nature portfolio

Peer Review File

Photonic Mixture-of-Experts for scalable multi-task on-chip optical neural networks

Corresponding Author: Professor Hongwei Chen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper proposed a method called Photonic Mixture-of-Expert (PMoE) to address the bottleneck of depth-dependent designs. The authors leverage multiple photonic experts' network to execute multi-task simultaneously. And the results show the effectiveness of this method in accuracy, efficiency and hardware overhead. The paper is generally written well and easy to follow. However, I do not recommend publishing the manuscript in its current form.

1. The concept of photonic Mixture-of-Expert has been proposed by [1-3]. Discussion should be included to demonstrate the advantages of the proposed method.
[1] Xu, B. et al, Optical Spiking Neurons Enable High-Speed and Energy-Efficient Optical Neural Networks, arXiv:2409.05726.
[2] Li, Chen. "Fully Optical Integrated Mixture-of-Experts System." Proceedings of the 2024 4th International Conference on Signal Processing and Communication Technology. 2024.
[3] Bernadskiy, Mikhail, et al. "Accelerating Frontier MoE Training with 3D Integrated Optics." 2025 IEEE Symposium on High-Performance Interconnects (HOTI). IEEE, 2025.
2. The section of result is thin. The authors should conduct more experiments using more complex datasets. The used datasets are not difficult to learn for single ONN.
3. Can some of experts be totally blocked? Such as blocking one or two experts, then test the accuracy. This will validate the contribution of different experts for specific tasks.
4. In Line 140, the authors claim that minimize the digital parameter, which is to minimize the model parameter counts or loss during training?
5. In Line 224, the authors claim that the data is loaded in parallel, however, in Fig. 1(b), the data is processed in sequence before routing.
6. In Line 249, the loading weights is reconfigurable, in Line 421, all optical and electronic parameters are fixed, please explain it.

Reviewer #2

(Remarks to the Author)
Manuscript ID: NCOMMS-26-016715
This manuscript proposes the Photonic Mixture-of-Experts architecture, presenting a commendable and timely solution to the scalability bottleneck in optical computing. By shifting to a “width-over-depth” paradigm, the authors demonstrate a scaling strategy that mitigates physical limitations inherent in deep meshes. The experimental results are compelling, demonstrating a multi-domain image classification accuracy of 97.1% with a compact computational footprint and a significant reduction in digital parameter overhead compared to discrete ONNs.
As the Mixture-of-Experts architecture has proven highly effective in scaling large language models, the translation of this concept to integrated photonics is innovative and sound. I recommend this paper for publication after considering the following suggestions:

1. I appreciate the authors' analysis of the diffractive computing throughputs and the comprehensive benchmarking against various metrics. The current experiment utilizes on-chip thermo-optic modulation (10 kHz), yet the computational density

(51.0 TOPS/mm2) and energy efficiency metrics are projected based on 10 GHz electro-optic modulators. The manuscript would benefit from a discussion regarding signal issues (e.g., crosstalk, insertion loss variations) that might arise when switching to high-speed junction-based modulators in such a dense diffractive layout.

2. Regarding the “task-specific weighted routing” mentioned in Fig. 4a, could the authors clarify the physical mechanism of the routing? Is it implemented by simply adjusting input power levels via thermo-optic modulators, or is there a dedicated optical switching stage? These details should be made clearer.

3. The authors advocate for a “width-over-depth” approach. However, width scaling faces its own bottlenecks, such as I/O count and optical splitting losses. Is there a theoretical limit to the number of experts before the signal-to-noise ratio degrades the classification accuracy, given the laser power splitting shown in Fig. 1? A comment on this trade-off would add depth to the discussion.

4. As the authors acknowledge, photonic components in current computing systems typically perform a linear operation. This implies that before the detection stage, the effective depth of the neural network remains 1 regardless of the number of optical layers. While linear transformations can perform powerful computation, it would be necessary to clarify the benefit of physical depth and the addition of diffractive layers.

5. Given the ultra-compact footprint (0.06 mm2) and the use of thermo-optic encoding, was there any observable thermal crosstalk between adjacent diffractive neurons or modulators? If so, how was it mitigated during the calibration process?

Minor Errors and Typos:
1. Fig. 2 Caption (Page 6, line 181): “The insect is the error distribution histogram” probably should be “The inset is”.
2. Page 9, line 267: “achieving a 67% reduction in compared to conventional” probably should be “reduction compared to”.
3. Page 9, line 294: “scale can be further extended by expand the number” probably should be “by expanding”.

Reviewer #3

(Remarks to the Author)
In their work „Photonic Mixture-of-Experts for large-scale on-chip optical neural networks“ the authors demonstrate a hybrid MoE network for MNIST/Fashion-MNIST/Extended-MNIST classification. The “MoE part” of the network is executed optically using an integrated, diffractive architecture operating at 10 kHz. Applying the concept of MoE to photonic processors is interesting, the experimental chip fabrication and packaging is impressive, and the results are well illustrated. However, the positioning of the work (and the advance with respect to the state of the art) is not fully clear and in its current form the manuscript contains not sufficiently backed up claims. Thus, major revisions are necessarily to be suitable for publications.
About the positioning of the work:
The authors make strong hardware related claims regarding “integration level / density” and “on-chip parallel kernels”, but do not provide the experimental evidence for a significant advance over the SOTA. For example, the system only operates at 10 kHz and the comparison to the SOTA does not have a sufficient depth. E.g. the work of Feldmann et al. cited in Fig 5 / Table 1 uses WDM in a crossbar array to compute several MVMs (different input vector encodings using the same physical matrix) in parallel, implementing a fundamentally different operation than the one presented in the paper.
Similarly, the authors do not provide an in-depth analysis of the photonic MoE idea that would present a significant advance with respect to the SOTA. The combination of MoE and analog processors has been already studied (e.g. Büchel, J., Vasilopoulos, A., Simon, W.A. et al. Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing. Nat Comput Sci 5, 13–26 (2025). https://doi.org/10.1038/s43588-024-00753-x). Implementing this concept in the photonic domain is interesting but then should be properly analyzed from a neural network perspective (especially benchmarking against arbitrary digital CNNs baselines is odd). Similarly, it’s not sufficiently supported why the combination with photonics is good (usually the goal of MoE is to reduce compute for large models, not parameters. In contrast, photonics usually struggles to provide the necessary number of programmable weights rather than providing sufficient compute).
Claims/Results in the paper that need to be backed up in more detail:
• “networks, featuring 18 parallel kernels within a compact computational footprint of 0.06 mm2“ This is misleading in the abstract as this implies system-level behavior but only considers the footprint of a small part of the system.
• “delivering superior accuracy and efficiency of 51.0 TOPS/mm2“. This statement in the introduction doesn’t reflect the actual performance, as the authors point out in their own discussion.
• The error plot Fig 2h indicates a significant nonsystematic error contribution. How does this effect the overall system?
• “representing a ~16% improvement over the digital counterpart“ It would make more sense to benchmark against the published SOTA for this task rather than some CNN trained by the authors. The baseline also seems to be quite low, given the comparable simple datasets.
• „This work charts a path for the future scalability of optical computing.“ Number of programmable weights is one of the largest challenges for scaling optical computing rather than compute. How would this approach solve this exactly? An in-depth scaling analysis would be good to back up this claim.
• „the intrinsic computational core area (0.06 mm2) and the total footprint (15 mm2) encompassing active modulation components.“ Those numbers should include the full systems, especially the I/O (e.g. DAC,ADC, TIA and datapaths) to be suitable for a comparison with the SOTA.
• Extrapolating from 10 kHz to 10 GHz needs more data points in between. Electronic noise characteristics and overall accuracy will be completely different.
• „power consumption for PMoE can be predicted as 0.628 W, resulting 81.3 TOPS/W/mm2“ The authors assume efficiency of 10 fJ/bit in the supplement (which should account for the full IO ?). How do the authors come up with the number

considering the full DAC, ADC, TIA chain?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thank you for your revised manuscript and the detailed response to my previous comments. I am pleased to see that all my initial concerns have been fully addressed. However, before final acceptance, there remain a few minor issues that should be corrected. Once these minor revisions are completed, the paper can be accepted for publication. Thank you for your efforts.

(1) In line 129 and line 393, it should be "l-th metaline layer ", not "i-th"?
(2) In Fig. 1(b), (c) and Fig. 4, g(n) is between input and "experts", however, which is between output and SDN in Fig. 1 (b). Besides, can the authors render the actual output channels of each expert? like 4, 6, 8.
(3) If feasible, please consider releasing the relevant code to enhance the reproducibility of your work. This is not a mandatory requirement, but would be greatly appreciated by the community.

Reviewer #2

(Remarks to the Author)
The authors have adequately addressed all the comments and revised the manuscript carefully. The quality of the paper meets the publication requirements. I recommend acceptance for publication.

Reviewer #3

(Remarks to the Author)
The authors have provided thoughtful and substantive replies to most of the problems raised and made revisions and supplements in response to the review comments, significantly improving the scientific rigor and clarity of the manuscript.
A minor comment: "Integration level" in Fig 5B is still misleading, as it implies different levels of integration (e.g. on-chip laser sources, on-chip I/O ...). Something like "Integration density" makes more sense given the unit /mm^2 .

Response Letter for NCOMMS-26-016715

We sincerely appreciate the reviewers' valuable time and constructive feedback, which have been pivotal in refining and strengthening our work. In this response letter, we address each remark point by point. Responses to comments are highlighted in blue, and corresponding changes to the manuscript are marked in red.

Summary of the main revisions:

1. We have added discussions of related MoE, analog computing, and optical MoE works. The revisions are reflected in the [Introduction] and the [References].
2. We have provided experiments on the expert activation strategies and a numerical evaluation on CIFAR-10 to further assess on a more complex dataset. The revisions are included in [Results], [Supplementary note 7 & 9] and [Fig. S6].
3. We have provided an experimentally anchored high-speed robustness evaluation to support the 10 GHz electro-optic projection. The revisions can be found in [Discussion], [Supplementary note 6], and [Fig. S5].
4. We have included a CNN-based electronic MoE baseline for a more controlled architecture-level comparison and strengthened the ablation analysis. The revisions can be found in [Discussion], [Supplementary note 8] and [Fig. 5 & S9].
5. We have clarified the motivation and positioning of the PMoE architecture. The revisions can be found in the [Abstract], [Introduction], and [Discussion].
6. We have revised the claims to clarify experimental thermo-optic results and projected high-speed electro-optic metrics. The revisions are reflected in the [Introduction], [Discussion], and [Supplementary note 10].
7. We have refined the footprint analysis and updated the power-consumption model with the included full I/O chains. The revisions can be found in [Introduction], [Supplementary note 10 & 11] and [Table S1 & S2].
8. We have clarified the role of physical depth and nonsystematic errors in the optical front-end. The revisions appear in [Supplementary note 2 & 4].
9. We have thoroughly proofread and revised the manuscript and Supplementary Information for clarity, consistency, and precision.

The point-by-point responses to all review comments are presented below. All modifications have been highlighted in the revised manuscript and Supplementary Information.

Point-by-Point Responses to Reviewer #1's Comments

R1-0. *This paper proposed a method called Photonic Mixture-of-Expert (PMoE) to address the bottleneck of depth-dependent designs. The authors leverage multiple photonic experts' network to execute multi-task simultaneously. And the results show the effectiveness of this method in accuracy, efficiency and hardware overhead. The paper is generally written well and easy to follow. However, I do not recommend publishing the manuscript in its current form.*

A1-0. We sincerely thank the reviewer for the thorough evaluation and the accurate summary of our work. We are encouraged by the your recognition of the core concept of our architecture and the proposed scaling paradigm. We also appreciate the constructive comments and insightful suggestions provided. These valuable inputs have helped us improve the clarity, rigor, and completeness of the manuscript. We have carefully considered all comments and revised the manuscript accordingly. We hope that these revisions have adequately addressed your concerns and improved the manuscript to a standard suitable for publication. Below, we provide a detailed point-by-point response.

R1-1. *The concept of photonic Mixture-of-Expert has been proposed by [1-3]. Discussion should be included to demonstrate the advantages of the proposed method.*
[1] Xu, B. et al, Optical Spiking Neurons Enable High-Speed and Energy-Efficient Optical Neural Networks, arXiv:2409.05726.
[2] Li, Chen. "Fully Optical Integrated Mixture-of-Experts System." Proceedings of the 2024 4th International Conference on Signal Processing and Communication Technology. 2024.
[3] Bernadskiy, Mikhail, et al. "Accelerating Frontier MoE Training with 3D Integrated Optics." 2025 IEEE Symposium on High-Performance Interconnects (HOTI). IEEE, 2025.

A1-1. We sincerely thank you for highlighting these highly relevant and interesting recent works. We have carefully studied these references. While the overarching concept of "MoE" also appears in these papers, our proposed Photonic Mixture-of-Experts (PMoE) architecture differs in physical implementation, computing paradigm, and level of experimental validation.

Per your constructive suggestion, we have incorporated a discussion to explicitly demonstrate the advantages of our PMoE method compared to these prior works:

1. Comparison with Xu et al. [1] (Free-space vs. on-chip integrated):

Xu et al. proposed an Optically Parallel Mixture of Experts (OPMOE) based on a free-space spatial light modulator (SLM) and optical spiking neurons. Their approach operates in a free-space optical setup and combines multiple classifiers through an Error-Correcting Output Codes (ECOC)-based framework.

- Difference from our work: In contrast, our PMoE is implemented as a fully on-chip integrated architecture with a compact computational footprint. Furthermore, rather

than relying on a fixed ensemble decoding process, our method realizes task-specific weighted routing, selectively allocating optical power according to different task demands.

2. Comparison with Li [2] (simulation vs. physical experimental demonstration):

Li proposed a fully optical integrated MoE system combining Mach-Zehnder Interferometers (MZIs) as gates and subwavelength units (SWUs) as experts. However, this work is based on theoretical design and numerical simulation without physical hardware demonstration.

- Difference from our work: In contrast, our work not only establishes the PMoE framework, but also provides a physically validated on-chip implementation of a diffractive Mixture-of-Experts optical computing system through chip fabrication, packaging, and experimental evaluation on multiple datasets. In addition, our work investigates the physical impact of error accumulation and motivates a width-over-depth scaling paradigm for optical neural networks.

3. Comparison with Bernadskiy et al. [3] (optical interconnects vs. computing):

Bernadskiy et al. focused on utilizing 3D integrated optics (e.g., optical interposers) to scale the communication bandwidth among conventional electronic GPUs for training large digital MoE models.

- Difference from our work: Our proposed PMoE instead represents an optical computing architecture in which the expert inference process is carried out directly in the analog optical domain, rather than using optics only for communication and interconnects in an otherwise digital computing system.

To clarify the positioning of our work and explicitly state its advantages, we have added a statement with these references in the Introduction section of the revised manuscript and included [1-3] in the reference list.

Relevant changes made:

- Revised manuscript (Introduction, Page 2, Lines 79-87): "Drawing inspiration from the digital MoE, we propose a width-over-depth paradigm that adapts this architecture to the optical domain. Related efforts have also explored the combination of MoE with analog processors [41]. By routing diverse inputs into multiple photonic experts, this approach offers a robust pathway to circumvent the depth-scaling parameter bottleneck and efficiently execute diverse multi-task workloads without reconfiguring physical weights. Although recent works have explored the synergy between optics and the MoE concept, ranging from free-space optical classifier ensembles [42] and theoretical simulations of hybrid systems [43] to deploying optical interconnects for electronic GPUs [44], an experimentally validated on-chip photonic MoE computing architecture remains lacking."

- New added Reference:
  "42. Xu, B. *et al.* Optical Spiking Neurons Enable High-Speed and Energy-

Efficient Optical Neural Networks. *arXiv preprint arXiv:2409.05726* (2024).
43. Li, C. in *Proceedings of the 2024 4th International Conference on Signal Processing and Communication Technology.* 260-265.
44. Bernadskiy, M. *et al.* in *2025 IEEE Symposium on High-Performance Interconnects (HOTI).* 25-35 (IEEE)."

R1-2. *The section of result is thin. The authors should conduct more experiments using more complex datasets. The used datasets are not difficult to learn for single ONN.*

A1-2. We sincerely thank you for this important comment. We agree that the originally presented datasets are relatively simple from the perspective of single-task image classification, and that additional evaluation on a more complex dataset can further strengthen the Results section.

We would like to clarify that the primary goal of this work is not to pursue the highest classification accuracy on a single benchmark, but to experimentally validate the core functionality and scaling strategy of the proposed PMoE architecture in a controlled multi-domain setting. In particular, our work focuses on demonstrating that multiple photonic expert networks can be integrated on a single chip, coordinated through task-specific routing, and combined with a shared electronic backend to process diverse tasks efficiently without reconfiguring the photonic expert weights. From this perspective, the selected datasets, namely MNIST, Fashion-MNIST, and Extended-MNIST, provide three representative domains with different data distributions and classification characteristics, which are suitable for validating the key architectural features of PMoE, including multi-domain processing, expert collaboration, and backend parameter sharing.

In addition, the chip used in this work was physically designed and optimized for this multi-domain experimental configuration, including the allocation of different output-channel numbers for different expert networks. Therefore, the current chip-level experiments are intended to provide a hardware-grounded validation of the proposed architecture under a realistic and controlled task setting, rather than to establish state-of-the-art performance on individually difficult datasets.

Following the reviewer's suggestion, we further supplemented the study with an additional numerical evaluation on a more complex dataset, CIFAR-10, using the PMoE framework. In this supplementary study, we investigated the classification accuracy as the number of expert networks increases, which surpass the accuracy of single ONN while the total parameter count growing moderately, as shown in **Fig. R1-1**. The results show that, for a more complex visual task, increasing the number of experts still leads to performance improvement, further supporting the scalability of the proposed width-over-depth PMoE architecture beyond the handwritten-image datasets used in the chip experiment. We note that this added result is presented as a supplementary numerical validation, while the main text remains focused on the experimentally demonstrated chip-level multi-domain processing.

Taken together, we believe that the existing chip experiments are sufficient to

support the main conclusion of the manuscript, namely that PMoE enables efficient multi-domain optical computing with reduced backend parameter overhead, and that the newly added CIFAR-10 dataset evaluation further strengthens the generality of the proposed scaling strategy on a more complex task.

We clarified in the main text that the selected datasets are used to validate the PMoE architecture in a controlled multi-domain hardware setting, and we added a brief statement referring to a new Supplementary note, where additional numerical results on CIFAR-10 are provided to further assess the scalability of the PMoE framework on a more complex dataset.

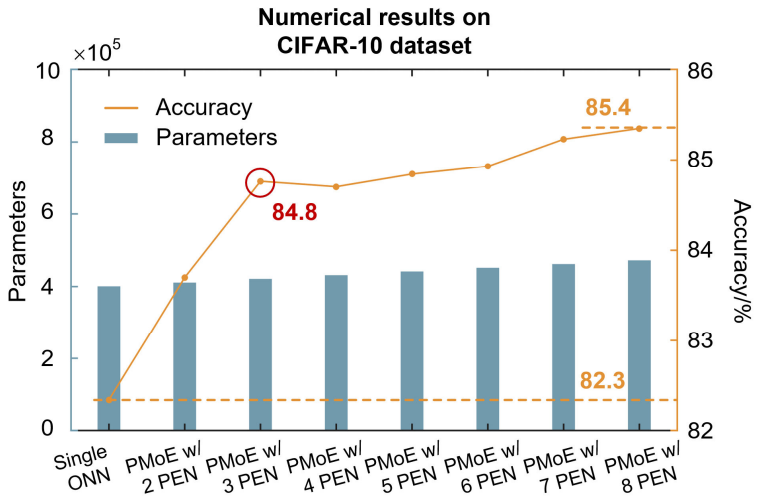


Fig. R1-1. Numerical results on the CIFAR-10 dataset using the PMoE framework with different numbers of experts. The classification accuracy improves as the number of PENs increases, while the total parameter count grows moderately, further supporting the scalability of the PMoE architecture on a more complex visual task.

Relevant changes made:

- Revised manuscript (Results, Page 7, Lines 232-235): "After finalizing the overall architecture of the sparse PMoE network, we conducted experimental validation of its multi-domain processing capability. The selected datasets were used to validate the PMoE architecture in a controlled multi-domain hardware setting, where distinct domains with different data distributions can be processed within a unified chip framework."

- Revised manuscript (Results, Page 9, Lines 282-289): "Furthermore, by increasing the number of simultaneously activated experts from a single PEN (hard-routing) to multiple PENs (weighted-routing), the performance can be improved through collaborative expert processing without increasing the digital parameter overhead, as shown in Fig. 4i. Notably, this performance gain with increasing expert number is also supported by an additional numerical study on the more complex CIFAR-10 dataset, where the classification accuracy improves as more experts are introduced (Supplementary note 7). These results further support that PMoE offers a scalable and efficient paradigm for on-chip optical processors."

- Revised Supplementary Information (Note 7, Page 9):

Supplementary note 7: Impact of the number of activated experts and routing strategies

To systematically evaluate the contribution of individual PENs and further validate the necessity of the multi-expert architecture, we investigated how the PMoE performance evolves with different routing strategies and different numbers of activated experts. As illustrated in Fig. S6a, a conventional single ONN is typically optimized for a single task, whereas the proposed PMoE architecture integrates multiple experts on a single chip and supports multi-domain processing through routing and a shared backend network. By selectively activating or blocking specific experts, the PMoE can transition from a sparse expert setting to a denser collaborative regime with increasing numbers of active experts.

For the chip-level experiments, we evaluated the PMoE under three representative routing configurations. In the hard-routing case, only one PEN is activated for each input sample, corresponding to the expert with the highest task-specific routing weight, while the remaining experts are completely blocked. In the weighted-routing with 2 PENs case, the expert with the lowest routing weight is blocked and only the top two experts are activated. In the weighted-routing with 3 PENs case, all three experts are activated simultaneously according to their respective analog routing weights.

The experimental results on the MNIST, Fashion-MNIST, and Extended-MNIST datasets are summarized in Fig. S6b. As the number of activated experts increases, the average classification accuracy improves steadily from 94.8% under hard-routing, to 95.7% with 2 PENs, and further to 97.1% when all 3 PENs are activated. This result indicates that while a task-specific single expert is already capable of basic feature extraction, collaborative processing with multiple experts can further enhance the representational capacity of the system. Notably, compared with the electronic counterpart, which achieves an average accuracy of 91.2%, the PMoE shows clear performance advantages under all routing configurations. Moreover, the weighted-routing PMoE also outperforms the reference scheme of using three single ONNs for three tasks, which achieves an average accuracy of 95.0% but requires three times the parameter cost. These results confirm that the PMoE architecture can improve multi-domain classification performance while maintaining a shared backend structure.

