Efficient Quantum Vision Transformers

Vinay C Gogineni[1], Swapnil Mishra[2], Tridib Dey[2], Subrahmanyam Mula[2], and Esmaeil S Nadimi[1]

[1]Applied AI an Data Science, MMMI, University of Southern Denmark, [2]Electrical Engineering, IIT Palakkad

Research Question

How can Vision Transformer architectures be redesigned to operate efficiently in quantum domain, while maintaining competitive accuracy and scalability?

Contributions

- Developed Quantum Fourier Transform-based Vision Transformer (QFT-ViT), designed specifically for efficient vision tasks on quantum hardware.
- Introduced a QFT-based token mixer that enables fast global token mixing with poly-logarithmic time complexity.
- Incorporated adaptive frequency filtering in the quantum spectral domain to balance accuracy, scalability, and resource efficiency.

Proposed QFT-based QVITs

Our QFT-ViT architecture replaces the resource-heavy self-attention mechanism with a quantum-native token mixing module based on the QFT. The QFT converts an input quantum state $|j\rangle$ into a Fourier basis as follows:

$$\text{QFT}\,|j\rangle = \frac{1}{\sqrt{M}} \sum_{k=0}^{M-1} e^{2\pi i jk/M}\,|k\rangle \tag{1}$$

In the QFT-ViT encoder, image token embeddings are first encoded into a quantum state. After applying the QFT, a learnable Parameterized Quantum Circuit (PQC) acts as an adaptive filter in the frequency (spectral) domain, enabling the model to focus on task-relevant features:

$$|\Psi_{\text{filtered}}\rangle = U_{\text{filter}}(\boldsymbol{\phi})\,|\Psi_{\text{spectral}}\rangle \tag{2}$$

An Inverse QFT then transforms the filtered state back to the token domain for measurement. This quantum token mixer, combined with a hybrid Quantum MLP, is integrated within a standard residual block. This design avoids non-native quantum operations and exploits the QFT's efficient $\mathcal{O}((\log n)^2)$ complexity.

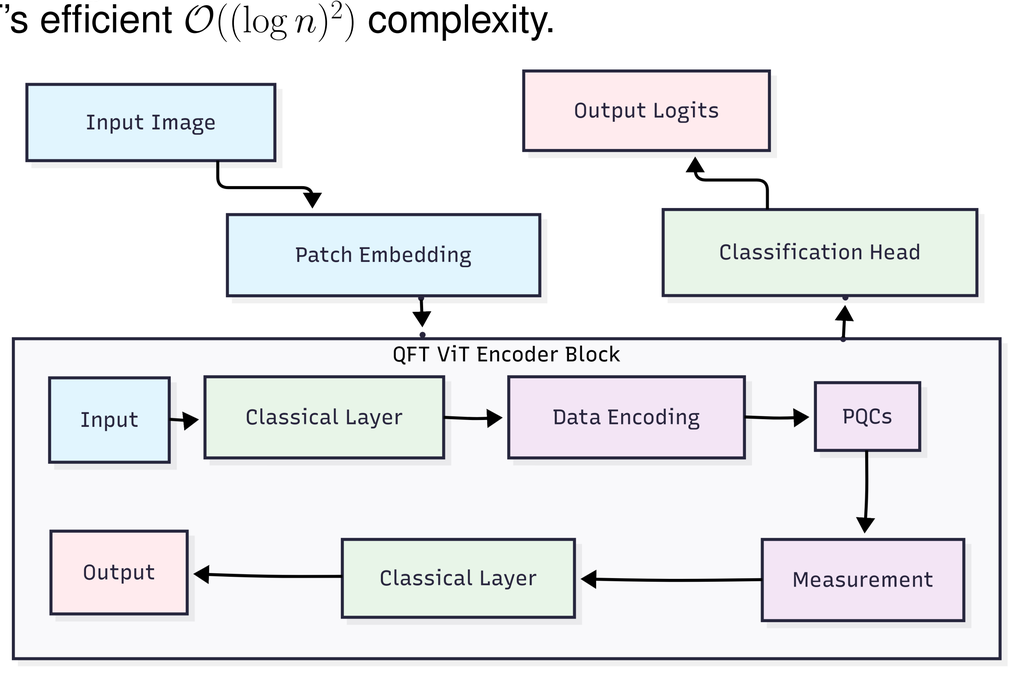
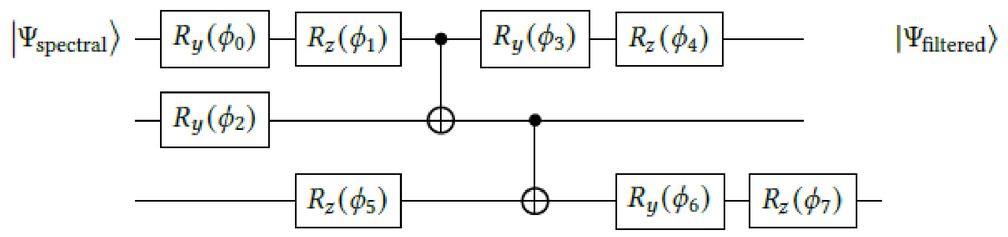
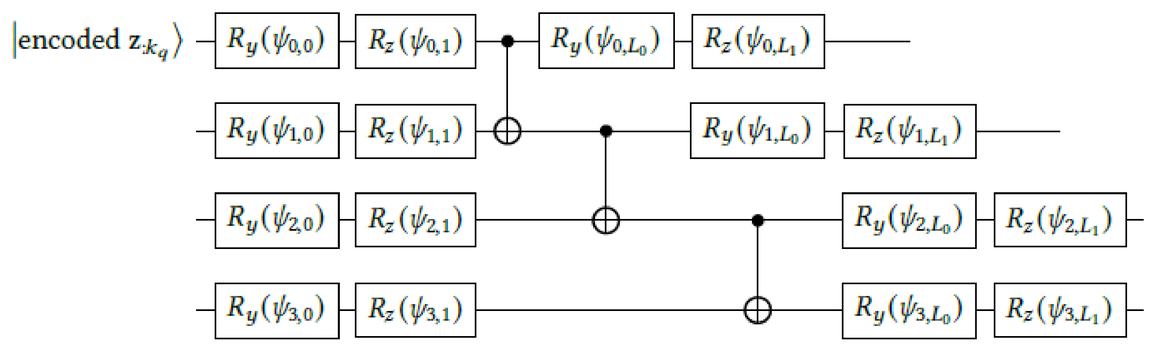


