Hybrid-QVLM: A Quantum-Classical Framework for Enhanced Vision-Language Modeling

Abstract

As Large Vision-Language Models (LVLMs) scale toward trillion-parameter architectures, the computational cost of training and inference has become a primary bottleneck for sustainable AI development. We propose Hybrid-QVLM, a framework that explores the integration of parameterized quantum circuits (PQCs) into sparse classical backbones as a potential pathway toward more efficient architectures. Our framework introduces a Quantum CNN (QCNN) for vision-side feature extraction and Position-Aware Quantum Mixing (PAQM) to encode token relationships through quantum entanglement. To maintain practical viability, we employ Multi-head Latent Attention (MLA) and a Sparse Mixture-of-Experts (MoE) router, achieving a model with 571.33 M total parameters and sparsity-driven activation of only 288.06 M parameters per forward pass (49.6% sparsity ratio). Experimental evaluation on VQA v2 yields a validation accuracy of 72.20% and a CIDEr score of 0.66, while achieving 91.5% accuracy on Food-101 and 1.80 CIDEr on DiffusionDB, with an average MoE inference latency of ~24ms. While further research is needed to fully understand the interplay between quantum components and classical deep learning, this work opens a promising avenue for investigating hybrid quantum-classical architectures in the NISQ era. As quantum hardware continues to mature, such approaches may offer complementary strategies to conventional scaling paradigms, potentially contributing to more sustainable AI development.