We further observed that the performance gain brought by activating more experts depends on the intrinsic complexity of the task. In particular, the Fashion-MNIST task shows the most pronounced improvement when transitioning from hard-routing to weighted-routing (Fig. 4i), indicating that more complex visual patterns benefit more strongly from collaborative expert processing.

To further assess whether this trend extends beyond handwritten-image datasets, we additionally performed a numerical study on the more complex CIFAR-10 dataset using the PMoE framework. The corresponding results are shown in Fig. S6c. Starting from a single-ONN baseline accuracy of 82.3%, the PMoE performance improves as the number of experts increases, reaching 84.8% with 3 PENs and further to 85.4% with 8 PENs. Meanwhile, the total parameters increase only moderately with the number of experts. Although this result is numerical rather than chip-level, it provides additional evidence that the proposed width-over-depth PMoE strategy remains effective for more

complex visual tasks and that increasing the number of experts can consistently improve performance under the same architectural framework.

These experimental and numerical results demonstrate that the number of activated experts is an important factor in PMoE. This trend is observed not only in the chip-level multi-domain experiments but also in the more complex CIFAR-10 numerical study, further supporting the scalability and generality of the PMoE architecture.

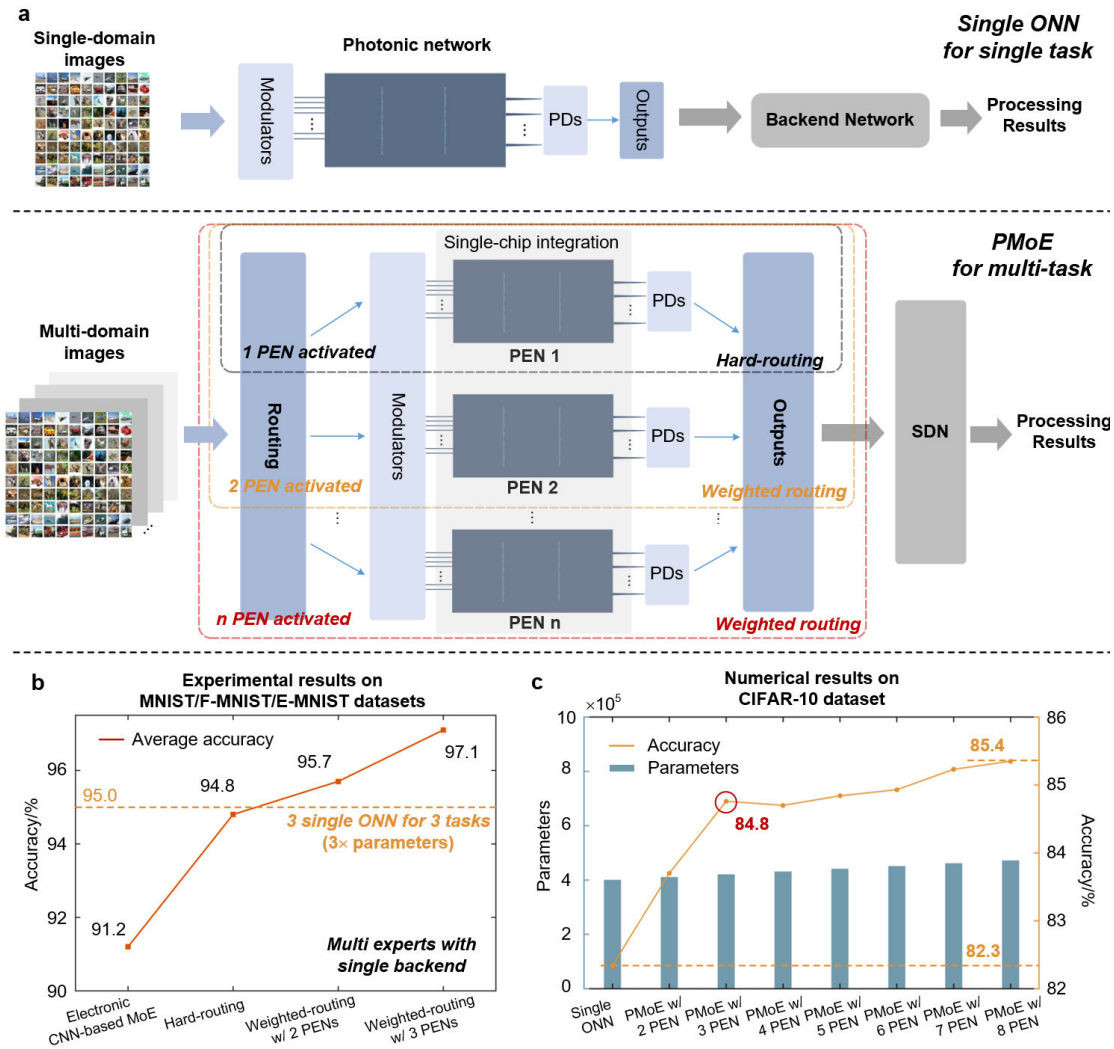


Fig. S6.

Impact of the number of activated experts on PMoE performance. a, Schematic comparison between a conventional single optical neural network (ONN) for a single task and the proposed PMoE architecture with different activated photonic expert networks (PENs) for multi-task processing. **b,** Experimental results on the MNIST, Fashion-MNIST, and Extended-MNIST datasets under different expert-activation configurations. **c,** Numerical results on the CIFAR-10 dataset using the PMoE framework with different numbers of experts.

R1-3. *Can some of experts be totally blocked? Such as blocking one or two experts, then test the accuracy. This will validate the contribution of different experts for specific tasks.*

A1-3. We sincerely appreciate your insightful suggestion. To answer this question directly, experts can indeed be completely blocked, and we have already conducted this ablation experiment in the original manuscript. The corresponding results were initially presented in the original Fig. 4i and are now more explicitly detailed in **Fig. R1-2b** of this response.

We acknowledge that this ablation analysis was not sufficiently emphasized or clearly explained in the original text. In essence, varying the number of activated experts corresponds to the transition from the hard-routing strategy to weighted-routing with multiple experts, while keeping the digital backend parameter overhead basically unchanged.

Specifically, we experimentally evaluated the PMoE under distinct routing configurations with different numbers of activated PENs, as illustrated in **Fig. R1-3a**:

Hard-routing (1 PEN activated): For a given input sample, only the single optical expert with the highest task-specific routing weight is activated, while the other experts are completely blocked.

Weighted-routing with 2 PENs: The expert with the lowest routing weight is blocked, and only the top two experts with the highest task-specific weights are activated.

Weighted-routing with 3 PENs: All the three assigned experts are activated simultaneously according to their respective analog routing weights.

Our experimental results show that increasing the number of simultaneously activated experts consistently improves the classification accuracy, and that the magnitude of the improvement depends on the complexity of the task. As shown in **Fig. R1-2b** and **Fig. R1-3b**, when transitioning from hard-routing to weighted-routing, the system exhibits accuracy improvements across all domains. Notably, the more complex Fashion-MNIST (F-MNIST) task shows the most pronounced gain compared with MNIST and Extended-MNIST (E-MNIST).

Furthermore, **Fig. R1-3c** shows the progression of the average accuracy across all three datasets. The average accuracy increases from 94.8% under hard-routing to 95.7% with 2 PENs, and further to 97.1% when all 3 PENs are activated through full weighted-routing. This trend indicates that, while a task-specific single expert can already perform basic feature extraction, collaborative processing with multiple experts further enhances the representational capacity of the system. These results support the effectiveness of the PMoE architecture for handling complex and diverse data distributions in multi-task scenarios.

To make this ablation study clearer and to explicitly validate the contribution of different experts as suggested by the reviewer, we have revised the corresponding description in the manuscript and added a dedicated Supplementary note with further numerical study on CIFAR-10 datasets in the revised Supplementary Information.

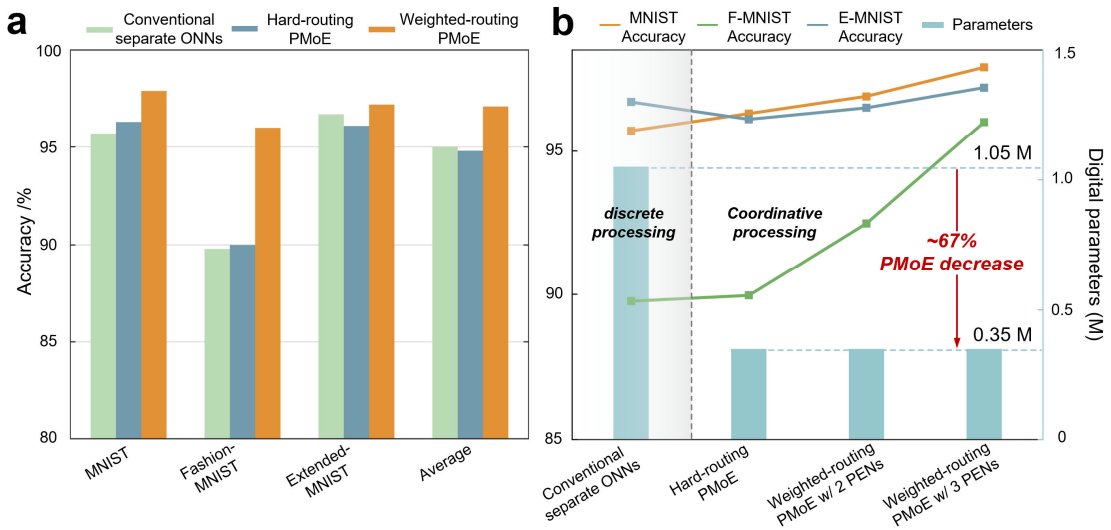


Fig. R1-2. a, Performance benchmarking comparing conventional separate ONNs, hard-routing PMoE, and weighted-routing PMoE. **b,** Comparison of PMoE with different routing strategies and conventional ONNs. the PMoE architecture maintains superior accuracy while reducing the required digital parameters by approximately 67%.

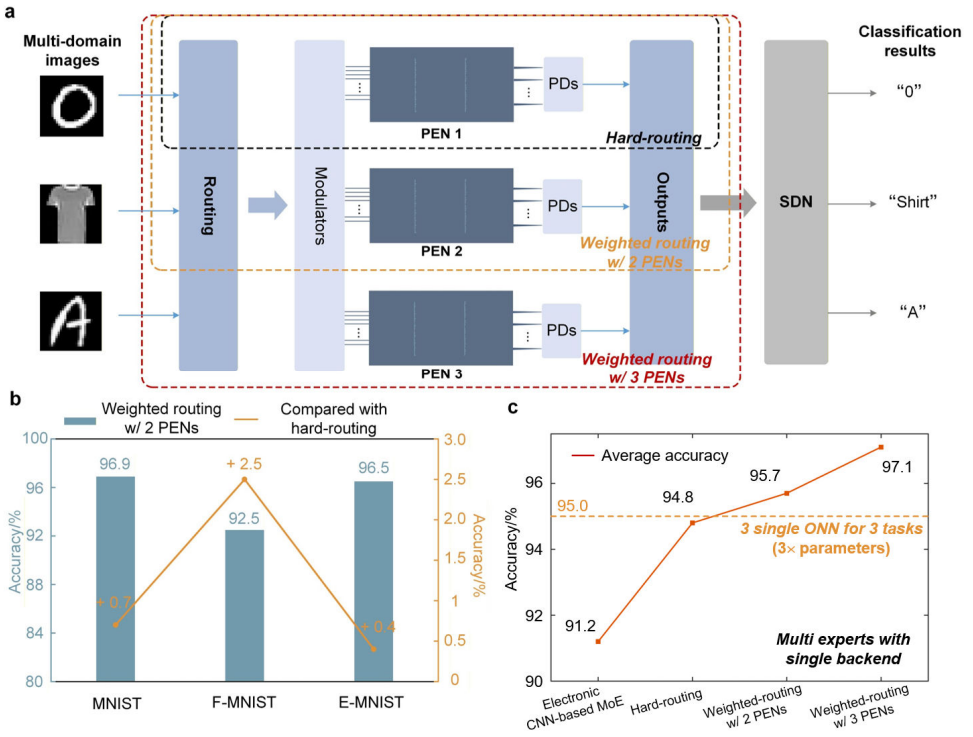


Fig. R1-3. a, Schematic illustration of the routing strategies. The system transitions from hard-routing (activating only the top 1 PEN) to weighted-routing with 2 PENs (activating the top 2 PENs), and finally to full weighted-routing with 3 PENs. **b,** Task-specific classification accuracies when employing the weighted-routing strategy with 2 PENs with accuracy improvement for each dataset compared to the baseline hard-routing strategy. **c,** The average classification accuracy across the multi-domain datasets under different configurations. The Weighted-routing schemes outperform the conventional ONN baseline (utilizes 3× the digital parameters).

Relevant changes made:

- Revised manuscript (Results, Page 9, Lines 280-289): "By engaging multiple experts through weighted fusion, this strategy further expands the effective network scale and parallelism for each task, outperforming the hard-routing baseline. Furthermore, by increasing the number of simultaneously activated experts from a single PEN (hard-routing) to multiple PENs (weighted-routing), the performance can be improved through collaborative expert processing without increasing the digital parameter overhead, as shown in Fig. 4i. Notably, this performance gain with increasing expert number is also supported by an additional numerical study on the more complex CIFAR-10 dataset, where the classification accuracy improves as more experts are introduced (Supplementary note 7). These results further support that PMoE offers a scalable and efficient paradigm for on-chip optical processors."

- Revised Supplementary Information (Note 7, Page 9):

Supplementary note 7: Impact of the number of activated experts and routing strategies

To systematically evaluate the contribution of individual PENs and further validate the necessity of the multi-expert architecture, we investigated how the PMoE performance evolves with different routing strategies and different numbers of activated experts. As illustrated in Fig. S6a, a conventional single ONN is typically optimized for a single task, whereas the proposed PMoE architecture integrates multiple experts on a single chip and supports multi-domain processing through routing and a shared backend network. By selectively activating or blocking specific experts, the PMoE can transition from a sparse expert setting to a denser collaborative regime with increasing numbers of active experts.

For the chip-level experiments, we evaluated the PMoE under three representative routing configurations. In the hard-routing case, only one PEN is activated for each input sample, corresponding to the expert with the highest task-specific routing weight, while the remaining experts are completely blocked. In the weighted-routing with 2 PENs case, the expert with the lowest routing weight is blocked and only the top two experts are activated. In the weighted-routing with 3 PENs case, all three experts are activated simultaneously according to their respective analog routing weights.

The experimental results on the MNIST, Fashion-MNIST, and Extended-MNIST datasets are summarized in Fig. S6b. As the number of activated experts increases, the average classification accuracy improves steadily from 94.8% under hard-routing, to 95.7% with 2 PENs, and further to 97.1% when all 3 PENs are activated. This result indicates that while a task-specific single expert is already capable of basic feature extraction, collaborative processing with multiple experts can further enhance the representational capacity of the system. Notably, compared with the electronic counterpart, which achieves an average accuracy of 91.2%, the PMoE shows clear performance advantages under all routing configurations. Moreover, the weighted-routing PMoE also outperforms the reference scheme of using three single ONNs for three tasks, which achieves an average accuracy of 95.0% but requires three times the parameter cost. These results confirm that the PMoE architecture can improve multi-

domain classification performance while maintaining a shared backend structure.

We further observed that the performance gain brought by activating more experts depends on the intrinsic complexity of the task. In particular, the Fashion-MNIST task shows the most pronounced improvement when transitioning from hard-routing to weighted-routing (Fig. 4i), indicating that more complex visual patterns benefit more strongly from collaborative expert processing.

To further assess whether this trend extends beyond handwritten-image datasets, we additionally performed a numerical study on the more complex CIFAR-10 dataset using the PMoE framework. The corresponding results are shown in Fig. S6c. Starting from a single-ONN baseline accuracy of 82.3%, the PMoE performance improves as the number of experts increases, reaching 84.8% with 3 PENs and further to 85.4% with 8 PENs. Meanwhile, the total parameters increase only moderately with the number of experts. Although this result is numerical rather than chip-level, it provides additional evidence that the proposed width-over-depth PMoE strategy remains effective for more complex visual tasks and that increasing the number of experts can consistently improve performance under the same architectural framework.

These experimental and numerical results demonstrate that the number of activated experts is an important factor in PMoE. This trend is observed not only in the chip-level multi-domain experiments but also in the more complex CIFAR-10 numerical study, further supporting the scalability and generality of the PMoE architecture.

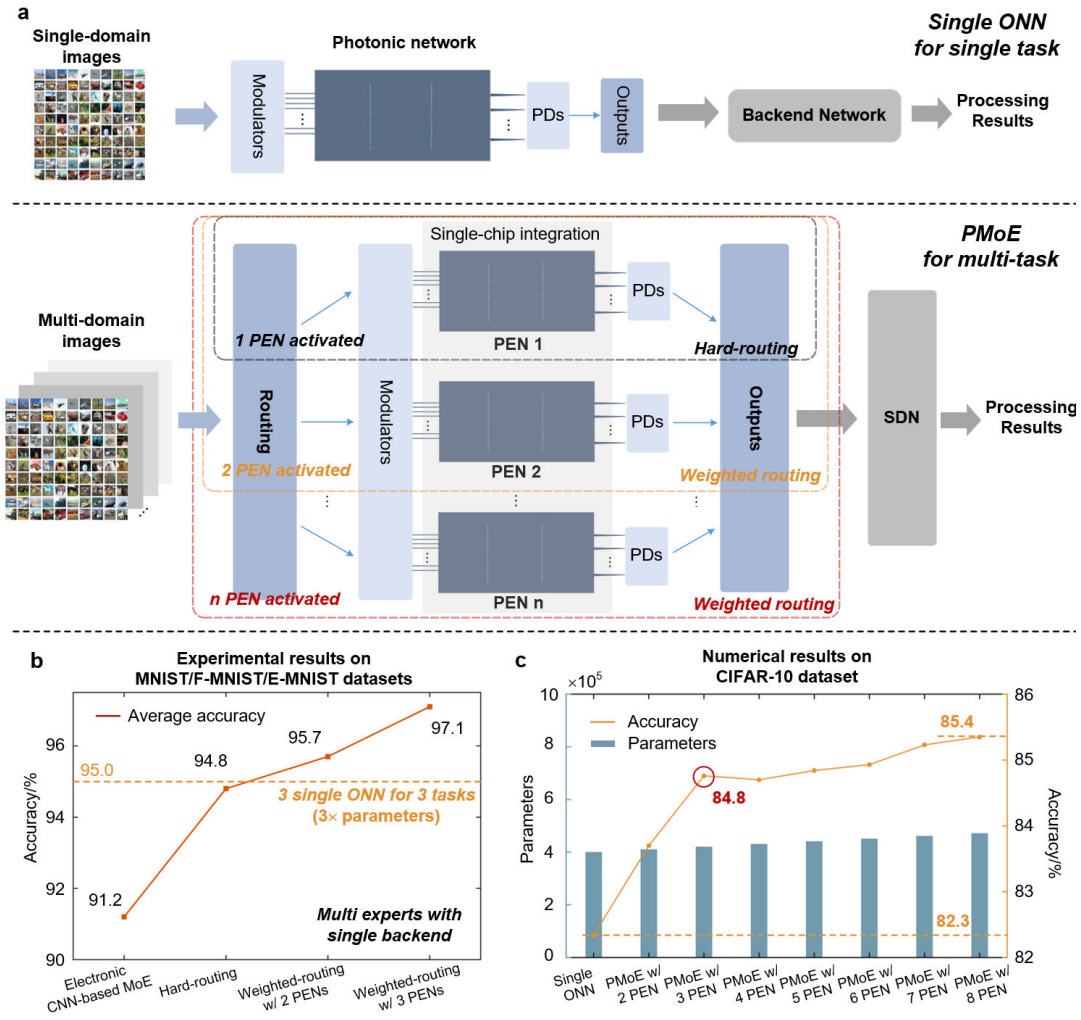


Fig. S6.
Impact of the number of activated experts on PMoE performance. a, Schematic comparison between a conventional single optical neural network (ONN) for a single task and the proposed PMoE architecture with different activated photonic expert networks (PENs) for multi-task processing. **b,** Experimental results on the MNIST, Fashion-MNIST, and Extended-MNIST datasets under different expert-activation configurations. **c,** Numerical results on the CIFAR-10 dataset using the PMoE framework with different numbers of experts.

R1-4. *In Line 140, the authors claim that minimize the digital parameter, which is to minimize the model parameter counts or loss during training?*

A1-4. We thank the reviewer for pointing out this ambiguity in our phrasing. To clarify, "minimizing the digital parameter" in Line 140 refers specifically to minimizing the model parameter count of the shared backend digital network while maintaining high classification accuracy.

In the Photonic Mixture-of-Experts (PMoE) architecture designed for multi-task learning, the size (i.e., parameter count) of the shared backend digital network serves as an optimizable structural hyperparameter. As illustrated in **Fig. R1-4** (Fig. 3d in the manuscript), the parameter count of the FC layer, which constitutes the major part of the backend digital network, scales with the number of neurons. Meanwhile, the average test accuracy initially improves and then saturates around a neuron count of q = 128, which is the structural configuration selected in our design.

Our primary objective is to minimize the digital computational overhead as much as possible while maintaining high classification accuracy across all tasks. By shifting the multi-domain feature extraction burden to the optical front-end experts, the PMoE framework reduces the reliance on backend digital parameters without sacrificing multi-task performance.

To avoid confusion for readers, we have revised the sentence in the manuscript to explicitly state "minimize the digital model parameter counts while maintaining high classification accuracy."

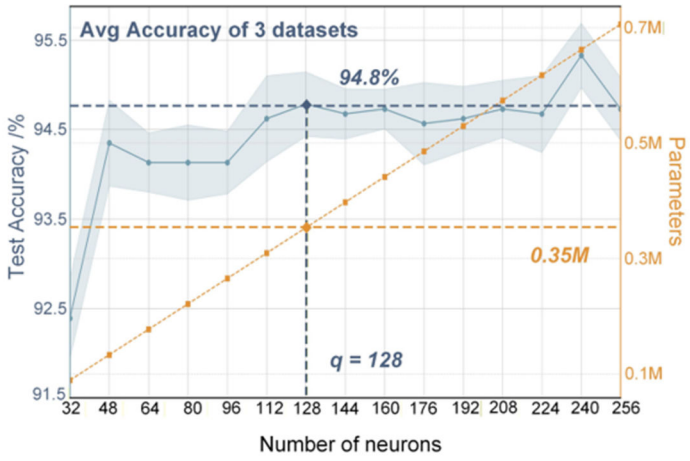


Fig. R1-4. The curve of the average test accuracy (blue line), total parameter count (orange dashed line), and the number of neurons in the backend fully connected layer. The optimal configuration is identified at q=128.

