate in push-pull mode to encode both magnitude and phase of each operand. For calibration, we first characterized each MZI by fitting its transmission curve using random input vectors. The functionality of the DDot unit for performing full-range multiplication of two operands was experimentally demonstrated, as shown in Fig. 2(a). To further evaluation, we implemented an image fusion task for brain computed tomography (CT) and magnetic resonance imaging (MRI), exploiting their complementary contrast, with CT emphasizing bone/calcification and MRI highlighting soft tissue/lesions. The Hadamard product between the images and dynamic masks is implemented on the DoPTC (Fig. 2(b)), and the fused images exhibit the expected features with low error relative to the digital reference.

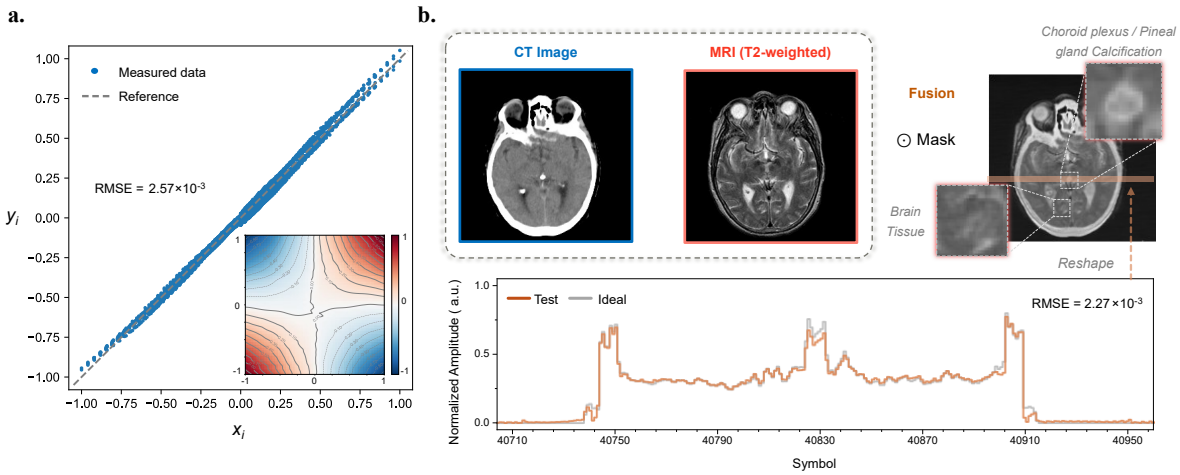


Fig. 2: Experimental results for full-range multiplication and image fusion. (a) On-chip measurement of the product of two input signals; the inset shows the corresponding 2D contour map of the multiplication output versus the two input operands. (b) Measured and ideal waveforms for the CT–MRI image fusion.

## 4. Conclusion

In this work, we demonstrated a coherent DoPTC capable of dynamic, full-range matrix multiplications for Transformer self-attention. By exploiting compute-in-interference and intra-core optical broadcasting, the proposed architecture provides a compact, hardware-efficient, and scalable photonic accelerator. This work is supported by the Air Force Office of Scientific Research and the Multidisciplinary University Research Initiative (MURI).

## References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems **30** (2017).
2. S. Ning, H. Zhu, C. Feng, J. Gu, Z. Jiang, Z. Ying, J. Midkiff, S. Jain, M. H. Hlaing, D. Z. Pan *et al.*, "Photonic-electronic integrated circuits for high-performance computing and ai accelerators," Journal of Lightwave Technology (2024).
3. H. Zhu, J. Gu, H. Wang, Z. Jiang, Z. Zhang, R. Tang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan, "Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, 2024), pp. 686–703.