Figure 1: Workflow of QFT-ViT



(a) Adaptive quantum filter $U_{\text{filter}}(\phi)$ for the QFT Token mixer. It uses layers of single-qubit rotations and entangling CNOT gates to process the spectral states.



(b) PQC for $U_{\text{qmlp}}(\psi)$ quantum MLP. It processes encoded features using learnable single-qubit rotations and a linear chain of entangling gates to enhance feature transformation.

Figure 2: PQCs used in the encoder block.

Experimental Results

We evaluated our QFT-ViT against the standard ViT, a Hybrid Quantum ViT, and the classical FFT-ViT on the CIFAR-10, MEDMNIST, and OASIS datasets. The key metrics were test accuracy and computational cost (FLOPs), with all experiments simulated classically.

Table 1: Performance comparison on CIFAR-10.

Model	Test Acc. (%)	FLOPs (G)
ViT	80.50	0.19
Hybrid Quantum ViT	81.30	0.21
FFT-ViT	81.40	0.09
QFT-ViT (Ours)	**82.00**	**0.13**

The results on CIFAR-10 show that QFT-ViT achieves the highest test accuracy while maintaining a low computational footprint, significantly outperforming the standard ViT in efficiency. The model also demonstrated strong, competitive performance on the MEDMNIST and OASIS medical imaging benchmarks, confirming the robustness and generalization of our quantum-native approach.

Discussion & Conclusion

Our QFT-ViT balances efficiency and accuracy by replacing the resource intensive self-attention with a QFT-based token mixer, reducing complexity from quadratic to poly-logarithmic $\mathcal{O}((\log n)^2)$. Adaptive spectral filtering via a learnable PQC enables effective feature extraction while being quantum hardware-friendly. The design is inherently compatible with quantum hardware, demonstrating a promising path toward scalable quantum vision models.

References

1. Jacob Fein-Ashley, "The FFT Strikes Back: An Efficient Alternative to Self-Attention," *arXiv preprint arXiv:2502.18394*, 2025.
2. G. He, J. Tian, J.-B. Liu, and G.-L. Long, "Quantum Vision Transformers," *arXiv preprint arXiv:2209.08167*, 2022.
3. M. Schuld and F. Petruccione, *Supervised Learning with Quantum Computers*, Springer, 2018.

HAMLET August 20 - 22, 2025 Copenhagen, Denmark
How to Apply Machine Learning to Experimental & Theoretical PHYSICS