Relevant changes made:

- Revised manuscript (Results, Page 5, Lines 152-154): "The SDN includes a shared convolutional block and a shared fully connected (FC) layer, which is optimized for multi-domain processing while minimizing the digital parameter count and maintaining high classification accuracy."

R1-5. *In Line 224, the authors claim that the data is loaded in parallel, however, in Fig. 1(b), the data is processed in sequence before routing.*

A1-5. We sincerely appreciate your careful observation. We agree that our original phrasing could lead to ambiguity, and that the meaning of "parallel" should be clarified more explicitly.

To clarify, our original statement that "data is loaded in parallel" does not mean that multiple different datasets or multiple distinct input samples are fed into the system simultaneously at the initial stage. We acknowledge that this description was inappropriately placed in the section discussing the hard-routing strategy, which may have caused the confusion. As correctly illustrated in Fig. 1(b), individual input samples (e.g., images) are fed into the system in sequence.

The term "parallel" in our original text was intended to describe the physical hardware operation under the weighted-routing strategy for a single input sample. Once one image sample is fed into the system, its corresponding optical signals are distributed to and processed by multiple optical experts in parallel. Therefore, the parallelism occurs spatially across multiple experts during the signal loading and computing stages, rather than temporally across different input samples.

To avoid potential confusion for readers, we have revised the manuscript accordingly. Specifically, we removed the "parallel" description from the hard-routing section and added the following clarifying sentence in the weighted-routing section: "The signal of each image sample is modulated in parallel to the multiple PENs via weighted routing."

Relevant changes made:

- Revised manuscript (Results, Page 7, Line 232-233): "After finalizing the overall architecture of the sparse PMoE network, we conducted experimental validation of multi-domain processing capability."
- Revised manuscript (Results, Page 7, Line 238-239): "During the experiments, the PMoE performed data loading and processing across domains."
- Revised manuscript (Results, Page 9, Lines 261-263): "The signal of each image sample is modulated and routed in parallel to the multiple PENs under the

weighted-routing strategy. During the signal modulation, distinct task-specific pretrained weight coefficients are applied across the experts."

R1-6. *In Line 249, the loading weights is reconfigurable, in Line 421, all optical and electronic parameters are fixed, please explain it.*

A1-6. We sincerely appreciate the reviewer for highlighting this apparent contradiction. We agree that placing these two statements together without further clarification could cause confusion, and we have therefore revised the relevant description.

To clarify, these two statements refer to different components of the PMoE system and to different stages of operation.

When we stated in Line 421 that "all optical and electronic parameters are fixed," we were referring specifically to the intrinsic learned parameters of the core computational network itself. This includes the physical structure of the diffractive optical experts and the parameters of the unified digital backend. Once the PMoE system is trained and deployed, this core computational network remains frozen and unchanged.

By contrast, the "reconfigurable loading weights" mentioned in Line 249 refer only to the task-specific routing weights applied at the input stage. Depending on the task being executed, the input signal is distributed to different experts with different power ratios (i.e., the loading weights). These routing weights are not part of the fixed core network parameters; rather, they act as task-dependent input controls that activate the appropriate experts before the actual network computation.

To make this physical and functional distinction explicit and to avoid potential misunderstanding, we have revised the corresponding statements in the manuscript. Specifically, we now refer to these quantities as "reconfigurable task-specific input routing weights," and we clarify that "the intrinsic parameters of the core optical and electronic computational network are fixed."

Relevant changes made:

- Revised manuscript (Abstract, Page 1, Lines 30-32): "By dynamically routing inputs to these experts, the PMoE efficiently executes diverse multi-task workloads without altering the physical optical weights."
- Revised manuscript (Results, Page 9, Lines 266-267): "Notably, switching between distinct tasks requires only the reconfiguration of task-specific input routing weights."
- Revised manuscript (Methods, Page 13, Lines 429-430): "Once the network is converged and the chip is fabricated, the intrinsic parameters of the core optical and electronic computational network are fixed."

Point-by-Point Responses to Reviewer #2's Comments

R2-0. *This manuscript proposes the Photonic Mixture-of-Experts architecture, presenting a commendable and timely solution to the scalability bottleneck in optical computing. By shifting to a "width-over-depth" paradigm, the authors demonstrate a scaling strategy that mitigates physical limitations inherent in deep meshes. The experimental results are compelling, demonstrating a multi-domain image classification accuracy of 97.1% with a compact computational footprint and a significant reduction in digital parameter overhead compared to discrete ONNs.*

A2-0. We sincerely thank the reviewer for the positive evaluation and for recognizing the significance of our architecture. We appreciate your recognition of the proposed "width-over-depth" scaling strategy as a useful approach to addressing the physical bottlenecks of optical computing. We also thank for your constructive comments on the physical implementation details. We have carefully considered each point and revised the manuscript accordingly. Please find our detailed point-by-point responses below.

R2-1. *I appreciate the authors' analysis of the diffractive computing throughputs and the comprehensive benchmarking against various metrics. The current experiment utilizes on-chip thermo-optic modulation (10 kHz), yet the computational density (51.0 TOPS/mm2) and energy efficiency metrics are projected based on 10 GHz electro-optic modulators. The manuscript would benefit from a discussion regarding signal issues (e.g., crosstalk, insertion loss variations) that might arise when switching to high-speed junction-based modulators in such a dense diffractive layout.*

A2-1. We thank the reviewer for this insightful comment. Regarding the challenges in high-speed implementation, we believe that one key issue is the link-loss and signal-to-noise-ratio budget of the system. High-speed operation significantly reduces the integration time per bit and therefore requires sufficient optical power to remain above the sensitivity threshold of high-speed photodetectors. While optical or electrical amplification can compensate for part of the link loss, it may also introduce amplified spontaneous emission or thermal noise, thereby degrading the system SNR. We therefore emphasize that reducing insertion loss and maintaining signal integrity are important prerequisites for migrating from thermo-optic operation to high-speed electro-optic implementation.

We also agree that the practical signal integrity under high-speed electro-optic deployment should be supported experimentally rather than inferred only from throughput estimates. To address this point, we conducted a preliminary high-speed characterization based on an integrated diffractive-computing chain comprising on-chip high-speed modulators, the diffractive computing core, and on-chip high-speed photodetectors. Using this test vehicle, we characterized the end-to-end optical-computing response at 1, 2, 5, 8, and 10 GHz. As shown in **Fig. R2-1a**, the output feature maps remain highly consistent with the simulated target across the tested baud rates, and the measured 10 GHz temporal waveform is shown in **Fig. R2-1b**. Quantitatively, the signal quality degrades gradually with baud rate, with the SNR decreasing from 16.61 dB at 1 GHz to 14.61 dB at 10 GHz, while the MSE increases

from 7.0×10^{-3} to 1.15×10^{-2}, as shown in **Fig. R2-1c**.

These measurements provide an experimental basis for evaluating the combined signal degradation that may arise under dense high-speed integrated operation, including the aggregate effects of waveform distortion, insertion-loss-induced amplitude variation, and receiver noise. While the present experiment does not separately quantify each contribution, such as crosstalk or insertion-loss variation, it shows that the end-to-end degradation remains gradual up to 10 GHz and stays within a range that is tolerable for downstream processing.

In addition, to directly connect this high-speed characterization to the PMoE application, we performed an experimentally anchored robustness evaluation on the multi-domain classification task, as shown in **Fig. R2-2.** Specifically, we introduced into the PMoE experimental feature maps the same level of signal degradation measured from the high-speed chain at different baud rates and then re-evaluated the multi-domain classification accuracy. The results show that the average accuracy decreases only slightly from 97.1% in the original thermo-optic experiment to 96.8% under the 10 GHz matched-degradation condition, indicating that the PMoE network remains robust to the experimentally observed high-speed signal-quality reduction.

In response to this comment, we have added a brief discussion in the main text and a new Supplementary note 6 in the Supplementary Information, together with Fig. S5, to present the high-speed robustness evaluation and its impact on multi-domain classification accuracy.

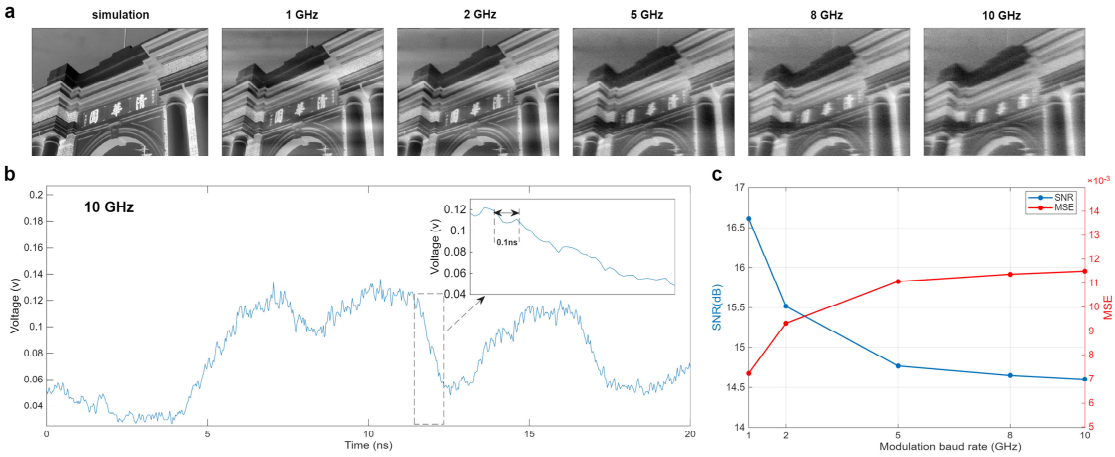


Fig. R2-1. Experimental characterization of a high-speed integrated diffractive-computing chain. a, Simulated and experimentally measured output feature maps at 1, 2, 5, 8, and 10 GHz. **b,** Measured temporal output waveform at 10 GHz. Inset: enlarged view showing a 0.1 ns symbol interval. **c,** Measured signal quality versus modulation baud rate, showing the corresponding SNR and MSE.

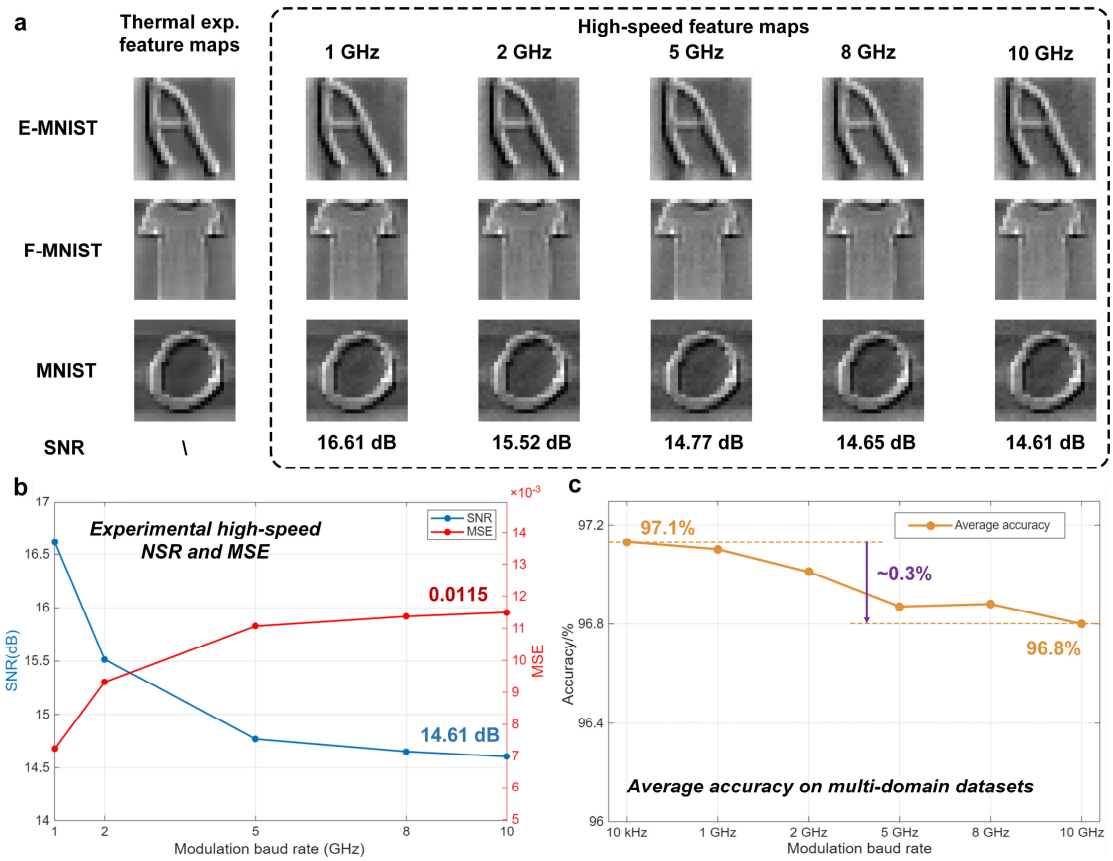


Fig. R2-2. Experimentally anchored evaluation of PMoE robustness under high-speed deployment conditions. a, Thermo-optic experimental feature maps and the corresponding feature maps with matched degradation under 1, 2, 5, 8, and 10 GHz conditions. **b,** SNR and MSE used to emulate the experimentally measured high-speed signal degradation. **c,** Average classification accuracy of the PMoE system on the multi-domain datasets under matched high-speed degradation conditions.

Relevant changes made:

- Revised manuscript (Discussion, Page 10, Lines 338-341): "We further evaluated the PMoE architecture under high-speed degradation conditions. The average classification accuracy decreases only slightly from 97.1% to 96.8% under the 10 GHz matched condition (Fig. S5), providing preliminary support for high-speed electro-optic implementations."

- Revised Supplementary Information (Note 6, Pages 7-8):

Supplementary note 6: Evaluation of PMoE under high-speed deployment conditions

To assess the feasibility of high-speed deployment, we further evaluated the PMoE architecture under experimentally matched high-speed signal degradation conditions. Since the current PMoE chip is driven by thermo-optic modulators, direct full-system measurements at GHz rates are not available on the present platform. Therefore, we performed a preliminary high-speed characterization on an integrated optical diffractive-computing chain and used the experimentally measured signal degradation to construct a physically grounded evaluation for the PMoE task.

As shown in Fig. S5a, we introduced matched degradation into the experimentally obtained PMoE feature maps according to the signal quality measured at different modulation baud rates (1, 2, 5, 8, and 10 GHz). Representative feature maps for Extended-MNIST, Fashion-MNIST, and MNIST remain visually consistent under all tested conditions, although the image quality gradually decreases as the baud rate increases. Fig. S5b shows the introduced degradation levels. The corresponding signal-to-noise ratio (SNR) decreases from 16.61 dB at 1 GHz to 14.61 dB at 10 GHz, while the mean squared error (MSE) increases to 0.0115 at 10 GHz.

Using these experimentally matched degradation levels, we further evaluated the multi-domain classification accuracy of the PMoE. With the same network structure and training setup, the average accuracy shows only a slight reduction, decreasing from 97.1% in the original thermo-optic experiment to 96.8% under the 10 GHz matched condition, corresponding to a drop of about 0.3 percentage points, as shown in Fig. S5c. These results indicate that the PMoE network for multi-domain processing remains robust to the level of signal degradation observed in high-speed deployment, and provide preliminary support for future implementations for high-throughput optical computing.

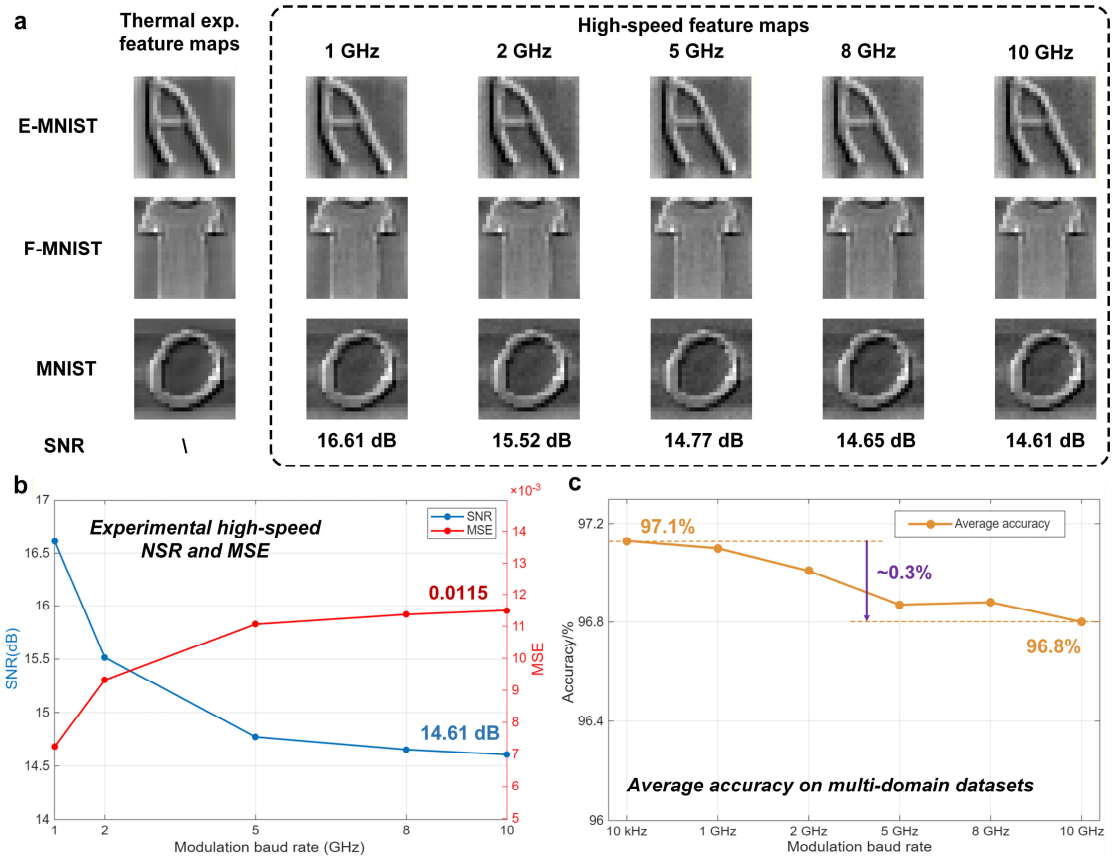


Fig. S5.

Experimentally anchored evaluation of PMoE robustness under high-speed deployment conditions. a, Thermo-optic experimental feature maps and the corresponding feature maps with matched degradation under 1, 2, 5, 8, and 10 GHz conditions. **b,** SNR and MSE used to emulate the experimentally measured high-speed signal degradation. **c,** Average classification accuracy of the PMoE system on the multi-

domain datasets under matched high-speed degradation conditions.

R2-2. *Regarding the "task-specific weighted routing" mentioned in Fig. 4a, could the authors clarify the physical mechanism of the routing? Is it implemented by simply adjusting input power levels via thermo-optic modulators, or is there a dedicated optical switching stage? These details should be made clearer.*

A2-2. We thank the reviewer for the insightful comment. We agree that the physical mechanism of the routing process was not sufficiently explained in the original manuscript, and we appreciate the opportunity to clarify it. In the current experimental setup, the "task-specific weighted routing" is implemented by adjusting the input optical power delivered to each expert, rather than by using a separate dedicated optical switching stage. In this sense, the routing in our present system should be understood as task-specific power weighting at the input stage, rather than physical path switching inside the optical network.

Specifically, the current implementation works as follows. We first identify the linear modulation regime of the on-chip thermo-optic modulators by applying a fixed bias voltage. The input data are then mapped to the corresponding driving voltages. To impose the task-specific routing weights for the three experts, these voltage signals are passed through a parallel voltage-divider circuit. This electrical configuration directly scales the driving voltages of the modulators and therefore adjusts the input optical power delivered to each expert. In this way, the required task-specific weighting is realized in the present PMoE system.

We also agree with the reviewer that, as the system scales, more sophisticated optical routing mechanisms would be beneficial. In future scaled-up versions of PMoE, the routing could be implemented more efficiently in the optical domain. For example, a "modulate-then-split" scheme could be adopted, in which a single set of modulators is followed by tunable optical splitters (e.g., cascaded Mach-Zehnder interferometers) to distribute weighted optical power among experts. Alternatively, dedicated high-efficiency optical switch arrays could also be used to realize more advanced routing and weighting functions.

We have revised the manuscript to clearly describe the current physical routing mechanism and to briefly discuss the possibility of dedicated optical routing stages in future implementations.

Relevant changes made:

- Revised manuscript (Results, Page 9, Lines 263-266): "This task-specific weighted-routing is implemented by adjusting the input optical power levels with a parallel voltage divider. This configuration directly scales the driving voltages of the on-chip thermo-optic modulators, thereby adjusting the input optical power delivered to each expert."
- Revised manuscript (Discussion, Page 10, Lines 326-330): "Future implementations can further extend this scaling range by adopting advanced

dynamic optical routing mechanisms, such as modulate-then-split configurations using tunable optical splitters or the integration of highly efficient optical switch arrays. These routing schemes, coupled with sparsity-aware network algorithms, can further improve the scalability of the PMoE architecture."

- Revised Supplementary Information (Note 9, Page 11): "The task-specific weighted-routing is implemented by adjusting the input optical power levels with a parallel voltage divider, which are input into one arm of the MZI modulator in the PMoE chip via metal wire leads."

R2-3. *The authors advocate for a "width-over-depth" approach. However, width scaling faces its own bottlenecks, such as I/O count and optical splitting losses. Is there a theoretical limit to the number of experts before the signal-to-noise ratio degrades the classification accuracy, given the laser power splitting shown in Fig. 1? A comment on this trade-off would add depth to the discussion.*

A2-3. We greatly appreciate the reviewer's insightful comment. We agree that width scaling is subject to important physical trade-offs, particularly those associated with optical splitting loss and the signal-to-noise ratio (SNR) budget. While the "width-over-depth" strategy alleviates the limitations of deep optical networks, such as cumulative loss, error accumulation, and active-control complexity, its scalability is still constrained by available optical power, insertion loss, and detector sensitivity.

Based on the reviewer's suggestion, we have added a quantitative link-budget analysis to estimate this practical limit. Assuming a realistic on-chip laser input power of 15 dBm, a photodetector sensitivity threshold of −20 dBm, and an estimated total loss of 15 dB for each expert subnetwork, the remaining optical power budget available for splitting is 20 dB. Under these assumptions, this corresponds to an order-of-magnitude estimate of approximately 100 concurrently active experts. We have added this discussion to the revised manuscript to make the physical trade-off more explicit.

Importantly, because the MoE architecture relies on sparse task-specific routing rather than activating all subnetworks simultaneously, the total number of physical experts integrated on a chip can exceed the number of concurrently active experts. In this sense, the concurrent-activation limit imposed by direct optical splitting does not necessarily define the total expert capacity of the architecture, but rather the number of experts that can be efficiently activated at the same time under the present link-budget assumptions.

We also agree that more advanced optical routing mechanisms would be beneficial for further extending this scaling range. In future implementations, direct power splitting can be replaced or supplemented by more efficient routing schemes, such as a modulate-then-split architecture using tunable optical splitters, or the integration of highly efficient optical switch arrays. These approaches, especially when combined with sparsity-aware routing algorithms, can further improve the scalability of the PMoE framework.

To explicitly address this important trade-off and quantify the practical scaling pathway, we have added a dedicated paragraph in the Discussion section of the revised manuscript.

Relevant changes made:

- Revised manuscript (Discussion, Page 10, Lines 318-326): "While the width-over-depth scaling paradigm offers significant advantages in mitigating the limitations of deep optical neural networks, width scaling via optical splitting faces its own physical constraints. Assuming a typical on-chip laser input power of 15 dBm, a photodetector sensitivity threshold of −20 dBm, and an estimated total loss of 15 dB per expert subnetwork, the remaining optical power budget available for splitting is 20 dB. Under an equal-power splitting condition, the allowable number of concurrently active optical experts is on the order of 100. Because the MoE architecture relies on sparse task-specific routing rather than activating all subnetworks simultaneously, the total number of physical experts integrated on the chip can exceed this concurrent-activation estimate."
- Revised manuscript (Discussion, Page 10, Lines 326-330): "Future implementations can further extend this scaling range by adopting advanced dynamic optical routing mechanisms, such as modulate-then-split configurations using tunable optical splitters or the integration of highly efficient optical switch arrays. These routing schemes, coupled with sparsity-aware network algorithms, can further improve the scalability of the PMoE architecture."

R2-4. *As the authors acknowledge, photonic components in current computing systems typically perform a linear operation. This implies that before the detection stage, the effective depth of the neural network remains 1 regardless of the number of optical layers. While linear transformations can perform powerful computation, it would be necessary to clarify the benefit of physical depth and the addition of diffractive layers.*

A2-4. We sincerely appreciate your important theoretical comment. We agree that, in the absence of optical nonlinearity within the diffractive layers, the entire optical front-end before detection is still mathematically equivalent to a single linear transformation. In this sense, the functional depth of the optical front-end remains effectively 1 prior to photodetection.

However, the benefit of physical depth lies not in introducing additional nonlinearity before detection, but in enlarging the set of linear transformations that can be practically realized under the physical constraints of diffractive optics. In a diffractive network, the target transformation matrix is not implemented directly. Instead, it is synthesized through the cascaded multiplication of propagation matrices and phase-modulation diagonal matrices. Therefore, multiple diffractive layers provide a more flexible constrained factorization of the target linear operator, increasing the design degrees of freedom and improving the fidelity with which complex linear mappings can be approximated.

As discussed in previous works [1–4], although the overall optical operation remains linear, increasing the number of diffractive layers generally improves the optimization flexibility and the representation capability of the physically implementable linear transformation. In other words, a multi-layer diffractive structure does not change the fact that the system is linear before detection, but it can substantially improve how well the desired linear operator is realized in practice. This trend has been reported in prior literature, as illustrated in **Fig. R2-3** [1], **Table R2-1** [2], and **Fig. R2-4** [3].

In our implementation, each diffractive expert functions as a multi-core linear convolution layer. As detailed in our previous analysis of optical convolution architectures, the final performance is closely related to structural hyperparameters, particularly the number of diffractive layers and the neuron count per layer. Guided by these observations, we balanced matrix-mapping fidelity against optical implementation complexity and selected a 2-layer diffractive expert with 50 neurons per layer in the present PMoE design.

To make this design rationale clearer to readers, we have added a more explicit explanation of the role of physical depth and the corresponding parameter selection in the revised Supplementary Information.

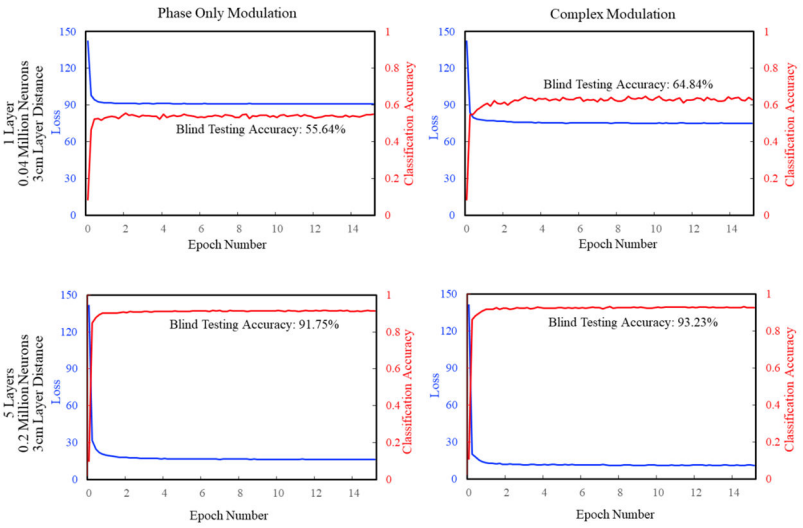


Fig. R2-3. MNIST training convergence plots of a phase-only modulation D^2NN (left column) and a complex valued (i.e., phase and amplitude) modulation D^2NN (right column) as a function of the number of diffractive layers (N = 1 and 5) and the number of neurons used in the network by Lin et al. [1].

Table. R2-1. Prediction accuracy of various on-chip DONNs on the iris blind test sets by Fu et al. [2].

DONN with m hidden layers (DONN-Im)	Accuracy
On-chip DONN-I1	86.7%
On-chip DONN-I2	86.7%
On-chip DONN-I3	90%
On-chip DONN-I4	90%
On-chip DONN-I5	90%

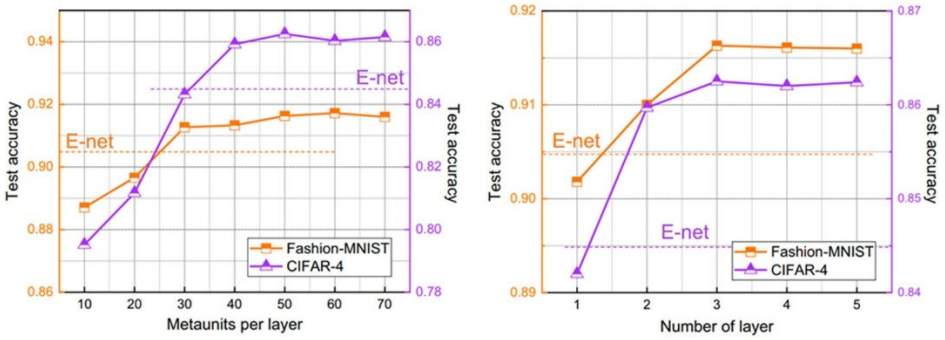


Fig. R2-4. Classification performance evaluations on both data sets with respect to two main physical parameters of OCU: the number of metaunit per layer and the number of the exploited metaline layer by Huang et al. [3].

Reference:

1. X. Lin et al., All-optical machine learning using diffractive deep neural networks. Science 361, 1004-1008 (2018).

2. T. Fu et al., Photonic machine learning with on-chip diffractive optics. Nature Communications 14, 70 (2023).

3. Y. Huang et al., Sophisticated deep learning with on-chip optical diffractive tensor processing. Photonics Research 11, 1125 (2023).

4. W. Liu et al., Ultra-compact multi-task processor based on in-memory optical computing. Light: Science & Applications 14, 134 (2025).

Relevant changes made:

- Revised Supplementary Information (Note 2, Page 3): “In our implementation, each diffractive expert functions as a multi-core linear convolution layer. Although the overall optical front-end before detection is mathematically equivalent to a linear transformation, the number of diffractive layers remains important because the target operator must be realized through a cascaded factorization of propagation and phase-modulation matrices. As a result, the final performance is closely related to structural hyperparameters, particularly the number of diffractive layers and the neuron count per layer. Guided by these considerations, we balanced matrix-mapping fidelity against optical implementation complexity and selected a 2-layer diffractive expert with 50 neurons per layer in the PMoE design.”

R2-5. *Given the ultra-compact footprint (0.06 mm2) and the use of thermo-optic encoding, was there any observable thermal crosstalk between adjacent diffractive neurons or modulators? If so, how was it mitigated during the calibration process?*

A2-5. We thank the reviewer for raising this practical concern regarding thermal crosstalk, which is indeed an important issue in thermo-optic devices. In the present PMoE chip, the diffractive neurons themselves are passive and do not rely on local thermal tuning. Therefore, the main concern for thermal crosstalk is associated with the adjacent thermo-optic modulators rather than the passive diffractive neurons. Based on the stability of the calibrated modulation response and the good agreement between the numerical and experimental optical feature maps (**Fig. R2-5**), we did not observe obvious performance degradation attributable to thermal crosstalk under the operating conditions used in this work.

We suppressed thermal crosstalk through a combination of device layout and thermal-management strategies:

1. Low power operation: The electrical power applied to the on-chip modulators was kept relatively low, thereby reducing heat generation at the source.

2. Spatial isolation: The device layout provides sufficient physical spacing between adjacent phase shifters to reduce thermal interaction.

3. Active thermal management: The chip was tightly bonded to a highly thermally conductive metal sub-mount and operated with a thermistor-monitored TEC feedback loop. This configuration provides efficient heat dissipation and helps maintain a stable chip temperature during operation.

During calibration, the chip temperature was first stabilized by the TEC feedback loop, and the modulator response curves were measured under this stabilized condition before mapping the input data to the corresponding driving voltages. This procedure suppresses thermal drift during calibration and helps ensure that any residual thermal interaction remains small during operation. Under these conditions, no additional crosstalk-compensation step was required in our experiments.

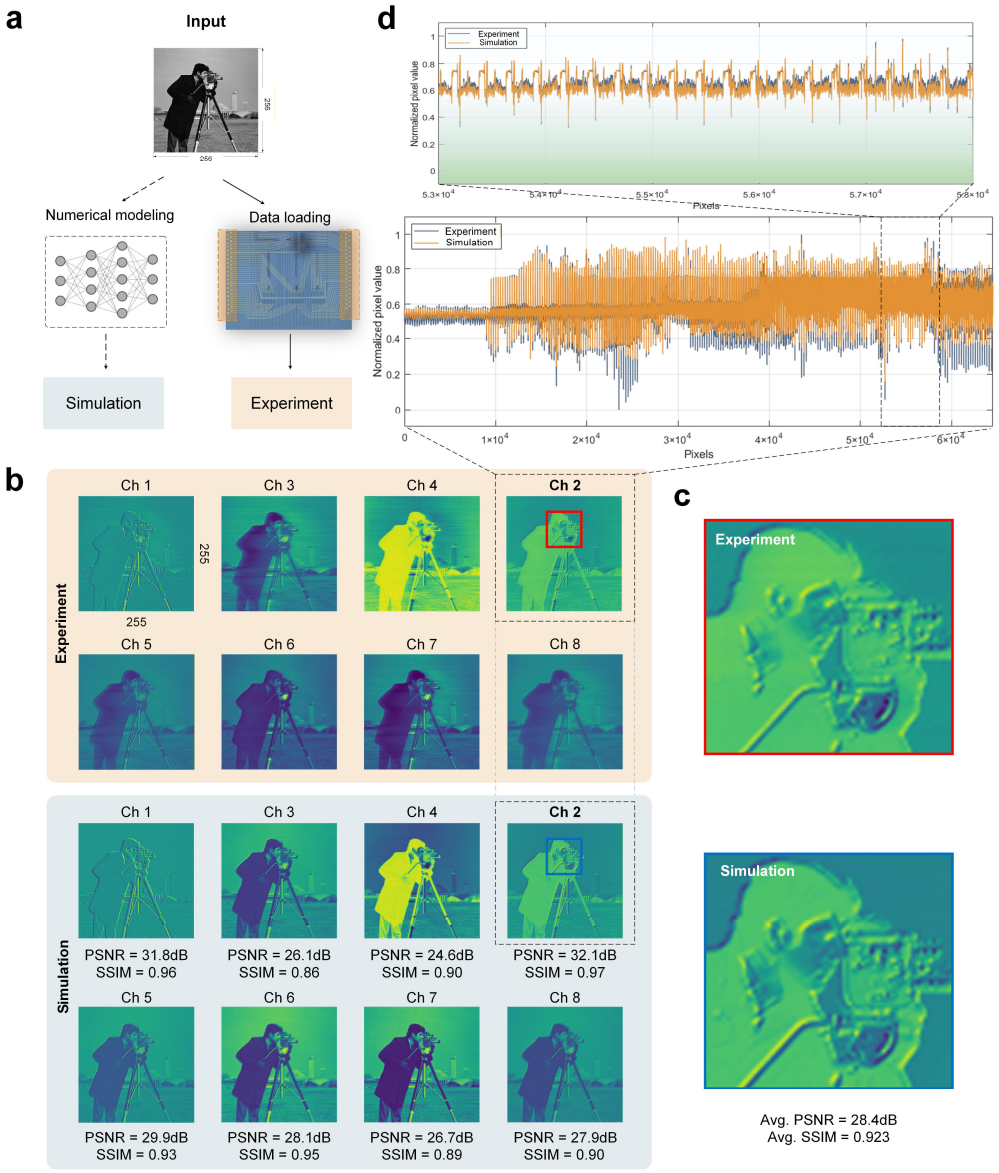


Fig. R2-5. Experimental validation of the 8-channel photonic expert network. a, Comparison diagram of the numerical simulation and the 8-ch PEN outputs. **b,** Output feature maps between the numerical simulation and the 8-ch PEN. **c,** Detailed comparison of a specific region. **d,** Pixel-wise comparison between the experimental and simulated results.

R2-6. *Fig. 2 Caption (Page 6, line 181): "The insect is the error distribution histogram" probably should be "The inset is".*

A2-6. We appreciate your careful reading. We apologize for this typographical error and have corrected "insect" to "inset" in the caption of Fig. 2 in the revised manuscript.

Relevant changes made:

- Revised manuscript (Fig. 2, Page 6, Lines 192-193): "The inset is the error distribution histogram with RMSE of 0.042."

R2-7. *Page 9, line 267: "achieving a 67% reduction in compared to conventional" probably should be "reduction compared to".*

A2-7. We appreciate your careful reading. We have removed the extra word and corrected the phrase to "…reducing the number of network parameters by 67%." in the revised manuscript.

Relevant changes made:

- Revised manuscript (Results, Page 9, Line 296-298): "Compared with separate electronic or optical task-specific networks on the same multi-domain datasets, PMoE achieved higher accuracy while reducing the number of network parameters by ~67%."

R2-8. *Page 9, line 294: "scale can be further extended by expand the number" probably should be "by expanding".*

A2-8. We appreciate your careful reading. We have revised the text to "…can further improve the scalability of the PMoE architecture." in the revised manuscript.

Relevant changes made:

- Revised manuscript (Discussion, Page 10, Line 329-330): "These routing schemes, coupled with sparsity-aware network algorithms, can further improve the scalability of the PMoE architecture."

Point-by-Point Responses to Reviewer #3's Comments

R3-0. *In their work, Photonic Mixture-of-Experts for large-scale on-chip optical neural networks, the authors demonstrate a hybrid MoE network for MNIST/Fashion-MNIST/Extended-MNIST classification. The "MoE part" of the network is executed optically using an integrated, diffractive architecture operating at 10 kHz. Applying the concept of MoE to photonic processors is interesting, the experimental chip fabrication and packaging is impressive, and the results are well illustrated. However, the positioning of the work (and the advance with respect to the state of the art) is not fully clear and in its current form the manuscript contains not sufficiently backed up claims. Thus, major revisions are necessarily to be suitable for publications.*

A3-0. We sincerely thank the reviewer for taking the time to review our manuscript and for recognizing the interest of applying the MoE concept to photonic processors, as well as the experimental chip fabrication, packaging effort, and presentation of the results. We also appreciate the constructive feedback regarding the positioning of our work with respect to the state of the art and the need to better support several of our claims. In response, we have made substantial revisions to the manuscript to clarify the scope and contribution of the work, strengthen the supporting analysis, and improve the overall rigor of the presentation. Please find our detailed point-by-point responses below.

R3-1. *About the positioning of the work:*
The authors make strong hardware related claims regarding "integration level / density" and "on-chip parallel kernels", but do not provide the experimental evidence for a significant advance over the SOTA. For example, the system only operates at 10 kHz and the comparison to the SOTA does not have a sufficient depth. E.g. the work of Feldmann et al. cited in Fig 5 / Table 1 uses WDM in a crossbar array to compute several MVMs (different input vector encodings using the same physical matrix) in parallel, implementing a fundamentally different operation than the one presented in the paper.

A3-1. We sincerely appreciate your critical and insightful evaluation. We agree that clarifying the operating speed, the definition of integration-level metrics, and the scope of our SOTA comparison is essential for accurately positioning the novelty and contribution of the PMoE architecture.

1. Clarification on the integration level and computational density:

In Fig. 5 (**Fig. R3-1** in this response), the "**integration level**" is intended to describe a spatial structural metric of the computational core, rather than a full-system performance metric. In this sense, it reflects the fundamentally physical footprint required for the inner optical computing core itself, independent of the modulation rate. Our main claim here is that, by introducing MoE-style width expansion into diffractive optical cores, we can integrate 18 coordinated equivalent convolution kernels within a passive computational-core area of 0.06 mm^2 for multi-domain processing. We agree, however, that this metric should not be interpreted as a complete system-level comparison, and we have revised the manuscript accordingly to clarify this scope.

In Table 1, by contrast, the "**Computational density**" (TOPS/mm²) is a spatiotemporal throughput-related metric. We agree that the current PMoE experiment operates at 10 kHz due to the use of thermo-optic modulators for signal loading. Therefore, the 51.0 TOPS/mm² value should not be interpreted as directly measured performance of the present thermo-optic PMoE chip. Instead, it represents a projected high-speed electro-optic metric under 10 GHz modulation. Since high-speed electro-optic modulators are mature [1-2]; these are available technologies and commonly used for scaling projection in previous optical computing literatures [3-4]. We have revised the Introduction, Discussion, and Supplementary note 10 to explicitly distinguish the experimentally measured 10 kHz performance from the projected 10 GHz electro-optic performance. In addition, as discussed in our response **A3-9**, we added an experimentally anchored high-speed robustness evaluation to provide intermediate-rate support for this projection.

2. The unified MVM abstraction and the calculation of the parallel kernels metric:

The reviewer raised a valid point regarding SOTA architectures that compute several matrix-vector multiplications (MVMs) in parallel, suggesting that they implement different physical operations and dataflow mechanisms. We agree that architectures such as WDM crossbars, MRR arrays, MZI meshes, and diffractive processors are not physically identical. However, from a foundational mathematical perspective, their passive optical computing cores can be commonly represented as an end-to-end linear MVM when treated at the input-output port level. Specifically, for a fixed set of input and output ports, the operation can be written as $y = \boldsymbol{K}x$, x is the input vector, $\boldsymbol{K}$ is the equivalent optical weight matrix implemented by the hardware, and y is the output vector. Therefore, our comparison is performed under this MVM-level abstraction, rather than assuming that all platforms share the same physical mechanism or dataflow.

Building on this MVM abstraction, the "**On-chip parallel kernels**" metric (the horizontal axis in **Fig. R3-1b**) represents the equivalent mathematical decomposition of the simultaneously available optical weight matrix $\boldsymbol{K}$ into multiple localized vector inner products, which spatially correspond to parallel convolution kernels. Since a 3×3 convolution kernel corresponds to a 9-dimensional vector inner product, an optical matrix with N_{in} input channels and N_{out} simultaneous output channels can be normalized as $N_{in}N_{out}/9$ equivalent 3×3 parallel kernels. For WDM or multi-channel platforms, only physically simultaneous channels are counted, while temporal reuse or sequential loading is not included in this metric.

For instance, in the WDM-based crossbar architecture [5], the authors explicitly state that their implementation uses "four image kernels of size 3×3 (corresponding to a 9×4 filter matrix)". Following the same equivalent-kernel normalization, our PMoE experts with matrix dimensions of 9×4, 9×6, and 9×8 correspond to 4, 6, and 8 equivalent 3×3 parallel kernels, respectively, yielding a total of 18 equivalent parallel kernels. As theoretically derived in Supplementary note 1 (Eq. S1), the spatial convolution decomposition of the PMoE is formulated as:

$$\hat{O}=\begin{bmatrix} O_1 \\ O_2 \\ \vdots \\ O_m \end{bmatrix}=\begin{bmatrix} k_{11} & k_{21} & \cdots & k_{m,1} \\ k_{12} & k_{22} & \cdots & k_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1,H^2} & k_{2,H^2} & \cdots & k_{m,H^2} \end{bmatrix}^T \circ \begin{bmatrix} i_{11} & i_{12} & \cdots & i_{1,N^2} \\ i_{21} & i_{22} & \cdots & i_{2,N^2} \\ \vdots & \vdots & \ddots & \vdots \\ i_{H^2,1} & i_{H^2,2} & \cdots & i_{H^2,N^2} \end{bmatrix} = K \circ \hat{I}$$

where m represents the number of spatial output ports, and $\boldsymbol{K}$ denotes the equivalent multi-convolutional kernel matrix. $\hat{I}$ is the input image patch matrix, where each column represents a flattened $H\times H$ (3×3 in this work) image patch processed sequentially from an $N\times N$ image. O_m represents the corresponding convolution result output from the m-th spatial port. In the PMoE, the three parallel photonic expert networks (PENs) are configured with 9 input ports and 4, 6, 8 output ports, respectively, yielding a total of 18 independent parallel optical kernels. Based on this principle, we compared the PMoE with other SOTA platforms under the same standard of inner parallel kernels in **Fig. R3-1b**.

To ensure this positioning and mathematical equivalence are more transparent, we have revised the Discussion and Supplementary Information to make comparison under a unified MVM framework. Furthermore, we incorporated 10 GHz theoretical extrapolations into the Introduction and Discussion to rigorously clarify the architecture's intrinsic spatiotemporal potential. Finally, detailed equation descriptions and step-by-step capacity calculations were added to the Supplementary Information.

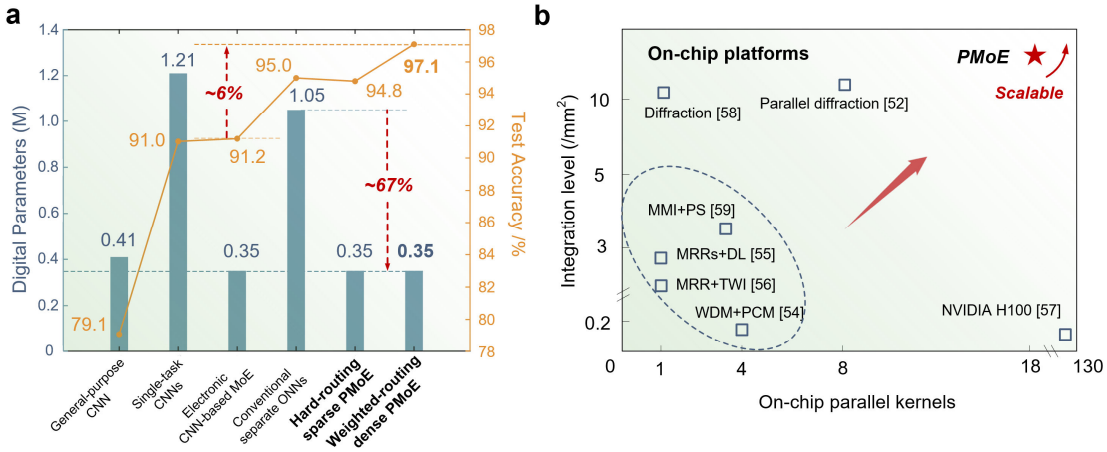


Fig. R3-1. Comparison of multi-domain and on-chip computing platforms. a, Experimental comparison between PMoE and the other multi-domain platforms. PMoE reduces the digital parameter count by ~67% compared with conventional separate ONNs and improves the average accuracy by ~6% compared with an electronic CNN-based MoE with a similar architectural design. **b,** Comparison of PMoE with other on-chip computing platforms in terms of computational-core integration level and equivalent parallel kernels.

Reference:

1. C. Han et al., Slow-light silicon modulator with 110-GHz bandwidth. Science Advances 9, eadi5339 (2023).

2. M. Xu et al., High-performance coherent optical modulators based on thin-film lithium niobate platform. Nature communications 11, 3911 (2020).

3. C. Wang, Y. Cheng, Z. Xu, Q. Dai, L. Fang, Diffractive tensorized unit for million-TOPS general-purpose computing. Nature Photonics, 1-10 (2025).

4. J. Cheng et al., Multimodal deep learning using on-chip diffractive optics with in situ training capability. Nature Communications 15, 6189 (2024).

5. J. Feldmann et al., Parallel convolutional processing using an integrated photonic tensor core. Nature 589, 52-58 (2021).

Relevant changes made:

- Revised manuscript (Introduction, Page 3, Lines 105-109): "Facilitating substantial width-wise scalability, the PMoE experimentally demonstrates efficient multi-domain optical processing, while reaching a projected computational density of 51.0 TOPS/mm² under 10 GHz electro-optic modulation."
- Revised manuscript (Discussion, Page 9, Lines 302-310): "To provide a normalized architectural comparison across different computing platforms, we evaluate the compared computing cores under a common matrix-vector multiplication (MVM) abstraction. At the input-output port level, these linear computing cores can be represented as an equivalent weight matrix. Under this abstraction, the integration level (/mm²) and the number of parallel kernels serve as crucial metrics for evaluating the physical parameter density and scalability. As shown in Fig. 5b, these on-chip parallel kernels are derived through the equivalent mathematical decomposition of the global weight matrix into localized spatial convolutions. These results support the feasibility of the MoE-based width-scaling strategy for improving structural parameter capacity and network scalability in optical systems."
- Revised manuscript (Discussion, Page 10, Lines 331-338): "In the experimental demonstration, we utilized on-chip thermo-optic modulation for signal loading at up to 10 kHz, corresponding to an experimental computational capacity of 3.06 MOPS and a unit computational capacity of 51 MOPS/mm^2 based on the 0.06 mm^2 computational-core area. For high-speed electro-optic implementation, the projected computational capacity of PMoE would reach 3.06 TOPS, with a unit computational capacity of 51 TOPS/mm^2 under 10 GHz modulation (Supplementary note 10). Including the full transmit/receive chain in the system model, the power consumption for PMoE can be estimated as 4.94 W, resulting in 10.3 TOPS/W/mm^2 under the same computational-core area (Table S1)."
- Revised manuscript (Discussion, Page 10, Lines 338-341): "We further evaluated the PMoE architecture under high-speed degradation conditions. The average classification accuracy decreases only slightly from 97.1% to 96.8% under the 10 GHz matched condition (Fig. S5), providing preliminary support for high-speed electro-optic implementations."

- Revised Supplementary Information (Note 1, Page 2): "where m represents the number of spatial output ports, and $\boldsymbol{K}$ denotes the equivalent multi-convolutional kernel matrix. $\hat{I}$ is the input image patch matrix, where each column represents a flattened $H\times H$ (3×3 in this work) image patch processed sequentially from an $N\times N$ image. O_m represents the corresponding convolution result output from the m-th spatial port."

- Revised Supplementary Information (Note 6, Pages 7–8):

  We added a high-speed robustness evaluation to support the 10 GHz electro-optic projection. Experimentally measured signal degradation at different baud rates was introduced into the PMoE feature maps, and the corresponding multi-domain classification accuracy was evaluated under matched high-speed degradation conditions.

- Revised Supplementary Information (Note 10, Pages 14-15): "The PMoE chip we fabricated includes 9 input channels corresponding to a 3×3 processing core, along with 3 diffractive experts. Photon-based computing systems face significant challenges in achieving non-linear operations, so regardless of scale and depth, the processing can ultimately be characterized as a linear operation. This is the rationale for our focus on expanding the network's width, rather than its depth. Unlike previous methods that represent computation as a series of fully connected layers, for the linear optical front-end, the input-output relation should be represented as an end-to-end linear transformation when treated at the black-box level. Under this abstraction, each PEN can be expressed as an equivalent set of convolutional kernels with dimensions 3×3×4, 3×3×6, and 3×3×8, respectively, corresponding to a total of 18 equivalent parallel kernels for the PMoE system."

- Revised Supplementary Information (Note 10, Page 15): "Consequently, under the 10 kHz thermo-optic modulation, the computational capacity of the PMoE chip is 3.06 MOPS, corresponding to 3.06 MOPS/0.06 mm^2 = 51 MOPS/mm^2 based on the computational-core area. Assuming the integration of plasma-dispersion-based silicon modulators with a modulation rate of 10 GHz, the computational density is estimated to reach 51.0 TOPS/mm^2 based on the 0.06 mm^2 computational-core area, or 0.12 TOPS/mm^2 under an estimated total footprint of 25 mm^2."

R3-2. *Similarly, the authors do not provide an in-depth analysis of the photonic MoE idea that would present a significant advance with respect to the SOTA. The combination of MoE and analog processors has been already studied (e.g. Büchel, J., Vasilopoulos, A., Simon, W.A. et al. Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing. Nat Comput Sci 5, 13–26 (2025). https://doi.org/10.1038/s43588-024-00753-x). Implementing this concept in the photonic domain is interesting but then should be properly analyzed from a neural network perspective (especially benchmarking against arbitrary digital CNNs baselines is odd). Similarly, it's not sufficiently supported why the combination with photonics is good (usually the goal of MoE is to reduce compute for large models, not parameters. In contrast, photonics usually struggles to provide the necessary number*

of programmable weights rather than providing sufficient compute).

A3-2. We sincerely appreciate your insightful comments. We agree that the motivation for combining MoE with photonic hardware differs from that in digital large-scale models. We also thank the reviewer for directing us to the important work by Büchel et al., which demonstrates the synergy between analog in-memory computing and MoE architectures.

Following your constructive suggestions, we have revised the manuscript to provide a clearer analysis of the architectural motivation of PMoE and a more appropriate benchmarking methodology.

1. Empowering optical neural networks via the MoE architecture:

The reviewer raised an important point: in digital large models, MoE is usually introduced to reduce compute, whereas photonic computing is often limited by the difficulty of scaling programmable optical weights rather than by insufficient compute throughput. We agree with this assessment, and we have revised the manuscript to clarify that this distinction is precisely the motivation for our PMoE architecture.

Currently, conventional optical neural networks face critical bottlenecks in **scaling effective parameter scalability** and **handling multi-domain tasks** [1–4]. Expanding a single optical network through depth encounters inherent linearity, error accumulation, and increased active-control complexity. In addition, scaling fully active programmable optical weights, such as those based on MZIs or microring resonators, typically requires dense electrical control, calibration, and stabilization. For multi-domain processing, conventional approaches often require either multiple independent hardware networks or reconfiguration of network parameters, which increases deployment and control complexity.

The PMoE is introduced to alleviate these bottlenecks through a width-over-depth strategy. Instead of increasing the depth of a single optical network, we integrate multiple passive, pre-programmed photonic expert networks and coordinate them through a small number of task-specific routing weights. In this way, the architecture increases the effective parameter capacity for multi-task optical inference without requiring active programming of all optical weights during operation. More specifically, although the diffractive weights are not dynamically reprogrammable after fabrication, coordinating multiple fixed photonic experts through lightweight routing provides a practical route to expand the effective parameter and multi-task capacity while avoiding the active-control burden associated with fully programmable optical weights.

2. Rigorous benchmarking and the rationale for our baselines:

Regarding the baseline selection, we agree that benchmarking only against an arbitrary digital CNN would be insufficient from a neural-network perspective. Therefore, we have revised the comparison and ablation analysis to clarify the role of each baseline. The purpose of these baselines is not to claim state-of-the-art classification accuracy on MNIST, Fashion-MNIST, or Extended-MNIST as standalone

tasks, but to evaluate the effectiveness of the PMoE architecture under a controlled multi-domain setting.

Specifically, we use identical datasets and training/test splits, and compare PMoE with several reference architectures under similar architectural constraints and comparable parameter budgets (**Fig. R3-2a**). These include separate task-specific CNNs, a unified CNN baseline, conventional separate ONNs, and a newly added CNN-based electronic MoE baseline. The CNN-based electronic MoE follows the same overall multi-expert/shared-backend design principle as PMoE: task-specific expert convolutional layers are used in the front-end, while most backend parameters are shared across the three tasks, as detailed in our response **A3-6**.

This newly added electronic MoE provides a more appropriate architecture-level electronic baseline than the original auxiliary digital CNN. It achieves an average classification accuracy of 91.2%, confirming that the expert-based front-end and shared-backend design is itself effective for multi-domain classification. Compared with this electronic MoE counterpart, the final PMoE achieves about 6 percentage points higher average accuracy, while maintaining the same overall multi-expert/shared-backend design principle (**Fig. R3-2b**). This comparison supports that the performance advantage of PMoE does not arise merely from using an MoE-style architecture, but from combining this architecture with the photonic expert front-end.

3. State-of-the-art of MoE combined with optics:

We fully acknowledge that the integration of MoE with analog computing and optical systems is an emerging and active direction. As now discussed in the revised manuscript, recent works have explored MoE-related optical or analog systems, including analog in-memory computing with MoE, free-space optical classifier ensembles, theoretical simulations of fully optical MoE systems, and optical interconnects for electronic GPU-based MoE training.

Our work is positioned differently. Rather than using optics only as an interconnect or presenting a simulation-only MoE design, we experimentally demonstrate an on-chip PMoE architecture in which the photonic expert front-end performs optical feature extraction for multi-domain inference. The key point is not that MoE itself is new, but that the MoE design philosophy is mapped onto the physical strengths of passive diffractive photonics: high-density and scalable pre-programmed optical weights, and parallel expert-wise processing for multi-task inference.

To make this positioning clearer, we have revised the Abstract, Introduction, Discussion, and Supplementary note 8. We now explicitly clarify the physical motivation for photonic MoE, cite related analog/optical MoE works, and expand the ablation analysis to include the controlled CNN-based electronic MoE baseline.

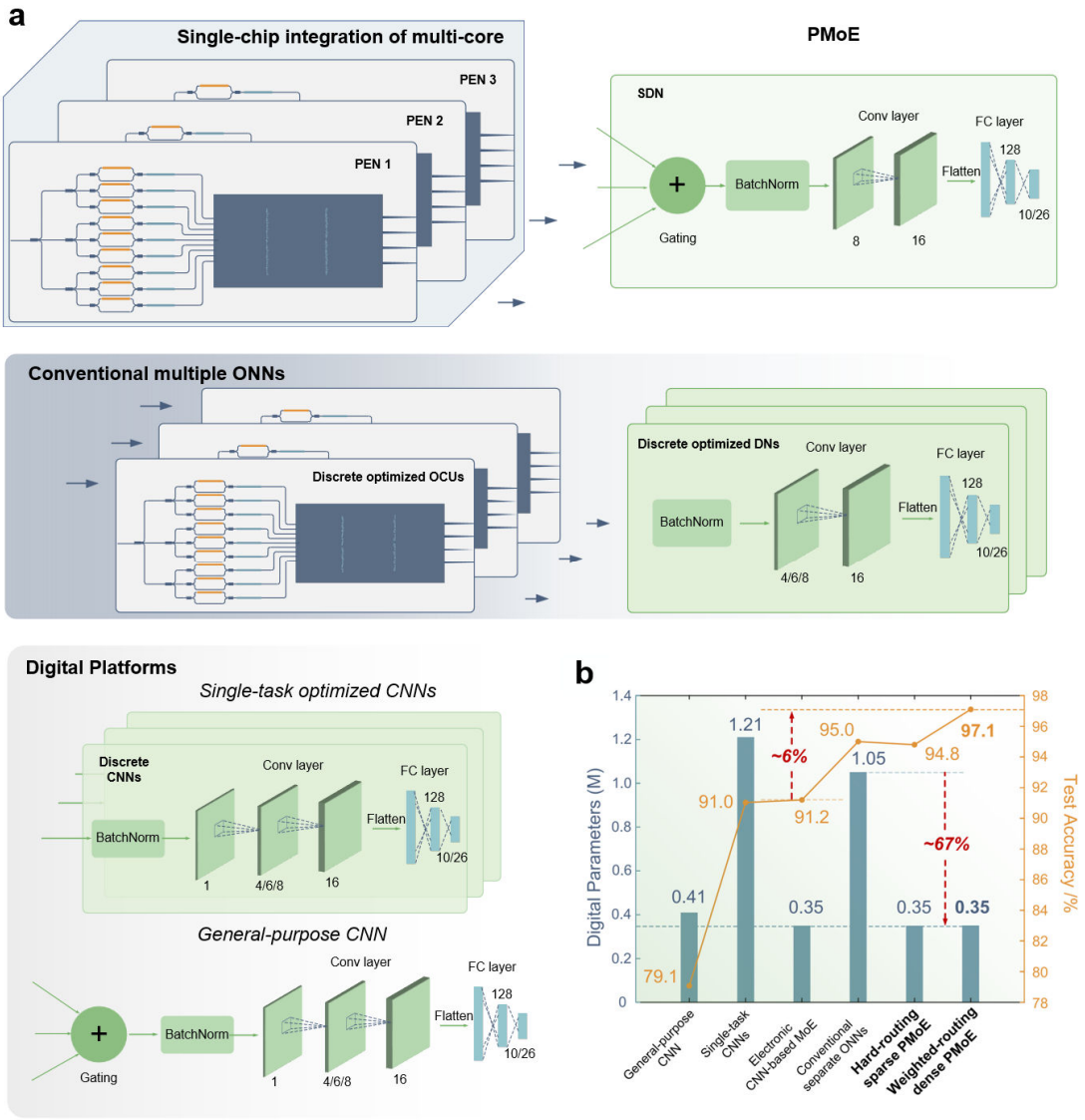


Fig. R3-2. Benchmarking of the PMoE architecture against conventional optical and digital platforms. a, Schematic comparisons of the evaluated network architectures. To ensure a fair ablation study, the baseline digital platforms (single-task and general-purpose CNNs) and the conventional optical platform are constructed with identical layer configurations as the PMoE. **b,** Experimental comparison between PMoE and the other multi-domain platforms. PMoE reduces the digital parameter count by ~67% compared with conventional separate ONNs and improves the average accuracy by ~6% compared with an electronic CNN-based MoE with a similar architectural design.

Reference:

1. O. Kulce et al., All-optical information-processing capacity of diffractive surfaces. Light: Science & Applications 10, 25 (2021).

2. H. Zhou et al., Photonic matrix multiplication lights up photonic accelerator and beyond. Light: Science & Applications 11, 30 (2022).

3. W. Liu et al., Ultra-compact multi-task processor based on in-memory optical computing. Light: Science & Applications 14, 134 (2025).

4. N. Wu, et al., Intelligent nanophotonics: when machine learning sheds light. eLight 5, 5 (2025).

Relevant changes made:

- Revised manuscript (Abstract, Page 1, Lines 24-27): "Photonic computing offers an energy-efficient, high-bandwidth platform for artificial intelligence (AI) but currently faces scalability bottlenecks stemming from depth-dependent designs, linear optical structures, and intrinsic optical losses, along with high hardware and reconfiguration costs for multi-domain processing."

- Revised manuscript (Abstract, Page 1, Lines 30-32): "By dynamically routing inputs to these experts, the PMoE efficiently executes diverse multi-task workloads without altering the physical optical weights."

- Revised manuscript (Introduction, Page 2, Lines 78-89): "To circumvent these limitations while capitalizing on the intrinsic efficiency and parallelism of photonics, we introduce a parameter scaling and multi-task processing method. Drawing inspiration from the digital MoE, we propose a width-over-depth paradigm that adapts this architecture to the optical domain. Related efforts have also explored the combination of MoE with analog processors [41]. By routing diverse inputs into multiple photonic experts, this approach offers a robust pathway to circumvent the depth-scaling parameter bottleneck and efficiently execute diverse multi-task workloads without reconfiguring physical weights. Although recent works have explored the synergy between optics and the MoE concept, ranging from free-space optical classifier ensembles [42] and theoretical simulations of hybrid systems [43] to deploying optical interconnects for electronic GPUs [44], an experimentally validated on-chip photonic MoE computing architecture remains lacking."

- Revised manuscript (Introduction, Page 3, Lines 105-109): "Facilitating substantial width-wise scalability, the PMoE experimentally demonstrates efficient multi-domain optical processing, while reaching a projected computational density of 51.0 TOPS/mm² under 10 GHz electro-optic modulation. This parallel on-chip framework paves the way for large-scale optical neural networks and next-generation photonic AI processors."

- Revised manuscript (Discussion, Page 9, Lines 292-301): "For a rigorous ablation, we compared the PMoE against multiple optical and electronic baselines under the same multi-domain setting, with comparable network scale and similar architectural constraints (Supplementary note 8). Fig. 5a shows the controlled comparison curves between PMoE and other platforms in terms of average classification accuracy and the number of parameters in the electronic network. Compared with separate electronic or optical task-specific networks on the same multi-domain datasets, PMoE achieved higher accuracy while reducing the number

of network parameters by ~67%. Compared to an electronic CNN-based MoE counterpart with the same overall design principle, the PMoE achieves a ~6% improvement in average accuracy. These ablation results further support that the proposed PMoE architecture effectively leverages optical parallelism while maintaining strong multi-domain accuracy and parameter efficiency."

- Revised manuscript (Discussion, Page 9, Lines 310-315): "Crucially, this MoE-driven width-over-depth strategy is compatible with passive diffractive experts, which provide high-density pre-programmed optical weights without requiring individual active control during inference. Although these diffractive weights are not dynamically reprogrammable after fabrication, coordinating multiple fixed experts through lightweight routing provides a practical route to increase the effective parameter capacity for multi-domain optical inference."

- Revised manuscript (Discussion, Page 10, Lines 345-346): "In conclusion, we present a novel multi-core monolithic PMoE architecture designed to enable scalable optical multi-domain processing."

- Revised Supplementary Information (Note 8, Pages 10-11):

  We expanded the ablation discussion in this note to clarify the comparison principles, including the use of identical multi-domain datasets, comparable architectural constraints, and similar parameter budgets. We also added a CNN-based electronic MoE baseline with the same overall multi-expert/shared-backend design principle as PMoE.

- New added Reference:
  "41. Büchel, J. *et al.* Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing. *Nature Computational Science* **5**, 13-26 (2025).
  42. Xu, B. *et al.* Optical Spiking Neurons Enable High-Speed and Energy-Efficient Optical Neural Networks. *arXiv preprint arXiv:2409.05726* (2024).
  43. Li, C. in *Proceedings of the 2024 4th International Conference on Signal Processing and Communication Technology.* 260-265.
  44. Bernadskiy, M. *et al.* in *2025 IEEE Symposium on High-Performance Interconnects (HOTI).* 25-35 (IEEE)."

R3-3. *Claims/Results in the paper that need to be backed up in more detail: "networks, featuring 18 parallel kernels within a compact computational footprint of 0.06 mm2" This is misleading in the abstract as this implies system-level behavior but only considers the footprint of a small part of the system.*

A3-3. We sincerely thank you for pointing out this ambiguity. We agree that presenting the 0.06 mm² footprint in the Abstract without explicitly specifying that it refers to the computational core could be interpreted as implying a system-level footprint.

We would like to clarify that the 0.06 mm² value refers strictly to the intrinsic computational-core area of the PMoE chip, rather than the full system including the I/O

chain and peripheral components. In the revised manuscript and Supplementary Information, we now explicitly distinguish between the intrinsic computational-core footprint and the estimated full-system footprint including the full I/O chain.

Our rationale for highlighting the 0.06 mm^2 metric is that the computational core represents the physical footprint of the actual computing engine. Directly comparing the total system area across different hardware paradigms can be challenging due to high design variability. Furthermore, because many reported optical chips rely on external discrete electro-optic modulators and peripherals rather than monolithic integration, which are not included in the footprint benchmark [1-3], cross-platform comparisons of total area are inherently inconsistent and difficult. Therefore, to provide a more reliable comparison basis, we isolate the area of the computing core. At the same time, we agree that system-level footprint is also important. Therefore, following your suggestion, we have added a conservative projected full-system footprint estimate of approximately 25 mm^2 in the Supplementary Information, which includes the I/O chain, datapaths, and integrated PMoE components, as detailed in our response **A3-8**.

To avoid any potential misunderstanding, we have revised the Abstract, Introduction and Discussion of the manuscript to explicitly state that 0.06 mm^2 refers to the intrinsic computational-core footprint. We have also revised Supplementary note 10 and the corresponding comparison Table S2 to report the projected full-system footprint as a system-level reference.

Reference:

1. Z. Xu et al., Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. Science 384, 202-209 (2024).

2. J. Feldmann et al., Parallel convolutional processing using an integrated photonic tensor core. Nature 589, 52-58 (2021).

3. X. Wang et al., Integrated photonic encoder for low power and high-speed image processing. Nature Communications 15, 4510 (2024).

Relevant changes made:

- Revised manuscript (Abstract, Page 1, Lines 33-34): “featuring 18 parallel kernels within a compact intrinsic computational-core footprint of 0.06 mm^2.”
- Revised manuscript (Introduction, Page 3, Line 100): “Occupying a compact intrinsic computational-core footprint of 0.06 mm^2.”
- Revised manuscript (Discussion, Page 10, Lines 332-334): “corresponding to an experimental computational capacity of 3.06 MOPS and a unit computational capacity of 51 MOPS/mm^2 based on the 0.06 mm^2 computational-core area.”
- Revised Supplementary Information (Note 10, Page 15): “The computational core of PMoE chip has a dimension of 300 μm in total length and 74.5 μm in width of each PEN, resulting in a total footprint of 0.06 mm^2. Regarding the overhead of peripheral components, modern mixed-signal components (such as ADCs, DACs

and TIA) have become highly integrable. For instance, a 14 GHz DAC fabricated occupies a mere 0.011 mm^2, a 25 GHz ADC occupies only 0.23 mm^2 and a high-speed TIA occupies 0.54 mm^2. Given these advanced data and accounting for data paths, the projected total footprint of the integrated PMoE system, including the I/O chain, is conservatively estimated to be approximately 25 mm^2."

- Revised Supplementary Information (Note 10, Page 15): "Assuming the integration of plasma-dispersion-based silicon modulators with a modulation rate of 10 GHz, the computational density is estimated to reach 51.0 TOPS/mm^2 based on the 0.06 mm^2 computational-core area, or 0.12 TOPS/mm^2 under an estimated total footprint of 25 mm^2."

- Revised Supplementary Information (Note 10, Page 15): "However, directly comparing the total system area across disparate hardware paradigms is inherently inconsistent due to massive design variability. In electronic platforms like GPUs, the total die or board size is heavily dictated by macroscopic requirements, including DRAM placement, power delivery networks, and massive cooling solutions. Similarly, in optical computing, the total chip area is largely governed by subjective layout-dependent factors, such as waveguide routing, I/O pad spacing, and foundry-specific standard size limits. Furthermore, because many state-of-the-art optical architectures rely on external electro-optic modulators or bulk free-space optics rather than full monolithic integration, system-level area comparisons are often difficult to standardize. To provide a more consistent comparison metric, our evaluation primarily isolates the area of the intrinsic computational core. Unlike the highly variable peripheral layout, this diffractive core represents the fundamental physical footprint required for the actual optical processing, thereby serving as a stable and equitable basis for benchmarking structural computing density."

R3-4. *"delivering superior accuracy and efficiency of 51.0 TOPS/mm2." This statement in the introduction doesn't reflect the actual performance, as the authors point out in their own discussion.*

A3-4. We sincerely thank the reviewer for this rigorous scrutiny. We agree that the original wording could be read as not sufficiently distinguishing the experimentally demonstrated thermo-optic operation from the projected high-speed electro-optic performance. We have revised the Introduction, Discussion, and Supplementary note 10 to explicitly make this distinction while retaining the scalability implication of the PMoE architecture.

As you correctly noted, our experimental proof-of-concept uses on-chip thermo-optic modulation at up to 10 kHz. The 51.0 TOPS/mm^2 value is not a directly measured result of the present thermo-optic PMoE chip; rather, it is an estimated computational-density metric under projected 10 GHz high-speed electro-optic modulation. Current high-speed electro-optic modulators have been widely demonstrated [1–2], and similar high-speed assumptions have also been used in prior optical computing projections [3–

4]. Nevertheless, we agree that this distinction must be stated explicitly. Therefore, we have revised the Introduction, Discussion, Supplementary note 10, and Table S2 to clearly separate the experimentally measured 10 kHz performance from the projected 10 GHz electro-optic performance.

In addition, to better support the projected high-speed operation, we have added an experimentally anchored high-speed robustness evaluation, in which matched signal degradation measured at different baud rates is introduced into the PMoE feature maps. This added result provides preliminary support for the revised high-speed projection and helps distinguish estimated electro-optic performance from the directly measured thermo-optic results, as detailed in our response **A3-9.**

We fully agree with the need of distinguishing between experiments and theoretical projections. To maintain scientific rigor while avoiding under-representing the architecture's scalability, we have revised the Introduction and Discussion in the manuscript and supplementary note 10 to clarify the experimental performance and the theoretically projected computational density.

Reference:

1. C. Han et al., Slow-light silicon modulator with 110-GHz bandwidth. Science Advances 9, eadi5339 (2023).

2. M. Xu et al., High-performance coherent optical modulators based on thin-film lithium niobate platform. Nature communications 11, 3911 (2020).

3. C. Wang, Y. Cheng, Z. Xu, Q. Dai, L. Fang, Diffractive tensorized unit for million-TOPS general-purpose computing. Nature Photonics, 1-10 (2025).

4. J. Cheng et al., Multimodal deep learning using on-chip diffractive optics with in situ training capability. Nature Communications 15, 6189 (2024).

Relevant changes made:

- Revised manuscript (Introduction, Page 3, Lines 105-107): "Facilitating substantial width-wise scalability, the PMoE experimentally demonstrates efficient multi-domain optical processing, while reaching a projected computational density of 51.0 TOPS/mm^2 under 10 GHz electro-optic modulation."
- Revised manuscript (Discussion, Page 10, Lines 331-338): "In the experimental demonstration, we utilized on-chip thermo-optic modulation for signal loading at up to 10 kHz, corresponding to an experimental computational capacity of 3.06 MOPS and a unit computational capacity of 51 MOPS/mm^2 based on the 0.06 mm^2 computational-core area. For high-speed electro-optic implementation, the projected computational capacity of PMoE would reach 3.06 TOPS, with a unit computational capacity of 51 TOPS/mm^2 under 10 GHz modulation (Supplementary note 10). Including the full transmit/receive chain in the system model, the power consumption for PMoE can be estimated as 4.94 W, resulting in 10.3 TOPS/W/mm^2 under the same computational-core area (Table S1)."

- Revised manuscript (Discussion, Page 10, Lines 338-341): “We further evaluated the PMoE architecture under high-speed degradation conditions. The average classification accuracy decreases only slightly from 97.1% to 96.8% under the 10 GHz matched condition (Fig. S5), providing preliminary support for high-speed electro-optic implementations.”

- Revised Supplementary Information (Note 6, Pages 7–8):

  We added a high-speed robustness evaluation to support the 10 GHz electro-optic projection. Experimentally measured signal degradation at different baud rates was introduced into the PMoE feature maps, and the corresponding multi-domain classification accuracy was evaluated under matched high-speed degradation conditions.

- Revised Supplementary Information (Note 10, Page 15): “Consequently, under the 10 kHz thermo-optic modulation, the computational capacity of the PMoE chip is 3.06 MOPS, corresponding to 3.06 MOPS/0.06 mm^2 = 51 MOPS/mm^2 based on the computational-core area. Assuming the integration of plasma-dispersion-based silicon modulators with a modulation rate of 10 GHz, the computational density is estimated to reach 51.0 TOPS/mm^2 based on the 0.06 mm^2 computational-core area, or 0.12 TOPS/mm^2 under an estimated total footprint of 25 mm^2.”

R3-5. *The error plot Fig 2h indicates a significant nonsystematic error contribution. How does this effect the overall system?*

A3-5. We sincerely thank you for this keen observation. We agree that Fig. 2h exhibits a certain degree of nonsystematic error, with an RMSE of 0.042. These deviations primarily originate from the inherent physical imperfections of analog optical computing architectures. Specifically, they may be attributed to fabrication tolerances during the etching of the sub-wavelength metalines, possible residual thermal drift or crosstalk in the on-chip thermo-optic modulators, and unavoidable noise in input signal loading and output measurement.

Regarding its effect on the overall system, we would like to clarify that these errors have a limited impact on the final classification performance, as the PMoE still achieves an average accuracy of 97.1%. This robustness mainly comes from two aspects:

1. Preservation of structural features: While analog noise introduces deviations in absolute intensity values, the diffractive optical core preserves the relative structural features well. As evaluated in Fig. 2g and Supplementary note 4, the experimental feature maps maintain a high average Structural Similarity Index (SSIM) of 0.923 compared to the ideal simulations. For multi-domain classification tasks, capturing these relative structural features, such as edges and contours, is more important than strict pixel-wise absolute intensity matching.

2. Robustness of the hybrid opto-electronic architecture: The PMoE utilizes a hybrid hardware framework. The frontend optical variations are subsequently processed by the backend Shared Digital Network (SDN). Deep neural networks,

particularly the backend fully connected and convolutional layers, are generally tolerant to moderate, largely non-systematic analog perturbations [1–2]. The digital backend therefore helps classify inputs based on the preserved feature distributions, reducing the influence of frontend hardware-induced nonsystematic errors.

Following your constructive feedback, we have expanded Supplementary note 4 in the revised Supplementary Information to explicitly discuss the origins of these nonsystematic errors and how the hybrid architecture mitigates their impact on the overall system performance.

Reference:

1. Y. Chen et al., All-analog photoelectronic chip for high-speed vision tasks. Nature 623, 48-57 (2023).

2. T. Fu et al., Optical neural networks: progress and challenges. Light: Science & Applications 13, 263 (2024).

Relevant changes made:

- Revised manuscript (Results, Page 7, Lines 202-203): "Although minor errors exist, the structural features are well preserved, and the final classification performance remains stable."
- Revised Supplementary Information (Note 4, Page 5): "Furthermore, as indicated in Fig. 2h (main text), a certain degree of error (RMSE = 0.042) exists in the experimental outcomes. These deviations primarily originate from the physical imperfections inherent to analog optical computing architectures, such as fabrication tolerances during the etching of the sub-wavelength metalines, possible residual thermal drift or crosstalk in the on-chip thermo-optic modulators, and noise from input signal loading and output measurement. However, these errors have a limited impact on the final classification performance. While analog noise introduces deviations in absolute intensity values, the diffractive optical core preserves the relative structural features well, as evidenced by the high average SSIM of 0.923. In addition, the PMoE utilizes a hybrid opto-electronic framework. The backend network is generally tolerant to moderate, largely non-systematic analog perturbations. The SDN therefore acts as an algorithmic buffer, classifying inputs based on the preserved feature distributions and reducing the impact of frontend hardware-induced nonsystematic errors."

R3-6. *"representing a ~16% improvement over the digital counterpart" It would make more sense to benchmark against the published SOTA for this task rather than some CNN trained by the authors. The baseline also seems to be quite low, given the comparable simple datasets.*

A3-6. We thank you for this important comment. We agree that the digital CNN baseline shown in the original manuscript should not be interpreted as the published state of the art for MNIST, Fashion-MNIST, or Extended-MNIST, and that a more appropriate

neural-network baseline is needed for a fairer comparison with PMoE.

Our intention in the original manuscript was not to claim state-of-the-art classification accuracy on these datasets as standalone tasks. Rather, we aimed to validate the effectiveness "MoE for photonics" architecture, with a particular focus on multi-domain processing and width-wise scalability in optical neural networks. In this context, the digital CNN baseline was introduced as a controlled reference under the same multi-domain setting, with comparable network scale and similar architectural constraints to the PMoE system, as shown in **Fig. R3-2**. Since our work focuses on expert-specific front-end feature extraction together with a shared backend, direct comparison to single-task SOTA results is not strictly apples-to-apples. Those results are typically obtained using fully electronic models optimized for a single dataset, without the same multi-domain setting, shared-backend or parameter count constraints.

Following your suggestion, we have revised the manuscript to clarify the setup of the different baselines and to strengthen the architecture-level comparison. More importantly, we now include a controlled electronic CNN-based MoE baseline with the same overall design philosophy as PMoE. In this model, the front-end expert convolutional layers are responsible for capturing task-specific differences, while most of the backend parameters are shared across the three classification tasks, consistent with the PMoE framework. As shown in **Fig. R3-3**, this electronic CNN-based MoE achieves test accuracies of 95.8%, 84.3%, and 92.1% on MNIST, Fashion-MNIST, and Extended-MNIST, respectively, corresponding to an average accuracy of 91.2%. Compared with the ordinary digital CNN baseline, this result confirms that the expert-based front-end together with a shared-backend design is itself an effective strategy for multi-domain classification under constrained parameters.

Under this more appropriate controlled comparison, PMoE still achieves a higher average accuracy (~6%) than the electronic CNN-based MoE counterpart, while maintaining the same overall multi-expert/shared-backend design philosophy. We therefore regard this electronic CNN-based MoE as a more relevant electronic baseline for PMoE than the original auxiliary digital CNN. A possible reason for the remaining performance gap is that the photonic expert front-end implements a richer complex-valued optical transformation, which may provide more expressive feature extraction than a purely electronic convolutional counterpart under the same overall architecture.

We have updated the comparison figures and related descriptions (~16% compared with the digital counterpart => ~6% compared with electronic CNN-based MoE counterpart) in the manuscript, and expanded the ablation note in the revised Supplementary Information to provide a clearer explanation of the different baselines, including the added CNN-based electronic MoE and its corresponding figure.

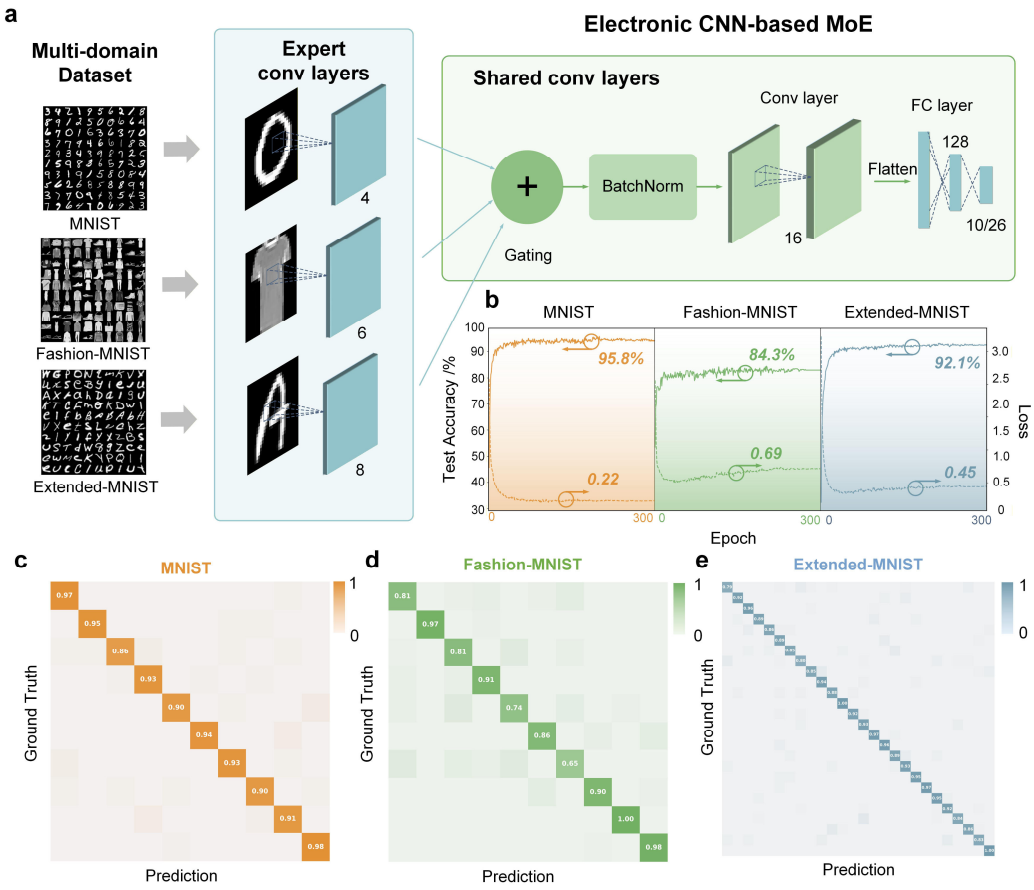


Fig. R3-3. Controlled electronic CNN-based MoE baseline for multi-domain classification. a, Architecture of the electronic CNN-based MoE counterpart, in which task-specific expert convolutional layers are combined with a largely shared backend network. **b,** Test accuracy and loss curves for MNIST, Fashion-MNIST, and Extended-MNIST. **c–e,** Confusion matrices of the electronic CNN-based MoE baseline on the MNIST, Fashion-MNIST, and Extended-MNIST test sets, respectively.

Relevant changes made:

- Revised manuscript (Discussion, Page 9, Lines 292-301): "For a rigorous ablation, we compared the PMoE against multiple optical and electronic baselines under the same multi-domain setting, with comparable network scale and similar architectural constraints (Supplementary note 8). Fig. 5a shows the controlled comparison curves between PMoE and other platforms in terms of average classification accuracy and the number of parameters in the electronic network. Compared with separate electronic or optical task-specific networks on the same multi-domain datasets, PMoE achieved higher accuracy while reducing the number of network parameters by ~67%. Compared to an electronic CNN-based MoE counterpart with the same overall design principle, the PMoE achieves a ~6% improvement in average accuracy. These ablation results further support that the proposed PMoE architecture effectively leverages optical parallelism while maintaining strong multi-domain accuracy and parameter efficiency."

- Revised manuscript (Fig. 5, Page 12):

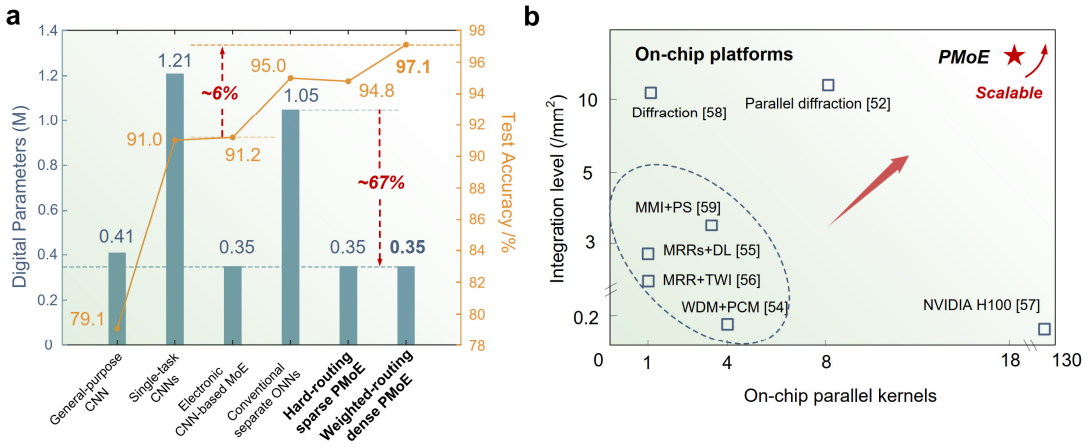


Fig. 5.

Comparison of multi-domain and on-chip computing platforms. a, Experimental comparison between PMoE and the other multi-domain platforms. PMoE reduces the digital parameter count by ~67% compared with conventional separate ONNs and improves the average accuracy by ~6% compared with an electronic CNN-based MoE with a similar architectural design. **b,** Comparison of PMoE with other on-chip computing platforms in terms of computational-core integration level and equivalent parallel kernels.

- Revised Supplementary Information (Note 8, Pages 10-11): "To further highlight the performance of the PMoE system, we conducted comparative evaluations against multiple electronic and optical baselines under the same multi-domain dataset setting. The purpose of these ablation studies is not to benchmark against the published single-task state of the art, but to evaluate the effectiveness of the PMoE architecture under controlled and comparable conditions. Therefore, the electronic CNN baselines were introduced as a controlled reference under the same multi-domain setting, with comparable network scale and similar architectural constraints to the PMoE system."

- Revised Supplementary Information (Note 8, Page 11): "For the baseline of three single-task, two-stage electronic CNNs, each model was trained individually on its corresponding dataset. This baseline reflects the conventional strategy of solving multi-domain processing by deploying multiple task-specific networks. Under this setting, PMoE achieves higher average accuracy while requiring a much smaller total parameter count in the backend, because it uses a shared downstream network rather than three fully separate task-specific models. The training dynamics and confusion matrices for each single-task CNN are shown in Fig. S8a to c."

- Revised Supplementary Information (Note 8, Page 11): "To provide a more appropriate architecture-level electronic baseline, we further constructed an electronic CNN-based MoE counterpart following the same overall design philosophy as PMoE. As shown in Fig. S9a, this model adopts task-specific expert convolutional layers in the front-end, while most of the backend parameters remain

shared across the three classification tasks. This baseline was designed with the same overall two-stage architecture and a comparable parameter budget, so that the comparison isolates the role of the photonic expert front-end rather than being dominated by differences in model scale. The corresponding training curves are shown in Fig. S9b, and the confusion matrices on multi-domain tasks are given in Fig. S9c to e. This result confirms that the expert-based front-end together with a shared backend is an effective design for multi-domain classification under constrained parameter budgets. Under this more controlled architecture-level comparison, PMoE achieves a higher average accuracy (~6%) than the electronic CNN-based MoE counterpart. A possible reason for this remaining performance gap is that the photonic expert front-end implements a richer complex-valued optical transformation, which may provide more expressive feature extraction than a purely electronic convolutional counterpart under the same overall architecture."

- Revised Supplementary Information (Fig. S9, Page 14):

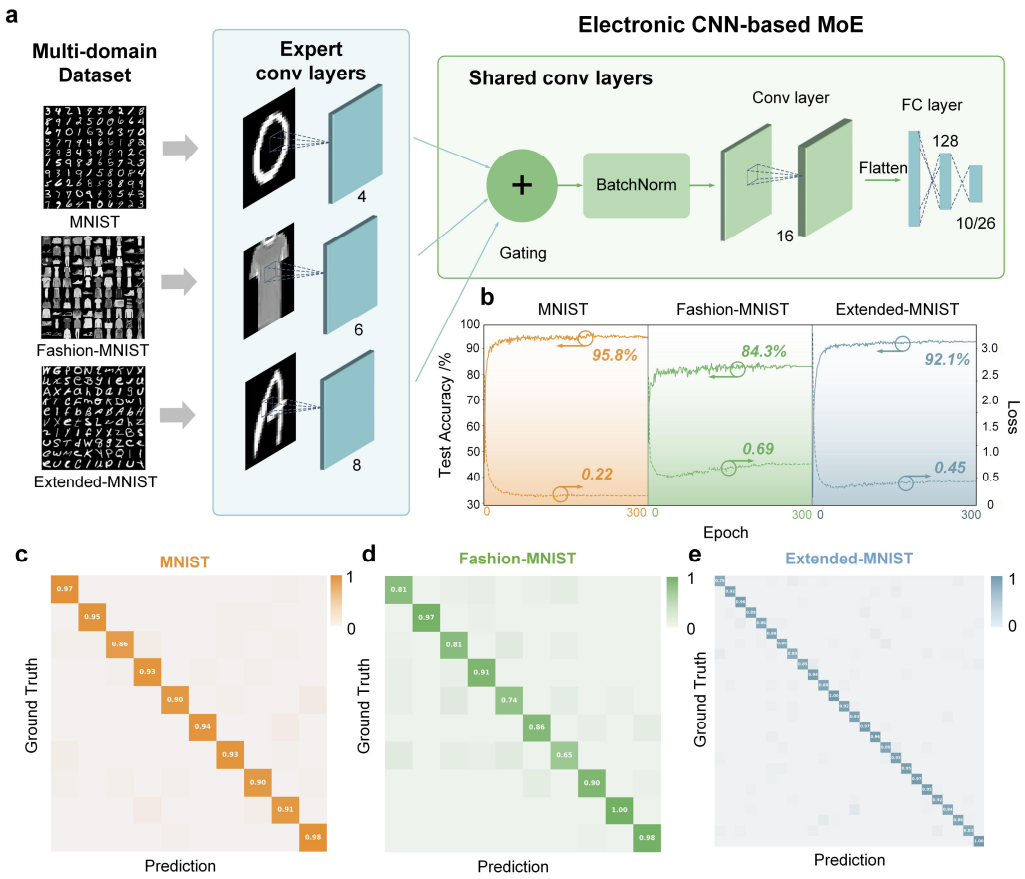


Fig. S9.
Controlled electronic CNN-based MoE baseline for multi-domain classification. a, Architecture of the electronic CNN-based MoE counterpart, in which task-specific expert convolutional layers are combined with a largely shared backend network. **b,** Test accuracy and loss curves for MNIST, Fashion-MNIST, and Extended-MNIST. **c–e,** Confusion matrices of the electronic CNN-based MoE baseline on the multi-domain datasets.

R3-7. *"This work charts a path for the future scalability of optical computing." Number of programmable weights is one of the largest challenges for scaling optical computing rather than compute. How would this approach solve this exactly? An in-depth scaling analysis would be good to back up this claim.*

A3-7. We sincerely appreciate the reviewer for highlighting this fundamental issue. We agree that the limited number of programmable optical weights, rather than pure compute throughput alone, is one of the most critical bottlenecks for scaling optical computing. In particular, active programmable weights based on discrete components like MZIs or microring resonators usually require dense electrical control, calibration, and stabilization, which makes large-scale parameter expansion challenging. Our PMoE does not claim to make every optical weight dynamically programmable after fabrication. Instead, it alleviates the effective parameter-capacity bottleneck by using high-density passive, pre-programmed diffractive optical weights coordinated by a small number of task-specific routing weights for optical multi-task inference.

To explicitly clarify the core motivation and the rationale behind introducing MoE to the optical domain, we would like to highlight the structural synergy between the MoE paradigm and optical hardware:

1. The physical bottleneck of optics: Conventional AI models typically scale by increasing depth. However, directly applying this "depth-scaling" strategy to optical neural networks encounters major physical constraints. Cascading deep optical layers leads to accumulated insertion loss, noise and error accumulation, and increasing control complexity. Moreover, scaling the number of fully active programmable optical weights introduces substantial electrical-control and calibration overhead. Therefore, optical networks are physically constrained not only by computation throughput, but also by the difficulty of scaling controllable optical parameters.

2. The MoE synergy (passive width-over-depth scaling): The MoE architecture changes how the model capacity is expanded: it increases width by partitioning the model into multiple parallel expert subnetworks. This algorithmic structure is well aligned with the physical strengths of diffractive photonics. Compared with MZI- or microring-based active programmable weights, diffractive optical weights can be densely integrated as passive pre-programmed structures without requiring individual active control during inference. Although these diffractive weights are not dynamically reprogrammable after fabrication, they provide a highly compact way to expand the number of available optical parameters. In PMoE, multiple such fixed optical expert networks are coordinated through lightweight task-specific routing, allowing the system to adapt to different tasks without reprogramming all optical weights.

This decoupling between static high-density optical feature extraction and dynamic lightweight routing is central to PMoE. It allows the architecture to expand its effective multi-domain parameter capacity in width while avoiding the control burden that would arise from tuning a very large number of active optical weights. Therefore, the PMoE provides a practical route to increase the effective accessible parameter capacity for

multi-domain optical inference.

To support this claim quantitatively, we added a physical scaling analysis in the revised Discussion section. The scaling limit is determined by the optical link budget. Specifically:

1. Assuming a typical on-chip laser input power of 15 dBm and a photodetector sensitivity threshold of −20 dBm to maintain an acceptable signal-to-noise ratio;
2. Given an estimated loss of approximately 15 dB for each diffractive expert, the remaining allowable power budget for equal-power splitting is 20 dB;
3. Under these assumptions, the optical equal-power splitting budget supports on the order of 100 concurrently active optical experts.

This estimate should be interpreted as an order-of-magnitude link-budget analysis rather than a universal hard limit. It indicates that width scaling through parallel expert activation can substantially expand the number of simultaneously accessible optical weights without increasing the depth of any single optical subnetwork. Moreover, because MoE relies on sparse task-specific routing rather than activating all subnetworks simultaneously, the total number of physical experts fabricated on a chip can exceed the concurrently active expert count, especially when combined with advanced dynamic optical routing or switch arrays in future implementations.

We thank the reviewer for prompting us to clarify this important distinction. In the revised manuscript, we now explicitly distinguish programmable optical weights from effective parameter-capacity scaling through passive pre-programmed experts and lightweight routing. We also added an analysis to clarify the scaling limitation of concurrently active experts. These revisions better support the positioning of PMoE as a practical width-scaling pathway for multi-domain optical inference.

Relevant changes made:

- Revised manuscript (Results, Page 3, Lines 115-119): “It uses a small number of task-specific routing weights to coordinate multiple passive, pre-programmed optical expert networks. This decoupling between dynamic routing and static optical feature extraction increases the effective parameter capacity for multi-domain processing without requiring active programming of optical weights during operation.”
- Revised manuscript (Discussion, Page 9, Lines 310-317): “Crucially, this MoE-driven width-over-depth strategy is compatible with passive diffractive experts, which provide high-density pre-programmed optical weights without requiring individual active control during inference. Although these diffractive weights are not dynamically reprogrammable after fabrication, coordinating multiple fixed experts through lightweight routing provides a practical route to increase the effective parameter capacity for multi-domain optical inference. With appropriate routing and weighting mechanisms, this broader width-over-depth design principle may also inform future photonic architectures beyond diffractive systems.”

- Revised manuscript (Discussion, Page 10, Lines 318-330): “While the width-over-depth scaling paradigm offers significant advantages in mitigating the limitations of deep optical neural networks, width scaling via optical splitting faces its own physical constraints. Assuming a typical on-chip laser input power of 15 dBm, a photodetector sensitivity threshold of −20 dBm, and an estimated total loss of 15 dB per expert subnetwork, the remaining optical power budget available for splitting is 20 dB. Under an equal-power splitting condition, the allowable number of concurrently active optical experts is on the order of 100. Because the MoE architecture relies on sparse task-specific routing rather than activating all subnetworks simultaneously, the total number of physical experts integrated on the chip can exceed this concurrent-activation estimate. Future implementations can further extend this scaling range by adopting advanced dynamic optical routing mechanisms, such as modulate-then-split configurations using tunable optical splitters or the integration of highly efficient optical switch arrays. These routing schemes, coupled with sparsity-aware network algorithms, can further improve the scalability of the PMoE architecture.”

R3-8. *“the intrinsic computational core area (0.06 mm2) and the total footprint (15 mm2) encompassing active modulation components.” Those numbers should include the full systems, especially the I/O (e.g. DAC, ADC, TIA and datapaths) to be suitable for a comparison with the SOTA.*

A3-8. We sincerely appreciate the reviewer for highlighting the importance of peripheral overhead in area benchmarking. We agree that, for a more complete comparison with the state-of-the-art, the footprint should not be limited to the intrinsic computational core, but should also account for the I/O chain and other supporting components. Following this suggestion, we have refined our benchmarking approach to explicitly report both the intrinsic computational-core footprint and a conservative projected full-system footprint that includes the I/O chain and datapaths.

1. Full-system footprint estimation including peripheral components:

To address your concern, we estimated the area contribution of the main peripheral components required for a projected high-speed PMoE implementation, including DACs, ADCs, TIAs, and datapaths. Based on representative state-of-the-art mixed-signal components, a 14 GHz 8-bit DAC occupies approximately 0.011 mm^2 [1], a 25 GHz ADC occupies approximately 0.23 mm^2 [2], and a high-speed TIA occupies approximately 0.54 mm^2 [3]. Using these component-level estimates together with the projected datapath overhead, we conservatively estimate the total footprint of the integrated PMoE system to be approximately 25 mm^2. We have added this estimate to the revised Supplementary Information and comparison table as a supplementary system-level reference.

2. Role of core-area and full-system-area metrics:

We agree that the full-system footprint is important for system-level comparison. At the same time, we retain the computational-core area as an important structural

metric, because it reflects the physical footprint of the optical computing engine itself and is less dependent on layout, packaging, and peripheral integration choices. For electronic GPUs, the total board or die area is heavily dictated by subjective design choices regarding power management modules, DRAM placement, and massive cooling solutions. Similarly, for optical computing chips, the total area is severely influenced by subjective layout decisions, such as waveguide routing, component packing density, and foundry-specific standard size limits. Furthermore, because many reported optical architectures rely on external discrete modulators or peripheral devices rather than full monolithic integration, cross-platform comparisons of total area including the I/O chain are often difficult to standardize [4-5]. Therefore, to provide a more reliable comparison basis, we isolate the area of the computing core, which represents the fundamental architectural footprint required for the actual processing.

In the revised manuscript, we therefore provide these two quantities clearly: the core area is used for intrinsic structural density and cross-platform comparison, while the full-system footprint is provided as an estimated system-level perspective.

3. Revised benchmark of intrinsic core computing density (area):

As the computing core is the fundamental engine of any architecture, a smaller core generally indicates higher intrinsic computing density and scalability. To ensure an objective structural comparison, we benchmark the efficiency of the computing cores. In our revised comparison table, the computing density of state-of-the-art electronic GPUs and optical computing chips is estimated based strictly on their arithmetic core region. For GPUs, it typically occupies ~20% of the chip area [6]. Furthermore, following your constructive suggestion, we have added the benchmarks under the conservative total footprint estimation of 25 mm^2. This refined comparison in **Table R3-1** shows that our PMoE architecture retains a notable advantage in core computing density, while also providing a reference under the estimated full-system footprint.

4. Revised benchmark of end-to-end system efficiency (energy):

While area is benchmarked primarily at the core level to evaluate structural scalability and area efficiency, we fully agree that energy must be evaluated at the system level. To address the concern, we have evaluated the total energy consumption of all utilized components (including advanced ADCs, DACs, and TIAs) for our energy efficiency metric. As detailed in **Table R3-2**, this comprehensive estimation yields a realistic total system power of 4.94 W. This provides an objective comparison of the total energy required to support the computing throughput, resulting in a full-system energy efficiency of 10.3 TOPS/W/mm^2 (or 24.7 GOPS/W/mm^2 under the estimated 25 mm^2 total footprint).

To accurately reflect these system-level evaluations, we have revised the theoretical basis of the computational density in the Supplementary Information. To ensure a coherent presentation and overall consistency throughout the paper, we have synchronized and consolidated the relevant descriptions in the Discussion section of the main text and Table 1, with the Supplementary Information note 10 and 11.

Table R3-1: Comparison of computational footprint, density, and energy efficiency for PMoE and other on-chip computing architectures.

Works	**Technique**	**Computational footprint (mm^2) [1]**	**Computational density ($TOPS/mm^2$) [2]**	**Energy efficiency ($GOPS/W/mm^2$)**
Feldmann. et al	WDM+PCM	6.07	0.02	20.74
Fu. et al	Diffraction	0.3	0.4*	1.47*
Dong. et al	Weighted Matrix	38.37	0.42	26.1
Huang. et al	Parallel diffraction	0.08	17*	62.9*
Hong. et al	MZI	20.44	0.06	59.2
Ahmed. et al	MZI	349	0.19	5.01
Xu. et al	Diffraction+MZI	28.66	0.28	15.7
Nvidia H100 PCIe	MOSFET	145	6.07	8.52
This work	**Diffraction-based PMoE**	**0.06 (~25[#])**	**51.0*** (0.12[#])	**10.3×10^3*** (24.7[#])

1 Footprint only encompasses the intrinsic computational-core area.
2 Computational density is evaluated based on peak throughput.
* Estimated under 10 GHz modulation.
[#] Total footprint estimation including full I/O chain, datapaths and integrated PMoE.

Table R3-2: Benchtop and estimated power consumption of the full PMoE system.

Components		**Benchtop Power (W)**	**Estimated Power (W)**
External laser source		5 (peripheral)	0.16
Digital backend network		25 (peripheral)	0.1
Tx chain	DACs	\	0.45
	Modulators	0.49 (on-chip)	0.02*
Rx chain	TIAs	\	1.08
	ADCs	\	2.92
	PDs	9 (peripheral)	0.2**
TEC		0.005	
Total power consumption		39.5	4.94

* 10 fJ/bit energy efficiency for modulators with 10 GHz frequency is taken for the estimation.
** 1.12 A/W Ge photodetector with receiving power of -20 dBm is taken for the estimation.

Reference:

1. P. Caragiulo et al., 2020 IEEE Symposium on VLSI Circuits. (IEEE, 2020), pp. 1-2.

2. Y. Duan et al., A 12.8 GS/s time-interleaved ADC with 25 GHz effective resolution

bandwidth and 4.6 ENOB. IEEE Journal of Solid-State Circuits 49, 1725-1738 (2014).

3. B. Sedighi, & J. C. Scheytt, Low-power SiGe BiCMOS transimpedance amplifier for 25-GBaud optical links. *IEEE Transactions on Circuits and Systems II: Express Briefs* **59**, 461-465 (2012).

4. Z. Xu et al., Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. Science 384, 202-209 (2024).

5. J. Feldmann et al., Parallel convolutional processing using an integrated photonic tensor core. Nature 589, 52-58 (2021).

6. NVIDIA, NVIDIA H100 Tensor Core GPU Architecture (NVIDIA Corporation, 2022).

Relevant changes made:

- Revised Supplementary Information (Note 10, Page 15): "The computational core of PMoE chip has a dimension of 300 μm in total length and 74.5 μm in width of each PEN, resulting in a total footprint of 0.06 mm^2. Regarding the overhead of peripheral components, modern mixed-signal components (such as ADCs, DACs and TIA) have become highly integrable. For instance, a 14 GHz DAC fabricated occupies a mere 0.011 mm^2, a 25 GHz ADC occupies only 0.23 mm^2 and a high-speed TIA occupies 0.54 mm^2. Given these advanced data and accounting for data paths, the projected total footprint of the integrated PMoE system, including the I/O chain, is conservatively estimated to be approximately 25 mm^2."
- Revised Supplementary Information (Note 10, Page 15): "However, directly comparing the total system area across disparate hardware paradigms is inherently inconsistent due to massive design variability. In electronic platforms like GPUs, the total die or board size is heavily dictated by macroscopic requirements, including DRAM placement, power delivery networks, and massive cooling solutions. Similarly, in optical computing, the total chip area is largely governed by subjective layout-dependent factors, such as waveguide routing, I/O pad spacing, and foundry-specific standard size limits. Furthermore, because many state-of-the-art optical architectures rely on external electro-optic modulators or bulk free-space optics rather than full monolithic integration, system-level area comparisons are often difficult to standardize. To provide a more consistent comparison metric, our evaluation primarily isolates the area of the intrinsic computational core. Unlike the highly variable peripheral layout, this diffractive core represents the fundamental physical footprint required for the actual optical processing, thereby serving as a stable and equitable basis for benchmarking structural computing density."
- Revised Supplementary Information (Note 10, Page 15): "Consequently, under the 10 kHz thermo-optic modulation, the computational capacity of the PMoE chip is 3.06 MOPS, corresponding to 3.06 MOPS/0.06 mm^2 = 51 MOPS/mm^2 based on the computational-core area. Assuming the integration of plasma-dispersion-based

silicon modulators with a modulation rate of 10 GHz, the computational density is estimated to reach 51.0 TOPS/mm^2 based on the 0.06 mm^2 computational-core area, or 0.12 TOPS/mm^2 under an estimated total footprint of 25 mm^2."

- Revised Supplementary Information (Table S2, Page 18):

Table S2: Comparison of computational footprint, density, and energy efficiency for PMoE and other on-chip computing architectures.

Works	**Technique**	**Computational footprint (mm^2) [1]**	**Computational density (TOPS/mm^2) [2]**	**Energy efficiency (GOPS/W/mm^2)**
Feldmann. et al	WDM+PCM	6.07	0.02	20.74
Fu. et al	Diffraction	0.3	0.4*	1.47*
Dong. et al	Weighted Matrix	38.37	0.42	26.1
Huang. et al	Parallel diffraction	0.08	17*	62.9*
Hong. et al	MZI	20.44	0.06	59.2
Ahmed. et al	MZI	349	0.19	5.01
Xu. et al	Diffraction+MZI	28.66	0.28	15.7
Nvidia H100 PCIe	MOSFET	145	6.07	8.52
This work	**Diffraction-based PMoE**	**0.06 (~25[#])**	**51.0*** (0.12[#])	**10.3×10^3*** (24.7[#])

1 Footprint only encompasses the intrinsic computational-core area.
2 Computational density is evaluated based on peak throughput.
* Estimated under 10GHz modulation.
[#] Total footprint estimation including full I/O chain, datapaths and integrated PMoE.

R3-9. *Extrapolating from 10 kHz to 10 GHz needs more data points in between. Electronic noise characteristics and overall accuracy will be completely different.*

A3-9. We thank the reviewer for this important comment. We agree that a direct extrapolation from the present 10 kHz thermo-optic experiment to 10 GHz electro-optic operation would be insufficient without intermediate experimental evidence. To address this concern, we conducted a preliminary high-speed experiment based on a dedicated integrated diffractive-computing chain that comprises on-chip high-speed modulators, the diffractive computing core, and on-chip high-speed photodetectors. Using this test vehicle, we characterized its response for an input image signal at multiple intermediate baud rates, namely 1, 2, 5, 8, and 10 GHz, as shown in **Fig. R3-4a and b**. Quantitatively, the experimental signal quality degrades gradually with baud rate, with the SNR decreasing from 16.61 dB at 1 GHz to 14.61 dB at 10 GHz, while the MSE increases from 7.0 × 10^{-3} to 1.15 × 10^{-2}, as shown in **Fig. R3-4c**.

Based on these measurements, we further evaluated how the measured degradation would affect the accuracy of PMoE in the multi-domain task. As shown in **Fig. R3-5a**, we introduced matched degradation into the experimentally obtained PMoE feature maps according to the signal quality measured at different modulation baud rates (1, 2,

5, 8, and 10 GHz). Representative feature maps for Extended-MNIST, Fashion-MNIST, and MNIST remain visually consistent under all tested conditions, although the image quality gradually decreases as the baud rate increases. **Fig. R3-5b** shows the introduced degradation levels, with SNR and MSE corresponding to the high-speed experiment.

Using these experimentally matched degradation levels, we further evaluated the multi-domain classification accuracy of the PMoE system. With the same network structure and training setup, the average accuracy shows only a slight reduction, decreasing from 97.1% in the original thermo-optic experiment to 96.8% under the 10 GHz matched condition, corresponding to a drop of about 0.3 percentage points, as shown in **Fig. R3-5c**. These results indicate that the PMoE architecture for multi-domain processing remains robust to the level of signal degradation observed in high-speed deployment, and provide preliminary support for future electro-optic implementations for high-throughput optical computing.

Taken together, these results provide an experimentally constrained bridge between the currently demonstrated 10 kHz thermo-optic PMoE system and a prospective high-speed electro-optic implementation. Although this does not constitute a full-system 10 GHz demonstration of the present PMoE chip, it provides intermediate experimental evidence and an experimentally anchored evaluation of task-level robustness, which together support the revised high-speed projection. Following your suggestion, we have added a brief discussion in the main text and a new Supplementary note 6 in the Supplementary Information, together with Fig. S5, to present the high-speed robustness evaluation and its impact on multi-domain classification accuracy.

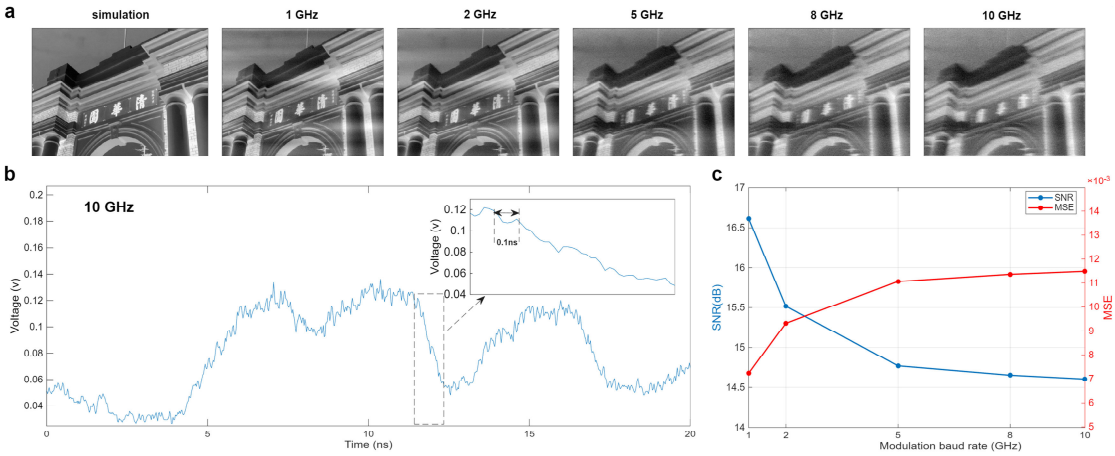


Fig. R3-4. Experimental characterization of the high-speed integrated diffractive-computing chain. a, Simulated and experimentally measured output feature maps at 1, 2, 5, 8, and 10 GHz. **b,** Measured temporal output waveform at 10 GHz. Inset: enlarged view showing a 0.1 ns symbol interval. **c,** Measured signal quality versus modulation baud rate, showing the corresponding SNR and MSE.

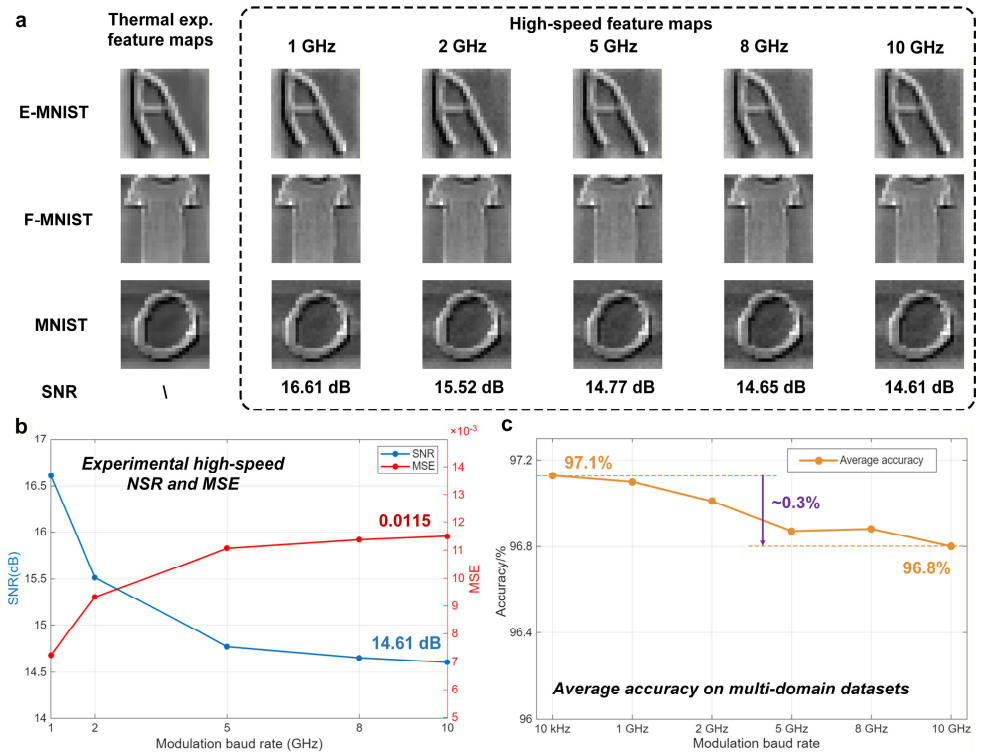


Fig. R3-5. Experimentally anchored evaluation of PMoE robustness under high-speed deployment conditions. a, Original thermal experimental feature maps and the corresponding feature maps with matched degradation under 1, 2, 5, 8, and 10 GHz conditions. **b,** SNR and MSE used to emulate the experimentally measured high-speed signal degradation. **c,** Average classification accuracy of the PMoE system on the multi-domain datasets under matched high-speed degradation conditions.

Relevant changes made:

- Revised manuscript (Discussion, Page 10, Lines 338-341): "We further evaluated the PMoE architecture under high-speed degradation conditions. The average classification accuracy decreases only slightly from 97.1% to 96.8% under the 10 GHz matched condition (Fig. S5), providing preliminary support for high-speed electro-optic implementations."

- Revised Supplementary Information (Note 6, Pages 7-8):

Supplementary note 6: Evaluation of PMoE under high-speed deployment conditions

To assess the feasibility of high-speed deployment, we further evaluated the PMoE architecture under experimentally matched high-speed signal degradation conditions. Since the current PMoE chip is driven by thermo-optic modulators, direct full-system measurements at GHz rates are not available on the present platform. Therefore, we performed a preliminary high-speed characterization on an integrated optical diffractive-computing chain and used the experimentally measured signal degradation to construct a physically grounded evaluation for the PMoE task.

As shown in Fig. S5a, we introduced matched degradation into the experimentally obtained PMoE feature maps according to the signal quality measured at different modulation baud rates (1, 2, 5, 8, and 10 GHz). Representative feature maps for

Extended-MNIST, Fashion-MNIST, and MNIST remain visually consistent under all tested conditions, although the image quality gradually decreases as the baud rate increases. Fig. S5b shows the introduced degradation levels. The corresponding signal-to-noise ratio (SNR) decreases from 16.61 dB at 1 GHz to 14.61 dB at 10 GHz, while the mean squared error (MSE) increases to 0.0115 at 10 GHz.

Using these experimentally matched degradation levels, we further evaluated the multi-domain classification accuracy of the PMoE. With the same network structure and training setup, the average accuracy shows only a slight reduction, decreasing from 97.1% in the original thermo-optic experiment to 96.8% under the 10 GHz matched condition, corresponding to a drop of about 0.3 percentage points, as shown in Fig. S5c. These results indicate that the PMoE network for multi-domain processing remains robust to the level of signal degradation observed in high-speed deployment, and provide preliminary support for future implementations for high-throughput optical computing.

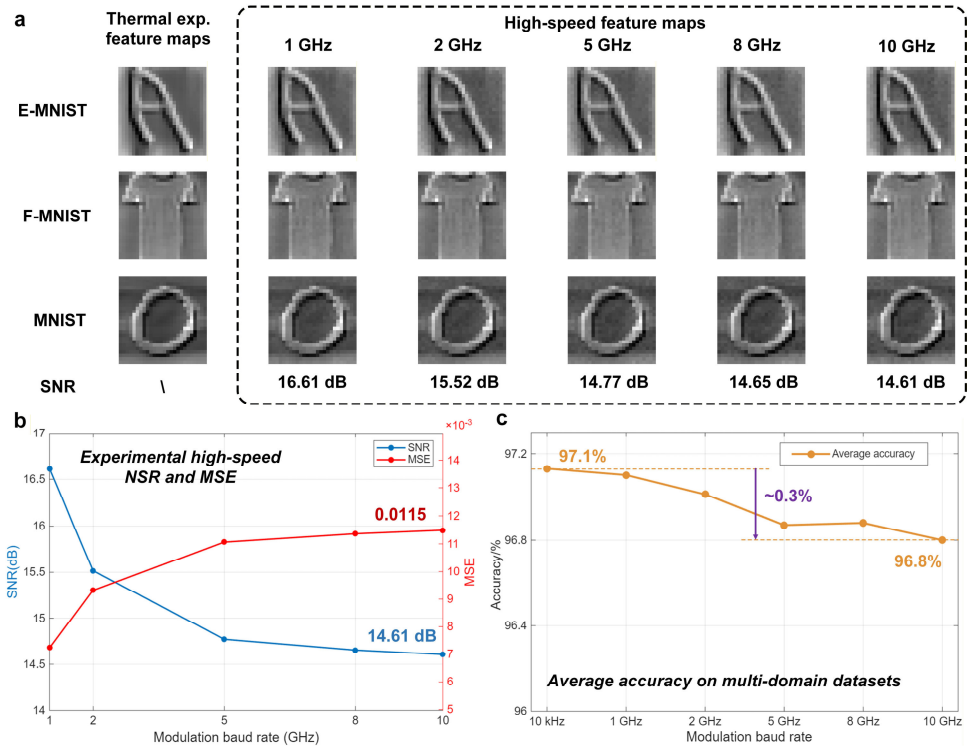


Fig. S5.

Experimentally anchored evaluation of PMoE robustness under high-speed deployment conditions. a, Thermo-optic experimental feature maps and the corresponding feature maps with matched degradation under 1, 2, 5, 8, and 10 GHz conditions. **b,** SNR and MSE used to emulate the experimentally measured high-speed signal degradation. **c,** Average classification accuracy of the PMoE system on the multi-domain datasets under matched high-speed degradation conditions.

R3-10. *"power consumption for PMoE can be predicted as 0.628 W, resulting 81.3 TOPS/W/mm2" The authors assume efficiency of 10 fJ/bit in the supplement (which should account for the full IO?). How do the authors come up with the number considering the full DAC, ADC, TIA chain?*

A3-10. We sincerely thank you for raising this important question. We agree that, in integrated opto-electronic computing systems, the peripheral electrical I/O, especially the DACs, ADCs, and TIAs, can dominate the total system power. Therefore, the 10 fJ/bit should not be interpreted as the energy efficiency of the full I/O chain and energy efficiency should be evaluated at a comprehensive full-system level as noted in **A3-8**.

To answer your question directly, the 10 fJ/bit value refers only to the dynamic energy consumption of the high-speed electro-optic modulators themselves, rather than the complete transmit/receive chain. In the original manuscript, our 0.628 W estimate did not make this distinction sufficiently clear and did not fully account for the realistic power of the full-bandwidth DAC/ADC/TIA chain.

Following your suggestion, we revised the power estimation model to include the complete I/O chain for a projected high-speed integrated implementation. With this revised model, the total system power is estimated to be **4.94 W**, detailed in **Table R3-3**. Specifically, the updated estimate includes:

1. **Transmitter chain (Tx):** On-chip Modulators (10 fJ/bit, ~0.02 W) + High-speed DACs (~0.45 W, allocating ~50 mW each) [1] for 9 input channels (each PEN).
2. **Receiver chain (Rx):** Photodetectors (~0.2 W) + High-speed TIAs (~1.08 W, allocating ~60 mW each) [2] + High-speed ADCs (~2.92 W, allocating ~162 mW each) [3] for 18 output channels (4, 6, 8 for the three PENs).
3. **Laser source:** ~0.16 W (assuming a packaged telecom laser with 32 mW output and a 20% wall-plug efficiency).
4. **Backend processing:** Lightweight digital backend network (~0.1 W) [4].
5. **TEC:** Thermal stabilization (~0.005 W).

Under the projected 10 GHz throughput, this revised full-system estimate corresponds to an energy efficiency of 10.3 TOPS/W/mm² when normalized to the 0.06 mm² computational-core area. We now explicitly state in the manuscript and Supplementary Information that this is a full-system estimate for a high-speed electro-optic implementation. For completeness, we also provide the corresponding footprint-normalized value under an estimated total system footprint in Supplementary note 11.

To improve transparency, we have revised the Discussion, Supplementary note 11, and Table S1 to explicitly distinguish the experimental benchtop power consumption figure from the full-system high-speed projected power estimate, and to provide the updated I/O-chain breakdown. Accordingly, we have also made following revisions in the Discussion of the manuscript.

Table R3-3: Benchtop and estimated power consumption of the full PMoE system.

Components	Benchtop Power (W)	Estimated Power (W)
External laser source	5 (peripheral)	0.16

Digital backend network		25 (peripheral)	0.1
Tx chain	DACs	\	0.45
	Modulators	0.49 (on-chip)	0.02*
Rx chain	TIAs	\	1.08
	ADCs	\	2.92
	PDs	9 (peripheral)	0.2**
TEC		0.005	
Total power consumption		39.5	4.94

* 10 fJ/bit energy efficiency for modulators with 10 GHz frequency is taken for the estimation.

** 1.12 A/W Ge photodetector with receiving power of -20 dBm is taken for the estimation.

Reference:

1. P. Caragiulo et al., 2020 IEEE Symposium on VLSI Circuits. (IEEE, 2020), pp. 1-2.

2. B. Sedighi, & J. C. Scheytt, Low-power SiGe BiCMOS transimpedance amplifier for 25-GBaud optical links. *IEEE Transactions on Circuits and Systems II: Express Briefs* **59**, 461-465 (2012).

3. Y. Duan et al., A 12.8 GS/s time-interleaved ADC with 25 GHz effective resolution bandwidth and 4.6 ENOB. IEEE Journal of Solid-State Circuits 49, 1725-1738 (2014).

4. Y. Chen et al., All-analog photoelectronic chip for high-speed vision tasks. Nature 623, 48-57 (2023).

Relevant changes made:

- Revised manuscript (Discussion, Page 10, Lines 333-337): "For high-speed electro-optic implementation, the projected computational capacity of PMoE would reach 3.06 TOPS, with a unit computational capacity of 51 TOPS/mm^2 under 10 GHz modulation (Supplementary note 10). Including the full transmit/receive chain in the system model, the power consumption for PMoE can be estimated as 4.94 W, resulting in 10.3 TOPS/W/mm^2 under the same computational-core area (Table S1)."
- Revised Supplementary Information (Note 11, Page 15): "The power consumption of the PMoE chip mainly comes from: external laser source, on-chip modulator array, off-chip current source, photodetector array, temperature electrical controller (TEC), digital backend and data controllers (e.g. DACs, ADCs)."
- Revised Supplementary Information (Note 11, Page 16): "Together, the full-system experimental benchtop power consumption is approximately 39.5 W."
- Revised Supplementary Information (Note 11, Page 17): "While benchtop devices

were utilized as transmitters and receivers in the experiment, a practical deployment would employ compact, advanced ADCs and DACs. Current state-of-the-art high-speed DACs with a 14 GHz bandwidth consume approximately 50 mW [8], and ADCs with a 25 GHz bandwidth consume approximately 162 mW [9]. Consequently, the estimated total power consumption for the integrated DACs and ADCs is approximately 3.34 W. A high-speed TIAs can be estimated as approximately 60 mW [10], with the total TIA power consumption of 1.08 W."

- Revised Supplementary Information (Note 11, Page 17): "Taking together, the full-system estimated power consumption for high-speed projection is approximately 4.94 W. We conclude the power consumption of the PMoE system in Table S1."
- Revised Supplementary Information (Note 11, Page 17): "The PMoE chip performs 306 operations per inference as calculated above, resulting a total system level ECPO under 10 GHz modulation of 4.94 W/(10 GHz × 306 OPs) = 1.61 pJ/OP and 1.61 pJ/OP × 0.06 mm^2 = 96.9 fJ·mm^2/OP under a 0.06 mm^2 core area. That is, the energy efficiency of the PMoE chip under 10 GHz modulation prediction can be calculated as 10.3 TOPS/(W · mm^2) and 24.7 GOPS/W/mm^2 under an estimated total footprint of 25 mm^2."
- Revised Supplementary Information (Table S1, Page 18):

Table S1: Benchtop and estimated power consumption of the full PMoE system.

Components		Benchtop Power (W)	Estimated Power (W)
External laser source		5 (peripheral)	0.16
Digital backend network		25 (peripheral)	0.1
Tx chain	DACs	\	0.45
	Modulators	0.49 (on-chip)	0.02*
Rx chain	TIAs	\	1.08
	ADCs	\	2.92
	PDs	9 (peripheral)	0.2**
TEC		0.005	
Total power consumption		39.5	4.94

* 10 fJ/bit energy efficiency for modulators with 10 GHz frequency is taken for the estimation.

** 1.12 A/W Ge photodetector with receiving power of -20 dBm is taken for the estimation.

Response Letter for NCOMMS-26-016715A

We sincerely appreciate the reviewers' continued evaluation of our revised manuscript and their positive assessment of the revision. We also thank the reviewers for pointing out the remaining minor issues, which have helped us further improve the precision and consistency of the final version. In this response letter, we address each remaining comment point by point. Responses to comments are highlighted in blue, and corresponding changes to the manuscript are marked in red.

Summary of the main revisions:

1. We have corrected the notation error in the description of the metaline layer index. The revisions are reflected in the [Results] and [Methods] sections.
2. We have revised Fig. 1 to make the placement of the task-specific routing weights consistent across the figures.
3. We have revised the terminology in Fig. 5b from "Integration level" to "Integration density" to avoid confusion with system-level integration. The corresponding figure caption and related text have also been updated.
4. We have made the relevant PMoE architecture code, source data and experimental output data/ raw input datasets publicly available on GitHub to improve reproducibility. The corresponding information has been added to the [Data availability] and [Code availability] section.
5. We have thoroughly proofread the manuscript and Supplementary Information again to ensure consistency, clarity, and precision.

The point-by-point responses to all remaining review comments are presented below.

Point-by-Point Responses to Reviewer #1's Comments

R1-0. *Thank you for your revised manuscript and the detailed response to my previous comments. I am pleased to see that all my initial concerns have been fully addressed. However, before final acceptance, there remain a few minor issues that should be corrected. Once these minor revisions are completed, the paper can be accepted for publication.* Thank you for your efforts.

A1-0. We sincerely thank the reviewer for the positive evaluation of our revised manuscript and for recognizing that the concerns raised in the previous round have been fully addressed. We are greatly encouraged by your supportive assessment and appreciate the remaining minor comments. We have carefully addressed these minor issues in the revised manuscript. We hope that the current revision now meets the standard for publication. Below, we provide our point-by-point responses to the remaining comments.

R1-1. *In line 129 and line 393, it should be "l-th metaline layer ", not "i-th"?*

A1-1. We thank the reviewer for the careful reading and for pointing out this notation error. We agree that the correct notation should be the *l*-th metaline layer, rather than the *i*-th metaline layer, because *l* denotes the layer index in our formulation. We have corrected this notation in both locations in the revised manuscript.

Relevant changes made:

- Revised manuscript (Results, Page 3, Line 129): "where $\Delta\phi_v^{(l)}$ is the phase shift induced by the *v*-th photonic neuron of the *l*-th metaline layer".
- Revised manuscript (Methods, Page 12, Line 393): "where $\Delta\phi_v^{(l)}$ is the phase shift induced by the *v*-th photonic neuron of the *l*-th metaline layer".

R1-2. *In Fig. 1(b), (c) and Fig. 4, g(n) is between input and "experts", however, which is between output and SDN in Fig. 1 (b). Besides, can the authors render the actual output channels of each expert? like 4, 6, 8.*

A1-2. We thank the reviewer for the careful observation. We agree that the placement of g(n) in the previous figures was not fully consistent and could cause confusion regarding the physical meaning of the routing operation.

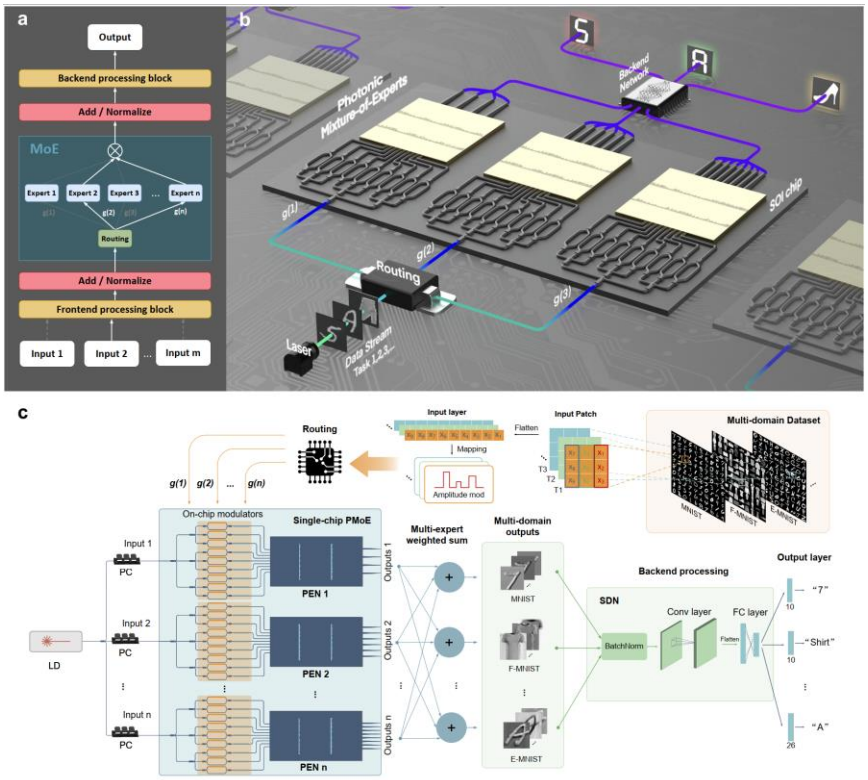
To clarify, g(n) represents the task-specific routing/loading weights applied at the input stage. In the current PMoE implementation, these weights are used to adjust the input optical power delivered to different photonic expert networks, thereby controlling the contribution of each expert for a given task. Therefore, g(n) should be placed between the input modulation stage and the photonic experts.

Following your suggestion, we have revised Fig. 1b to place g(n) consistently at the input-to-expert routing stage. We have also updated the figure to explicitly render

the actual output-channel numbers of the three experts, namely 4, 6, and 8 output channels, respectively. These revisions make the task-specific routing mechanism and the heterogeneous expert-output dimensions clearer.

Relevant changes made:

- Revised Fig. 1: repositioned g(n) to the input-to-expert routing stage to consistently indicate task-specific input routing weights. We also explicitly labeled the actual output-channel numbers of the three photonic experts as 4, 6, and 8, respectively.



R1-3. *If feasible, please consider releasing the relevant code to enhance the reproducibility of your work. This is not a mandatory requirement, but would be greatly appreciated by the community.*

A1-3. We thank the reviewer for this valuable suggestion. We agree that releasing the relevant code can improve the reproducibility of this work and benefit the broader community. Following your suggestion, we have made the source data and code publicly available on Github. The repository includes the source data and experimental results of the manuscript and code for the PMoE training and evaluation. We have also added the corresponding link in the revised manuscript.

Relevant changes made:

- Revised manuscript, Data availability section: "The code used for the PMoE network and data processing is available in the GitHub repository at https://github.com/Liuwc01/Photonic-Mixture-of-Experts."

Point-by-Point Responses to Reviewer #2's Comments

R2-0. *The authors have adequately addressed all the comments and revised the manuscript carefully. The quality of the paper meets the publication requirements. I recommend acceptance for publication.*

A2-0. We sincerely thank the reviewer for the positive evaluation of our manuscript. We greatly appreciate your careful assessment during the review process and are encouraged that the revised manuscript now meets the publication requirements. Thank you again for your valuable comments and constructive suggestions, which helped us improve the quality, clarity, and rigor of the work.

Point-by-Point Responses to Reviewer #3's Comments

R3-0. *The authors have provided thoughtful and substantive replies to most of the problems raised and made revisions and supplements in response to the review comments, significantly improving the scientific rigor and clarity of the manuscript.*

A3-0. We sincerely thank the reviewer for the positive evaluation of our revised manuscript and for recognizing the improvements in scientific rigor and clarity. We greatly appreciate your thoughtful and constructive comments throughout the review process, which have helped us refine the positioning, benchmarking, physical analysis, and presentation of the PMoE architecture. We have carefully considered the remaining comment and made corresponding revisions to further improve the manuscript. Below, we provide our point-by-point response.

R3-1. *A minor comment: "Integration level" in Fig 5B is still misleading, as it implies different levels of integration (e.g. on-chip laser sources, on-chip I/O ...). Something like "Integration density" makes more sense given the unit /mm^2 .*

A3-1. We thank the reviewer for this helpful suggestion. We agree that the term "Integration level" could be misleading, as it may imply the degree of system-level integration, such as on-chip light sources, I/O interfaces, or peripheral electronics. In Fig. 5b, our intended metric is the spatial density of equivalent computing kernels within the computational core, as indicated by the unit /mm^2. Therefore, "Integration density" is a more accurate and less ambiguous term.

Following your suggestion, we have revised the label in Fig. 5b from "Integration level" to "Integration density". We also updated the corresponding figure caption and related text to ensure consistent terminology throughout the manuscript.

Relevant changes made:

- Revised Fig. 5b: changed "Integration level" to "Integration density".
- Revised Fig. 5 caption: changed "computational-core integration level" to "computational-core integration density".
- Revised manuscript, Discussion: replaced "integration level" with "integration density" where this metric is discussed.